Université de Montréal

FACTS-ON : Fighting Against Counterfeit Truths in Online social Networks : fake news, misinformation and disinformation

par

Sabrine Amri

Département d’informatique et de recherche opérationnelle Université de Montréal
Faculté des arts et des sciences

Thèse présentée en vue de l’obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Informatique

18 mars 2024

Université de Montréal

Faculté des arts et des sciences

Cette thèse intitulée

FACTS-ON : Fighting Against Counterfeit Truths in Online social Networks : fake news, misinformation and disinformation

présentée par

Sabrine Amri

a été évaluée par un jury composé des personnes suivantes :

Margarida Carvalho
(présidente-rapporteure)

Esma Aïmeur
(directrice de recherche)

Bang Liu
(membre du jury)

Kimiz Dalkir
(examinatrice externe)

Juliette De Maeyer
(représentante du doyen de la FESP)

Résumé

L’évolution rapide des réseaux sociaux en ligne (RSO) représente un défi significatif dans l’identification et l’atténuation des fausses informations, incluant les fausses nouvelles, la désinformation et la mésinformation. Cette complexité est amplifiée dans les environnements numériques où les informations sont rapidement diffusées, nécessitant des stratégies sophistiquées pour différencier le contenu authentique du faux. L’un des principaux défis dans la détection automatique de fausses informations est leur présentation réaliste, ressemblant souvent de près aux faits vérifiables. Cela pose de considérables défis aux systèmes d’intelligence artificielle (IA), nécessitant des données supplémentaires de sources externes, telles que des vérifications par des tiers, pour discerner efficacement la vérité. Par conséquent, il y a une évolution technologique continue pour contrer la sophistication croissante des fausses informations, mettant au défi et avançant les capacités de l’IA.

En réponse à ces défis, ma thèse introduit le cadre FACTS-ON (Fighting Against Counterfeit Truths in Online Social Networks), une approche complète et systématique pour combattre la désinformation dans les RSO. FACTS-ON intègre une série de systèmes avancés, chacun s’appuyant sur les capacités de son prédécesseur pour améliorer la stratégie globale de détection et d’atténuation des fausses informations. Je commence par présenter le cadre FACTS-ON, qui pose les fondements de ma solution, puis je détaille chaque système au sein du cadre :

EXMULF (Explainable Multimodal Content-based Fake News Detection) se concentre sur l’analyse du texte et des images dans les contenus en ligne en utilisant des techniques multimodales avancées, couplées à une IA explicable pour fournir des évaluations transparentes et compréhensibles des fausses informations.

En s’appuyant sur les bases d’EXMULF, MythXpose (Multimodal Content and Social Context-based System for Explainable False Information Detection with Personality Prediction) ajoute une couche d’analyse du contexte social en prédisant les traits de personnalité des utilisateurs des RSO, améliorant la détection et les stratégies d’intervention précoce contre la désinformation.

ExFake (Explainable False Information Detection Based on Content, Context, and External Evidence) élargit encore le cadre, combinant l’analyse de contenu avec des insights du contexte social et des preuves externes. Il tire parti des données d'organisations de vérification des faits réputées et de comptes officiels, garantissant une approche plus complète et fiable de la détection de la désinformation. La méthodologie sophistiquée d'ExFake évalue non seulement le contenu des publications en ligne, mais prend également en compte le contexte plus large et corrobore les informations avec des sources externes crédibles, offrant ainsi une solution bien arrondie et robuste pour combattre les fausses informations dans les réseaux sociaux en ligne.

Complétant le cadre, AFCC (Automated Fact-checkers Consensus and Credibility) traite l'hétérogénéité des évaluations des différentes organisations de vérification des faits. Il standardise ces évaluations et évalue la crédibilité des sources, fournissant une évaluation unifiée et fiable de l'information.

Chaque système au sein du cadre FACTS-ON est rigoureusement évalué pour démontrer son efficacité dans la lutte contre la désinformation sur les RSO. Cette thèse détaille le développement, la mise en œuvre et l'évaluation complète de ces systèmes, soulignant leur contribution collective au domaine de la détection des fausses informations. La recherche ne met pas seulement en évidence les capacités actuelles dans la lutte contre la désinformation, mais prépare également le terrain pour de futures avancées dans ce domaine critique d'étude.

Mots clés : Réseaux Sociaux en Ligne, Fausses Informations, Fausses Nouvelles, Désinformation, Mésinformation, Intelligence Artificielle (IA), FACTS-ON, EXMULF, MythXpose, ExFake, AFCC, Analyse de Contenu Multimodal, IA Explicable, Vérification des Faits, Détection des Fausses Nouvelles.

Abstract

The rapid evolution of online social networks (OSN) presents a significant challenge in identifying and mitigating false information, which includes Fake News, Disinformation, and Misinformation. This complexity is amplified in digital environments where information is quickly disseminated, requiring sophisticated strategies to differentiate between genuine and false content. One of the primary challenges in automatically detecting false information is its realistic presentation, often closely resembling verifiable facts. This poses considerable challenges for artificial intelligence (AI) systems, necessitating additional data from external sources, such as third-party verifications, to effectively discern the truth. Consequently, there is a continuous technological evolution to counter the growing sophistication of false information, challenging and advancing the capabilities of AI.

In response to these challenges, my dissertation introduces the FACTS-ON framework (Fighting Against Counterfeit Truths in Online Social Networks), a comprehensive and systematic approach to combat false information in OSNs. FACTS-ON integrates a series of advanced systems, each building upon the capabilities of its predecessor to enhance the overall strategy for detecting and mitigating false information. I begin by introducing the FACTS-ON framework, which sets the foundation for my solution, and then detail each system within the framework:

EXMULF (Explainable Multimodal Content-based Fake News Detection) focuses on analyzing both text and image in online content using advanced multimodal techniques, coupled with explainable AI to provide transparent and understandable assessments of false information.

Building upon EXMULF's foundation, MythXpose (Multimodal Content and Social Context-based System for Explainable False Information Detection with Personality Prediction) adds a layer of social context analysis by predicting the personality traits of OSN users, enhancing the detection and early intervention strategies against false information.

ExFake (Explainable False Information Detection Based on Content, Context, and External Evidence) further expands the framework, combining content analysis with insights from social context and external evidence. It leverages data from reputable fact-checking organizations and official social accounts, ensuring a more comprehensive and reliable approach to the detection of false information. ExFake's sophisticated methodology not only evaluates the content of online posts but also considers the broader context and corroborates information with external, credible sources, thereby offering a well-rounded and robust solution for combating false information in online social networks.

Completing the framework, AFCC (Automated Fact-checkers Consensus and Credibility) addresses the heterogeneity of ratings from various fact-checking organizations. It standardizes these ratings and assesses the credibility of the sources, providing a unified and trustworthy assessment of information.

Each system within the FACTS-ON framework is rigorously evaluated to demonstrate its effectiveness in combating false information on OSN. This dissertation details the development, implementation, and comprehensive evaluation of these systems, highlighting their collective contribution to the field of false information detection. The research not only showcases the current capabilities in addressing false information but also sets the stage for future advancements in this critical area of study.

Keywords: Online Social Networks, False Information, Fake News, Disinformation, Misinformation, Artificial Intelligence (AI), FACTS-ON Framework, EXMULF, MythXpose, ExFake, AFCC, Multimodal Content Analysis, Explainable AI, Fact-Checking, Fake News Detection.

Contents

List of Tables

List of Figures

List of Acronyms and Abbreviations

AFCC	Automatic Fact-checkers' Consensus and Credibility Assessment System
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
ExFake	Explainable False Information Detection Based on Content, Context, and External Evidence
EXMULF	EXplainable MULtimodal Content-based Fake news Detection System
FACTS-ON	Fighting Against Counterfeit Truths in Online Social Networks
IFCN	International Fact-Checking Network
LDA	Latent Dirichlet Allocation
LIME	Local Interpretable Model-agnostic Explanations

MythXpose	Multimodal Content and Social Context-based System for Explainable False Information Detection with Personality Prediction
NER	Named Entity Recognition
NLI	Natural Language Inference
NLP	Natural Language Processing
OSN	Online Social Networks
SBERT	Sentence-BERT
VilBERT	Vision-and-Language BERT
XAI	Explainable Artificial Intelligence

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude envers ma directrice de thèse, Esma Aïmeur, pour son soutien indéfectible, son expertise et son dévouement tout au long de ce travail. Je suis également reconnaissante envers les membres de Jury, Professeur Margarida Carvalho, Professeur Bang Liu, Professeur Kimiz Dalkir pour le temps et les efforts qu'ils ont consacrés à juger mon travail.

Je tiens à exprimer ma profonde gratitude envers mon mari Ezaki Ibrahmi, qui a été une source de soutien et d'encouragement tout au long de mes études doctorales. Sa patience, son soutien et son amour inconditionnel ont été une source de force pour moi dans les moments difficiles. Il a été un véritable partenaire dans cette aventure, m'aidant à équilibrer mes responsabilités professionnelles et personnelles, et me soutenant dans toutes les étapes de cette recherche.

Je tiens à le remercier du fond du cœur pour tout ce qu'il a fait pour moi et pour notre famille tout au long de cette aventure. Je ne pourrais pas avoir accompli cela sans son amour et son soutien constants.

Un merci tout spécial à mon fils Ilyes, petit rayon de soleil de deux ans, qui, avec ses sourires et ses câlins, a apporté de la joie et de la légèreté dans les moments les plus chargés. Son innocence et sa joie de vivre m'ont souvent rappelé l'importance de l'équilibre entre le travail et la vie de famille. Ilyes, tu es une source d'inspiration et de bonheur inestimable dans ma vie.

Merci Zaki, je t'aime.

Je voudrais également remercier ma famille et mes amis pour leur soutien et leur encouragement constants tout au long de mes études. Leurs encouragements ont été un moteur pour moi dans les moments difficiles.

Enfin, je tiens à remercier tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Merci infiniment à tous et à toutes.

Introduction

Research context

In the age of digital communication and the rapid dissemination of information, online social networks (OSN) have become pervasive platforms for sharing news and information. However, such widespread access to information has also given rise to a pressing challenge. Namely, the proliferation of *counterfeit truths*, including fake news, misinformation, and disinformation. These misleading and false narratives have the potential to cause significant societal, political, and economic disruptions. Addressing this challenge is not only an academic task but also a vital need to preserve the integrity of our information ecosystem.

Counterfeit truths refer to information that is deliberately fabricated or manipulated to mimic the appearance of truth, despite being false or misleading. This concept includes a spectrum of deceptive information practices, from fabricated stories and manipulated media to intentionally misleading narratives. The proliferation of such counterfeit truths can have far-reaching consequences, impacting public opinion, influencing elections, and undermining trust in reputable sources of information.

Before delving into the specifics of this challenge, it is important to clarify a key aspect of the terminology used in this dissertation. Throughout the document, I will use the term "Twitter" to refer to what is currently known as ‘X’. This decision is made for clarity and ease of comprehension, as the term "Twitter" is more familiar to a wider audience and is historically associated with the platform. All mentions of "Twitter" should thus be understood as references to the platform currently called ‘X’. Additionally, the terms "fake news" and "false information" will be used interchangeably, as "fake news" is more commonly recognized, but "false information" is a broader term encompassing fake news, misinformation, and disinformation. The common attribute of both "fake news" and "false information" is their lack of authenticity.

The phenomenon of fake news, misinformation, and disinformation encompasses a spectrum of deceptive practices that range from fabricated stories and manipulated media to intentionally misleading narratives. The consequences of these practices can be far-reaching, impacting public opinion, influencing elections, and undermining trust in reputable sources of information. This multifaceted challenge demands a comprehensive and interdisciplinary approach that leverages both technological innovations and theoretical insights.

The primary goal of identifying fake news, regardless of whether it is mis or disinformation, is to ensure the dependability and trustworthiness of the information that is being circulated on online social networks. By actively addressing this issue, work can be advanced toward creating a more informed society that values truthfulness and accuracy in digital interactions. Consequently, my primary motivations can be outlined as follows:

- The identification of fake news on social networks burgeoning research area that is currently receiving considerable attention.
- The detection of fake news on social media is still in its early stages, and many challenging issues require more thorough investigation.
- It is essential to explore potential research directions that can enhance the detection and mitigation of fake news.
- The dynamic nature of fake news propagation through social networks further complicates matters. False information can rapidly spread and affect a large number of users in a short time.

Problem statement

Fake news, disinformation and misinformation have become such a scourge that Marcia McNutt, president of the National Academy of Sciences of the United States, is quoted to have said (making an implicit reference to the Covid-19 pandemic) "Misinformation is worse than an epidemic: It spreads at the speed of light throughout the globe, and can prove deadly when it reinforces misplaced personal bias against all trustworthy evidence" in a joint statement of the National Academies[1] posted on 15 July 2021. Indeed, although online social networks, also called social media, have improved the ease with which real-time information is broadcast, its popularity and its massive use have expanded the spread of fake news by increasing the speed and scope at which it can spread. Fake news may refer to the manipulation of information that can be carried out through the production of false information, or the distortion of true information. However, that does not mean that this problem is only created with social media. A long time ago there were rumours in the

[1] `https://www.nationalacademies.org/news/2021/07/as-surgeon-general-urges-whole-of-society-effort-to-fight-health-misinformation-the-work-of-the-national-academies-helps-foster-an-evidence-based-information-environment`, last access date: 30-12-2023.

traditional media that Elvis was not dead[2], that the Earth was flat[3], that aliens had invaded us[4], etc.

Therefore, social media has become nowadays a powerful source for fake news dissemination [242, 254]. According to Pew Research Center's analysis of news use across social media platforms, in 2020 about half of American adults get news on social media at least sometimes[5], while in 2018 only one-fifth of them say they often get news via social media[6].

Hence, fake news can have a significant impact on society as manipulated and false content is easier to generate and harder to detect [147] and as disinformation actors change their tactics [147, 178]. In 2017, Snow predicted in the MIT *Technology Review* [265] that most individuals in mature economies will consume more false than valid information by 2022.

Recent news on the Covid-19 pandemic, which has flooded the web and created panic in many countries, has been reported as fake[7]. For example, holding your breath for ten seconds to one minute is not a self-test for Covid-19[8] (see Figure 1). Similarly, online posts claiming to reveal various "cures" for Covid-19 such as eating boiled garlic or drinking chlorine dioxide (which is an industrial bleach), were verified[9] as fake and in some cases as dangerous and will never cure the infection.

Social media outperformed television as the major news source for young people of the UK and US[10]. Moreover, as it is easier to generate and disseminate news online than with traditional media or face to face, large volumes of fake news are produced online for many reasons [254]. Furthermore, it has been reported in a previous study about the spread of online news on Twitter [294] that the spread of false news online is six times faster than truthful content and that 70% of the users could not distinguish real from fake news [294] due to the attraction of the novelty of the latter [41]. It was determined that falsehood spreads significantly farther, faster, deeper and more broadly than the truth in all categories of information, and the effects are more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information [294].

[2] `https://time.com/4897819/elvis-presley-alive-conspiracy-theories/`, last access date: 30-12-2023.

[3] `https://www.therichest.com/shocking/the-evidence-15-reasons-people-think-the-earth-is-flat/`, last access date: 30-12-2023.

[4] `https://www.grunge.com/657584/the-truth-about-1952s-alien-invasion-of-washington-dc/`, last access date: 30-12-2023.

[5] `https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/`, last access date: 30-12-2023.

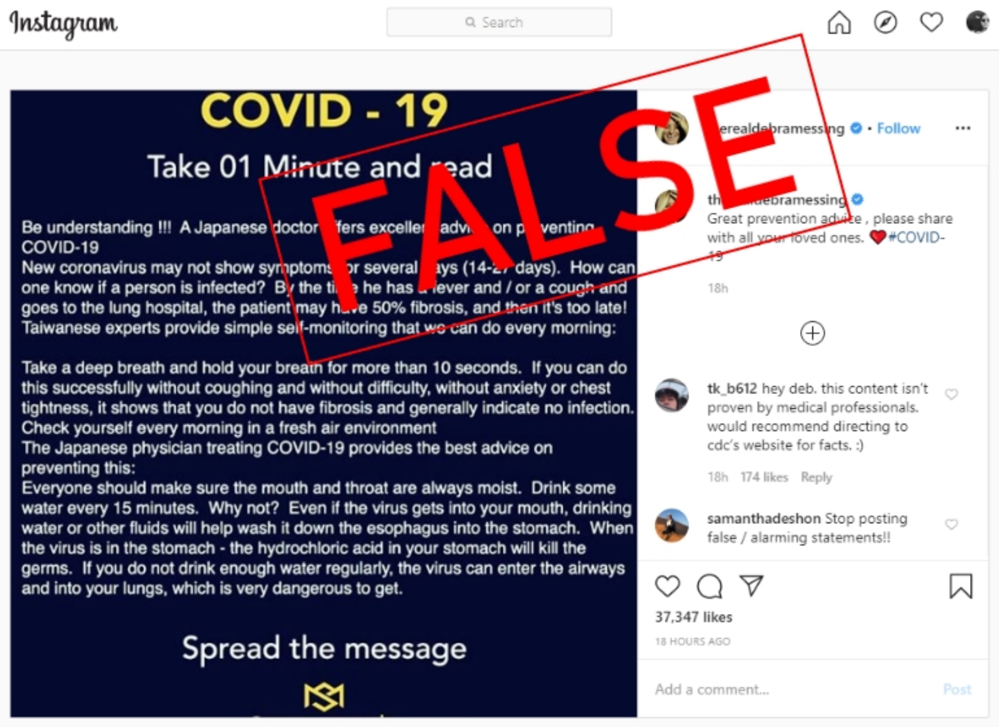
[6] `https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/`, last access date: 30-12-2023.

[7] `https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-fake-news-disinformation-rumours-hoaxes`, last access date: 30-12-2023.

[8] `https://www.factcheck.org/2020/03/viral-social-media-posts-offer-false-coronavirus-tips/`, last access date: 30-12-2023.

[9] `https://www.factcheck.org/2020/02/fake-coronavirus-cures-part-2-garlic-isnt-a-cure/`, last access date: 30-12-2023.

[10] `https://www.bbc.com/news/uk-36528256`, last access date: 30-12-2023.



FALSE

COVID - 19

Take 01 Minute and read

Be understanding !!! A Japanese doctor offers excellent advice on preventing COVID-19
New coronavirus may not show symptoms for several days (14-27 days). How can one know if a person is infected? By the time he has a fever and / or a cough and goes to the lung hospital, the patient may have 50% fibrosis, and then it's too late! Taiwanese experts provide simple self-monitoring that we can do every morning:

Take a deep breath and hold your breath for more than 10 seconds. If you can do this successfully without coughing and without difficulty, without anxiety or chest tightness, it shows that you do not have fibrosis and generally indicate no infection. Check yourself every morning in a fresh air environment
The Japanese physician treating COVID-19 provides the best advice on preventing this:
Everyone should make sure the mouth and throat are always moist. Drink some water every 15 minutes. Why not? Even if the virus gets into your mouth, drinking water or other fluids will help wash it down the esophagus into the stomach. When the virus is in the stomach - the hydrochloric acid in your stomach will kill the germs. If you do not drink enough water regularly, the virus can enter the airways and into your lungs, which is very dangerous to get.

Spread the message

therealdebramessing · Follow

therealdebramessing Great prevention advice , please share with all your loved ones. ❤#COVID-19

18h

tk_b612 hey deb. this content isn't proven by medical professionals. would recommend directing to cdc's website for facts. :)

18h 174 likes Reply

samanthadeshon Stop posting false / alarming statements!!

37,347 likes

18 HOURS AGO

Figure 1 – Fake news example about a self-test for Covid-19
source: `https://cdn.factcheck.org/UploadedFiles/Screenshot031120_false.jpg`, last access date: 30-12-2023.

Over 1 million tweets were estimated to be related to fake news by the end of the 2016 US presidential election [11]. In 2017 in Germany a government spokesman affirmed: "We are dealing with a phenomenon of a dimension that we have not seen before", referring to an unprecedented spread of fake news on social networks [12]. Given the strength of this new phenomenon, fake news has been chosen as the word of the year by the Macquarie dictionary both in 2016 [13] and in 2018 [14] as well as by the Collins dictionary in 2017 [15, 16]. Since 2020, the new term "infodemic" was coined, reflecting widespread researchers' concern [18, 107, 116, 178, 243] about the proliferation of misinformation linked to the Covid-19 pandemic.

The Gartner Group's top strategic predictions for 2018 and beyond included the need for IT leaders to quickly develop Artificial Intelligence (AI) algorithms to address counterfeit

[11] `https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory`, last access date: 30-12-2023.

[12] `https://www.theguardian.com/world/2017/jan/09/germany-investigating-spread-fake-news-online-russia-election`, last access date: 30-12-2023.

[13] `https://www.macquariedictionary.com.au/resources/view/word/of/the/year/2016`, last access date: 30-12-2023.

[14] `https://www.macquariedictionary.com.au/resources/view/word/of/the/year/2018`, last access date: 30-12-2023.

[15] `https://apnews.com/article/47466c5e260149b1a23641b9e319fda6`, last access date: 30-12-2023.

[16] `https://blog.collinsdictionary.com/language-lovers/collins-2017-word-of-the-year-shortlist/`, last access date: 30-12-2023.

reality and fake news [17]. However, fake news identification is a complex issue. Snow [265] questioned the ability of AI to win the war against fake news. Similarly, other researchers concurred that even the best AI for spotting fake news is still ineffective [18]. Besides, recent studies have shown that the power of AI algorithms for identifying fake news is lower than its ability to create it [204].

Consequently, automatic fake news detection remains a huge challenge, primarily because the content is designed to closely resemble the truth in order to deceive users, and as a result, it is often hard to determine its veracity by AI alone. Therefore, it is crucial to consider more effective approaches to solve the problem of fake news in social media.

Research objectives and contributions

The goal of this dissertation is to contribute to the development of effective strategies for combating fake news, misinformation, and disinformation on online social networks. My research is rooted in the integration of content, context, and external evidence analyses, complemented by advanced explainability techniques. These contributions are derived from a series of research papers and a submitted work that collectively delve into the multifaceted nature of the fake news landscape.

This dissertation seeks to achieve the following key objectives:

(1) Comprehensive Understanding of the Field: Conduct a thorough systematic review of existing literature on fake news, disinformation, and misinformation to establish a foundational understanding of the landscape.

(2) Multimodal Content Analysis: Develop a multimodal content-based fake news detection system, leveraging a range of textual and visual (i.e., image) features to enhance accuracy.

(3) Social Context Integration: Investigate the role of social context in the propagation of false information and design a framework that integrates contextual cues for improved detection. This includes considering contextual cues available prior to the spread of fake content, such as Online Social Network (OSN) users, who are key entities in OSNs. Particularly, their past sharing behaviour, personality traits, as well as date and time, are crucial social context-based information available in the early stages of online content dissemination.

(4) External Evidence Integration: Leverage trusted external entities, such as established fact-checking organizations and official sources (i.e., official social accounts), to assess content credibility.

[17] https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond/, last access date: 30-12-2023.

[18] https://www.technologyreview.com/s/612236/even-the-best-ai-for-spotting-fake-news-is-still-terrible/, last access date: 30-12-2023.

(5) Explainable Detection: Integrate explainability techniques to provide transparent and interpretable insights into the decision-making processes of fake news detection models.

(6) Consensus Inference: Propose a novel approach for automatic fact-checkers' consensus inference and credibility assessment to analyze and unify fact-checkers' diverse rating labels and decisions, and assess their credibility. This approach not only consolidates varying evaluations for a unified news credibility consensus but also computes the credibility of the fact-checkers themselves, thereby enhancing trust-based fake news detection.

Addressing these challenges requires innovative and interdisciplinary approaches that integrate advanced techniques with theoretical insights. This dissertation embarks on a comprehensive exploration of combating fake news and misinformation, presenting a series of contributions that collectively emphasize the power of a multifaceted, multimodal, and explainable approach.

The multifaceted nature of fake news necessitates a nuanced understanding of its different forms, propagation mechanisms, and sociopolitical implications. My initial contribution, "Fake News, Disinformation and Misinformation in Social Media: A Review", surveys the existing literature to elucidate the intricate distinctions between these categories of false information. This review provides a solid foundation for subsequent research, highlighting the need for comprehensive detection methods that account for the various aspects of false information.

"The Scourge of Online Deception in Social Networks" delves deeper into the landscape of online deception, analyzing the factors that contribute to the proliferation of false information. Through a meticulous examination of psychological and sociotechnical mechanisms, this work underscores the urgency of developing strategies that encompass both content and context dimensions.

Building upon these insights, I introduce "EXMULF: an EXplainable MUltimodal content-based Fake news detection system." This contribution leverages the power of multimodal analysis, incorporating textual and visual features to enhance the accuracy of fake news detection. Notably, the system integrates explainability techniques to provide transparent insights into the decision-making process, fostering user trust and understanding.

Continuously advancing on this trajectory, "MythXpose: Multimodal Content and Social Context-based System for Explainable False Information Detection with Personality Prediction" extends and refines the capabilities established by EXMULF. It strategically combines content analysis with social context-based information, thereby synergistically enhancing the ability to discern deceptive content in the context of users' online behaviours and personality traits. Importantly, MythXpose ensures transparency through the integration of explainability mechanisms.

Further extending my research, "ExFake: Towards an Explainable Fake News Detection Based on Content, Social Context and External Evidence Information" integrates external evidence and social context cues with content-based detection. This approach goes beyond analyzing textual content alone and takes into account the contextual factors surrounding information dissemination. It also integrates evidence from trusted fact-checkers and official sources (i.e., official social accounts), offering a comprehensive perspective on fake news detection. Explainability mechanisms are also employed, ensuring that detection outcomes remain interpretable and accessible to users.

Concluding my research, I present "AFCC: Towards an Automatic Fact-Checkers' Consensus Inference and Credibility Assessment for Trust-based Fake News Detection." This innovative approach harnesses consensus inference and credibility assessment to establish a trust-based framework for detecting fake news. By aggregating automated fact-checkers' judgments, AFCC can enhance detection outcomes' reliability, contributing to a more trustworthy information environment.

Dissertation outline

The remainder of this dissertation is organized as follows:

— Chapter 1 "Fake News, Disinformation and Misinformation in Social Media: Related Concepts and Challenges Review": In this chapter, I delve into the intricate web of Fake News, Disinformation, and Misinformation within the realm of social media. I comprehensively review the related concepts and challenges that characterize this phenomenon. By exploring the nuances of these terms and understanding their distinct implications, I lay a solid foundation for the subsequent chapters that delve into detection methods and frameworks.

— Chapter 2: "Fake News, Disinformation and Misinformation in Social Media: Detection Methods and Used Techniques Review": Building upon the conceptual groundwork established in the previous chapter, I embark on an exploration of various detection methods and techniques employed to counteract fake news, disinformation, and misinformation on social media platforms. This chapter delves into the arsenal of tools and methodologies that researchers and practitioners have employed to identify and mitigate the spread of deceptive content within the digital landscape.

— Chapter 3: "FACTS-ON in Theory: Combating False Information in Online Social Networks": At the heart of my dissertation lies the innovative FACTS-ON framework. In this chapter, I present the conceptual architecture that stands as the cornerstone of my approach to tackling the challenges posed by deceptive content. I elucidate the intricacies of this framework, detailing its modules, functionalities, and the rationale behind its design. Through this chapter, readers gain an in-depth understanding of how FACTS-ON functions as a comprehensive solution to address the dissemination of counterfeit truths.

— Chapter 4: "Explainable Multimodal Content-based Fake News Detection: EXMULF": As I progress, I delve into the specifics of my approach with the EXMULF system. This chapter focuses on the explainable multimodal content-based fake news detection strategy. I unravel the mechanisms and techniques employed within EXMULF, providing a transparent and comprehensible overview of how it harnesses various modalities to identify misleading content while offering explanations for its decisions.

— Chapter 5: "Multimodal Content and Social Context-based System for Explainable False Information Detection with Personality Prediction: MythXpose": Continuing my exploration, this chapter delves into the intricacies of the MythXpose system, strategically combining a content-based module (i.e., EXMULF, introduced in Chapter 4) with social context-based information module (i.e., PERSONA: Personality-Based Evaluation for Reliable Social Online News Analysis). The PERSONA module, designed to assess the personality traits of OSN users, enriches the capability for explainable fake news detection by providing valuable insights into user behaviours and tendencies. This fusion enhances the understanding of deceptive content in the realm of social media.

— Chapter 6: "Explainable False Information Detection based on Content, Context and External Evidence: ExFake": Expanding the horizon of my explorations, this chapter introduces ExFake, a system that delves into fake news detection by harnessing the power of content, context, and external evidence. I dissect the intricacies of this method, showcasing how it amalgamates multiple dimensions of information to form a holistic understanding, which empowers the identification and interpretation of deceptive content.

— Chapter 7: "Fact-checkers' Consensus Inference and Credibility Assessment for Trust-based Fake News Detection: AFCC": Trust and credibility are pivotal in combating false information. In this chapter, I introduce AFCC, a framework that leverages the collective insight of fact-checkers for trust-based fake news detection. I delve into the mechanics of this approach, illustrating how it synthesizes multiple perspectives to offer a nuanced understanding of content authenticity and reliability.

— Chapter 8: "Conclusion and future work": In this chapter, I synthesize the knowledge garnered across the preceding chapters. In this concluding chapter, I reflect on the contributions made by my research and the impact it holds on the realm of fake news detection. Furthermore, I outline potential avenues for future exploration, signalling the continuous evolution of strategies to counter the dissemination of deceptive content.

List of publications

In this section, I present the list of the research articles produced as part of this dissertation.

Journal papers

(1) Esma Aïmeur, Sabrine Amri, and Gilles Brassard. **"Fake news, disinformation and misinformation in social media: a review."** Social Network Analysis and Mining 13, no. 1 (2023): 30. doi: 10.1007/s13278-023-01028-5.

(2) Sabrine Amri, Esma Aimeur. **"AFCC: Towards an Automatic Fact-checkers' Consensus Inference and Credibility Assessment for Trust-based Fake News Detection"** Submitted to the International Journal of Information Technology.

0.0.1. Conference papers

(1) Sabrine Amri, Henri-Cedric Mputu Boleilanga, Esma Aimeur. **"ExFake: Towards an Explainable Fake News Detection Based on Content and Social Context Information."** International Conference on Security & Management (SAM'23). Accepted.

(2) Sabrine Amri, Dorsaf Sallami, and Esma Aïmeur. **"Exmulf: an explainable multimodal content-based fake news detection system."** In International Symposium on Foundations and Practice of Security, pp. 177-187. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-031-08147-7_12.

(3) Esma Aïmeur, Hicham Hage and Sabrine Amri, **"The Scourge of Online Deception in Social Networks"**, 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 1266-1271, doi: 10.1109/CSCI46756.2018.00244.

Chapter 1

Fake News, Disinformation and Misinformation in Social Media: Related Concepts and Challenges Review

1.1. Introduction

This chapter focuses primarily on understanding the "fake news" problem, its related concepts, challenges, and root causes. The subsequent chapter will be dedicated to reviewing the state-of-the-art methods for automatic fake news detection and mitigation in online social networks, as addressed by researchers. The main contributions of this chapter are summarized below:

- Providing an overview of the general context from which the fake news problem emerged (i.e., online deception).
- Reviewing existing definitions of fake news, identifying the most commonly used terms and features for defining fake news, and categorizing related works accordingly.
- Proposing a classification of fake news typology based on various categorizations reported in the literature.
- Highlighting the most challenging factors that hinder researchers from proposing effective solutions for automatic fake news detection in social media.
- Presenting and categorizing representative studies in the domain of automatic fake news detection and mitigation on online social networks, including the key methods and techniques used to generate detection models.

1.2. Review methodology

This section introduces the systematic review methodology employed for this chapter and the subsequent one. It begins with the formulation of research questions, which guide the selection of relevant research literature. Subsequently, it outlines the various sources of information, along with the search and inclusion/exclusion criteria applied to select the final set of papers.

1.2.1. Research questions formulation

The research scope, research questions, and inclusion/exclusion criteria were established following an initial evaluation of the literature. The following research questions were formulated and addressed:

— RQ1: what is fake news in social media? how is it defined in the literature, and what are its related concepts and different types?

— RQ2: what are the existing challenges and issues related to fake news?

— RQ3: which will be addressed in the subsequent chapter: What are the available approaches and techniques used to perform fake news detection in social media?

1.2.2. Sources of information

A broad search was conducted for journal and conference research articles, books, and magazines as sources of data to extract relevant articles. The main sources of scientific databases and digital libraries were used in the search, such as Google Scholar[1], IEEE Xplore[2], Springer Link[3], ScienceDirect[4], Scopus[5], ACM Digital Library[6]. Additionally, most of the related high-profile conferences such as WWW, SIGKDD, VLDB, ICDE, and others were screened to identify recent work.

1.2.3. Search criteria

The research was focused over a period of ten years, ensuring that about two-thirds of the research papers considered were published in or after 2019. This approach was adopted to uncover the latest strategies for identifying counterfeit truths. By emphasizing recent literature, the research aligns with the latest challenges and innovations in the field, ensuring that the findings are relevant and contribute effectively to understanding and mitigating the impact of counterfeit truths in the digital age.

Additionally, a set of keywords was defined to search the aforementioned scientific databases, as the focus was on reviewing the current state-of-the-art in addition to the challenges and future directions. The set of keywords includes the following terms: fake news, disinformation, misinformation, information disorder, social media, detection techniques, detection methods, survey, and literature review.

[1] https://scholar.google.ca/, last access date: 30-12-2023.
[2] https://ieeexplore.ieee.org/, last access date: 30-12-2023.
[3] https://link.springer.com/, last access date: 30-12-2023.
[4] https://www.sciencedirect.com/, last access date: 30-12-2023.
[5] https://www.scopus.com/, last access date: 30-12-2023.
[6] https://www.acm.org/digital-library, last access date: 30-12-2023.

1.2.4. Study selection, exclusion and inclusion criteria

To retrieve relevant research articles, based on the identified sources of information and search criteria, a systematic keyword-based search was carried out by posing different search queries, as shown in Table 1.

Table 1 – List of keywords for searching relevant articles

Keywords
Fake news + social media
Fake news + disinformation
Fake news + misinformation
Fake news + information disorder
Fake news + survey
Fake news + detection methods
Fake news + literature review
Fake news + detection techniques
Fake news + detection + social media
Disinformation + misinformation + social media

A primary list of articles was discovered. On the obtained initial list of studies, a set of inclusion/exclusion criteria presented in Table 2 was applied to select the appropriate research papers. The inclusion and exclusion principles were applied to determine whether a study should be included or not.

After reading the abstracts, some articles that did not meet the criteria were excluded. The most significant research was chosen to aid in understanding the field. Upon a complete review of the articles, 68 research papers that discuss the definition of the term fake news and its related concepts were found (see Table 4). The remaining papers were used to understand the field, reveal the challenges, review the detection techniques, and discuss future directions.

1.3. A brief introduction of online deception

The Cambridge Online Dictionary defines Deception as "*the act of hiding the truth, especially to get an advantage*". Deception relies on peoples' trust, doubt and strong emotions that may prevent them from thinking and acting clearly [6]. It is also defined in previous work [6] as the process that undermines the ability to consciously make decisions and take convenient actions, following personal values and boundaries. In other words, deception gets

Table 2 – Inclusion and exclusion criteria

Inclusion criterion	**Exclusion criterion**
Peer-reviewed and written in the English language.	Articles in a different language than English.
Clearly describes fake news, misinformation, and disinformation problems in social networks.	Does not focus on fake news, misinformation, or disinformation problem in social networks.
Written by academic or industrial researchers. High number of citations. Recent articles only (last ten years).	Short papers, posters, or similar.
In the case of equivalent studies, the one published in the highest-rated journal or conference is selected to sustain a high-quality set of articles on which the review is conducted. Articles that propose methodologies, methods, or approaches for fake news detection online social networks.	Articles not following these inclusion criteria.

people to do things they would not otherwise do. In the context of online deception, several factors need to be considered: the deceiver, the purpose or aim of the deception, the social media service, the deception technique and the potential target [6, 110].

Researchers are working on developing new ways to protect users and prevent online deception [6]. Due to the sophistication of attacks, this is a complex task. Hence, malicious attackers are using more complex tools and strategies to deceive users. Furthermore, the way information is organized and exchanged in social media may lead to exposing OSN users to many risks [5].

In fact, this field is one of the recent research areas that need collaborative efforts of multidisciplinary practices such as psychology, sociology, journalism, computer science as well as cyber-security and digital marketing (which are not yet well explored in the field of dis/mis/mal-information but relevant for future research). Moreover, Ismailov et al. [126] analyzed the main causes that could be responsible for the efficiency gap between lab results and real-world implementations.

In this dissertation, reviewing the state-of-the-art in online deception is not within the scope of work. However, I think it is crucial to note that fake news, misinformation and disinformation are indeed parts of the larger landscape of online deception [110].

1.4. Fake news, the modern-day problem

Fake news have existed for a very long time, much before its wide circulation became facilitated by the invention of the printing press [7]. For instance, Socrates was condemned to death more than twenty-five hundred years ago under the fake news that he was guilty of impiety against the pantheon of Athens and corruption of the youth [8]. A Google Trends Analysis of the term “fake news” reveals an explosion in popularity around the time of the 2016 US presidential election [9]. Fake news detection is a problem that has recently been addressed by numerous organizations, including the European Union [10] and NATO [11].

In this section, an overview of the fake news definitions as provided in the literature is first presented. The terms and features used in these definitions are identified, and the definitions are then classified based on them. Following this, a fake news typology based on distinct categorizations is proposed, and the most cited forms of one specific fake news category (i.e., the intent-based fake news category) are defined and compared.

1.4.1. Definitions of fake news

"Fake news" is defined in the Collins English Dictionary as false and often sensational information disseminated under the guise of news reporting [12], yet the term has evolved over time and has become synonymous with the spread of false information [63].

The first definition of the term *fake news* was provided by Allcott and Gentzkow [9] as news articles that are intentionally and verifiably false and could mislead readers. Then, other definitions were provided in the literature, but they all agree on the *authenticity* of fake news to be false (i.e., being non-factual). However, they disagree on the inclusion and exclusion of some related concepts such as *satire*, *rumours*, *conspiracy theories*, *misinformation* and *hoaxes* from the given definition. More recently, Nakov [191] reported that the term fake news started to mean different things to different people, and for some politicians, it even means "news that I do not like".

Hence, there is still no agreed definition of the term "fake news". Moreover, many terms and concepts in the literature refer to fake news [3, 9, 11, 25, 44, 53, 62, 80, 103, 132, 144, 153, 187, 189, 191, 212, 229, 242, 254, 255, 271, 285, 308, 328, 332], disinformation [33, 124, 135, 138, 148, 168, 248, 255, 267], misinformation [178, 209, 239, 241, 255, 310], malinformation [52,

[7] https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535, last access date: 30-12-2023.

[8] https://en.wikipedia.org/wiki/Trial_of_Socrates, last access date: 30-12-2023.

[9] https://trends.google.com/trends/explore?hl=en-US&tz=-180&date=2013-12-06+2018-01-06&geo=US&q=fake+news&sni=3, last access date: 30-12-2023.

[10] https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation, last access date: 30-12-2023.

[11] https://www.nato.int/cps/en/natohq/177273.htm, last access date: 30-12-2023.

[12] https://www.collinsdictionary.com/dictionary/english/fake-news, last access date: 30-12-2023.

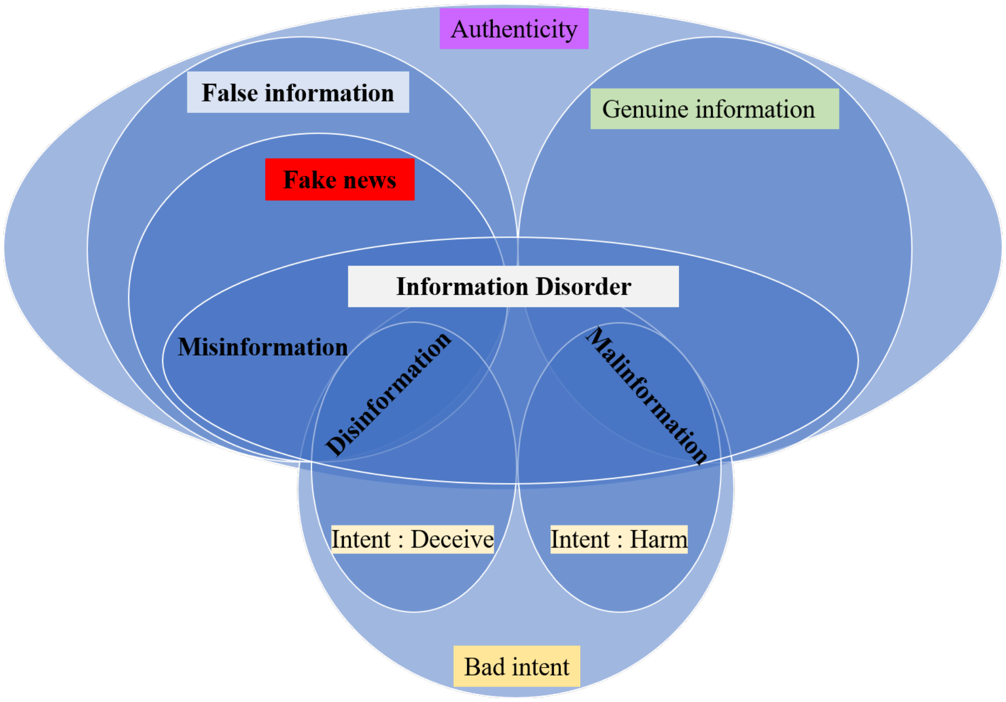


Figure 1 – Modelling of the relationship between terms related to fake news

65, 255], false information [105, 109, 147], information disorder [71, 255, 306, 307], information warfare [103] and information pollution [174].

There is also a remarkable amount of disagreement over the classification of the term fake news in the research literature as well as in policy [66, 82, 83]. Some consider fake news as a type of misinformation [11, 53, 75, 108, 125, 144, 209, 212, 241, 242, 262], others consider it as a type of disinformation [28, 29, 45, 66, 80, 141, 191, 254, 255, 271, 279], while others associate the term with both disinformation and misinformation [9, 52, 65, 103, 114, 212, 271, 308, 313, 328]. Alternatively, some prefer to differentiate fake news from both terms [25, 44, 82, 83, 132, 187, 229, 332].

The existing terms can be separated into two groups. The first group represents the general terms, which are *information disorder*, *false information* and *fake news*, each of which includes a subset of terms from the second group. The second group represents the elementary terms, which are *misinformation*, *disinformation* and *malinformation*. The literature agrees on the definitions of the latter group, but there is still no agreed-upon definition of the first group. In Figure 1 the relationship between the most used terms in the literature is modelled.

The terms most used in the literature to refer to, categorize and classify fake news can be summarized and defined as shown in Table 3. This table captures the similarities and shows the differences between the different terms based on two common key features, which are the intent and the authenticity of the news content. The intent feature refers to the intention behind the term that is used (i.e., whether or not the purpose is to mislead or cause harm), whereas the authenticity feature refers to its factual aspect. (i.e., whether the content is verifiably false or not, which is labelled as genuine in the second case). Some of these terms are explicitly used to refer to fake news (i.e., disinformation, misinformation and false information) while others are not (i.e., malinformation). In the comparison table, the empty grey cell denotes that the classification does not apply.

Table 3 – A comparison between used terms based on intent and authenticity

Term	**Definition**	**Intent**	**Authenticity**
False information	Verifiably false information.	▨	False
Misinformation	False information that is shared without the intention to mislead or to cause harm.	Not to mislead.	False
Disinformation	False information that is shared to intentionally mislead.	To mislead.	False
Malinformation	Genuine information that is shared with an intent to cause harm.	To cause harm.	Genuine

In Figure 2, the different features used in the literature to define fake news (i.e., intent, authenticity and knowledge) are identified. Hence, some definitions are based on two key features, which are *authenticity and intent* (i.e., news articles that are intentionally and verifiably false and could mislead readers). However, other definitions are based on either authenticity *or* intent. Other researchers categorize false information on the web and social media based on its intent and *knowledge* (i.e., when there is a single ground truth). In Table 4, the existing fake news definitions are classified based on the used *term* and the used *features*. In the classification, the references in the cells refer to the research study in which a fake news definition was provided, while the empty grey cells denote that the classification does not apply.

1.4.2. Fake news typology

Various categorizations of fake news have been provided in the literature. Two major categories of fake news can be distinguished based on the studied perspective (i.e., intention or content) as shown in Figure 3. However, the proposed fake news typology is not about detection methods and it is not exclusive. Hence, a given category of fake news can be described based on both perspectives (i.e., intention and content) at the same time. For instance, satire, which falls under intent-based fake news, can contain text and/or multimedia content types of data (e.g., headline, body, image, video), aligning it with content-based fake news as well, and so on.

Table 4 – Classification of fake news definitions based on the used term and features

	Fake news	Misinformation	Disinformation	False information	Malinformation	Information disorder
Intent and authenticity	Shu et al. [254], Sharma et al. [242], Mustafaraj and Metaxas [189], Klein and Wueller [144], Potthast et al. [212], Allcott and Gentzkow [9], Zhou and Zafarani [332], Zhang and Ghorbani [328], Conroy et al. [62], Celliers and Hattingh [53], Nakov [191], Shu et al. [255], Tandoc et al. [271], Abu Arqoub et al.[3], Molina et al.[187], de Cock Buning [66], Meel et al.[174]	Wu et al. [310], Shu et al. [255], Islam et al. [125], Hameleers et al.[114]	Kapantai et al. [138], Shu et al. [248], Shu et al. [255], Kumar et al. [148], Jungherr and Schroeder [135], Starbird et al. [267], de Cock Buning [66], Bastick [29], bringula et al.[45], Tsang[279], Hameleers et al.[114], Wu et al.[313]		Shu et al. [255], Di Domenico et al.[75], Dame[65]	Wardle and Derakhshan [307], Wardle [306], Derakhshan and Wardle [71], Shu et al. [255]
Intent or authenticity	Jin et al. [132], Rubin et al. [229], Balmas [25], Brewer et al. [44], Egelhofer and Lecheler [80], Lazer et al. [153], Allen et al. [11], Guadagno and Guttieri [103], van der Linden et al.[285], ERGA[82]	Pennycook and Rand [209], Shao et al. [239], Shao et al. [241], Micallef et al. [178], Ha et al.[108], Singh et al.[262], Wu et al.[313]	Marsden et al. [168], Ireton and Posetti [124], ERGA[83], Baptista et al.[28]	Habib et al. [109]	Carmi et al. [52]	
Intent and knowledge	Weiss et al. [308]		Bhattacharjee et al. [33], Khan et al.[141]	Kumar and Shah [147], Guo et al. [105]		

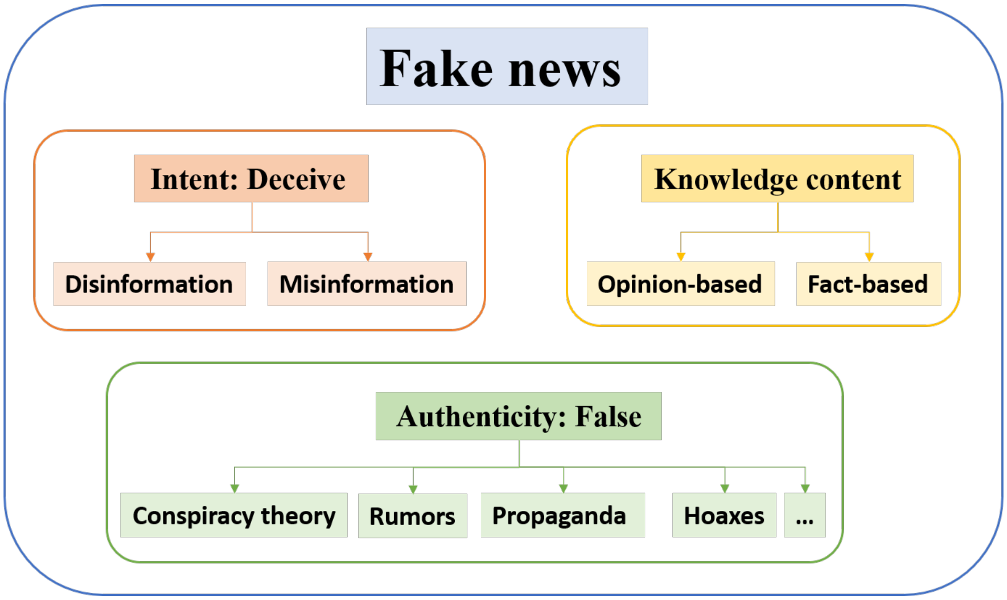


Figure 2 – The features used for fake news definition

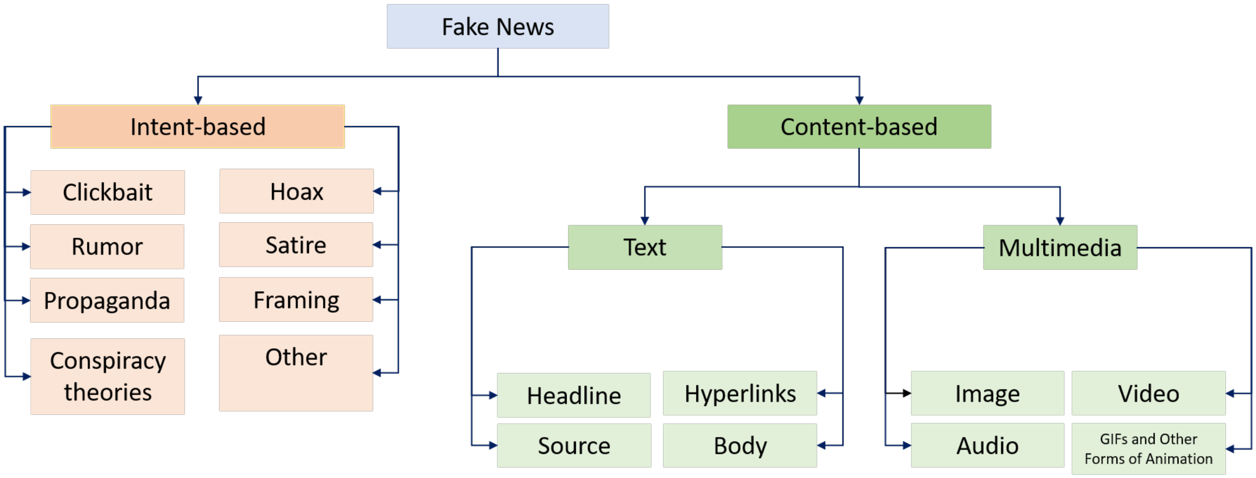


Figure 3 – Fake news typology

Most researchers classify fake news based on the intent [39, 61, 147, 148, 254, 305, 322] (see Subsection 1.4.2.2). However, other researchers [24, 81, 89, 117, 169, 202, 317] focus on the content to categorize types of fake news through distinguishing the different formats and content types of data in the news (e.g., text and/or multimedia).

Recently, another classification was proposed by Zhang and Ghorbani [328]. It is based on the combination of content and intent to categorize fake news. They distinguish physical news content and non-physical news content from fake news. Physical content consists of the carriers and format of the news, and non-physical content consists of the opinions, emotions, attitudes and sentiments that the news creators want to express.

1.4.2.1. **Content-based fake news category:** According to researchers of this category [24, 81, 89, 117, 169, 202, 317], forms of fake news may include false text such as hyperlinks or embedded content; multimedia such as false videos [69], images [169, 244], audios [69] and so on. Moreover, there is also multimodal content [248] that is fake news articles and posts composed of multiple types of data combined together. For example, a fabricated image along with a text related to the image [248]. In this category of fake news forms, I can mention as examples deepfake videos [317] and GAN-generated fake images [329], which are artificial intelligence-based machine-generated fake content that is hard for unsophisticated social network users to identify.

The effects of these forms of fake news content vary on the credibility assessment as well as sharing intentions which influences the spread of fake news on OSN. For instance, people with little knowledge about the issue compared to those who are strongly concerned about the key issue of fake news tend to be easier to convince that the misleading or fake news is real, especially when shared via a video modality as compared to the text or the audio modality [69].

1.4.2.2. **Intent-based fake news category:** The most often mentioned and discussed forms of fake news according to researchers in this category include but are not restricted to *clickbait*, *hoax*, *rumour*, *satire*, *propaganda*, *framing*, *conspiracy theories* and others. In the following subsections, these types of fake news, as defined in the literature, are explained, and a brief comparison between them is undertaken, as depicted in Table 5. The following are the most cited forms of intent-based types of fake news and their comparison is based on what is suspected to be the most common criteria mentioned by researchers.

— **Clickbait.** Clickbait refers to misleading headlines and thumbnails of content on the web [322] that tend to be fake stories with catchy headlines aimed at enticing the reader to click on a link [61]. This type of fake news is considered to be the least severe type of false information because if a user reads/views the whole content, it is possible to distinguish if the headline and/or the thumbnail were misleading [322]. However, the goal behind using clickbait is to increase the traffic to a website [322].

— **Hoax.** A hoax is a false [333] or inaccurate [322] intentionally fabricated [61] news story used to masquerade the truth [333] and is presented as factual [322] to deceive the public or audiences [61]. This category is also known either as half-truth or factoid stories [322]. Popular examples of hoaxes are stories that report the false death of celebrities [322] and public figures [61]. Recently hoaxes about the Covid-19 have been circulating through social media.

— **Rumour.** The term rumour refers to ambiguous or never confirmed claims [322] that are disseminated with a lack of evidence to support them [242]. This kind of information is widely propagated on OSN [322]. However, they are not necessarily false and may turn out to be true [333]. Rumours originate from unverified sources but may be true or false or remain unresolved [333].

- **Satire.** Satire refers to stories that contain a lot of irony and humor [322]. It presents stories as news that might be factually incorrect, but the intent is not to deceive but rather to call out, ridicule, or expose behaviour that is shameful, corrupt, or otherwise "bad" [97]. This is done with a fabricated story or by exaggerating the truth reported in mainstream media in the form of comedy [61]. The intent behind satire seems kind of legitimate and many authors (such as Wardle [305]) do include satire as a type of fake news as there is no intention to cause harm but it has the potential to mislead or fool people.

  Also, Golbeck et al. [97] mention that there is a spectrum from fake to satirical news that they found to be exploited by many fake news sites. These sites used disclaimers at the bottom of their webpages to suggest they were "satirical" even when there was nothing satirical about their articles, to protect them from accusations of being fake. The difference with a satirical form of fake news is that the authors or the host present themselves as a comedian or as an entertainer rather than a journalist informing the public [61]. However, most audiences believed the information passed in this satirical form because the comedian usually projects news from mainstream media and frames them to suit their program [61].

- **Propaganda.** Propaganda refers to news stories created by political entities to mislead people. It is a special instance of fabricated stories that aim to harm the interests of a particular party and, typically, has a political context [322]. Propaganda was widely used during both World Wars [61] as well as during the Cold War [322]. It is a consequential type of false information as it can change the course of human history (e.g., by changing the outcome of an election) [322]. States are the main actors of propaganda. Recently, propaganda has been used by politicians and media organizations to support a certain position or view [61]. Online astroturfing can be an example of the tools used for the dissemination of propaganda. It is a covert manipulation of public opinion [206] that aims to make it seem that many people share the same opinion about something. Astroturfing can affect different domains of interest, based on which online astroturfing can be mainly divided into political astroturfing, corporate astroturfing and astroturfing in e-commerce or online services [165]. Propaganda types of fake news can be debunked with manual fact-based detection models such as the use of expert-based fact-checkers [61].

- **Framing.** Framing refers to employing some aspect of reality to make content more visible while the truth is concealed [61] to deceive and misguide readers. People will understand certain concepts based on the way they are coined and invented. An example of framing was provided by Collins et al. [61]: "suppose a leader X says "I will neutralize my opponent" simply meaning he will beat his opponent in a given election. Such a statement will be framed such as "leader X threatens to kill Y" and this framed statement provides a total misrepresentation of the original meaning.

- **Conspiracy theories.** Conspiracy theories refer to the belief that an event is the result of secret plots generated by powerful conspirators. Conspiracy belief refers to people's adoption and belief of conspiracy theories, and it is associated with psychological, political and social factors [78]. Conspiracy theories are widespread in contemporary democracies [269], and they have major consequences. For instance, lately and during the Covid-19 pandemic, conspiracy theories have been discussed from a public health perspective [12, 90, 176].

1.4.2.3. **Comparison between most popular intent-based types of fake news:** Following a review of the most popular intent-based types of fake news, a comparison is made as shown in Table 5. It is based on the most common criteria mentioned by researchers in their definitions. The criteria used for this comparison are listed below.

- the intent behind the news, which refers to whether a given news type was mainly created to intentionally deceive people or not (e.g., humor, irony, entertainment, etc.);
- the way that the news propagates through OSN, which determines the nature of the propagation of each type of fake news and this can be either fast or slow propagation;
- the severity of the impact of the news on OSN users, which refers to whether the public has been highly impacted by the given type of fake news; the mentioned impact of each fake news type is mainly the proportion of the negative impact;
- and the goal behind disseminating the news, which can be to gain popularity for a particular entity (e.g., political party), for profit (e.g., lucrative business), or other reasons such as humour and irony in the case of satire, spreading panic or anger, and manipulating the public in the case of hoaxes, made-up stories about a particular person or entity in the case of rumours, and misguiding readers in the case of framing.

However, the comparison provided in Table 5 is deduced from the studied research papers, it reflects my point of view and is not based on empirical data.

I suspect that the most dangerous types of fake news are the ones with high intention to deceive the public, fast propagation through social media, high negative impact on OSN users, and complicated hidden goals and agendas. However, while the other types of fake news are less dangerous, they should not be ignored.

Moreover, it is important to highlight that the existence of the overlap in the types of fake news mentioned above has been proven, thus it is possible to observe false information that may fall within multiple categories [322]. Here, two examples by Zannettou et al. [322] are provided to better understand possible overlaps: 1) a rumour may also use clickbait techniques to increase the audience that will read the story; and 2) propaganda stories, as a special instance of a framing story.

Table 5 – A comparison between the different types of intent-based fake news

	Intent to deceive	**Propagation**	**Negative Impact**	**Goal**
Clickbait	High	Slow	Low	Popularity, Profit
Hoax	High	Fast	Low	Other
Rumour	High	Fast	High	Other
Satire	Low	Slow	Low	Popularity, Other
Propaganda	High	Fast	High	Popularity
Framing	High	Fast	Low	Other
Conspiracy theory	High	Fast	High	Other

1.5. Challenges related to fake news detection and mitigation

To alleviate fake news and its threats, it is crucial to first identify and understand the factors involved that continue to challenge researchers. Thus, the main question is to explore and investigate the factors that make it easier to fall for manipulated information. Despite the tremendous progress made in alleviating some of the challenges in fake news detection [242, 248, 328, 332], much more work needs to be accomplished to address the problem effectively.

In this section, several open issues that make fake news detection in social media a challenging problem are discussed. These issues can be summarized as follows: content-based issues (i.e., deceptive content that resembles the truth very closely), contextual issues (i.e., lack of user awareness, social bots spreaders of fake content, and OSN's dynamic natures that leads to the fast propagation) as well as the issue of existing datasets (i.e., there still no one size fits all benchmark dataset for fake news detection). These various aspects have been proven [254] to have a great impact on the accuracy of fake news detection approaches.

1.5.1. Content-based issue, deceptive content

Automatic fake news detection remains a huge challenge, primarily because the content is designed in a way that it closely resembles the truth. Besides, most deceivers choose their words carefully and use their language strategically to avoid being caught. Therefore, it is often hard to determine its veracity by AI without the reliance on additional information from third parties such as fact-checkers.

Abdullah-All-Tanvir et al. [2] reported that fake news tends to have more complicated stories and hardly ever make any references. It is more likely to contain a greater number of words that express negative emotions. This makes it so complicated that it becomes impossible for a human to manually detect the credibility of this content. Therefore, detecting fake news on social media is quite challenging. Moreover, fake news appears in multiple types and forms, which makes it hard and challenging to define a single global solution able to capture and deal with the disseminated content. Consequently, detecting false information is not a straightforward task due to its various types and forms [322].

1.5.2. Contextual issues

Contextual issues are challenges that are suspected to not be related to the content of the news but rather are inferred from the context of the online news post (i.e., humans are the weakest factor due to lack of user awareness, social bots spreaders, dynamic nature of online social platforms and fast propagation of fake news).

1.5.2.1. **Humans are the weakest factor due to the lack of awareness:** Recent statistics [13] show that the percentage of unintentional fake news spreaders (people who share fake news without the intention to mislead) over social media is five times higher than intentional spreaders. Moreover, another recent statistic [14] shows that the percentage of people who were confident about their ability to discern fact from fiction is ten times higher than those who were not confident about the truthfulness of what they were sharing. As a result, a lack of human awareness about the ascent of fake news can be deduced.

Public susceptibility and lack of user awareness [242] have always been the most challenging problems when dealing with fake news and misinformation. This is a complex issue because many people believe almost everything on the Internet and the ones who are new to digital technology or have less expertise may be easily fooled [79].

Moreover, it has been widely proven [79, 177] that people are often motivated to support and accept information that goes with their preexisting viewpoints and beliefs, and reject information that does not fit in as well. Hence, Shu et al. [254] illustrate an interesting correlation between fake news spread and psychological and cognitive theories. They further suggest that humans are more likely to believe information that confirms their existing views

[13] https://www.statista.com/statistics/657111/fake-news-sharing-online/, last access date: 30-12-2023.

[14] https://www.statista.com/statistics/657090/fake-news-recogition-confidence/, last access date: 30-12-2023.

and ideological beliefs. Consequently, they deduce that humans are naturally not very good at differentiating real information from fake information.

Recent research by Giachanou et al. [94] studies the role of personality and linguistic patterns in discriminating between fake news spreaders and fact-checkers. They classify a user as a potential fact-checker or a potential fake news spreader based on features that represent users' personality traits and linguistic patterns used in their tweets. They show that leveraging personality traits and linguistic patterns can improve the performance in differentiating between checkers and spreaders.

Furthermore, several researchers studied the prevalence of fake news on social networks during [9, 27, 102, 104] and after [93] the 2016 US presidential election and found that individuals most likely to engage with fake news sources were generally conservative-leaning, older, and highly engaged with political news.

Metzger et al. [177] examine how individuals evaluate the credibility of biased news sources and stories. They investigate the role of both cognitive dissonance and credibility perceptions in selective exposure to attitude-consistent news information. They found that online news consumers tend to perceive attitude-consistent news stories as more accurate and more credible than attitude-inconsistent stories.

Similarly, Edgerly et al. [79] explore the impact of news headlines on the audience's intent to verify whether given news is true or false. They concluded that participants exhibit higher intent to verify the news only when they believe the headline to be true, which is predicted by perceived congruence with preexisting ideological tendencies.

Luo et al. [161] evaluate the effects of endorsement cues in social media on message credibility and detection accuracy. Results showed that headlines associated with a high number of likes increased credibility, thereby enhancing detection accuracy for real news but undermining accuracy for fake news. Consequently, they highlight the urgency of empowering individuals to assess both news veracity and endorsement cues appropriately on social media.

Moreover, misinformed people are a greater problem than uninformed people [146], because the former hold inaccurate opinions (which may concern politics, climate change, and medicine) that are harder to correct. Indeed, people find it difficult to update their misinformation-based beliefs even after they have been proven to be false [88]. Moreover, even if a person has accepted the corrected information, his/her belief may still affect their opinion [196].

Falling for disinformation may also be explained by a lack of critical thinking and the need for evidence that supports information [22, 290]. However, it is also possible that people choose misinformation because they engage in directionally motivated reasoning [22, 88]. Online clients are normally vulnerable and will, in general, perceive web-based networking media as reliable, as reported by Abdullah-All-Tanvir et al. [1], who propose to mechanize fake news recognition.

It is worth noting that in addition to bots causing the outpouring of the majority of the misrepresentations, specific individuals are also contributing a large share of this issue [1]. Furthermore, Vosoughi et al. [294] found that contrary to conventional wisdom, robots have accelerated the spread of real and fake news at the same rate, implying that fake news spreads more than the truth because humans, not robots, are more likely to spread it. In this case, verified users and those with numerous followers were not necessarily responsible for spreading misinformation about the corrupted posts [1].

Viral fake news can cause much havoc in our society. Therefore, to mitigate the negative impact of fake news it is important to analyze the factors that lead people to fall for misinformation and to further understand why people spread fake news [56]. Measuring the accuracy, credibility, veracity and validity of news content can also be a key countermeasure to consider.

Mirhoseini et al. [181] conducted a study to investigate the reasons behind people's belief in fake news. The study consisted of two experiments. The first experiment used behavioural and neurophysiological tools to test two competing theories in the disinformation literature, while the second experiment was an online survey that provided participants with feedback on their performance halfway through the survey. The results indicated a correlation between the belief in fake news and the lack of analytical thinking and actively open-minded thinking (AOT).

1.5.2.2. **Social bots spreaders:** Several authors [32, 240, 247, 254, 256] have also shown that fake news is likely to be created and spread by non-human accounts with similar attributes and structure in the network, such as social bots [87]. Bots (short for software robots) have existed since the early days of computers. A social bot is a computer algorithm that automatically produces content and interacts with humans on social media, trying to emulate and possibly alter their behaviour [87]. Although they are designed to provide a useful service, they can be harmful, for example when they contribute to the spread of unverified information or rumours [87]. However, it is important to note that bots are simply tools created and maintained by humans for some specific hidden agendas.

Social bots tend to connect with legitimate users instead of other bots. They try to act like a human with fewer words and fewer followers on social media. This contributes to the forwarding of fake news [130]. Moreover, there is a difference between bot-generated and human-written clickbait [154].

Many researchers have addressed ways of identifying and analyzing possible sources of fake news spread in social media. Recent research by Shu et al. [248] describes social bots' use of two strategies to spread low-credibility content. First, they amplify interactions with content as soon as it is created to make it look legitimate and to facilitate its spread across social networks. Next, they try to increase public exposure to the created content and thus boost its perceived credibility by targeting influential users who are more likely to believe disinformation in the hope of getting them to "repost" the fabricated content. They further discuss the social bot detection systems taxonomy proposed by Ferrara et al. [87] which divides bot detection methods into three classes: (1) graph-based, (2) crowdsourcing and (3) feature-based social bot detection methods.

Similarly, Shao et al. [240] examine social bots and how they promote the spread of misinformation through millions of Twitter posts during and following the 2016 US presidential campaign. They found that social bots played a disproportionate role in spreading articles from low-credibility sources by amplifying such content in the early spreading moments and targeting users with many followers through replies and mentions to expose them to this content and induce them to share it.

Ismailov et al. [126] assert that the techniques used to detect bots depend on the social platform and the objective. They note that a malicious bot designed to make friends with as many accounts as possible will require a different detection approach than a bot designed to repeatedly post links to malicious websites. Therefore, they identify two models for detecting malicious accounts, each using a different set of features. Social context models achieve detection by examining features related to an account's social presence including features such as relationships to other accounts, similarities to other users' behaviours, and a variety of graph-based features. User behaviour models primarily focus on features related to an individual user's behaviour, such as frequency of activities (e.g., number of tweets or posts per time interval), patterns of activity and clickstream sequences.

Therefore, it is crucial to consider bot detection techniques to distinguish bots from normal users to better leverage user profile features to detect fake news.

However, there is also another "bot-like" strategy that aims to massively promote disinformation and fake content on social platforms, which is called bot farms or troll farms. It is not social bots, but it is a group of organized individuals engaging in trolling or bot-like promotion of narratives in a coordinated fashion [306] hired to massively spread fake news or any other harmful content. A prominent troll farm example is the Russia-based Internet Research Agency (IRA), which disseminated inflammatory content online to influence the outcome of the 2016 U.S. presidential election [15]. As a result, Twitter suspended accounts connected to the IRA and deleted 200,000 tweets from Russian trolls [128]. Another example to mention in this category is review bombing [188]. Review bombing refers to coordinated groups of people massively performing the same negative actions online(e.g., dislike, negative review/comment) on an online video, game, post, product, etc., in order to reduce its aggregate review score. The review bombers can be both humans and bots coordinated in order to cause harm and mislead people by falsifying facts.

[15] https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731, last access date: 30-12-2023.

1.5.2.3. **Dynamic nature of online social platforms and fast propagation of fake news:** Karishma et al. [242] affirm that the fast proliferation of fake news through social networks makes it hard and challenging to assess the information's credibility on social media. Similarly, Qian et al. [217] assert that fake news and fabricated content propagates exponentially at the early stage of its creation and can cause a significant loss in a short amount of time [91] including manipulating the outcome of political events [32, 158].

Moreover, while analyzing the way sources and promoters of fake news operate over the web through multiple online platforms, Zannettou et al. [322] discovered that false information is more likely to spread across platforms (18% appearing on multiple platforms) compared to real information (11%).

Furthermore, recently Shu et al. [255] attempted to understand the propagation of disinformation and fake news in social media and found that such content is produced and disseminated faster and easier through social media because of the low barriers that prevent doing so. Similarly, Shu et al. [253] studied hierarchical propagation networks for fake news detection. They performed a comparative analysis between fake and real news from structural, temporal and linguistic perspectives. They demonstrated the potential of using these features to detect fake news and they showed their effectiveness for fake news detection as well.

Lastly, Abdullah-All-Tanvir et al. [2] note that it is almost impossible to manually detect the sources and authenticity of fake news effectively and efficiently, due to its fast circulation in such a small amount of time. Therefore, it is crucial to note that the dynamic nature of the various online social platforms, which results in the continued rapid and exponential propagation of such fake content, remains a major challenge that requires further investigation while defining innovative solutions for fake news detection.

1.5.3. Datasets issue

The existing approaches lack an inclusive dataset with derived multidimensional information to detect fake news characteristics to achieve higher accuracy of machine learning classification model performance [197]. These datasets are primarily dedicated to validating the machine learning model and are the ultimate frame of reference to train the model and analyze its performance. Therefore, if a researcher evaluates their model based on an unrepresentative dataset, the validity and efficiency of the model become questionable when it comes to applying the fake news detection approach in a real-world scenario.

Moreover, several researchers [205, 214, 258, 303] believe that fake news is diverse and dynamic in terms of content, topics, publishing methods and media platforms, and sophisticated linguistic styles geared to emulate true news. Consequently, training machine learning models on such sophisticated content requires large-scale annotated fake news data that is difficult to obtain [258].

Therefore, datasets are also a great topic to work on to enhance data quality and have better results while defining any solutions. Adversarial Learning techniques (e.g., GAN, SeqGAN) can be used to provide machine-generated data that can be used to train deeper models and build robust systems to detect fake examples from the real ones. This approach can be used to counter the lack of datasets and the scarcity of data available to train models.

1.6. Conclusion

In this chapter, I introduced the general context of the fake news problem as one of the major issues of the online deception problem in online social networks. Based on reviewing the most relevant state-of-the-art, I summarized and classified existing definitions of fake news as well as its related terms. I also listed various typologies and existing categorizations of fake news such as intent-based fake news including clickbait, hoax, rumour, satire, propaganda, conspiracy theories, and framing as well as content-based fake news including text and multimedia-based fake news and in the latter I can tackle deepfake videos and GAN generated fake images. I discussed the major challenges related to fake news detection and mitigation in social media including the deceptive nature of the fabricated content, the lack of human awareness in the field of fake news, the non-human spreaders issue (e.g., social bots), the dynamicity of such online platforms, which results in a fast propagation of fake content and the quality of existing datasets, which still limits the efficiency of the proposed solutions.

The upcoming chapter will delve into the state-of-the-art in fake news detection, focusing on the detection methods and techniques, along with an exploration of how they can be made more transparent and interpretable through explainability approaches. That chapter aims to present a comprehensive overview of the current strategies being employed in the detection of fake news.

Chapter 2

Fake News, Disinformation and Misinformation in Social Media: Detection Methods and Used Techniques Review

2.1. Introduction

This chapter builds upon the foundational study presented in the previous chapter, which focused on "fake news" and its related concepts and challenges. It is dedicated to a comprehensive examination of the state-of-the-art methods in fake news detection within online social networks, with a particular emphasis on recent advancements in the fields of explainability and multi-modality, as addressed by researchers in this domain. The key contributions of this chapter are outlined below:

— An exhaustive exploration of the state-of-the-art in automatic fake news detection and mitigation on online social networks. This includes an in-depth review of the key methods and techniques used to generate detection models, with particular attention to recent advancements in explainability and multi-modal approaches.

— A critical analysis of the major challenges and limitations that encounter existing fake news detection methods in online social networks. This analysis also examines how recent innovations in explainability and multi-modal methods are addressing these challenges.

Integral to this chapter is a discussion of my contributions to the field. This discussion highlights how my work addresses the identified challenges within these approaches and moves forward with the current state of fake news detection. It places particular focus on the innovative aspects of my research, demonstrating how these methods not only counter false information but also strengthen user trust in a digital environment full of complexities. Through this comprehensive analysis, the chapter aims to provide an understanding of both the current methodologies in fake news detection and the promising directions for future research and innovation.

2.2. Fake news detection literature review

Fake news detection in social networks is still in the early stage of development and there are still challenging issues that need further investigation. This has become an emerging research area that is attracting huge attention.

The fake news problem has been addressed by researchers from various perspectives related to different topics. These topics include, but are not restricted to, *social science studies*, which investigate why and who falls for fake news [14, 22, 30, 103, 181, 210, 268, 308], Whom to trust and how perceptions of misinformation and disinformation relate to media trust and media consumption patterns [114], how fake news differs from personal lies [57, 84], examine how can the law regulate digital disinformation and how governments can regulate the values of social media companies that themselves regulate disinformation spread on their platforms [7, 48, 168, 235, 287, 289, 297], and argue the challenges to democracy [135]; *Behavioural interventions studies*, which examine what literacy ideas mean in the age of dis/mis- and malinformation [52], investigate whether media literacy helps identification of fake news [134] and attempt to improve people's news literacy [19, 65, 92, 113, 134, 180, 190] by encouraging people to pause to assess credibility of headlines [86], promote civic online reasoning [172, 173] and critical thinking [162], together with evaluations of credibility indicators [35, 43, 58, 60, 152, 177, 193, 194, 201, 207, 208, 240, 246]; as well as *social media-driven studies*, which investigate the effect of signals (e.g., sources) to detect and recognize fake news [21, 37, 76, 115, 127, 195, 244, 279, 291, 295, 319] and investigate fake and reliable news sources using complex networks analysis based on search engine optimization metric [170].

The impacts of fake news have reached various areas and disciplines beyond online social networks and society [92] such as economics [59, 99, 145], psychology [227, 228, 286], political science [9, 27, 45, 102, 104, 223, 283, 285], neuroscience [182], health science [13, 18, 72, 84, 116, 178, 208, 226, 243, 304], environmental science (e.g., climate-change) [155, 162, 163, 278], etc.

Interesting research has been carried out to review and study the fake news issue in online social networks. Some focus not only on fake news, but also distinguish between fake news and rumour [39, 174], while others tackle the whole problem, from characterization to processing techniques [105, 254, 332]. However, they mostly focus on studying approaches from a machine learning perspective [39, 272], data mining perspective [254], crowd intelligence perspective [105], or knowledge-based perspective [332]. Furthermore, most of these studies ignore at least one of the mentioned perspectives and in many cases, they do not cover other existing detection approaches using methods such as blockchain and fact-checking as well as analysis on metrics used for Search Engine Optimization [170]. However, in this chapter and to the best of my knowledge, I cover all the approaches used for fake news detection. Indeed, I investigate the proposed solutions from broader perspectives (i.e., the detection techniques that are used as well as the different aspects and types of the information used).

Therefore, in this dissertation, I am highly motivated by the following facts. First, fake news detection on social media is still in the early age of development, and many challenging issues remain that require deeper investigation. Hence, it is necessary to discuss potential research directions that can improve fake news detection and mitigation tasks. However, the dynamic nature of fake news propagation through social networks further complicates matters [242]. False information can easily reach and impact a large number of users in a short time [91, 217]. Moreover, fact-checking organizations cannot keep up with the dynamics of propagation as they require human verification, which can hold back a timely and cost-effective response [142, 230, 251].

There are various research studies on fake news detection in online social networks. Few of them have focused on the automatic detection of fake news using artificial intelligence techniques. In this section, the existing approaches used in automatic fake news detection and the techniques that have been adopted are reviewed. Then, a critical discussion, built on a primary classification scheme based on a specific set of criteria, is also emphasized.

2.2.1. Categories of fake news detection

In this section, I give an overview of most of the existing automatic fake news detection solutions adopted in the literature. A recent classification by Sharma et al. [242] uses three categories of fake news identification methods. Each category is further divided based on the type of existing methods (i.e., content-based, feedback-based and intervention-based methods). However, a review of the literature on fake news detection in online social networks shows that the existing studies can be classified into broader categories based on two major aspects that most authors inspect and make use of to define an adequate solution. These aspects can be considered as major sources of extracted information used for fake news detection and can be summarized as follows: the content-based (i.e., related to the content of the news post) and the contextual aspect (i.e., related to the context of the news post).

Consequently, the studies reviewed can be classified into three different categories based on the two aspects mentioned above (the third category is hybrid). As depicted in Figure 1, fake news detection solutions can be categorized as news content-based approaches, social context-based approaches (which can be further divided into network and user-based approaches), and hybrid approaches. The latter combines both content-based and contextual approaches to define the solution.

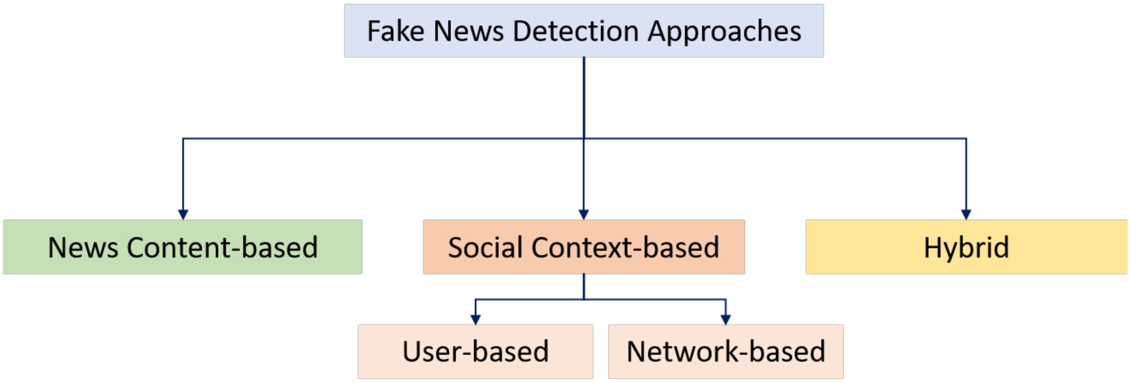


Figure 1 – Classification of fake news detection approaches

2.2.1.1. **News content-based category:** News content-based approaches are fake news detection approaches that use content information (i.e., information extracted from the content of the news post) and that focus on studying and exploiting the news content in their proposed solutions. Content refers to the body of the news, including source, headline, text and image-video, which can reflect subtle differences.

Researchers of this category rely on content-based detection cues (i.e., text and multimedia-based cues), which are features extracted from the content of the news post. Text-based cues are features extracted from the text of the news, whereas multimedia-based cues are features extracted from the images and videos attached to the news. Figure 2 summarizes the most widely used news content representation (i.e., text and multimedia/images) and detection techniques (i.e., machine learning (ML), deep Learning (DL), natural language processing (NLP), fact-checking, crowdsourcing (CDS) and blockchain (BKC)) in news content-based category of fake news detection approaches. Most of the reviewed research works based on news content for fake news detection rely on the text-based cues [1, 2, 23, 119, 120, 139, 140, 164, 197, 200, 288, 300, 330] extracted from the text of the news content including the body of the news and its headline. However, a few researchers such as Vishwakarma et al. [291] and Amri et al. [16] try to recognize text from the associated image.

Most researchers of this category rely on artificial intelligence (AI) techniques (such as ML, DL and NLP models) to improve performance in terms of prediction accuracy. Others use different techniques such as fact-checking, crowdsourcing and blockchain. Specifically, the AI and ML-based approaches in this category are trying to extract features from the news content, which they use later for content analysis and training tasks. In this particular case, the extracted features are the different types of information considered to be relevant for the analysis. Feature extraction is considered as one of the best techniques to reduce data size in automatic fake news detection. This technique aims to choose a subset of features from the original set to improve classification performance [320].

While contributing to this group of approaches (i.e., content-based detection), my work in **EXMULF**[16] stands out for its unique approach. Unlike researchers at the time, I utilized multiple aspects for content-based detection, namely, both text and the associated image simultaneously. Specifically, I presented an explainable multimodal content-based fake news detection system. This system focuses on analyzing the veracity of information based on both its textual content and the associated image simultaneously, while also incorporating an Explainable AI (XAI) assistant. To the best of my knowledge, this study represents the first attempt to provide a fully explainable multimodal content-based fake news detection system at that time. This was achieved by employing Latent Dirichlet Allocation (LDA) topic modelling, Vision-and-Language BERT (VilBERT), and Local Interpretable Model-agnostic Explanations (LIME) models. The experiments conducted on two real-world datasets demonstrated the significance of learning the connection between two content-based modalities (i.e., text and image), resulting in an accuracy that surpassed that of 10 state-of-the-art fake news detection models.

Table 1 lists the distinct features and metadata, as well as the used datasets in the news content-based category of fake news detection approaches.

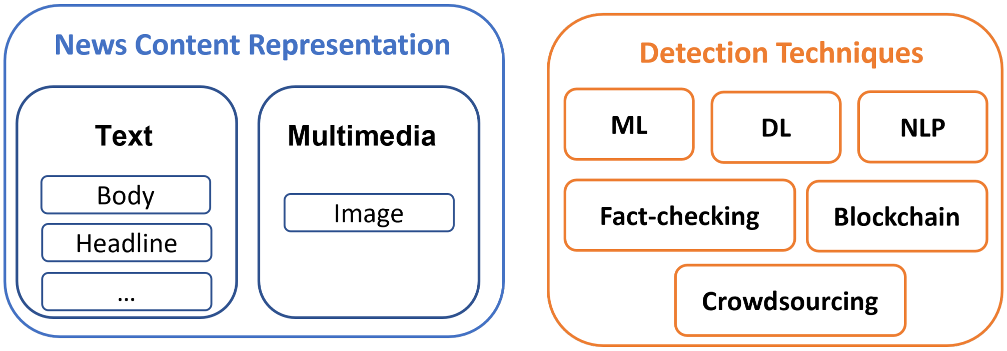


Figure 2 – News content-based category: news content representation and detection techniques

2.2.1.2. **Social context-based category:** Unlike news content-based solutions, the social context-based approaches capture the skeptical social context of the online news [328] rather than focusing on the news content. The social context-based category contains fake news detection approaches that use contextual aspects (i.e., information related to the context of the news post). These aspects are based on social context and they offer additional information to help detect fake news. They are the surrounding data outside of the fake news article itself, where they can be an essential part of automatic fake news detection. Some useful examples of contextual information may include: checking if the news itself and the source that published it are credible, checking the date of the news or the supporting resources, and checking if any other online news platforms are reporting the same or similar stories [328].

Table 1 – The features and datasets used in the news content-based approaches

Feature and metadata	Datasets	Reference
Average sentence word count, stop word frequency, and sentiment rate.	Getting real about fake news[a], Gathering mediabiasfactcheck[b], KaiDMML FakeNewsNet[c], Real news for Oct-Dec 2016[d].	Kapusta et al. [139]
Title, body, and label length distribution.	News trends, Kaggle, Reuters.	Kaur et al. [140]
Sociolinguistic, historical, cultural, ideological and syntactical features.	FakeNewsNet.	Vereshchaka et al. [288]
Term frequency.	BuzzFeed political news, Random political news, ISOT fake news.	Ozbay et al. [200]
Statement, speaker, context, labeling, and justification.	PolitiFact , LIAR[e].	Wang [300]
Spatial word proximity, term relations, and latent term-article connections.	Kaggle fake news dataset[f].	Hosseinimotlagh and Papalexakis [120]
Word length and count in tweets.	Twitter dataset, Chile earthquake 2010 datasets.	Abdullah-All-Tanvir et al. [1]
Count of negatively emotive words.	Twitter dataset.	Abdullah-All-Tanvir et al. [2]
Labeled data.	BuzzFeed[g], PolitiFact[h].	Mahabub [164]
Headline-body correlation and article bias.	Kaggle: real_or_fake[i], Fake news detection[j].	Bahad et al. [23]
Historical data, content topic/sentiment, and semantic context.	Facebook dataset.	Del Vicario et al. [68]
Image text veracity and top 15 Google search result credibility.	Google images, the Onion, Kaggle.	Vishwakarma et al. [291]
Semantic-level features.	PolitiFact and BuzzFeed.	Zhou et al. [330]
Text and image topic modelling in online news.	Twitter dataset[k], Weibo[l].	Amri et al. [16]

[a] `https://www.kaggle.com/anthonyc1/gathering-real-news-for-oct-dec-2016`, last access date: 30-12-2023.

[b] `https://mediabiasfactcheck.com/`, last access date: 30-12-2023.

[c] `https://github.com/KaiDMML/FakeNewsNet`, last access date: 30-12-2023.

[d] `https://www.kaggle.com/anthonyc1/gathering-real-news-for-oct-dec-2016`, last access date: 30-12-2023.

[e] `https://www.cs.ucsb.edu/~william/data/liar_dataset.zip`, last access date: 30-12-2023.

[f] `https://www.kaggle.com/mrisdal/fake-news`, last access date: 30-12-2023.

[g] `https://github.com/BuzzFeedNews/2016-10-facebook-fact-check`, last access date: 30-12-2023.

[h] `https://www.politifact.com/subjects/fake-news/`, last access date: 30-12-2023.

[i] `https://www.kaggle.com/rchitic17/real-or-fake`, last access date: 30-12-2023.

[j] `https://www.kaggle.com/jruvika/fake-news-detection`, last access date: 30-12-2023.

[k] `https://github.com/MKLab-ITI/image-verification-corpus`, last access date: 30-12-2023.

[l] `https://drive.google.com/file/d/14VQ7EWPiFeGzxp3XC2DeEHi-BEisDINn/view`, last access date: 30-12-2023.

Social context-based aspects can be classified into two subcategories, user-based and network-based and they can be used for context analysis and training tasks in the case of AI and ML-based approaches. User-based aspects refer to information captured from OSN users such as user profile information [112, 130, 179, 197, 259, 304] and user behaviour [50] such as user engagement [130, 197, 256, 282], response [217, 326], and personality traits [49, 183, 199, 233, 245]. Meanwhile, network-based aspects refer to information captured from the properties of the social network where the fake content is shared and disseminated such as news propagation path [158, 309] (e.g., propagation times and temporal characteristics of propagation), diffusion patterns [179, 257] (e.g., number of retweets, shares), as well as user relationships [112, 130, 179, 184] (e.g., friendship status among users).

The major drawback of this category of approaches (i.e., social context-based approaches) is that they mostly rely on information that is only available after the false information has spread, such as user behaviour, including users' responses and engagement with the false content, and network-based patterns, such as false information propagation paths (e.g., spatiotemporal propagation patterns) and diffusion patterns (e.g., the number of retweets and shares). While contributing to this group of approaches (i.e., social context-based detection), my work in **ExFake** [15] and **MythXpose** focuses on using social context features that are available before the spread of false information. In ExFake, I utilize user historical sharing behaviour along with external evidence from third parties, such as reputable fact-checkers and trusted official sources (i.e., official social accounts) if already available, and in MythXpose, I predict user personality traits. Unlike researchers at the time, this approach allows for **proactive** detection and mitigation of false information in an attempt for **early** detection of false content. A *proactive* strategy in combating false information, shifting from reactive measures (which kick in after the spread) to preventive measures, aiming to stop the spread before it begins.

Figure 3 summarizes some of the most widely adopted social context representations as well as the most used detection techniques (i.e., AI, ML, DL, fact-checking and blockchain) in the social context-based category of approaches.

Table 2 lists the distinct features and metadata, the adopted detection cues as well as the used datasets in the context-based category of fake news detection approaches.

2.2.1.3. **Hybrid approaches:** Most researchers are focusing on employing a specific method rather than a combination of both content and context-based methods. This is because some of them [311] believe that there are still some challenging limitations in the traditional fusion strategies due to existing feature correlations and semantic conflicts. For this reason, some researchers focus on extracting content-based information while others are capturing some social context-based information for their proposed approaches.

Table 2 – The features, detection cues and datasets used in the social context-based approaches

Feature and metadata	Detection cues	Datasets	Reference
Users' sharing behaviours and profile features.	User-based: user profile information.	FakeNewsNet.	Shu et al. [259]
Users' trust level and profile features of "experienced" vs "naive" users.	User-based: user engagement.	FakeNewsNet, BuzzFeed, PolitiFact.	Shu et al. [256]
Users' replies and stances on fake content.	User-based: user response.	RumourEval, PHEME.	Zhang et al. [326]
Historical user responses to articles.	User-based: user response.	Weibo, Twitter dataset.	Qian et al. [217]
Speaker information, e.g., name, job, political party.	User-based: user profile information.	LIAR.	Wang et al. [301]
Latent relationships among users and influence of prestigious users.	Networks-based: user relationships.	Twitter15 and Twitter16[a].	Mishra [184]
Tri-relationships among publishers, news items, and users.	Networks-based: diffusion patterns.	FakeNewsNet.	Shu et al. [259]
Propagation paths of news stories from retweets.	Networks-based: news propagation path.	Weibo, Twitter15, Twitter16.	Liu and Wu [158]
Message propagation in social networks.	Networks-based: news propagation path.	Twitter dataset.	Wu and Liu [309]
Spatiotemporal information and user engagement patterns.	User-based: user engagement.	FakeNewsNet, PolitiFact, GossipCop, Twitter.	Nyow and Chua [197]
Source credibility, user characteristics, and social graph.	User and network-based: profile and relationships.	Ego-Twitter[b].	Hamdi et al. [112]
Followee/follower network and user similarities.	User and network-based: profile, engagement, and relationships.	FakeNewsNet.	Jiang et al. [130]
The relationship between critical thinking, media literacy, and fake news detection abilities.	User-based: Critical thinking dispositions, new media literacies.	Data gathered from 157 university students.	Ali Orhan [199]
The profiles of users, their social relations, and the way news spreads,	User and network-based: profiles, users' social relations, diffusion patterns.	FakeNewsNet.	Michail et al. [179]
Personality traits (neuroticism, openness, and extraversion).	User-based: Personality Traits.	Survey of 242 Shiraz University students.	Mirzabeigi et al. [199]

[a] `https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip`, last access date: 30-12-2023.

[b] `https://snap.stanford.edu/data/ego-Twitter.html`, last access date: 30-12-2023.

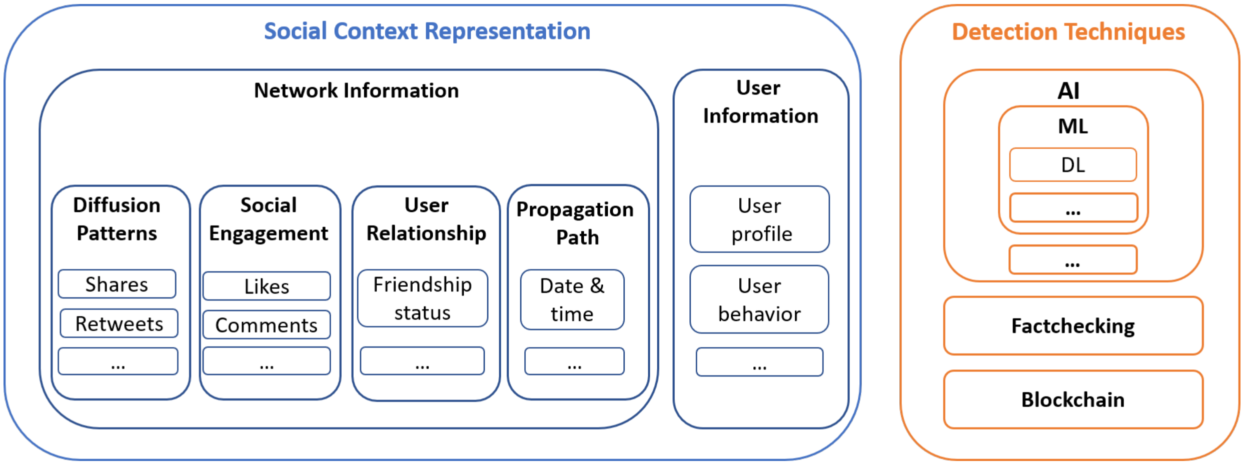


Figure 3 – Social context-based category: social context representation and detection techniques

However, it has proven challenging to successfully automate fake news detection based on just a single type of feature [230]. Therefore, recent directions tend to do a mixture by using both news content-based and social context-based approaches for fake news detection. This integrated strategy, which I employ in my systems **ExFake** [15] and **MythXpose**, combines various analytical perspectives (i.e., content analysis of news items including textual and visual elements, social context analysis based on OSN users' historical sharing behaviour and prediction of user personality traits, and the incorporation of external validation evidence from reputable fact-checkers and official sources such as official social accounts). MythXpose, for example, merges content analysis, encompassing both text and its associated image, with the prediction of OSN users' personality traits, a key aspect of social context analysis. Similarly, ExFake also integrates content and social context analysis, which includes the historical sharing behaviour of OSN users, along with leveraging insights from external sources (i.e., third parties), like reputable fact-checkers and trusted official sources (i.e., official social accounts). This **multifaceted** approach in both systems leads to a more comprehensive method of fake news detection. Additionally, both systems (i.e., ExFake and MythXpose) enhance transparency by providing clear explanations to OSN users about the detection process, fostering user understanding and trust.

Table 3 lists the distinct features and metadata as well as the used datasets in the hybrid category of fake news detection approaches.

2.2.2. Fake news detection techniques

Another vision for classifying automatic fake news detection is to look at techniques used in the literature. Hence, the classification of detection methods based on the techniques is divided into three groups:

Table 3 – The features and datasets used in the hybrid approaches

Feature and metadata	**Datasets**	**Reference**
News features: title, content, date, source, location.	SOT fake news dataset, LIAR dataset and FAKES dataset.	Elhadad et al. [81]
Spatiotemporal information, user Twitter profiles, engagement patterns.	FakeNewsNet, PolitiFact, GossipCop, Twitter.	Nyow and Chua [197]
Publisher domains and reputations, news terms and embeddings, shares, reactions, and comments.	BuzzFeed.	Xu et al. [314]
Tweeted content propagation, and discussion metrics.	Twitter dataset.	Aswani et al. [20]
News evolution features, user involvement features.	Twitter dataset.	Previti et al. [213]
Semantics and conflicts between posts and comments.	RumourEval, PHEME.	Wu and Rao [311]
Publisher information, user semantics/emotions, and latent news/comment representations.	Weibo.	Guo et al. [106]
Relationships between news, creators, and subjects.	PolitiFact.	Zhang et al. [325]
News source domains, author names.	George McIntire fake news dataset.	Deepak and Chitturi [67]
News content, social context, spatiotemporal data, and synthetic user engagement patterns.	FakeNewsNet.	Shu et al. [251]
News content, social reactions, post language, user dissemination, similarity, stance, sentiment, headlines, entities, sharing history, and comments.	SHPT, PolitiFact.	Wang, et al. [299]
News source, headline, author, publication time, and user interactions.	NELA-GT-2019, Fakeddit.	Raza and Ding [219]

- Human-based techniques: this category mainly includes the use of crowdsourcing and fact-checking techniques, which rely on human knowledge to check and validate the veracity of news content.
- Artificial Intelligence-based techniques: this category includes the most used AI approaches for fake news detection in the literature. Specifically, these are the approaches in which researchers use classical ML, deep learning techniques such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), as well as Natural Language Processing (NLP).

— Blockchain-based techniques: this category includes solutions using blockchain technology to detect and mitigate fake news in social media by checking source reliability and establishing the traceability of the news content.

2.2.2.1. **Human-based techniques:** One specific research direction for fake news detection consists of using human-based techniques such as crowdsourcing [178, 209] and fact-checking [58, 195, 292] techniques.

These approaches can be considered as low computational requirement techniques since both rely on human knowledge and expertise for fake news detection. However, fake news identification cannot be addressed solely through human force since it demands a lot of effort in terms of time and cost, and it is ineffective in terms of preventing the fast spread of fake content.

Crowdsourcing: Crowdsourcing approaches [142] are based on the "wisdom of the crowds" [61] for fake content detection. These approaches rely on the collective contributions and crowd signals [280] of a group of people for the aggregation of crowd intelligence to detect fake news [273] and to reduce the spread of misinformation on social media [178, 209].

Micallef et al. [178] highlight the role of the crowd in countering misinformation. They suspect that concerned citizens (i.e., the crowd), who use platforms where disinformation appears, can play a crucial role in spreading fact-checking information and in combating the spread of misinformation.

Recently Tchakounté et al. [273] proposed a voting system as a new method of binary aggregation of opinions of the crowd and the knowledge of a third-party expert. The aggregator is based on majority voting on the crowd side and weighted averaging on the third-party site.

Similarly, Huffaker et al. [123] propose a crowdsourced detection of emotionally manipulative language. They introduce an approach that transforms classification problems into a comparison task to mitigate conflation content by allowing the crowd to detect text that uses manipulative emotional language to sway users towards positions or actions. The proposed system leverages anchor comparison to distinguish between intrinsically emotional content and emotionally manipulative language.

La Barbera et al. [151] try to understand how people perceive the truthfulness of information presented to them. They collect data from US-based crowd workers, build a dataset of crowdsourced truthfulness judgments for political statements, and compare it with expert annotation data generated by fact-checkers such as PolitiFact.

Coscia and Rossi [64] introduce a crowdsourced flagging system that consists of online news flagging. The bipolar model of news-flagging attempts to capture the main ingredients that they observe in empirical research on fake news and disinformation.

Unlike the previously mentioned researchers who focus on news content in their approaches, Pennycook and Rand [209] focus on using crowdsourced judgments of the quality of news sources to combat social media disinformation.

Fact-checking: The fact-checking task is commonly manually performed by journalists to verify the truthfulness of a given claim. Indeed, fact-checking features are being adopted by multiple online social network platforms. For instance, Facebook [1] started addressing false information through independent fact-checkers in 2017, followed by Google [2] the same year. Two years later, Instagram [3] followed suit. However, the usefulness of fact-checking initiatives is questioned by journalists [4] as well as by researchers such as Andersen and Søe [17]. Therefore, work is being conducted to boost the effectiveness of these initiatives to reduce misinformation [58, 60, 195].

Most researchers use fact-checking websites (e.g., politifact.com [5], snopes.com [6], Reuters [7], etc.) as data sources to build their datasets and train their models. Therefore, in the following, specific examples of solutions that use fact-checking [292] to help build datasets that can be further used in the automatic detection of fake content are reviewed.

Yang et al. [316] use PolitiFact fact-checking website as a data source to train, tune, and evaluate their model named XFake, on political data. The XFake system is an explainable fake news detector that assists end-users in identifying news credibility. The fakeness of news items is detected and interpreted considering both content and contextual (e.g., statements) information (e.g., speaker).

Based on the idea that fact-checkers cannot clean all data, and it must be a selection of what "matters the most" to clean while checking a claim, Sintos et al. [264] propose a solution to help fact-checkers combat problems related to data quality (where inaccurate data leads to incorrect conclusions) and data phishing. The proposed solution is a combination of data cleaning and perturbation analysis to avoid uncertainties and errors in data and the possibility that data can be phished.

Tchechmedjiev et al. [274] propose a system named "ClaimsKG" as a knowledge graph of fact-checked claims aiming to facilitate structured queries about their truth values, authors, dates, journalistic reviews and other kinds of metadata. "ClaimsKG" designs the relationship

[1] `https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news`, last access date: 30-12-2023.

[2] `https://www.theguardian.com/technology/2017/apr/07/google-to-display-fact-checking-labels-to-show-if-news-is-true-or-false`, last access date: 30-12-2023.

[3] `https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram`, last access date: 30-12-2023.

[4] `https://www.wired.com/story/instagram-fact-checks-who-will-do-checking/`, last access date: 30-12-2023.

[5] `https://www.politifact.com/`, last access date: 30-12-2023.

[6] `https://www.snopes.com/`, last access date: 30-12-2023.

[7] `https://www.reutersagency.com/en/`, last access date: 30-12-2023.

between vocabularies. To gather vocabulary, a semi-automated pipeline periodically gathers data from popular fact-checking websites regularly.

2.2.2.2. **AI-based techniques:** Previous work by Yaqub et al. [318] has shown that people lack trust in automated solutions for fake news detection. However, work is already being undertaken to increase this trust, for instance by von der Weth [293].

Most researchers consider fake news detection as a classification problem and use artificial intelligence techniques, as shown in Figure 4. The adopted AI techniques may include machine learning ML (e.g., Naïve Bayes, Logistic Regression, Support Vector Machine SVM), deep learning DL (e.g., Convolutional Neural Networks CNN, Recurrent Neural Networks RNN, Long Short-Term Memory LSTM) and natural language processing NLP (e.g., Count vectorizer, TF IDF Vectorizer). Most of them combine many AI techniques in their solutions rather than relying on one specific approach.

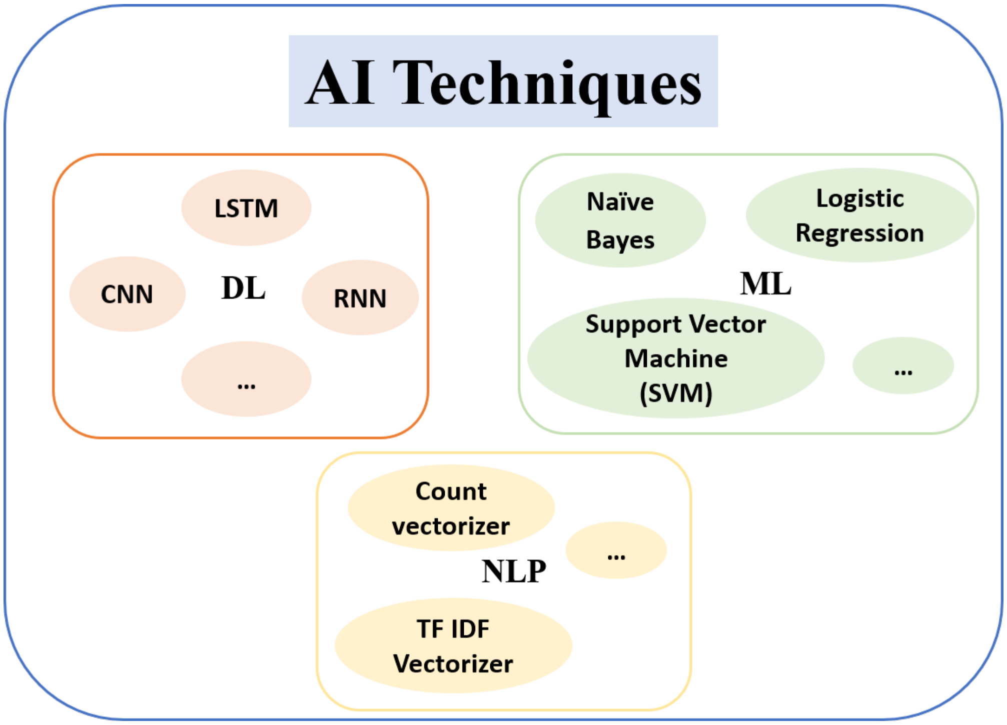


Figure 4 – Examples of the most widely used AI techniques for fake news detection

Many researchers are developing machine learning models in their solutions for fake news detection. Recently, deep neural network techniques are also being employed as they are generating promising results [125]. A neural network is a massively parallel distributed processor with simple units that can store important information and make it available for use [119]. Moreover, it has been proven [51] that the most widely used method for automatic detection of fake news is not simply a classical machine learning technique, but rather a fusion of classical techniques coordinated by a neural network.

Some researchers define purely machine learning models [20, 68, 81, 111, 262] in their fake news detection approaches. The more commonly used machine learning algorithms [1] for classification problems are Naïve Bayes, Logistic Regression and SVM.

Other researchers [98, 158, 184, 217, 300, 304, 325] prefer to do a mixture of different deep learning models, without combining them with classical machine learning techniques. Some even prove that deep learning techniques outperform traditional machine learning techniques [185]. Deep Learning is one of the most widely popular research topics in machine learning. Unlike traditional machine learning approaches, which are based on manually crafted features, deep learning approaches can learn hidden representations from simpler inputs both in context and content variations [39]. Moreover, traditional machine learning algorithms almost always require structured data and are designed to "learn" to act by understanding labelled data and then use it to produce new results with more datasets, which requires human intervention to "teach them" when the result is incorrect [203]. While deep learning networks rely on layers of artificial neural networks (ANN) and do not require human intervention, as multilevel layers in neural networks place data in a hierarchy of different concepts, which ultimately learn from their own mistakes [203]. The two most widely implemented paradigms in deep neural networks are Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN).

Still other researchers [1, 67, 137, 251, 304, 326] prefer to combine traditional machine learning and deep learning classification, models. Others combine machine learning and natural language processing techniques. A few combine deep learning models with natural language processing [288]. Some other researchers [4, 139, 200] combine natural language processing with machine learning models. Furthermore, others [1, 2, 23, 136, 140] prefer to combine all the previously mentioned techniques (i.e., ML, DL and NLP) in their approaches. Table 4, shows a comparison of the fake news detection solutions that I have reviewed based on their main approaches, the methodology that was used and the models.

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Del Vicario et al. [68]	An approach to analyze the sentiment associated with data textual content and add semantic knowledge to it.	ML	Linear Regression (LIN), Logistic Regression (LOG), Support Vector Machine (SVM) with Linear Kernel, K-Nearest Neighbors (KNN), Neural Network Models (NN), Decision Trees (DT).

Continued on next page

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Elhadad et al. [81]	An approach to select hybrid features from the textual content of the news, which they consider as blocks, without segmenting text into parts (title, content, date, source, etc.).	ML	Decision Tree, KNN, Logistic Regression, SVM, Naïve Bayes with n-gram, LSVM, Perceptron.
Aswani et al. [20]	A hybrid artificial bee colony approach to identify and segregate buzz in Twitter and analyze user-generated content (UGC) to mine useful information (content buzz/ popularity).	ML	KNN with artificial bee colony optimization.
Hakak et al. [111]	An ensemble of machine learning approaches for effective feature extraction to classify fake news.	ML	Decision Tree, Random Forest and Extra Tree Classifier.
Singh et al. [262]	A multimodal approach, combining text and visual analysis of online news stories to automatically detect fake news through predictive analysis to detect features most strongly associated with fake news.	ML	Logistic Regression, Linear Discrimination Analysis, Quadratic Discriminant Analysis, K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Classification and Regression Tree, and Random Forest Analysis.
Amri et al. [16]	An explainable multimodal content-based fake news detection system.	ML	Vision-and-Language BERT (VilBERT), Local Interpretable Model-Agnostic Explanations (LIME), Latent Dirichlet Allocation (LDA) topic modelling.

Continued on next page

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Wang et al. [301]	A hybrid deep neural network model to learn the useful features from contextual information and to capture the dependencies between sequences of contextual information.	DL	Recurrent and Convolutional Neural Networks (RNN and CNN).
Wang [300]	A hybrid convolutional neural network approach for automatic fake news detection.	DL	Recurrent and Convolutional Neural Networks (RNN and CNN).
Liu and Wu [158]	An early detection approach of fake news to classify the propagation path to mine the global and local changes of user characteristics in the diffusion path.	DL	Recurrent and Convolutional Neural Networks (RNN and CNN).
Mishra [184]	Unsupervised network representation learning methods to learn user (node) embeddings from both the follower network and the retweet network and to encode the propagation path sequence.	DL	RNN: (long short-term memory unit (LSTM)).

Continued on next page

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Qian et al. [217]	A Two-Level Convolutional Neural Network with User Response Generator (TCNN-URG) where TCNN captures semantic information from the article text by representing it at the sentence and word level. The URG learns a generative model of user responses to article text from historical user responses that it can use to generate responses to new articles to assist fake news detection.	DL	Convolutional Neural Network (CNN).
Zhang et al. [325]	Based on a set of explicit features extracted from the textual information, a deep diffusive network model is built to infer the credibility of news articles, creators and subjects simultaneously.	DL	Deep Diffusive Network Model Learning.
Goldani et al. [98]	A capsule networks (CapsNet) approach for fake news detection using two architectures for different lengths of news statements and claims that capsule neural networks have been successful in computer vision and are receiving attention for use in Natural Language Processing (NLP).	DL	Capsule Networks (CapsNet).

Continued on next page

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Wang et al. [301]	An automated approach to distinguish different cases of fake news (i.e., hoaxes, irony and propaganda) while assessing and classifying news articles and claims including linguistic cues as well as user credibility and news dissemination in social media.	DL, ML	Convolutional Neural Network (CNN), long Short-Term Memory (LSTM), logistic regression.
Abdullah-All-Tanvir et al. [1]	A model to recognize forged news messages from Twitter posts, by figuring out how to anticipate precision appraisals, in view of computerizing forged news identification in the Twitter dataset. A combination of traditional machine learning, as well as deep learning classification models, is tested to enhance the accuracy of prediction.	DL, ML	Naïve Bayes, Logistic Regression, Support Vector Machine, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM).
Kaliyar et al. [137]	An approach named (FNDNet) is based on the combination of the unsupervised learning algorithm GloVe and the deep convolutional neural network for fake news detection.	DL, ML	Deep Convolutional Neural Network (CNN), Global Vectors (GloVe).
Zhang et al. [326]	A hybrid approach to encode auxiliary information coming from people's replies alone in temporal order. Such auxiliary information is then used to update a priori belief generating a posteriori belief.	DL, ML	Deep Learning Model, Long Short-Term Memory Neural Network (LSTM).

Continued on next page

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Deepak and Chitturi [67]	A system that consists of live data mining in addition to the deep learning model.	DL, ML	Feedforward Neural Network (FNN) and LSTM Word Vector Model.
Shu et al. [251]	A multidimensional fake news data repository "FakeNewsNet" and conduct an exploratory analysis of the datasets to evaluate them.	DL, ML	Convolutional Neural Network (CNN), Support Vector Machines (SVMs), Logistic Regression (LR), Naïve Bayes (NB).
Vereshchaka et al. [288]	A sociocultural textual analysis, computational linguistics analysis, and textual classification using NLP, as well as deep learning models to distinguish fake from real news to mitigate the problem of disinformation.	DL, NLP	Short-Term Memory (LSTM), Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU).
Kapusta et al. [139]	A sentiment and frequency analysis using both machine learning and NLP in what is called text mining to process news content sentiment analysis and frequency analysis to compare basic text characteristics of fake and real news articles.	ML, NLP	The Natural Language Toolkit (NLTK), TF-IDF.

Continued on next page

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Ozbay and Alatas [200]	A hybrid approach based on text analysis and supervised artificial intelligence for fake news detection.	ML, NLP	Supervised algorithms: BayesNet, JRip, OneR, Decision Stump, ZeroR, Stochastic Gradient Descent (SGD), CV Parameter Selection (CVPS), Randomizable Filtered Classifier (RFC), Logistic Model Tree (LMT). NLP: TF weighting.
Ahmed et al. [4]	A machine learning and NLP text-based processing to identify fake news. Various features of the text are extracted through text processing and incorporated into classification.	ML, NLP	Machine learning classifiers (i.e., Passive-aggressive, Naïve Bayes and Support Vector Machine).
Abdullah-All-Tanvir et al. [2]	A hybrid neural network approach to identify authentic news on popular Twitter threads would outperform the traditional neural network architecture's performance. Three traditional supervised algorithms and two Deep Neural are combined to train the defined model. Some NLP concepts were also used to implement some of the traditional supervised machine learning algorithms over their dataset.	ML, DL, NLP	Traditional supervised algorithm (i.e., Logistic Regression, Bayesian Classifier and Support Vector Machine). Deep Neural Networks (i.e., Recurrent Neural Network, Long Short-Term Memory LSTM). NLP concepts such as Count vectorizer and TF IDF Vectorizer.

Continued on next page

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Kaur et al. [140]	A hybrid method to identify news articles as fake or real through finding out which classification model identifies false features accurately.	ML, DL, NLP	Neural Networks (NN) and Ensemble Models. Supervised Machine Learning Classifiers such as Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Linear Models. Term Frequency–Inverse Document Frequency (TF–IDF), Count-Vectorizer (CV), Hashing-Vectorizer (HV).
Kaliyar [136]	A fake news detection approach to classify the news article or other documents into certain or not. Natural language processing, machine learning and deep learning techniques are used to implement the defined models and to predict the accuracy of different models and classifiers.	ML, DL, NLP	Machine Learning Models: Naïve Bayes, K-nearest Neighbors, Decision Tree, Random Forest. Deep Learning Networks: Shallow Convolutional Neural Networks (CNN), Very Deep Convolutional Neural Network (VDCNN), Long Short-Term Memory Network (LSTM), Gated Recurrent Unit Network (GRU). A combination of Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM) and Convolutional Neural Network with Gated Recurrent Unit (CNN-LSTM).

Continued on next page

Table 4 – Comparison of AI-based fake news detection techniques

Reference	Approach	Method	Model
Mahabub [164]	An intelligent detection system to manage the classification of news as being either real or fake.	ML, DL, NLP	Machine Learning: Naïve Bayes, K-NN, SVM, Random Forest, Artificial Neural Network, Logistic Regression, Gradient Boosting, AdaBoost.
Bahad et al. [23]	A method based on a Bi-directional LSTM-recurrent neural network to analyze the relationship between the news article headline and article body.	ML, DL, NLP	Unsupervised Learning algorithm: Global Vectors (GloVe). Bi-directional LSTM-recurrent Neural Network.

2.2.2.3. **Blockchain-based techniques for Source reliability and traceability:** Another research direction for detecting and mitigating fake news in social media focuses on using blockchain solutions. Blockchain technology has recently attracted researchers' attention due to the interesting features it offers. Immutability, decentralization, tamperproof, consensus, record keeping and non-repudiation of transactions are some of the key features that make blockchain technology exploitable, not just for cryptocurrencies, but also to prove the authenticity and integrity of digital assets.

However, the proposed blockchain approaches are few in number and they are fundamental and theoretical approaches. Specifically, the solutions that are currently available are still in research, prototype, and beta testing stages [77, 274]. Furthermore, most researchers [47, 55, 133, 198, 216, 238, 266] do not specify which fake news type they are mitigating in their studies. They mention news content in general, which is not adequate for innovative solutions. For that, serious implementations should be provided to prove the usefulness and feasibility of this newly developing research vision.

Table 5 shows a classification of the reviewed blockchain-based approaches. In this classification, the following aspects are listed:

- The type of fake news that authors are trying to mitigate, which can be multimedia-based or text-based fake news.
- The techniques used for fake news mitigation, which can be either blockchain only or blockchain combined with other techniques such as AI, Data mining, Truth-discovery, Preservation metadata, Semantic similarity, Crowdsourcing, Graph theory and SIR model (Susceptible, Infected, Recovered).

— The feature that is offered as an advantage of the given solution (e.g., Reliability, Authenticity and Traceability). Reliability is the credibility and truthfulness of the news content, which consists of proving the trustworthiness of the content. Traceability aims to trace and archive the contents. Authenticity consists of checking whether the content is real and authentic.

A checkmark (✓) in Table 5 denotes that the mentioned criterion is explicitly mentioned in the proposed solution, while the empty grey cell for fake news type denotes that it depends on the case: the criterion was either not explicitly mentioned (e.g., fake news type) in the work or the classification does not apply (e.g., techniques/other).

Table 5 – A classification of popular blockchain-based approaches for fake news detection in social media

Reference	**Fake News Type**		**Techniques**	**Feature**
	Multimedia	Text		
Shae and Tsai [236]	✓	✓	AI	Reliability
Ochoa et al. [198]		✓	Data Mining, Truth-Discovery	Reliability
Huckle and White [122]	✓		Preservation Metadata	Reliability
Song et al. [266]				Traceability
Shang et al. [238]				Traceability
Qayyum, et al. [216]			Semantic Similarity	Reliability
Jing and Murugesan [133]			AI	Reliability
Buccafurri et al. [47]			Crowd-Sourcing	Reliability
Chen et al. [55]			SIR Model	Reliability
Hasan and Salah [117]	✓			Authenticity
Tchechmed-jiev et al. [274]			Graph theory	Reliability

2.3. Explainability and multi-modality

Several efforts have been made in the development of deep learning-based automatic fake news detection approaches. However, there has been little previous work going out of the black-box nature of such approaches and focusing on providing explanations to online social network users. Such explanations are crucial to reflect news credibility, raise OSN users' awareness and ultimately influence their behaviour in protecting the security and privacy of both individuals and society. Additionally, exploring the multimodal data available in news content is crucial for strengthening the explanations provided to OSN users, as well as for the early detection of fake news. Indeed, the content of the news is fully available in the early stages, unlike auxiliary information (e.g., social engagement, user response, propagation patterns) which can only be obtained after the news has spread.

This section builds on two existing research approaches. Firstly, it discusses methods for automatically detecting fake news by analyzing its multimodal content (i.e., text and image) without human intervention. Secondly, it explores how fake news classification models can be expanded using explainable AI (XAI) and visual analytics to help online social network (OSN) users understand how a particular classification result was achieved. In this section, a brief overview of relevant studies on multimodal content-based fake news detection and explainable fake news detection is provided.

2.3.1. Multimodal content-based fake news detection

Numerous studies in fake news detection started using visual information, as auxiliary information in their detection methods to infer the veracity of online news. They are named multimodal approaches since they analyze textual data and visual data extracted from the news content. Some of them focus on the correlation between the attached images and the credibility of the news text [95, 96, 149, 167, 175, 218, 263, 315, 324, 327, 331], while others only use one or the other data type [291, 321].

Xue et al. [315] propose a neural network approach for fake news detection named MCNN (Multimodal Consistency Neural Network), using a similarity measurement module to measure the similarity of multimodal data (text and images). Zeng et al. [324] define a fake news detection approach to comprehensively mine the semantic Correlations between text content and the attached image. Zhang et al. [327] propose an end-to-end model, named BERT-based domain adaptation neural network (BDANN) for multimodal fake news detection. Kumari et al. [149] propose an attention-based multimodal fake news detection framework named AMFB, with multimodal feature fusion that leverages information from text and image and tries to maximize the correlation between them to get the efficient multimodal shared representation. Mangal et al. [167] propose a fake news detection approach with the integration of embedded text cues and image features in which they extract text and objects available in the image and then check the similarity between them to find the fraud in a given information. Meel et al. [175] propose a multimodal fake news detection framework that unitedly exploits hidden pattern extraction capabilities from text and visual image features. Giachanou et al. [95, 96] propose multimodal multi-image systems that combine information from different modalities (textual, visual and semantic information) to detect fake news posted online. Singhal et al. [263] introduce a multimodal framework named SpotFake, which exploits both the textual and visual features of an article for fake news detection. Zhou et al. [331] propose a similarity-aware fake news detection method named SAFE, which investigates multimodal (textual and visual) information to recognize the falsity of news articles based on their text, images, or their « mismatches ». Qian et al. [218] proposes a hierarchical multimodal contextual attention network (HMCAN) for fake news detection by jointly modelling the multimodal context information (text and images) and the hierarchical semantics of text in a unified deep model. Yuan et al. [321] propose an approach named DAGA-NN to improve fake news detection with domain-adversarial and graph-attention neural networks. However, their approach is based on a text environment with multiple events/domains and not on multimodal data (text and image). Vishwakarma et al. [291] propose an approach to detect the veracity of information on various social media platforms available in the form of images. The veracity of image text is validated by exploring it on the web. Shah et al. [237] present a multimodal framework to detect fake news, without any further sub-task being taken into account, using a Cultural Algorithm with situational and normative knowledge.

In contrast, my system, "**EXMULF** (Explainable Multimodal Content-based Fake News Detection System)" [16], stands out as the first to use Vilbert for analyzing both image and text in fake news detection. This unique approach boosts both the accuracy and clarity in understanding how fake news is identified. Additionally, EXMULF was one of the initial systems to combine the multimodal detection method with clear explanations, making it a leader in offering a complete solution that is both multimodal and easily understandable for detecting fake news.

A comparison between these approaches with emphasis on the techniques and datasets used is provided in Table 6.

2.3.2. Explainable fake news detection

Although the progress in the detection of fake news has been significant, limited effort has been devoted to explainability. The integration of machine learning explanations is emerging as a promising direction to enhance transparency, particularly in applications like fake news detection within online social networks.

Machine learning explanations help to clarify the outcome of a machine learning model. It generates explanations and describes the reasoning behind the resulting decisions and predictions in order to enable users to understand how data is processed to reach a certain decision. For instance, the explainability in the detection of fake news consists in explaining the reasons why a given news has been detected as fake news. This prevents OSN users from further disseminating such fake content and thus limits the negative effects on the security of both individuals and society.

Explainable machine learning (ML) has emerged as a cutting-edge methodology in the detection of fake news. Significant works in this field, [150, 186, 215, 220] demonstrate the growing focus on integrating explainability into fake news detection models. Multiple

Table 6 – A comparison between the multimodal fake news detection approaches

Reference	Techniques used	Datasets used
Xue et al. [315]	BERT, ResNet50, cosine similarity	MCG-FNeWS, PolitiFact, Twitter
Zeng et al. [324]	VGG model, multimodal variational autoencoder	Twitter, Weibo
Zhang et al. [327]	BERT, VGG19	Twitter, Weibo
Kumari et al. [149]	ABS-BiLSTM, ABM-CNN–RNN, MFB	Twitter, Weibo
Mangal et al. [167]	VGG, Word2Vec, LSTM, cosine similarity	Collected 1000 images from Google, Kaggle, and onion for fake or real images with text
Meel et al. [175]	Hierarchical Attention Network (HAN), Caption and Headline matching (CHM), Noise Variance Inconsistency (NVI), Error Level Analysis (ELA)	Fake News Detection by Jruvika, All Data, Fake News Sample by Guilherme Pontes
Giachanou et al. [96]	BERT, VGG-16, cosine similarity	FakeNewsNet
Giachanou et al. [95]	Word2Vec, VGG19, LBP	MediaEval, PolitiFact, GossipCop
Singhal et al. [263]	BERT, VGG19	Twitter MediaEval, Weibo
Zhou et al. [331]	Text-CNN, Text-CNN, image2sentence, cosine similarity	PolitiFact, GossipCop
Qian et al. [218]	BERT, ResNet, attention mechanism	Twitter, Weibo
Yuan et al. [321]	BERT, VGG19, Bi-LSTM, Graph-attention layer	Twitter, Weibo
Vishwakarma et al. [291]	Optical Character Recognition (OCR), Web scraping	A dataset of thousands of images collected from Google Images, the Onion, and Kaggle
Shah et al. [237]	Sentiment Analysis, Cultural Algorithms (CA)	Twitter, Weibo
Amri et al. [16]	VilBERT, Latent Dirichlet Allocation (LDA), Visual Geometry Group Network VGGNet16, Similarity Analysis	Twitter, Weibo

researchers [34, 70, 160, 215, 220, 249, 260, 316] are trying to incorporate explainability in their prediction models for fake news detection tasks using multiple techniques and models. These techniques include but are not restricted to, neural network [249], co-Attention network [160], natural language processing (NLP) [70], network embedding learning [260], and Tsetlin machine (TM) [34]. These efforts underscore the importance of not only identifying false information but also providing transparent and understandable explanations for these detections.

In the realm of explainable detection, I have developed two distinct systems. The first, detailed in my paper "**EXMULF** (Explainable Multimodal Content-based Fake News Detection System)" [16], focuses on utilizing both textual and image-based information to provide explanations employing local interpretability (i.e., Local Interpretable Model-Agnostic Explanations (LIME)). This method significantly enhances the interpretability of fake news detection models, a fact that was substantiated by the experimental results.

In the second system, "**ExFake** (Explainable Fake News Detection Based on Content and Social Context and External Evidence Information)" [15], an explanation method inspired by Local Interpretable Model-agnostic Explanations (LIME) is featured. This approach is designed to provide clear and simple explanations to users for predictions in Natural Language Inference (NLI) tasks related to fake news detection. In ExFake, the explanations are designed to help users adopt certain behaviours on social networks by returning texts from trusted articles and legitimate tweets, highlighting words within these returned texts, and providing sources and publication timestamps. This strategy not only explains the prediction but also guides users in verifying the information before sharing it on online social networks.

A comparison between the reviewed approaches with emphasis on the techniques and datasets used is provided in Table 7.

Multiple studies on explainable machine learning are dedicated to investigating and evaluating existing fake news prediction models [8, 150, 186], including looking into which important features contribute to the models' prediction from the explainable machine learning perspective as shown in Table 8.

For instance, Alharbi et al. [8] evaluate the trustworthiness of three existing fake news detection models (i.e., DEFEND, TCNNURG, and HSFD) by using model-agnostic explainer (Captum, SHAP, and LIME,) to explain how the classification was made. Kurasinski et al. [150] claim that there is a lack of research concerning the explainability of a machine learning-based fake news detection model, while effort is mainly focused on its effectiveness. They investigate two classes of deep neural networks for the task of fake news detection (i.e., BiDir-LSTM-CNN, BERT), analyze them, and provide a deeper degree of explainability into the process. To explore how different types of explanations affect users in fake news detection, Mohseni et al. [186] designed four interpretable fake news detection algorithms (i.e., Bi-LSTM network, hierarchical attention network (HAN), Bi-LSTM teacher model with XGBoost student model, BiLSTM network with Word2Vec word embedding). Their designed algorithms are dedicated to evaluating model explanations from multiple perspectives (i.e., user engagement, mental model, trust, and performance measures). They report that adding

Table 7 – A comparison between the explainable fake news detection approaches

Reference	Approach	Techniques used	Datasets used
Shu et al. [249]	DEFEND	Attention neural network	PolitiFact, GossipCop
Reis et al. [220]	–	SHAP	BuzzFace
Yang et al. [316]	XFake	MIMIC, ATTN, PERT	An annotated benchmark dataset in the German language
Lu et al. [160]	GCAN	Co-Attention Network	Twitter datasets: Twitter15, Twitter16
Przybyła et al. [215]	–	Machine learning: linear method trained on stylometric features, a recurrent neural network method	Fake News Corpus dataset
Bhattarai et al. [34]	TM framework	Tsetlin Machine (TM)	PolitiFact, GossipCop
Denaux et al. [70]	–	NLP: semantic similarity and stance detection	Clef18, FakeNewsNet, coinform250
Silva et al. [260]	Propagation2Vec	Network embedding learning	PolitiFact, GossipCop
Amri et al. [16]	EXMULF	Local Interpretable Model-Agnostic Explanations (LIME)	Twitter, Weibo
Amri et al. [15]	ExFake	Word Importance Analysis	FakeNewsNet

explanations helped participants build appropriate mental models of the intelligent assistants in different conditions and adjust their trust accordingly.

Table 8 – Evaluation approaches

Reference	Evaluated Models	Used techniques	Used datasets
Alharbi et al. [8]	DEFEND, TCN-NURG, HSFD	Model-agnostic explainers (Captum, SHAP, LIME)	FakeNewsNet
Kurasinski et al. [150]	BiDir-LSTM-CNN, BERT	Summarization, Stemming, Lemmatization	Fake News Corpus dataset
Mohseni et al. [186]	Bi-LSTM, hierarchical attention network (HAN), Bi-LSTM with XGBoost, BiLSTM with Word2Vec	Attention mechanism	Snopes

2.4. Discussion

After reviewing the most relevant state-of-the-art for automatic fake news detection, a comprehensive classification is presented in Table 9. This classification is based on the detection aspects (i.e., content-based, contextual, or hybrid aspects) and the techniques used (i.e., AI, crowdsourcing, fact-checking, blockchain, or hybrid techniques). Hybrid techniques, in particular, refer to solutions that simultaneously combine different techniques from the previously mentioned categories (i.e., inter-hybrid methods), as well as techniques within the same class of methods (i.e., intra-hybrid methods), to define innovative solutions for fake news detection. A hybrid method is intended to bring together the best of both worlds.

Along with the classification, this discussion delves into the intricacies of each approach, exploring their advantages and challenges while also underscoring **my contributions** to these areas. It aims to encapsulate the nuances and innovations within each approach to fake news detection, reflecting my commitment to advancing this field with methods that are not only effective in detecting false information but also enhance user trust in a digitally complex information landscape.

2.4.1. News content-based methods

Most news content-based approaches to fake news detection view it as a classification problem, employing AI techniques such as classical machine learning (e.g., regression, Bayesian) and deep learning (e.g., neural methods such as CNN and RNN). These methods predominantly focus on text categorization, utilizing content features like words and hashtags for social media content classification, a fundamental task in social media mining. However, the main challenge in these approaches lies in feature extraction, deciding which features to use and how to reduce the data needed to train models for accurate results.

Researchers are motivated by the premise that news content is a central element in the deception process and provides a direct factor to analyze for predictive clues of deception. Yet, the complexity arises as fake news is often strategically created to mimic the truth, making it challenging, if not impossible, to solely rely on content for identifying useful features for detection. Furthermore, focusing only on news content overlooks the rich information and latent user intelligence in responses to previously disseminated articles, underscoring the importance of auxiliary information in effective fake news detection.

In this landscape, my work with the **EXMULF** system [16] introduces a significant advancement. Unlike conventional methods, EXMULF employs a multimodal strategy that simultaneously analyzes textual and visual elements of news content. This approach addresses the limitations of text-only analyses in detecting fake news, which often employs a combination of misleading text and images. The integration of Explainable AI (XAI) through Local Interpretable Model-agnostic Explanations (LIME) in EXMULF ensures transparent

Table 9 – Fake news detection approaches classification

	Artificial Intelligence			Crowdsourcing (CDS)	Blockchain (BKC)	Fact-checking	Hybrid
	ML	DL	NLP				
Content	Del Vicario et al. [68], Hosseinimotlagh and Papalexakis [120], Hakak et al. [111], Singh et al. [262], Amri et al. [16]	Wang [300], Hiriyannaiah et al. [119]	Zellers et al. [323]	Kim et al. [142], Tschiatschek et al. [280], Tchakounté et al. [273], Huffaker et al. [123], La Barbera et al. [151], Coscia and Rossi [64], Micallef et al. [178]	Song et al. [266]	Sintos et al. [264]	ML, DL, NLP: Abdullah-All-Tanvir et al. [2], Kaur et al. [140], Mahabub [164], Bahad et al. [23] Kaliyar [136] ML, DL: Abdullah-All-Tanvir et al. [1], Kaliyar et al. [137], Deepak and Chitturi [67] DL, NLP: Vereshchaka et al. [288] ML, NLP: Kapusta et al. [139], Ozbay and Alatas [200], Ahmed et al. [4] BKC, CDS: Buccafurri et al. [47]
Context		Qian et al. [217], Liu and Wu [158], Hamdi et al. [112], Wang et al. [301], Mishra [184]		Pennycook and Rand [209]	Huckle and White [122], Shang et al. [238]	Tchechmedjiev et al. [274]	ML, DL: Zhang et al. [326], Shu et al. [259], Shu et al. [256], Wu and Liu [309] BKC, AI: Ochoa et al. [198] BKC, SIR: Chen et al. [55]
Hybrid	Aswani et al. [20], Previti, et al. [213], Elhadad et al. [81], Nyow and Chua [197]	Ruchansky et al. [230], Wu and Rao [311], Guo et al. [106], Zhang et al. [325]	Xu et al. [314]		Qayyum et al. [216], Hasan and Salah [117], Tchechmedjiev et al. [274]	Yang et al. [316]	ML, DL: Shu et al. [251], Wang et al. [301] BKC, AI: Shae and Tsai [236], Jing and Murugesan [133]

and understandable decision-making processes. This is essential for building trust in AI systems, especially in sensitive areas like false information detection.

The use of advanced machine learning techniques such as Latent Dirichlet Allocation (LDA) and Vision-and-Language BERT (VilBERT) further differentiates EXMULF from standard techniques. These methods enable a more nuanced analysis, effectively overcoming the strategic deceptions often inherent in fake news content and surpassing the performance of 10 state-of-the-art models in real-world datasets. This comprehensive approach goes beyond the traditional perspective of fake news detection as a text categorization problem, incorporating both text and image analysis along with embedded explainability. EXMULF not only addresses the shortcomings of existing content-based methods but also sets a new standard in the field, emphasizing the need for accurate, reliable, and transparent news detection systems in today's digital information landscape.

2.4.2. Social context-based methods

Social context-based approaches in fake news detection explore data surrounding the news content, offering an effective alternative where content-based text classification may encounter limitations. These approaches typically focus on additional information from user behaviour and network diffusion patterns. However, most existing studies in this domain rely heavily on sophisticated machine learning techniques for feature extraction and often ignore the potential of other methods like web search and crowdsourcing, which can expedite early detection and identification of fake content.

In my work with systems like **ExFake** [15] and **MythXpose**, I've contributed to this category by focusing on social context features available before the spread of false information. ExFake leverages user historical sharing behaviour and incorporates external evidence from reputable third parties like fact-checkers and official sources (i.e., official social accounts), when available. MythXpose, conversely, focuses on predicting user *personality traits*. This *proactive* approach represents a significant shift from the predominantly reactive measures in existing methodologies, aiming to prevent the spread of false information rather than just responding to it post-factum.

The relevance of user *personality traits* in the context of false information is particularly noteworthy. Individuals participate in various stages of the false information lifecycle, from creation to dissemination [94]. The likelihood of users to accept or distribute false information is influenced by numerous factors, including their network characteristics, analytical reasoning, and cognitive abilities. Traditional methods of assessing personality traits often involve explicit questionnaires, but with advancements in Natural Language Processing (NLP), it is now possible to infer personality traits from user-generated text. Numerous studies treat personality detection as a classification task based on text and user-generated conversations [224].

Among the most widely recognized models for detecting personality traits is the Five-Factor Personality Model [261], also known as the Big Five. This model assesses human personality across five dimensions: Extroversion (EXT), Neuroticism (NEU), Agreeableness (AGR), Conscientiousness (CSN), and Openness (OPN). Incorporating this model into fake news detection, as done in MythXpose, provides a nuanced understanding of user behaviour and tendencies, contributing significantly to the early detection and prevention of the spread of false information. This integration of personality trait analysis with other social context-based methodologies in ExFake and MythXpose places my work at the forefront of innovative approaches in fake news detection.

2.4.3. Hybrid approaches

Hybrid approaches in fake news detection represent a significant advancement in the field, bringing together the strengths of both content-based and social context-based methodologies. These approaches aim to provide a comprehensive analysis of fake news by simultaneously considering the content of the news, such as text and images, and the contextual information based on user behaviour and network patterns on Online Social Networks (OSN). Despite their potential, hybrid models are inherently more complex, as they require the integration of diverse data types and analytical methods, a challenge highlighted in the survey by Bondielli et al. [39].

One of the primary complexities of hybrid approaches lies in data availability and feature selection. Effectively collating and integrating diverse datasets, which include both content information and user/network contextual data, poses a significant challenge. Moreover, selecting the most relevant and impactful features from these datasets to achieve precise results is critical. This complexity is not just in the volume of data but also in determining the most meaningful way to combine different types of information.

Another challenge in hybrid approaches is balancing the weight and relevance of content-based versus context-based information. This involves discerning which category of information provides the most significant insights for different scenarios of fake news detection. Deciding this balance requires continuous refinement and testing, as it varies depending on the nature of the news and the dynamics of the social networks through which it spreads.

In response to these challenges, my work in systems like **MythXpose** and **ExFake** demonstrates the potential of hybrid approaches. MythXpose, for example, combines advanced content analysis with user personality trait predictions, providing a nuanced view of how content interacts with user characteristics. Similarly, ExFake integrates historical user behaviour analysis with external evidence from trusted sources, offering a comprehensive approach to understanding and detecting fake news.

The future of hybrid approaches in fake news detection is promising. Their ability to provide a more holistic analysis by considering both the content of the news and its social context presents a powerful tool in the fight against false information. As the field evolves, improving data integration techniques and finding the optimal balance between different information sources will be key to enhancing the effectiveness of these systems.

Overall, hybrid approaches, while complex, offer a path toward more accurate and reliable fake news detection. By harnessing the combined power of content and context analyses, systems like MythXpose and ExFake pave the way for innovative solutions in this constantly evolving landscape.

2.4.4. Early detection

The rapid evolution and dissemination of fake news on OSN underscore the urgency and critical need for early detection mechanisms. This task, particularly challenging on dynamic social networks, is crucial in mitigating the spread of false information. Both news content-based and social context-based approaches encounter obstacles in achieving early detection of fake news.

Content analysis-based approaches, while somewhat more resilient in this regard, still struggle with limitations due to the scarcity of verifiable information in the initial phases of news spread. These methods often require a more substantial body of content for accurate verification, which is not always available in the early stages.

Yet, contextual analysis-based approaches frequently face more significant challenges in early detection, as they typically depend on data that emerges post-dissemination, such as social engagement metrics, user responses, and propagation patterns. This reliance often results in a reactive rather than *proactive* stance in identifying fake news.

To address these challenges, my systems, **EXMULF** [16], **ExFake** [15], and **MythXpose**, take innovative approaches to early detection. By employing a combination of advanced multimodal content analysis (including both textual and visual information), predictive analytics based on user historical sharing behaviour and personality traits, and integrating external verification from reputable fact-checkers and official sources (i.e., official social accounts), these systems can identify potential false information before it gains widespread traction. This multifaceted approach enables the detection of fake news in its nascent stages, leveraging available data more effectively and efficiently.

Furthermore, the integration of trusted human verification, alongside the strategic use of historical and real-time data, enhances the capability of these systems to detect fake content proactively. By doing so, my work contributes significantly to the field, moving beyond the limitations of traditional content and context-based methods and paving the way for early detection strategies in the fight against fake news.

2.4.5. Enhancing trust through explanations and consensus

In the domain of fake news detection, establishing trust is just as vital as ensuring the accuracy of detection. My contributions with systems like **EXMULF** [16], **ExFake** [15], and **MythXpose** focus heavily on this aspect, providing clear and understandable explanations for their detection decisions to OSN users. In addition to these systems, I have developed the Automated Fact-checkers Consensus and Credibility (**AFCC**) system, a distinct and innovative system dedicated to building *consensus* and assessing the *credibility* of fact-checkers. This combination of *explainability*, *consensus-building*, and *credibility assessment* ensures that the detection of false information is not only effective but also enhances the confidence and trust of the users these systems serve. Below, I detail the key strategies and features of these systems that contribute to building and maintaining user trust:

2.4.5.1. **Explanations in EXMULF, ExFake, and MythXpose:** These systems employ Explainable AI (XAI) techniques, such as Local Interpretable Model-agnostic Explanations (LIME), to offer users transparent and comprehensible explanations for AI decisions. This feature plays a crucial role in building user trust by demystifying (i.e., clarifying and making transparent) AI processes and enabling users to understand the rationale behind the identification of news as fake.

2.4.5.2. **Consensus and credibility assessment with AFCC:**. The AFCC system's primary function is to analyze and standardize verdicts (i.e., rating labels) from various fact-checking organizations. It evaluates the consensus among these sources and assigns credibility scores based on their historical accuracy and reliability. This ensures that the conclusions about the veracity of news are both unified and unbiased.

AFCC serves as a complementary tool to the direct fake news detection capabilities of EXMULF, ExFake, and MythXpose. While these systems engage in content analysis, social context scrutiny, and external evidence evaluation to detect fake news, AFCC introduces an additional verification layer. It focuses on the meta-analysis of outcomes from fact-checking organizations, thereby enhancing the overall trust in the detection process.

The integration of EXMULF, ExFake, and MythXpose with AFCC combines the explanatory capabilities of these systems with consensus and credibility assessments. This powerful combination empowers users with enhanced knowledge and transparency, enabling them to more critically assess news content.

The integration of explanatory features in EXMULF, ExFake, and MythXpose, combined with the consensus and credibility assessment capabilities of AFCC, ensures that users not only receive accurate and comprehensive information but also the information they can trust.

This fortifies the reliability of these systems, making them more effective in the fight against fake news.

2.5. Conclusion

In this chapter, I have provided a comprehensive review of the state-of-the-art in the automatic detection of fake news based on the categories of the adopted approaches (i.e., news content-based approaches, social context-based approaches, or hybrid approaches) and the techniques that are used (i.e., artificial intelligence-based methods; crowdsourcing, fact-checking, and blockchain-based methods; and hybrid methods). A comparative study between these diverse works was conducted to highlight their strengths and weaknesses.

This chapter has not only provided a comparative analysis of these diverse approaches but has also woven in the critical themes of explainability and multimodality. By delving into how classification models can be made more transparent through explainable AI (XAI), and how the analysis of multimodal content (i.e., text and images) enhances detection capabilities, this discourse has underscored the necessity of multi-faceted and interpretable solutions in the realm of fake news detection.

As I conclude this chapter, the insights garnered underscore the complex and multifaceted nature of fake news detection. The fight against fake news in online social networks demands a comprehensive approach that considers:

- The importance of integrating both content and contextual information, recognizing the synergy between the news material and user behaviour for enhanced detection.
- The significance of historical behaviour patterns of users, including their personality traits, which are crucial for the early detection and mitigation of fake news.
- The necessity of integrating external evidence sources, such as fact-checking organizations and official sources including official social accounts, to assess content credibility and enhance the accuracy of fake news detection.
- The need for transparent and interpretable solutions through explainability techniques, ensuring that the decision-making processes of fake news detection models can be easily understood by OSN users.
- The importance of consensus inference between distinct fact-checkers to enhance trust-based fake news detection by leveraging the collective wisdom of multiple fact-checkers regarding the same fact-checked news/claim.

These objectives collectively form a comprehensive strategy for addressing the complex challenges posed by fake news in online social networks. The subsequent chapters will build upon these foundations, introducing the *FACTS-ON* framework, which stands for **F**ighting **A**gainst **C**ounterfeit **T**ruths in **O**nline social **N**etworks, and showcasing how it addresses these challenges. The framework represents a *holistic* approach to combating fake news in online social networks, incorporating advanced techniques and methodologies to enhance the *accuracy*, *transparency*, and *effectiveness* of fake news detection.

FACTS-ON consists of different modules (i.e., the Analyser Module (AM), Content Module (CM), User Module (UM), Search Module (SM), and Explainable Decision Module (EDM)) implemented in various system components, namely *EXMULF*, *MythXpose*, *ExFake*, and *AFCC*. Each module is dedicated to a specific aspect of fake news detection, covering *explainable multimodal content-based* detection; *explainable multimodal content and social context-based* detection; *explainable content, social context, and external evidence-based* detection; as well as *fact-checkers' consensus inference and credibility assessment for trust-based* fake news detection, respectively. Building on the insights gained in this chapter, the following chapters will start by introducing a description of my solution, FACTS-ON, in the next chapter. Subsequently, each system component (i.e., EXMULF, MythXpose, ExFake, and AFCC) will be presented, explained, and evaluated in the following chapters, ensuring a comprehensive and detailed examination of the solution's functionality and effectiveness.

Chapter 3

FACTS-ON in Theory: Combating False Information in Online Social Networks

3.1. Introduction

This chapter introduces the FACTS-ON framework, an innovative and strategic solution developed to combat the spread of false information, such as fake news, misinformation, and disinformation, within online social networks (OSN). FACTS-ON, an acronym for **F**ighting **A**gainst **C**ounterfeit **T**ruths in **O**nline social **N**etworks, emerges as a holistic approach, inspired by the thorough study of state-of-the-art methods and motivated by the findings presented in the Discussion Section of Chapter 2. It goes beyond analyzing content and context. It uniquely integrates *multimodal content analysis* (of both *text* and *images*), *social-context analysis*, and *external verification* from *trusted sources*, including *fact-checking websites* and *official social accounts*. This pioneering combination has been proven to enhance accuracy and reliability, as evidenced by experimental results, with a focus on *explainability* to ensure transparency and user comprehension, influencing how OSN users interact with and share information.

This chapter offers a detailed overview of FACTS-ON, exploring its theoretical foundation, architectural principles, and various modules and components. This is followed by the technical background, elaborating on the artificial intelligence (AI) and machine learning (ML) tools integral to FACTS-ON. The chapter then transitions into a comparative analysis, highlighting FACTS-ON's unique contributions and advantages in the field of fake news detection.

The subsequent chapters will provide detailed insights into each system component of the FACTS-ON framework, elucidating the individual functionalities and contributions of each to the overarching strategy of FACTS-ON.

3.2. FACTS-ON general architecture

The detection of false information, such as fake news, misinformation, and disinformation in online social networks is a complex and multidimensional problem, particularly in today's era of information overload and digital manipulation. Addressing this multifaceted issue requires a multi-criteria approach that takes into account various factors including content analysis, contextual understanding and external evidence. To address this, the FACTS-ON framework was developed, following a thorough study of state-of-the-art methods and motivated by findings in the Discussion Section of Chapter 2. FACTS-ON employs a multimodal approach that not only considers textual content but also incorporates image-based information. It is then a multimodal framework integrating content, context, external evidence, and explanatory insights to enhance the detection and understanding of false information. FACTS-ON provides valuable explanations, which play a pivotal role in improving users' discernment of fake news. Such explanations can be crucial to reflect news credibility, raise OSN users' awareness and ultimately influence their behaviour in protecting the security and privacy of both individuals and society.

The fundamental modules of FACTS-ON are shown in Figure 1, which depicts the general architecture of the framework. FACTS-ON is composed of *five essential modules*: the **Analyzer Module (AM)**, **Content Module (CM)**, **Search Module (SM)**, **User Module (UM)**, and **Explainable Decision Module (EDM)**. This chapter provides detailed insights into each of these modules, elucidating their respective functionalities and contributions to the framework's overall objective.

FACTS-ON is a comprehensive multimodal framework comprising various components that employ diverse criteria for false information detection. It integrates content and context-based analyses, along with external evidence such as web search and human expertise, to deliver a thorough assessment of information veracity. Moreover, FACTS-ON provides explanatory insights to OSN users, attempting to aid their comprehension of why specific information is labelled as false, thus promoting a more discerning approach to news and information consumption in the future. To address these aspects effectively, FACTS-ON employs the following modules:

— **Analyzer Module (AM)**: The initial stage in detecting false information involves the extraction of pertinent data, a crucial process that equips the other modules with valuable insights and highlights nuanced distinctions in identifying misleading content.

— **Content Module (CM)**: Given that the content of online posts, encompassing both text and images, is often the primary carrier of false information, the Content Module (CM) specializes in the comprehensive analysis of this multimodal content. Recognizing the significance of both textual and visual elements within online posts as potential sources of deception, the CM is tailored to extract valuable insights from this

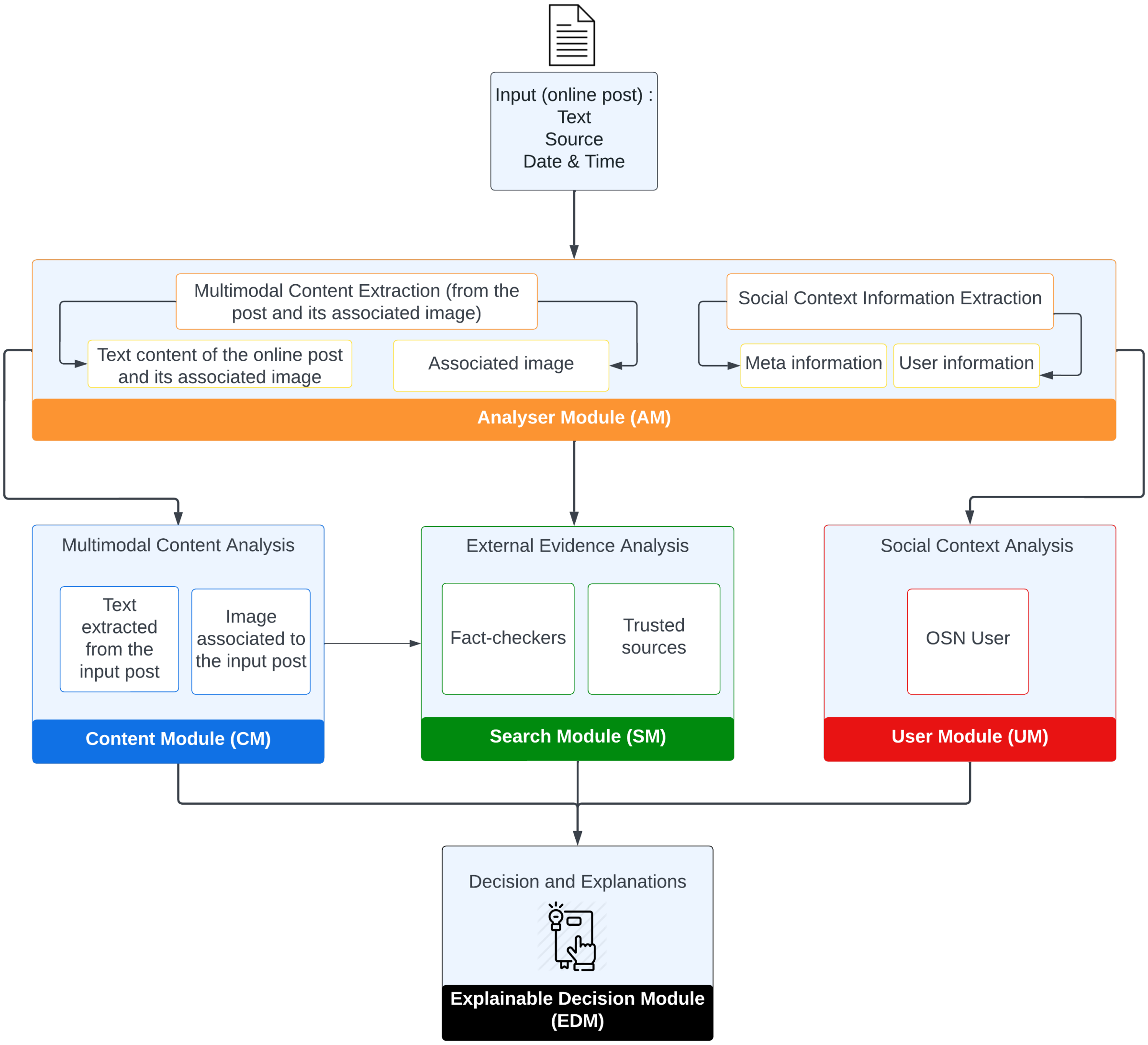

Input (online post) :
Text
Source
Date & Time
Multimodal Content Extraction (from the post and its associated image)
Text content of the online post and its associated image
Associated image
Social Context Information Extraction
Meta information
User information
Analyser Module (AM)
Multimodal Content Analysis
Text extracted from the input post
Image associated to the input post
Content Module (CM)
External Evidence Analysis
Fact-checkers
Trusted sources
Search Module (SM)
Social Context Analysis
OSN User
User Module (UM)
Decision and Explanations
Explainable Decision Module (EDM)

Figure 1 – FACTS-ON general architecture overview

combined content. Unlike auxiliary information (e.g., social engagement, user response, propagation patterns), this multimodal content is available in the *early stages* and can reveal critical insights. By analyzing the latent relationships embedded within the text and image-based content, the CM unleashes the full potential of content-based detection.

— **Search Module (SM)**: External evidence including web search and trusted fact-checkers plays a pivotal role in assessing content credibility. This is crucial because false information is often designed to closely resemble the truth, making it difficult to determine its veracity through artificial intelligence and machine learning techniques alone. Additional information from trusted third parties and human experts is necessary. However, relying only on fact-checkers is not enough. Therefore, it is crucial to apply web search and query trusted official sources (i.e., official social accounts) for fake news detection.

— **User Module (UM)**: The User Module is crucial for analyzing data from users who are involved in disseminating false information. This includes *early-stage* data like *historical sharing behaviour* and *personality traits*, which can be key for early detection of fake news. Research shows a strong link between OSN user activities and the spread of fake news, misinformation, and disinformation. By examining these early indicators, the UM aids in promptly identifying and mitigating the dissemination of false content, enhancing the framework's overall effectiveness in combating false information.

— **Explainable Decision Module (EDM)**: This module extends fake news classification models by incorporating explainable AI (XAI) and visual analytics to help OSN users understand the basis of classification results, facilitating explainable fake news detection.

To address the *multimodal content*-related aspect, the Content Module (CM) has been dedicated to the analysis of *text* and *image* content within online posts. The **contextual** aspect is managed by the User Module (UM), which analyzes information derived from *OSN users* who are engaged in disseminating the post. For the utilization of *external evidence*, the Search Module (SM) relies on links from *trusted entities* (i.e., official sources including official social accounts) and *reputable fact-checkers*. Ultimately, the Explainable Decision Module (EDM) delivers the final decision regarding the nature of the post, along with suitable *explanations* for OSN users.

FACTS-ON comprises *four integral system components*: **EXMULF**, **MythXpose**, **ExFake**, and **AFCC**, as illustrated in Figure 2. Each component specializes in a distinct form of analysis to detect false information. **EXMULF** conducts *multimodal content analysis*, examining both text and images. **MythXpose**, which synergizes EXMULF with the PERSONA module, performs both *content* and *social context analysis*; the latter assesses user personality traits to enhance detection accuracy. **ExFake** represents the culmination of the framework, integrating *content* analysis, *social context* insights, and *external evidence* from trusted entities (i.e., reputable *fact-checkers* and trusted *official sources*, namely the *official social accounts*) for a comprehensive approach. Lastly, **AFCC** focuses on *external evidence* analysis, specifically standardizing and assessing the *credibility* and *consensus* of various *fact-checking organizations*. The subsequent chapters provide a detailed exploration of these components and their contributions to the FACTS-ON framework.

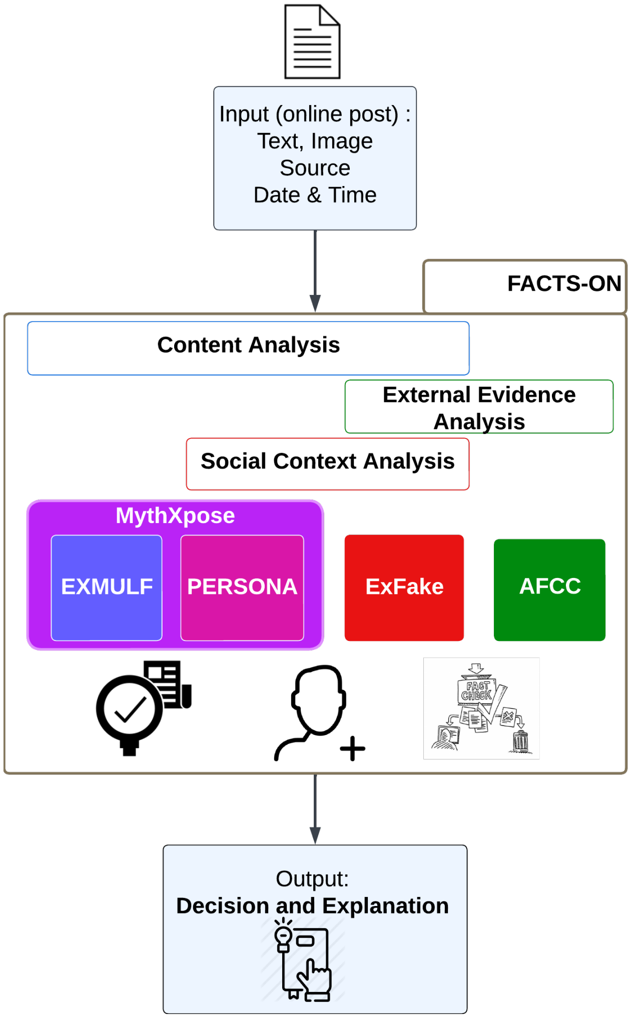


Figure 2 – FACTS-ON system components overview

3.3. FACTS-ON modules overview

3.3.1. Analyzer Module (AM)

Detecting fake news on social media requires handling both content-related information and social context information. My approach involves defining a module that can extract relevant patterns from large-scale social media data. This encompasses analyzing online post content (i.e., texts and images), user behaviour (notably their content sharing history and personality traits), and spatiotemporal details like the time and date of publication. These patterns are extracted from the online post and its surrounding social context data and are used in various modules of FACTS-ON.

The Analyzer Module (AM) is then motivated by the need to extract relevant information, which is the first step toward defining an efficient solution for fake news detection. The extracted data provides useful information that can help the other modules accomplish their tasks effectively and reflect subtle differences in fake news detection.

3.3.1.1. **Analyzer Module (AM) overview:** The task of processing online posts and extracting critical information is delegated to the Analyzer Module (AM). This module ensures the initial transformation of raw content into structured data that can be further analyzed by other components of the framework. The Analyzer Module (AM), as illustrated in its architectural representation in Figure 3, takes as input the online post, encompassing the news content, its source (e.g., user profile, specifically the profile of the user who shared the online post), and supplementary post-related details, including date and time, likes, shares, comments, and more. The Analyzer Module (AM) provides a comprehensive output that encompasses the following key elements:

(1) Text Content of the Online Post: The AM extracts the text content from the online post, including text from the post itself as well as any text found within the associated image.

(2) Associated Image: The output comprises the associated image related to the online post.

(3) User Information: The AM extracts user-centric information from the online post, focusing on historical sharing patterns and some profile information for user identification.

(4) Meta-Information: In the context of social media posts and digital content, meta-information refers to data that provides context about the primary content but is not part of the main message. This includes details like publication date and time, the digital format of the post, the device used for posting, and potentially the geographical location if available. These elements help in understanding the context and background of the post, which is vital in assessing its authenticity and relevance. *In FACTS-ON*, the analyzer module extracts additional valuable meta-information, specifically the post's publication date and time.

3.3.1.2. **Analyzer Module (AM) behaviour:** The Analyzer Module (AM) has a multi-faceted role within the FACTS-ON framework, encompassing two primary functions:

(1) Multimodal Content Extraction: Multimodal content extraction is useful in scenarios where textual information is embedded within multimedia content, such as images, or when a graphical image is associated with an online post. This capability enables the framework to extract meaningful information from diverse sources, enhancing its overall effectiveness in analyzing and understanding online content. Specifically, the AM is responsible for extracting content from online posts, which includes text from both the post itself and any associated images containing text, but only if applicable

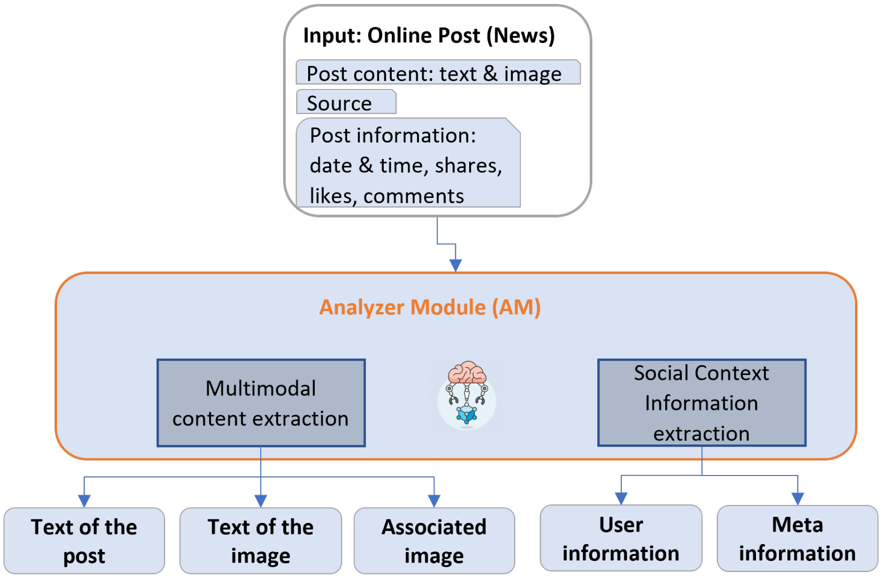

Input: Online Post (News)
Post content: text & image
Source
Post information: date & time, shares, likes, comments
Analyzer Module (AM)
Multimodal content extraction
Social Context Information extraction
Text of the post
Text of the image
Associated image
User information
Meta information

Figure 3 – Analyzer Module (AM) architecture overview

(i.e., this only applies if there is indeed text present in the associated images). This is performed by using machine learning and natural language processing techniques to extract meaningful text from different modalities and make it accessible and usable for analysis, search, topic modelling and other tasks. It also extracts the associated image, but this is done only if applicable and if the image consists of graphical content rather than text within an image. This means the extraction is carried out if there is an image associated with the online post. The AM is designed to handle both types of images: those containing primarily graphical content (i.e., images containing graphical content only without embedded text) and those with text embedded within them. Thus, regardless of whether the image is a graphical illustration or includes text content, the AM will extract it for further analysis.

(2) Social Context Information Extraction: Additionally, the AM extracts valuable social context information from online posts, such as user-related data and relevant metadata to improve detection performance. This encompasses details about the post's source (e.g., user profile), date and time of posting, engagement metrics (e.g., likes, shares, comments), and historical user behaviour. This social context information provides crucial contextual insights.

In summary, the Analyzer Module (AM) serves as a content-based data extraction and contextual analysis component within the FACTS-ON framework. It captures both text and image-based content from online posts, but only if applicable (i.e., if an image is indeed present in the online post). Additionally, it retrieves important social context information. The data it extracts is then used by other modules to perform their respective tasks, ultimately contributing to the framework's ability to detect false information in online social networks.

3.3.2. Content Module (CM)

False information can be difficult to identify, but the content of the post itself can provide important clues for detection. The content of the online post is a significant source for detection tasks, as it is available in the *early stages* and contains clues to distinguish between fake and real information. The Content Module (CM) is motivated by the *multifaceted* challenge of detecting false information, encompassing *text-based* and *multimodal* content. It recognizes the importance of analyzing both textual and visual elements, including associated images, to comprehensively assess the veracity of online posts. By combining these approaches, the CM aims to provide a more effective and thorough means of identifying deceptive content.

3.3.2.1. **Content Module (CM) overview:** The Content Module (CM) whose architecture is shown in Figure 4 is a crucial component of the framework designed for *multimodal content-based* false information detection, which includes addressing fake news, misinformation, and disinformation. It handles diverse inputs, encompassing both *textual* and *visual* (i.e., image) elements associated with the post. Additionally, the CM is responsible for extracting *named entities* from the content (i.e., Named Entities Recognition (NER)). Named entities refer to specific, identifiable names or terms within the text, such as the names of people, places, organizations, dates, and other proper nouns. These entities are crucial as they contribute to the overall assessment of potentially deceptive online content by the Search Module (SM). Identifying these named entities helps in contextualizing the content and provides significant insights into its authenticity.

The primary objective of the Content Module (CM) is to conduct a thorough *multimodal* analysis of online posts. This analysis encompasses the examination of the post, analyzing the text extracted from the associated image (if applicable), and evaluating the image associated with the post (also if applicable). Through this comprehensive approach, the CM effectively detects false information, playing a key role in the framework's capability to identify and address multimodal false information detection in online social networks.

3.3.2.2. **Content Module (CM) behaviour:** The Content Module (CM) in the FACTS-ON framework demonstrates a complex behaviour, characterized by several essential functions:

(1) Topic Modelling: CM initiates by conducting topic modelling on both the textual content extracted from the online post and the text found within its associated image (if applicable), and any associated graphical image. This process helps uncover underlying themes or topics within the content (i.e., in both the textual and visual content), offering a profound understanding of the post's overall context.

(2) Similarity Analysis: Following topic modelling, CM proceeds to assess the similarity between the topics derived from the textual content and those extracted from the associated graphical image. This comparison is instrumental in determining the

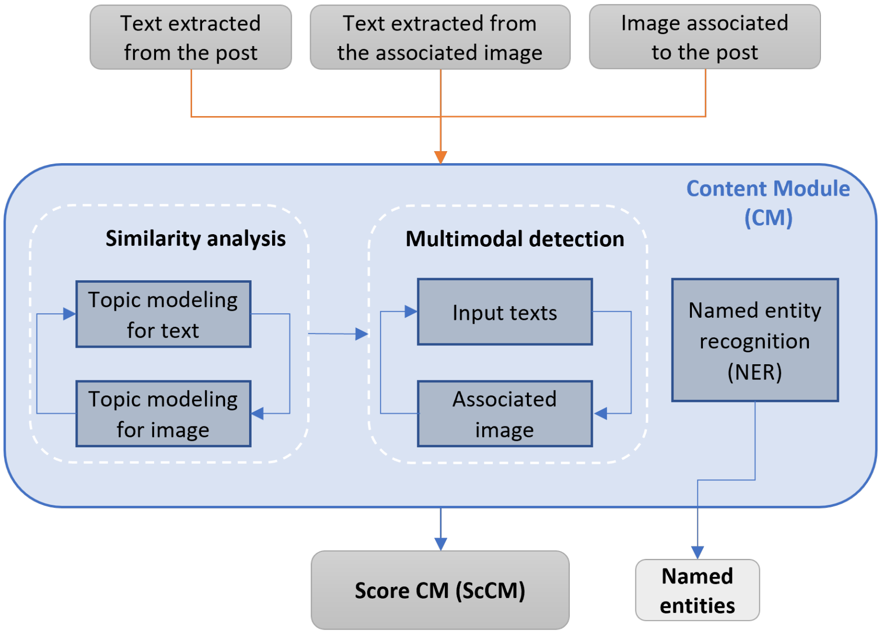


Figure 4 – Content Module (CM) architecture overview

alignment (i.e., correlations) or divergence (i.e., inconsistencies) between textual and image-based information, aiding in the identification of potentially false information.

(3) Multimodal Detection: CM employs Multimodal Detection techniques, leveraging both the input texts and the associated image. Analyzing the convergence or divergence between textual and visual content contributes to the overall evaluation of content authenticity. This step plays a pivotal role in detecting instances where text might contradict or misalign with the visual content.

(4) Named Entities Recognition (NER): One of the critical functions of CM is recognizing and identifying named entities within the content, such as names of individuals, organizations, or locations. This information is crucial for further analysis and verification by the Search Module (SM).

(5) Output Score: As a final output, CM calculates a Score CM "ScCM", representing the truthfulness percentage of the input post. This score quantifies the likelihood that the content is truthful, based on the comprehensive analysis of textual and visual elements. The ScCM score is presented as a percentage, with higher percentages indicating a greater probability of the content being genuine.

The culmination of these functions enables the CM to generate a Score for the input post, also referred to as the *ScCM*. This score is presented as a percentage, indicating the likelihood of the input post's authenticity. A score close to 0% suggests that the post is highly likely to be fake, whereas a score approaching 100% indicates a high probability of the post being genuine. The percentage-based score provides a more nuanced understanding of the post's credibility, reflecting a spectrum of authenticity rather than a binary classification.

3.3.3. User Module (UM)

Recently researchers [130, 256, 257, 259] have begun to explore the connection between user profiles and fake news. They aim to demonstrate how user profile characteristics can be used to identify fake news and validate the effectiveness of these characteristics by analyzing their importance in the classification task.

In line with this research, I propose an approach that integrates the analysis of user information, such as user historical sharing behaviour, and personality traits to enhance the early detection of fake news. By examining not only the types of news users typically share on social media but also their personality traits, this strategy can enhance early detection of fake news and may be a promising approach to identifying fake content. The rationale for this approach is grounded in the established link between user profile data and the detection of fake news [256, 259].

Existing detection methods predominantly depend on network-based information, such as user interactions with news articles on social networks. However, such data is only accessible postfactum, after content has already achieved wide circulation and been exposed to a large number of users. My method in the UM diverges from this by utilizing historical user behaviour towards prior articles and combining this with an assessment of users' personality traits. This dual analysis provides insights into users' tendencies to disseminate false information, thereby facilitating the identification of potential false information before it becomes widespread. By evaluating both historical user behaviour and inherent personality traits, FACTS-ON aims to predict and mitigate the spread of fake news more effectively than existing network-reliance techniques.

Analyzing the past social behaviour of users, such as the frequency and type of news they share, can offer valuable insights into their likelihood of sharing fake news. This methodology enhances the identification of fake news in its early stages, and it represents a promising approach to detecting false content. This approach is advantageous over existing methods that mostly depend on network-based information, which becomes available only after the widespread circulation of an article, as it considers valuable data and user intelligence hidden in their past behaviour towards false content.

The personality prediction task is added to the User Module (UM) to enhance its capabilities. This task involves analyzing the personality traits of users based on their shared content, providing a deeper understanding of their tendencies and motivations. It is particularly relevant when the same text is shared by multiple users, as it allows the UM

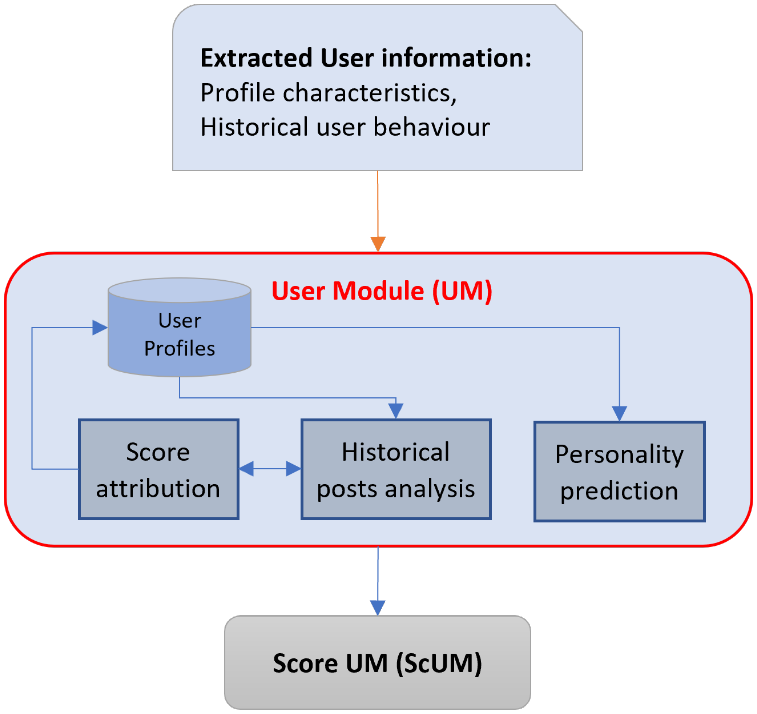


Figure 5 – User Module (UM) architecture overview

to discern personality differences that might influence the likelihood of spreading fake news, despite the shared content.

3.3.3.1. **User Module (UM) overview:** The task of detecting false information based on social context information, particularly on user information is assigned to an intelligent module called User Module (UM), as shown in Figure 5. Diverging from traditional methods that rely solely on user attributes like location, username, or job title, or on network-based information such as comments and propagation patterns, the UM adopts a distinctive perspective. It attributes a *user score* based on a comprehensive analysis of users' historical sharing behaviour across various dimensions. This user score reflects individual user attributes alongside their past engagement with online content, ultimately providing a more robust assessment of their potential involvement in disseminating false information. Additionally, the UM incorporates personality trait analysis to enhance its capability in the early detection of false information.

3.3.3.2. **User Module (UM) behaviour:** The User Module (UM) carries out three primary tasks as part of its behaviour. First, it analyzes the historical posting behaviour of users, second, it computes a user score based on this analysis, and third, it incorporates an individual personality trait analysis of the OSN user. These crucial functions play a pivotal role in the UM's ability to assess the likelihood of a user spreading false information.

(1) Historical Posts Analysis: The UM delves into the past posting behaviour (i.e., sharing behaviour) of users to identify patterns related to the veracity of their content. By analyzing the type of content a user has previously shared (real or fake), as well as

the frequency and ratio of genuine versus false content shared, the UM assigns a trustworthiness score to each user.

(2) User Score Attribution: The UM employs a unique scoring mechanism that considers multiple facets of user behaviour (i.e., posting history, content sharing patterns, frequency of posting). Namely, the posting history involves the nature of a user's posts, such as articles, images, videos, or comments. The content sharing patterns involve the type of content a user frequently shares (real or fake). The frequency of posting looks at how often a user posts. This holistic user score diverges from conventional approaches that rely solely on user attributes (e.g., location, username, or job title). It encapsulates the user's historical engagement with credible and deceptive content, offering a more nuanced assessment of their credibility.

(3) Personality prediction: The UM incorporates a personality prediction task, which involves analyzing the user's shared content to infer their personality traits. By leveraging advanced machine learning algorithms, this task assesses traits such as openness, conscientiousness, extraversion, agreeableness, and neuroticism, based on the user's posting on social media. This analysis provides deeper insights into the psychological profile of the user, aiding in understanding the motivations behind their content-sharing behaviour. The integration of these personality traits into the UM's evaluation process enhances FACTS-ON's ability to discern users who are more likely to spread fake news, based on their psychological predispositions. This component not only adds another layer to the user's credibility assessment but also enriches the UM's predictive capabilities, making it more promising in the early detection of false information.

(4) Output Score: Finally, UM calculates a Score UM "ScUM", quantifying the likelihood of a given OSN user sharing false content. This score is derived from an analysis of the user's historical sharing behaviour and personality traits. The ScUM score is presented as a percentage, with higher percentages indicating a greater likelihood of the user disseminating false information.

Through these functions, UM not only assesses historical user sharing behaviour and personality traits but also provides a quantifiable score (ScUM) indicating the propensity of users to spread false information. This score is calculated as a percentage, indicating the likelihood or propensity of users to spread false information. By offering a percentage-based score, the UM provides a more detailed assessment, recognizing the varying degrees of a user's inclination towards sharing false information. This comprehensive and nuanced approach is integral to the FACTS-ON framework, enhancing its capability in the early detection and mitigation of false information through a more precise evaluation of user behaviours.

3.3.4. Search Module (SM)

Automatic false information detection remains a huge challenge, primarily because the content is designed in a way that extremely resembles the truth to deceive users, and it is often hard to determine its veracity by artificial intelligence and machine learning techniques alone without additional information from trusted third parties. Hence, besides detecting fake information based on the content and social context, namely user behaviour, the intervention of trusted and skilled humans is required.

Therefore, some researchers are trying to overcome this issue by trying to bring human expertise into the challenging mission of fake news detection. Notable among these efforts are researchers [64, 123, 151, 273] that delve into the concept of "wisdom of the crowds", also known as "crowd intelligence" or "crowd signals" to identify and verify the authenticity of news content, social media posts, or other forms of online information. Crowdsourcing is a sourcing model in which individuals or organizations use contributions from Internet users also called "crowd workers" to obtain needed services or opinions. Crowd workers may come from various backgrounds and expertise levels, and they contribute their judgments on the accuracy and trustworthiness of the content. The process typically involves presenting crowd workers with pieces of information and asking them to determine whether the content is genuine or fake. Their assessments are then aggregated to make informed decisions about the accuracy of the content. Crowdsourcing-based false information detection can be conducted through dedicated platforms or online tools that facilitate the collection and analysis of crowd judgments.

However, this approach comes with challenges, including quality control, subjectivity, scalability, and biases among crowd workers. More specifically, crowdsourcing-based false information detection presents challenges stemming from the varying quality and subjectivity of contributions. Ensuring accurate and consistent assessments from diverse crowd workers can be complex, and potentially influenced by personal biases. While effective for small-scale tasks, scalability becomes an issue when dealing with the vast volume of online content. Additionally, the cost and time associated with crowdsourcing, especially for expert input, can be prohibitive. Limited expertise among crowd workers, susceptibility to adversarial attacks, and difficulties in labelling ambiguous content further add to the limitations. As disinformation tactics evolve, keeping up with changing trends becomes a constant challenge. Generalization to different contexts, privacy concerns, and the potential lack of context for accurate judgments contribute to the multifaceted limitations of this approach.

Indeed, several highly reputable *fact-checking organizations* are included in the International Fact-Checking Network's (IFCN) signatories list [1]. IFCN is an organization aiming to promote best practices in fact-checking and provides a place for collaboration between

[1] `https://ifcncodeofprinciples.poynter.org/signatories`, last access date: 11-09-2023.

fact-checkers worldwide which lists only trusted fact-checking websites that have high reputations (such as Africa Check, the Ferret Fact Service, factscan.ca, politifact.com, snopes.com, checkyourfact.com, truthorfiction.com, Reuters, etc.) who are working on verifying information and analyzing statements and claims on the web to mitigate the negative impact of fake news.

Therefore, I believe that highly *reputable fact-checking organizations*, along with *trusted official sources* on the web and in social networks (i.e., named entities) such as verified official social accounts and websites (e.g., newspaper or news magazine websites, TV and radio websites, Journalists websites and social accounts, celebrity official social accounts, governmental authorities' official websites and social accounts, etc.), can be used as *external evidence* to help check the credibility of a given news content. Specifically, today with the advance of the web and social media, it is possible to reach almost everybody out there and check the credibility of a given news directly from its "original" source. As a result, external evidence is utilized for fake news detection in the Search Module (SM) to verify whether the news in the online post has been *published* by a trusted official source or has been *verified* by a reputable fact-checking organization. Additionally, factors such as *analyzing the publication date* of news are employed to assess news credibility.

In other words, the main idea and motive behind the Search Module (SM) are to process the web, check the trustworthiness of the sources, and verify the news directly from its official source in case the news would not have been verified by fact-checkers. Thus, news credibility is analyzed based on *external news websites*, and not just based on social media websites.

Researchers have begun to utilize external evidence in their study of social media mining and fake news. To this end, they have employed various tools and techniques such as the Google search engine [171, 211], which is used to train models [186] and evaluate the credibility of claims. Web scraping techniques are also utilized to verify the accuracy of image text [291] and to search for information that can be presented in a more personalized way based on user preferences [74]. Additionally, some researchers leverage external knowledge from sources like Wikipedia to enhance their analyses [121, 277],while others use evidence from scientific articles to verify scientific claims' veracity [296].

In the FACTS-ON framework, the use of fact-checkers and official social accounts as external evidence offers distinct advantages over other methods like crowdsourcing. Fact-checkers provide reliable and expert-driven insights, ensuring consistency and standardization in information verification. This approach is time-efficient, significantly reducing the time taken to verify information compared to collecting and analyzing crowdsourced data. Moreover, leveraging official sources mitigates the risk of bias that can arise in crowdsourcing, where diverse crowd opinions may reflect personal biases. Additionally, the scalability of fact-checkers and official accounts allows for covering a wide range of topics more effectively than crowdsourcing.

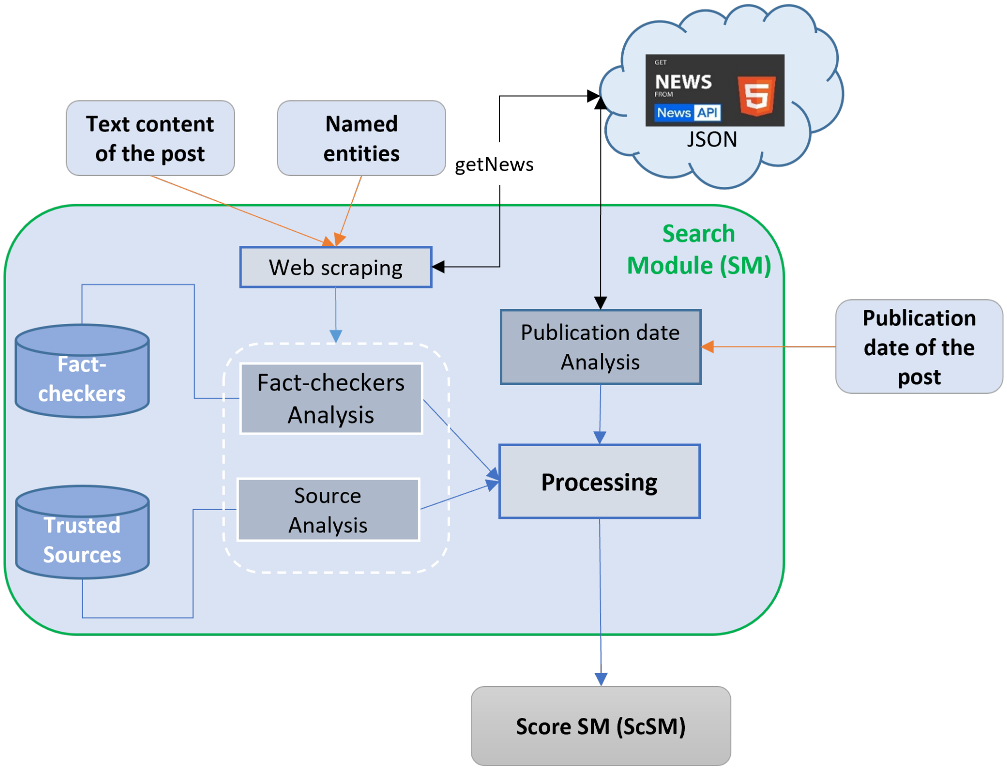


Figure 6 – Search Module (SM) architecture overview

Therefore, to detect fake news, I utilize *external evidence* in the Search Module (SM) to verify if the news has been published by a trustworthy *official source* or has been confirmed by a reliable *fact-checking* organization. Alongside external evidence, I also consider other factors, including the credibility analysis of news based on its *publication date*.

3.3.4.1. **Search Module (SM) overview:** The task of external evidence-based false information detection is deligated to an intelligent module called Search Module (SM) whose architecture is shown in Figure 6. This module focuses on assessing the credibility of an online post by utilizing a combination of three key factors: highly reputable fact-checking organizations, trusted sources, and publication dates.

3.3.4.2. **Search Module (SM) behaviour:** The Search Module (SM) employs a methodical approach to assess the credibility and nature of a given online post, primarily based on three crucial tasks: publication date analysis, source analysis, and fact-checking analysis. These tasks collectively inform the SM's decision score, guiding its determination of the content's authenticity.

(1) Publication Date Analysis: The SM initiates its evaluation by scrutinizing the publication date of the online post. Recognizing the significance of temporal context, this task aims to determine the relevance and timeliness of the content. Specifically, it checks whether the post's publication date aligns with current events or if it pertains to

historical information. The SM is diligent in its assessment to identify outdated posts, as re-sharing old or obsolete information can significantly impact the credibility of the online content. To ensure comprehensive coverage, the SM examines posts published on the specified date, allowing it to discern the temporal context and relevance of the content.

(2) Source Analysis: The source analysis task is centred around validating the credibility of the online post's origin. The SM seeks to ascertain whether the post has been disseminated by a trusted and reputable source. To achieve this, the module conducts an analysis that involves comparing the source of the online content with a database of recognized and respected sources. This dataset encompasses official websites representing various entities, including international organizations, government ministries, established media outlets, renowned news websites, and verified social media accounts. By cross-referencing the source against this repository of trusted sources, the SM determines whether the online post originates from a reliable and authoritative channel.

(3) Fact-Checkers Analysis: Fact-checkers analysis represents a pivotal task for the SM, focusing on the verification of online post content by trusted fact-checking organizations. To accomplish this, the module employs text similarity measurement techniques to compare the content of the online post with information provided on fact-checking websites. These fact-checking organizations maintain a reputation for their unwavering commitment to accuracy and integrity. By quantifying the similarity between the content under examination and the data within fact-checking databases, the SM assesses whether the online post has been independently verified and validated by a reputable fact-checking entity.

(4) Output Score: After conducting these analyses, the SM calculates a Score SM (ScSM). This score reflects the online post's credibility, considering its publication date, verification from trusted sources (such as official social accounts), and validation by fact-checkers. The ScSM score is presented as a percentage, with higher percentages indicating a greater probability of the content being genuine.

By combining the results of publication date analysis, source analysis, and fact-checking analysis, the SM provides a comprehensive evaluation of the content's authenticity. The ScSM serves as an indicator, categorizing the content as credible, or potentially false. This multifaceted approach enables the SM to effectively identify deceptive or misleading information and offers insights into the content's trustworthiness.

3.3.5. Explainable Decision Module (EDM)

The Explainable Decision Module (EDM) is a pivotal component within the FACTS-ON framework, designed to bring clarity and transparency to the decision-making process

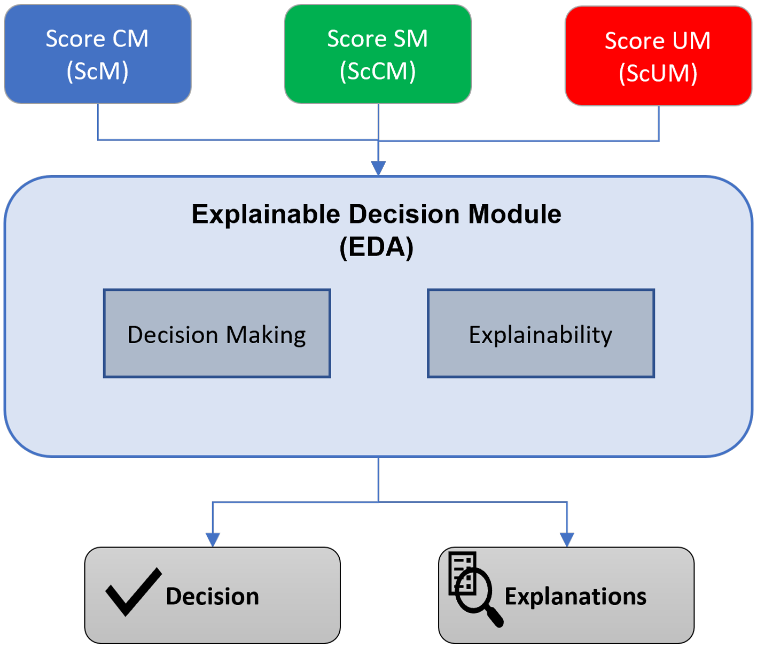


Figure 7 – Explainable Decision Module (EDM) architecture overview

surrounding the veracity of online content. It takes into account the assessments provided by other modules, namely the Content Module (CM), Search Module (SM), and User Module (UM). The EDM goes beyond delivering a final score decision, it also focuses on offering accessible and clear explanations to users on online social networks, aiming to influence their online behaviour positively.

3.3.5.1. **Explainable Decision Module (EDM) overview:** The EDM whose architecture is shown in Figure 7 plays a dual role within the FACTS-ON framework. Firstly, it conducts the decision-making task, which involves synthesizing the individual scores (ScCM, ScSM, ScUM) generated by the modules (CM, SM, UM) into a final decision score. This score serves as a critical indicator of the content's credibility. Secondly, the EDM undertakes an Explainability task, an innovative feature that sets FACTS-ON apart. Rather than merely highlighting specific words within the input text, the EDM provides comprehensive explanations derived from trusted fact-checkers and legitimate named entities. These explanations are designed to assist users in making informed decisions regarding the content they encounter on social networks.

3.3.5.2. **Explainable Decision Module (EDM) behaviour:** The Explainable Decision Module (EDM) is responsible for two core functions that are integral to its overall mission. These functions encompass Decision Making and Explainability, each playing a distinct yet interconnected role in enhancing the credibility assessment of online content.

(1) Decision Making: The EDM initiates its operation by receiving the individual scores (i.e., ScCM, ScUM, ScSM) from the modules (i.e., CM, UM, SM, respectively). These

scores, calculated as percentages, provide insights based on different criteria: the content's inherent credibility, the reliability of the user sharing the information, and external evidence verification. The EDM then synthesizes the input scores to formulate the final decision score. This score, also expressed as a percentage, offers a comprehensive assessment of the overall credibility of the content. It is instrumental in determining the nature of the content and discerning its likelihood of being truthful or potentially deceptive.

(2) Explainability: Simultaneously, the EDM conducts an Explainability task that aims to equip users with valuable insights. Instead of simple keyword highlighting, FACTS-ON's EDM presents users with excerpts from reputable fact-checking articles and references to credible named entities. These insights are embedded directly within the explanations, enhancing their reliability and relevance. Each explanation generated by the EDM includes the source of the returned text, ensuring users can verify the authenticity of the information. Additionally, publication dates and times are provided, aiding users in understanding the context and recency of the information. To further empower users, FACTS-ON offers direct links to the articles, allowing individuals to delve deeper into the content and verify its accuracy.

In essence, the EDM's unique approach promotes responsible online behaviour by encouraging users to verify the information before sharing it. By providing access to trusted sources and facilitating fact-checking, the EDM contributes to a safer and more informed online environment.

3.4. Technical background

This section provides an overview of the technical aspects of each system component within the FACTS-On framework, namely EXMULF, MythXpose, ExFake, and AFCC. Targeted at readers not deeply familiar with Artificial Intelligence (AI) and Machine Learning (ML), it offers clear explanations of the technologies used in each component of the framework. From advanced neural networks and Natural Language Processing (NLP) techniques to sophisticated algorithms for data analysis and interpretation, this section demystifies the complex systems underpinning FACTS-ON, elucidating how they contribute to effective fake news detection and analysis on social media.

3.4.1. EXMULF

EXMULF combines neural networks and image processing for analyzing multimodal content (text and images) in OSN posts.

3.4.1.1. **Vilbert (Vision-and-Language BERT):**. A neural network model that processes both images and textual data. It's designed to understand the context and relationship between visual elements and corresponding text, making it ideal for analyzing posts where the text and image content are intertwined.

3.4.1.2. **LDA (Latent Dirichlet Allocation):** LDA is a statistical model that identifies clusters of related words in documents. This technique is used for topic modelling. It helps in identifying the main topics or themes present in the textual content of a post. By analyzing the distribution of topics, LDA provides insights into the central themes or topics discussed.

3.4.1.3. **VGGNET16 (Visual Geometry Group Network 16):** A convolutional neural network model specifically designed for deep image analysis (i.e., image processing). VGGNET16 examines the visual aspects of posts, identifying patterns and features that are crucial for understanding the content and context of images.

3.4.2. MythXpose

MythXpose focuses on user behaviour and personality analysis, building on the capabilities of EXMULF.

3.4.2.1. **Big five personality prediction:** This model assesses a user's personality traits based on their online activities. It uses psychological profiling techniques to understand user behaviours and tendencies, which can be indicative of their likelihood to engage with or propagate false information. It analyzes how people interact on social media to infer their personality profile according to the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism).

- *Openness*: Individuals high in openness are typically imaginative, curious about the world, and open to new experiences. They often have a wide range of interests and are receptive to new ideas.
- *Conscientiousness*: This trait indicates a high level of organization, dependability, and discipline. People with high conscientiousness are usually meticulous, plan ahead, and are mindful of details and rules.
- *Extraversion*: Characterized by sociability and assertiveness, extroverted people are outgoing and thrive in social situations. They are often action-oriented and seek out social engagement.
- *Agreeableness*: This trait is marked by trust, kindness, and cooperativeness. Agreeable individuals are generally considerate, friendly, and willing to compromise, often prioritizing harmonious relationships.

— *Neuroticism*: High levels of neuroticism are associated with emotional instability, anxiety, and moodiness. Such individuals may experience frequent mood swings and stress, reacting more intensely to adverse events.

3.4.2.2. **Improved final classifier:** This is a neural network designed for intricate pattern recognition and decision-making. It comprises multiple layers that work together to analyze and interpret large sets of data, including fully connected layers, ReLU activations, and a sigmoid output function. This architecture allows the classifier to process and integrate intricate data patterns, making it capable of sophisticated pattern recognition and probabilistic predictions.

3.4.3. ExFake

ExFake employs NLP techniques for in-depth analysis of text in social media:

3.4.3.1. **Natural Language Processing (NLP):**. NLP is a field at the intersection of computer science, artificial intelligence, and linguistics. It is focused on enabling computers to understand, interpret, and respond to human language in a valuable way. NLP combines computational techniques with sophisticated models of language to process and analyze large amounts of natural language data. The ultimate goal of NLP is to build systems that can understand text and spoken words in much the same way humans do. NLP encompasses a range of techniques and tools, including text classification, sentiment analysis, language generation, and translation. In the context of the FACTS-ON framework, NLP is crucial for analyzing and interpreting the textual content in social media posts, enabling the system to extract meaningful insights, determine the sentiment of the text, and understand the nuances of human language.

3.4.3.2. **Named Entity Recognition (NER):**. Named Entity Recognition (NER) is a fundamental aspect of Natural Language Processing (NLP) that involves identifying and categorizing key entities within the text. In the context of NER, the SpaCy tool is utilized to extract and classify various named entities. SpaCy is proficient in recognizing a range of entity types and categorizing them into specific labels. These include GPE (Geopolitical Entities: geographic entities like countries, cities, states, etc.), PER (persons: named individuals or families), ORG (organizations: companies, agencies, institutions, etc.), DATE (dates), LOC (locations), and MONEY (monetary values). In ExFake, the primary focus is on three specific labels: *ORG*, *PER*, and *GPE*. This targeted approach in NER is crucial for extracting valuable information from large volumes of text, providing vital context, and enhancing the understanding of the content. By classifying these entities, NER becomes an indispensable tool in diverse NLP tasks like information retrieval, content classification, and knowledge extraction, contributing significantly to data mining, semantic search, and the broader field of natural language understanding.

3.4.3.3. **Text similarity and Natural Language Inference (NLI):**. These methods are crucial in linguistic analysis. They analyze the text to determine similarities with known information sources and infer logical relationships between statements. Text similarity measures how closely related two pieces of text are, essential in contexts like information verification. Conversely, NLI focuses on understanding the logical relationships between sentences, which is crucial for deducing the coherence and reliability of information. Both methods are vital for comprehending and verifying the content of social media posts.

To provide a foundational understanding, it is important to start with BERT (Bidirectional Encoder Representations from Transformers), a transformative model in natural language processing developed by Google. BERT's innovative approach lies in processing words in the context of all other words in a sentence, rather than sequentially, enabling it to capture nuanced contextual meanings effectively.

Expanding upon BERT's capabilities, *Sentence-BERT (SBERT)* is utilized for tasks such as *text similarity* and *NLI*. SBERT, a modification of the BERT model, is specifically optimized for generating sentence-level embeddings. It employs Siamese and triplet network structures, facilitating the production of semantically meaningful sentence embeddings. This feature is particularly vital for text similarity, allowing for the accurate measurement of relatedness between text segments. In the specific component (i.e., ExFake), SBERT's efficacy in *text similarity* is further enhanced through *fine-tuning* with the *STS benchmark* (Semantic Textual Similarity), a recognized standard for evaluating sentence similarity. For *NLI tasks*, which involve deducing logical sentence relationships, SBERT is fine-tuned using the *SNLI dataset* (Stanford Natural Language Inference dataset). These methodologies, leveraging the strengths of both BERT and SBERT, are integral to sophisticated text analysis, especially in applications aimed at authenticating and understanding the content in social media posts.

3.4.4. AFCC

AFCC uses a unique approach for evaluating information credibility through collective assessments from fact-checkers (i.e., consensus).

3.4.4.1. **Word2Vec and K-means clustering:** Word2Vec converts textual ratings into numerical vectors, making it easier to process and analyze. K-means clustering then groups these numerical values into clusters based on their similarities, aiding in identifying patterns (i.e., pattern recognition) and data categorization.

3.4.4.2. **Consensus and credibility assessment algorithms:** These algorithms evaluate the truthfulness of news based on the collective assessments of various fact-checkers (i.e., ratings or verdicts), offering a more nuanced understanding than simple majority voting. They help in determining the overall credibility of news content by evaluating the consensus among expert opinions.

3.4.4.3. **Majority voting:** Majority voting is a simple yet widely used method in decision-making processes, including in the context of fact-checking. In this approach, the verdict is based on the majority opinion among various fact-checkers. While straightforward, majority voting can be limited in situations where the number of fact-checkers or their expertise varies significantly. It may not fully capture the nuances in the credibility and reliability of different sources.

3.4.4.4. **Bayesian average:** The Bayesian Average offers a more refined and statistically robust approach compared to simple majority voting. The Bayesian average is a statistical technique used to derive a more accurate mean rating for each news item, especially when dealing with sparse data. This method combines the average rating of the news with the average rating across all news items, weighted by the number of ratings the specific news item has received. This approach helps in stabilizing the ratings, particularly in cases where there are few ratings, providing a more reliable and robust assessment.

The Bayesian Average in AFCC is utilized to moderate different ratings or opinions from various fact-checkers. Instead of a simple majority, this method can be adapted to consider the level of trust or credibility assigned to each fact-checker, thereby weighing their votes accordingly. This process ensures a more nuanced and fair assessment of the news content, based on the collective judgment of multiple trusted sources.

3.4.5. Role of fact-Checking organizations

In the context of this dissertation, the terms "***fact-checking organization***," "***fact-checker***," and "***fact-checking website***" are used *interchangeably*, all referring to entities crucial in verifying claims and news content.

Fact-checking organizations provide expert evaluations of news content, crucial for verifying the authenticity of information circulating on social media. These organizations, such as those aligned with the International Fact-Checking Network (IFCN), play a pivotal role in the ExFake and AFCC components of the FACTS-ON framework. The IFCN is a coalition of fact-checking organizations worldwide, committed to promoting best practices in fact-checking and upholding a Code of Principles that includes commitments to non-partisanship, fairness, transparency of sources, and transparency of funding and organization.

In ExFake, assessments and verified information from fact-checking organizations, including those adhering to the IFCN's standards, are vital in the analysis and validation of textual content. The application of NLP techniques such as Named Entity Recognition (NER) and

Natural Language Inference (NLI) is significantly strengthened by the input from these credible sources.

The AFCC component of FACTS-ON, in particular, integrates fact-checking organizations into its process of Consensus Inference and Credibility Assessment. This component is tailored for trust-based fake news detection, where the credibility of information and consensus among various fact-checkers are key determinants. AFCC employs advanced algorithms to analyze and synthesize evaluations from these organizations, assessing the overall trustworthiness and accuracy of news content. By considering the collective judgments of multiple fact-checkers, particularly those adhering to the IFCN's rigorous standards, the AFCC system can accurately infer the consensus about a news item's authenticity and the credibility of its sources. This process not only enhances the reliability of the information verification process but also builds a trust-based framework for assessing news content, leveraging the expertise and credibility of established fact-checking entities.

Integrating these expert evaluations, FACTS-ON harnesses both human expertise and AI analytics to enhance the accuracy and reliability of its fake news detection capabilities. Collaborating with fact-checkers, especially those part of the IFCN, ensures a more comprehensive and nuanced approach to verifying information's veracity in the digital space.

3.5. Comparative study

FACTS-ON is unique in its comprehensive integration of various aspects of false information detection. While other research in this field explores elements such as multimodal content analysis (combining text and image analysis), social-context analysis, and external verification from sources like fact-checkers, FACTS-ON distinctively combines all these elements. Additionally, it emphasizes explainability for transparency, aiming not only to detect false information but also to educate and potentially influence the sharing behaviour of OSN users. This multifaceted approach sets FACTS-ON apart in its strategy to combat false information in digital spaces.

Table 1 shows a comparison of FACT-ON with state-of-the-art approaches. It further delineates how FACTS-ON stands out, particularly in its holistic approach that combines content, context, external evidence, explainability, and multimodality. The comparison is based on various aspects, which are listed below.

- Content: This aspect evaluates the approach's reliance on the content of online posts for false information detection. It focuses on whether the method incorporates post content into its analysis.
- Context: Context refers to the consideration of factors surrounding an online post (i.e., the social context data). This aspect assesses whether the approach takes contextual elements into account, without evaluating the depth of their incorporation.

- External evidence: This aspect focuses on whether the approach integrates external sources of information, such as trusted fact-checkers and reputable sources, as part of the detection process.
- Multimodality: This aspect assesses whether the method can handle both textual and visual content within online posts, without evaluating the degree of success in combining information from images and text.
- Explainability: This aspect measures whether the framework provides clear and understandable explanations for its decisions regarding the truthfulness of online posts, without evaluating the quality or effectiveness of these explanations.

Table 1 – Comparison of FACTS-ON with state-of-the-art

	Aspect				
	Content	**Context**	**External evidence**	**Multimodality**	**Explainability**
Approaches in Chapter 2.Table 1	x				
Approaches in Chapter 2.Table 2		x			
Approaches in Chapter 2.Table 3	x	x			
Crowdsourcing: [64, 123, 151, 273] Web search: [74, 121, 171, 277, 291]			x		
Approaches in Chapter 2.Table 6	x			x	
Approaches in Chapter 2.Table 7					x
EXMULF	x			x	x
MythXpose	x	x		x	x
ExFake	x	x	x	x	x
FACTS-ON	**x**	**x**	**x**	**x**	**x**

In the FACTS-ON framework, my contributions significantly enhance the state-of-the-art in automatic fake news detection through an innovative approach integrating content, context, external evidence, explainability, and multimodality as discussed in the 2.4 Section of Chapter 2. FACTS-ON, with its distinct components like **EXMULF**, **MythXpose**, and **ExFake**, pioneers in analyzing both text and image-based content, while leveraging user personality traits and historical sharing behaviour for early detection. External evidence from fact-checkers and official sources is crucial in this framework, bolstered by the groundbreaking **AFCC** system that builds consensus and evaluates fact-checker credibility. This multifaceted strategy, coupled with transparent, user-centred explanations, places FACTS-ON at the forefront of combating false information in OSN.

Each of the compared approaches has its strengths and focus areas. FACTS-ON distinguishes itself by offering a holistic approach that integrates multiple aspects (Content, Context, External Evidence, Explainability, and Multimodality (including text and images)) within a single framework for false information detection, ensuring a more robust and comprehensive solution.

To the best of my knowledge, there may not be research that encompasses all the aspects of content, context, external evidence, explainability, and multimodality for false information detection in a single unified framework like FACTS-ON. Many researchers tend to focus on specific aspects or combinations of them rather than integrating all of them into a single framework.

3.6. Conclusion

This chapter culminates the overview of the FACTS-ON framework, an abstract and strategic solution meticulously designed to combat the spread of false information, encompassing fake news, misinformation, and disinformation, within online social networks (OSN). Throughout this chapter, I have delved into the intricate workings of each of its core modules: the Analyzer Module (AM), Content Module (CM), Search Module (SM), User Module (UM), and Explainable Decision Module (EDM). These modules collectively form the FACTS-ON framework, equipped with a multitude of features and functionalities, to tackle the pressing challenge of identifying and combating false information.

FACTS-ON's strength lies in its multifaceted approach. By examining the content, user behaviour, and external evidence, along with providing clear explanations, it offers a holistic solution for false information detection. The Analyzer Module (AM) efficiently extracts valuable information from online posts, including text and images, laying the foundation for subsequent analysis. The Content Module (CM) dives deep into the content's multimodal aspects, combining textual and visual analysis to assess the truthfulness of the post. Simultaneously, the User Module (UM) evaluates user historical behaviour, enhancing the framework's predictive capabilities.

The Search Module (SM) brings external evidence into the equation, cross-referencing posts with reputable fact-checking organizations and trusted sources. This external validation strengthens the decision-making process. Finally, the Explainable Decision Module (EDM) sets FACTS-ON apart with its unique approach to providing users with insightful explanations. Rather than simple keyword highlighting, it offers excerpts from trustworthy fact-checking articles and references to credible named entities, enriching the user experience and fostering responsible online behaviour.

FACTS-ON is the abstract representation of my solution, consisting of various systems such as EXMULF, MythXpose, EXFake, and AFCC, each embodying multiple modules integral to the framework. In the subsequent chapters, detailed insights into each of these system components of FACTS-ON will be provided, starting with the Explainable Multimodal Content-based Fake News Detection System: EXMULF in the next chapter. This in-depth exploration will highlight the specific functionalities and contributions of each module within these systems, further elucidating how they integrate and synergize within the overarching FACTS-ON framework.

As the first system component of FACTS-ON, EXMULF plays a vital role in demonstrating the framework's capacity for detecting fake news through a multimodal analysis, encompassing both text and image-based content. The upcoming analysis will reveal how EXMULF's innovative approach to combining and elucidating multimodal content is crucial in fortifying the overarching strategy of FACTS-ON in addressing multimodal false information in OSN.

Chapter 4

Explainable Multimodal Content-based Fake News Detection: EXMULF

4.1. Introduction

Following the comprehensive overview of the FACTS-ON framework in the preceding chapter, this chapter delves into the first key system component of the FACTS-ON framework: EXMULF (**EX**plainable **MUL**timodal **F**ake news detection). EXMULF [16] is a critical element of the FACTS-ON strategy, specifically designed to meet the challenge of detecting fake news through multimodal analysis of text and image content. This component not only represents the framework's commitment to advancing false information detection but also emphasizes transparency and user comprehension in combating deceptive content in online social networks (OSN).

This chapter introduces EXMULF as a system component containing three automated processes: 1) multimodal topic modelling, 2) multimodal content-based detection, and 3) multimodal explainability. These processes correspond to the abstract representation of FACTS-ON modules within EXMULF, including an Analyzer Module for data processing, a Multimodal Content Module for in-depth analysis, and a Multimodal Explainable Decision Module for clarity in explanations, as illustrated in Figure 1. In this chapter, the focus is placed on how EXMULF's integration of text and image analyses leads to more precise detection of false information and how its capability to provide multimodal explanations aligns with FACTS-ON's objectives, which emphasize advanced analytical techniques coupled with a strong focus on user-centric explainability.

EXMULF incorporates topic representation models, text classification models, image processing models, and explainable deep learning models. These components are essential for the multimodal detection of fake news. Table 1 presents a comparison of EXMULF with state-of-the-art methods in false information detection, highlighting its unique focus on multimodality, explainability, and reliance on news content features. In the context of

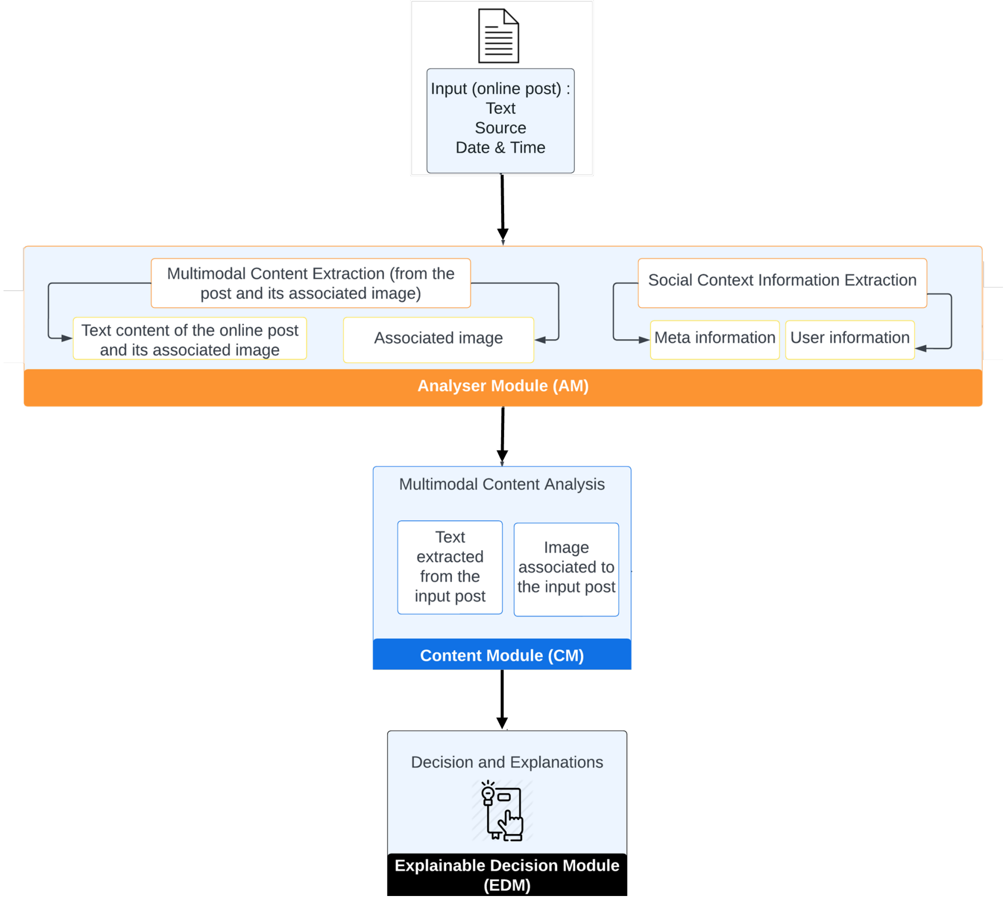

Input (online post) :
Text
Source
Date & Time
Multimodal Content Extraction (from the post and its associated image)
Text content of the online post and its associated image
Associated image
Social Context Information Extraction
Meta information
User information
Analyser Module (AM)
Multimodal Content Analysis
Text extracted from the input post
Image associated to the input post
Content Module (CM)
Decision and Explanations
Explainable Decision Module (EDM)

Figure 1 – FACTS-ON Modules Mapped to EXMULF. Diagram showing how EXMULF incorporates key FACTS-ON modules (i.e., Analyzer, Content, and Explainable Decision Modules).

multimodal and explainable fake news detection, EXMULF stands out as the first system to use Vilbert (Vision-and-Language BERT) for analyzing both image and text in fake news detection. This unique approach enhances both the accuracy of detection and clarity in understanding how fake news is identified, positioning EXMULF as a leader in providing a complete solution that is both multimodal and explainable for detecting fake news.

4.2. EXMULF system overview: explainable multimodal content-based fake news detection system

The architecture of EXMULF, depicted in Figure 2, consists of three core components:

(1) A topic modelling component,

(2) A multimodal content-based false information detection component (multimodal detector),

(3) A multimodal explainable detection component (multimodal explainer).

Table 1 – Overview of the state-of-the-art methods for false information detection

Approach	Multimodal	Explainable	News content
Shu et al. [249]		✓	
Reis et al. [220]		✓	
Yang et al. [316]		✓	
Lu et al. [160]		✓	
Przybyła et al. [215]		✓	✓
Bhattarai et al. [34]		✓	✓
Denaux et al. [70]		✓	✓
Silva et al. [260]		✓	
Xue et al. [315]	✓		✓
Zeng et al. [324]	✓		✓
Zhang et al. [327]	✓		✓
Kumari et al. [149]	✓		✓
Mangal et al. [167]	✓		✓
Meel et al. [175]	✓		✓
Giachanou et al. [96]	✓		✓
Giachanou et al. [95]	✓		✓
Singhal et al. [263]	✓		✓
Zhou et al. [331]	✓		✓
Qian et al. [218]	✓		
Yuan et al. [321]	✓		
Vishwakarma et al. [291]	✓		✓
Shah et al. [237]	✓		✓
EXMULF	✓	✓	✓

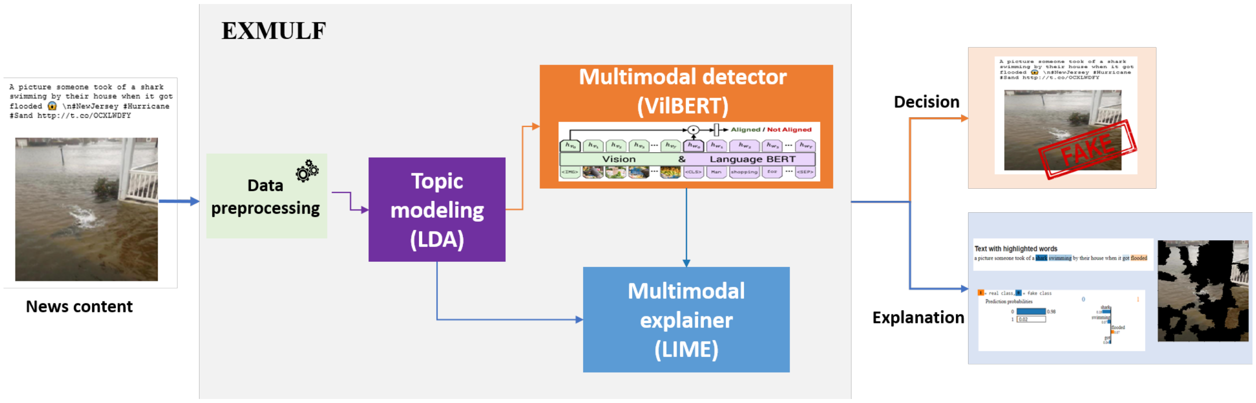

A picture someone took of a shark swimming by their house when it got flooded 😱 \n#NewJersey #Hurricane #Sand http://t.co/OCXLWDFY
News content
EXMULF
Data preprocessing
Topic modeling (LDA)
Multimodal detector (ViLBERT)
Aligned / Not Aligned
Vision & Language BERT
<IMG>
<CLS>
Man
shopping
for
<SEP>
Multimodal explainer (LIME)
Decision
FAKE
Explanation
Text with highlighted words
a picture someone took of a shark swimming by their house when it got flooded
real class
fake class
Prediction probabilities
0.98
0.02
shark
swimming
flooded
got

Figure 2 – The architecture of EXMULF

The adopted methodology, depicted in Figure 3, provides a clear overview of how EXMULF operates. It begins with taking as input the online post. The text within the post is extracted, and if applicable, text from the associated image is also retrieved. Both the text from the post and the image are processed for text analysis. Additionally, the associated image, even if it does not contain text, will be processed for image analysis.

Next, the obtained texts extracted from the multimodal data (i.e., text and image) are sent to the topic modelling component to detect topic similarity between them. If the captured topics differ, the input post content is deemed fake, and an explanation is given by the multimodal explainer component. If the topics are the same, the multimodal data is passed to the multimodal detector component to predict the post's veracity by analyzing the latent task-agnostic joint representations of the text and the associated image. The multimodal detector component processes these results to predict the veracity of the post content. Finally, the decision, prediction model, and extracted text and image are processed by the multimodal explainer component to generate relevant and interpretable explanations for OSN users.

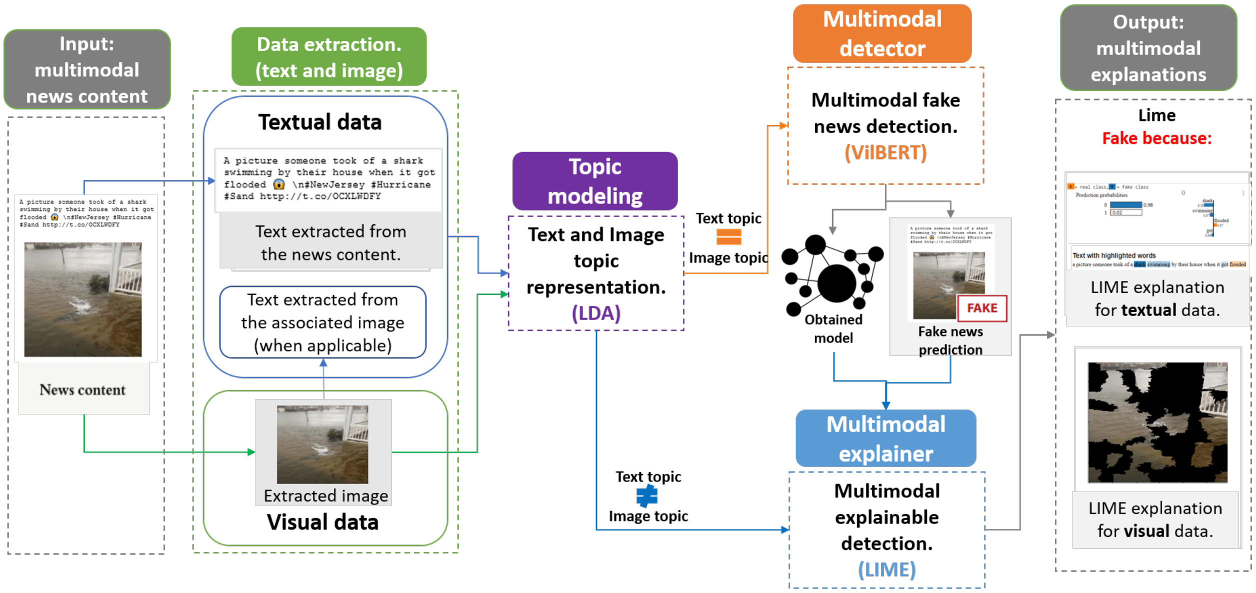

Input: multimodal news content
A picture someone took of a shark swimming by their house when it got flooded 😱 \n#NewJersey #Hurricane #Sand http://t.co/OCXLWDFY
News content
Data extraction. (text and image)
Textual data
A picture someone took of a shark swimming by their house when it got flooded 😱 \n#NewJersey #Hurricane #Sand http://t.co/OCXLWDFY
Text extracted from the news content.
Text extracted from the associated image (when applicable)
Extracted image
Visual data
Topic modeling
Text and Image topic representation. (LDA)
Text topic
Image topic
Text topic
Image topic
Multimodal detector
Multimodal fake news detection. (VilBERT)
Obtained model
FAKE
Fake news prediction
Multimodal explainer
Multimodal explainable detection. (LIME)
Output: multimodal explanations
Lime
Fake because:
1 = real class, 0 = fake class
Prediction probabilities
0
0.98
1
0.02
shark
swimming
flooded
got
Text with highlighted words
a picture someone took of a shark swimming by their house when it got flooded
LIME explanation for textual data.
LIME explanation for visual data.

Figure 3 – EXMULF methodology overview

4.2.1. Latent Dirichlet Allocation (LDA) topic Modelling

The topic representation models consist of topic modelling of both text and images within online posts, including scenarios where the images themselves contain text. The primary goal is to identify and analyze the coherence (consistency) between the topics presented in the text and those depicted or implied in the image. The LDA model systematically extracts topics from both the text and the image and then compares them to assess their similarity or disparity. A significant mismatch in topics could indicate that the post is potentially misleading. This not only aids in the accurate identification of fake news but also enhances transparency in the explanation of these findings to OSN users.

The topic modelling component is based on using the Latent Dirichlet Allocation (LDA) [36] which is a probabilistic modelling approach. This method makes it possible to create topic representations of texts in a corpus by identifying latent semantic structures in the text. An illustration of the LDA input/output workflow is presented in Figure 4[1].

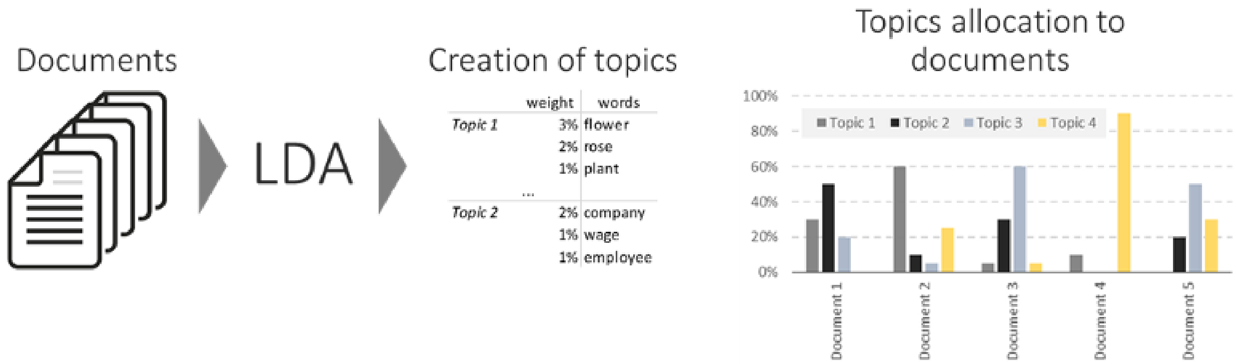


Figure 4 – Illustration of LDA input/output workflow

In EXMULF, LDA topic modelling is employed to identify the topics of both text and image of a given online post. If any inconsistency is detected between these topics, the system concludes that the news/post text and its associated image are not aligned. Consequently, the news is classified as fake, as it is believed to have been manipulated.

4.2.2. Multimodal content-based detection (multimodal detector)

The multimodal content-based detection model (multimodal detector) consists of detecting the veracity of a given post by analyzing the multimodal data available in its content, which includes the text body of the online post and its associated image. This approach is motivated by the fact that content is the primary factor in the deception process and is easily accessible and fully available for analysis in the early stages, making it ideal for the early detection of false information.

The multimodal detector employs VilBERT (Vision-and-Language BERT) [159], a model that learns joint representations of natural language and image content. VilBERT consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through co-attentional transformer layers. This enables sparse interaction through co-attention and allows for variable depths for each modality. The model is composed of repeated blocks of layers, denoted by dashed boxes with multiplier subscripts. Figure 5 illustrates the architecture of ViLBERT [159].

VilBERT is pre-trained on the conceptual captions dataset with two training objectives: masked multimodal learning and image text alignment prediction. The latter is what motivates

[1] `https://www.kdnuggets.com/2019/09/overview-topics-extraction-python-latent-dirichlet-allocation.html`, last access date: 18-09-2023.

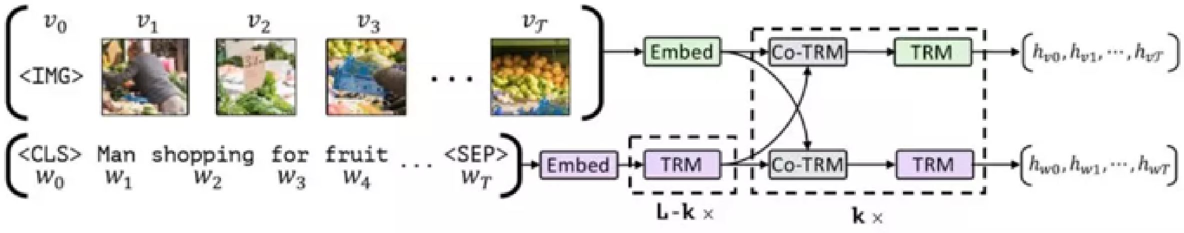

v_0
<IMG>
v_1
v_2
v_3
...
$v_{\mathcal{T}}$
Embed
Co-TRM
TRM
$\{h_{v0}, h_{v1}, \cdots, h_{v\mathcal{T}}\}$
<CLS> Man shopping for fruit ... <SEP>
w_0 w_1 w_2 w_3 w_4 w_T
Embed
TRM
Co-TRM
TRM
$\{h_{w0}, h_{w1}, \cdots, h_{wT}\}$
L-k ×
k ×

Figure 5 – ViLBERT model

the use of VilBERT in the multimodal detector component. ViLBERT is chosen for its high performance on a variety of visiolinguistic tasks, including visual question answering and image retrieval.

However, to apply the pre-trained VilBERT model in a multimodal false information detection/classification task, it was fine-tuned on the datasets to learn visually grounded language understanding in the fake news context. Specifically, the multi-task pre-trained model was used, and a linear classification layer of image and text representations was added to predict whether the news/online post is fake or real.

The multimodal detector component has two major tasks: text processing and image processing. These tasks are carried out in two separate streams, each with transformer blocks based on BERT [73] and co-attentive layers that enable the interaction between visual and textual modalities. In each co-attentive transformer layer, multi-head attention is computed in a similar way to a standard transformer block, with the exception that the visual modality handles the textual modality and vice versa.

In the text processing task, text tokens are generated from the BERT's tokenizer. In the image processing task, images are preprocessed in order to generate regional representations, including bounding boxes and regional features which are generated with a pre-trained object detection model (MaskRCNNn [118] in my case, unlike in the original VilBERT model [159] where the authors used Faster R-CNN to extract region features). It also encodes the spatial location of the regions. Regional image features and location features are then projected to the same dimension and summed to form the image embedding.

4.2.3. Multimodal explainable detection (multimodal explainer)

Explainable models consist of providing meaningful explanations that aim to let users build trust in the outcome so that they make use of the proposed systems [222]. Thus, these explanations help OSN users understand the decision made by the system and « why » given post is classified as fake. Consequently, it makes them aware of the danger of such content and influences their future behaviour. For instance, an OSN user who is convinced by the explanations provided as to why a given post is fake is unlikely to participate in its dissemination, support or recreation online.

Artificial intelligence applications require trust to aid in decision-making. Otherwise, their advice may be ignored due to a lack of trust. Specifically, if users do not trust a model or prediction, they will not use it [222]. In fact, end users will always prefer solutions that are easy to interpret and understand. Therefore, Explainable AI methods, such as LIME, help to understand how these models use complex mathematical decisions to get the corresponding predictions.

Ribeiro et al. [222] present LIME as an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model. The greatest assets of LIME are its accessibility and simplicity. LIME is a model agnosticism, which means that it can be used with any machine learning model, it provides explanations for almost any given model by treating it as a separate « black-box ». In addition, LIME gives local explanations, which are explanations for each observation instead of just the model itself. Furthermore, LIME is interpretable, it offers explanations based on the input features instead of abstract features.

In EXMULF, LIME is used to highlight the features, in both text and image input, that can help classify the news/post as fake or real. To achieve this, after obtaining the prediction of the multimodal detector, the text and image are separately analyzed.

4.3. Evaluation and discussion

In this section, the experimental details (i.e., the datasets and tools used), the interpretation of the results obtained, and a comparison of the proposed model with state-of-the-art methods are provided.

4.3.1. Datasets

Two publicly available real-world benchmark datasets were utilized for the experiments: Twitter [2] and Weibo [3]. Table 2 shows the distribution for both datasets after the preprocessing phase.

Table 2 – Statistics of the datasets used

Dataset	**Train**		**Test**	
	Fake	**Real**	**Fake**	**Real**
Twitter	6841	5009	2564	1217
Weibo	3748	3783	1000	996

[2] https://github.com/MKLab-ITI/image-verification-corpus

[3] https://drive.google.com/file/d/14VQ7EWPiFeGzxp3XC2DeEHi-BEisDINn/view

4.3.1.1. **Twitter dataset:** The Twitter dataset was released by Boididou et al. [38] as a part of Verifying Multimedia Use at MediaEval challenge. This dataset consists of two parts: a training set and a test set. Tweets in this dataset contain text, associated images, and contextual information.

The preprocessing steps were carried out as follows: First, instances that contained only text or image were removed, as the focus was on multimodal data. Then, for textual data, stop words, punctuation, symbols, and emojis were removed. Additionally, non-English text was translated into English using Google Translate. Since the images in the dataset had varying sizes, they were resized to match the input size of the neural network. Furthermore, text within the images (when applicable) was extracted using the pytesseract library in Python (Python-tesseract).

4.3.1.2. **Weibo dataset:** Curated by Jin et al. [131], consists of all the verified false rumour posts posted on the official rumour debunking system of Weibo (a micro-blogging website in China that encourages users to inform suspicious tweets) from May 2012 to January 2016.

Preprocessing for this dataset was conducted, with inspiration drawn from the preprocessing methods outlined by Wang et al. [302]. Duplicate images and odd-sized images were removed to ensure the dataset's integrity. The same preprocessing steps were applied to the text data as for the Twitter dataset, taking into account the Chinese language.

4.3.2. LDA topic modelling

Latent Dirichlet Allocation (LDA) is a topic model that can be used to assign text in a document to a certain subject. It generates a single topic per document model and a single word per topic model using Dirichlet distributions. Each text document in the collection is subjected to the LDA algorithm, which extracts a list of keywords. Documents are then grouped together in order to understand the recurrent keywords in the groupings of documents. These groups of recurrent keywords are therefore regarded as a topic shared by multiple papers in the collection.

The LDA topic modelling component measures the similarity between text and image topics of the online news. Therefore, in this section, I give experimental settings and results for each task separately.

4.3.2.1. **Topic modelling for textual Data:** After preprocessing the text (including Tokenization, removing stop words, lemmatization, and stemming), a dictionary is created to track the frequency of each word in the training set. Subsequently, the TF-IDF (term frequency-inverse document frequency) is computed to evaluate the significance of a term within a document relative to a collection or corpus. In this process, not only is the data and dictionary utilized, but the number of topics for training the base LDA model is also specified. The selection of 10 as the number of topics was made as it resulted in the highest coherence value, indicating its suitability for the task.

Various configurations were employed, and their details are presented in Table 3. The "Validation-set" denotes the used dataset, "topics" corresponds to the number of topics (K), "alpha" represents the Document-Topic Density hyperparameter, "beta" stands for Word-Topic Density hyperparameter, and "coherence" indicates the evaluation metric used to compare the model's performance under different hyperparameter settings.

These experiments were conducted sequentially, with one parameter being adjusted at a time while keeping the others constant. Additionally, the experiments were carried out using two distinct validation corpus sets, namely the "75% Corpus" and the "100% Corpus."

Table 3 – Topic modelling configuration

Validation-set	Topics	Alpha	Beta	Coherence
74% Corpus	2	0.01	0.01	0.402372
74% Corpus	2	0.01	0.31	0.379257
74% Corpus	2	0.01	0.61	0.378883
74% Corpus	2	0.01	0.91	0.389730
74% Corpus	2	0.01	symmetric	0.379257
....	....	....	....	....
100% Corps	10	asymmetric	0.01	0.491387
100% Corps	10	asymmetric	0.31	0.374487
100% Corps	10	asymmetric	0.61	0.408294
100% Corps	10	asymmetric	0.91	0.317167
100% Corps	10	asymmetric	symmetric	0.451740

The model was evaluated using topic coherence as an intrinsic evaluation metric. Topic coherence evaluates a single topic by quantifying the degree of semantic similarity among the high-scoring terms within that topic. This metric aids in distinguishing between semantically meaningful subjects and those that result from statistical inference artifacts.

4.3.2.2. **Topic modelling for image data:** Topic modelling for images presents a unique difficulty, requiring the interpretation of both visual and linguistic data, which are fundamentally different forms of information. To address this, the approach combines the use of Latent Dirichlet Allocation (LDA) and a pre-trained deep convolutional neural network model used for image recognition and classification known for its deep architecture consisting of 16 layers, the Visual Geometry Group Network 16 (VGGNet16) model. The LDA method is employed to extract topics from the captions associated with images, thereby handling the linguistic aspect. Concurrently, the pre-trained VGGNet16 model, known for its efficacy in image recognition tasks, is utilized to extract features from the images themselves.

In this process, a custom generator class is used to load the images and the topics derived from their captions. The images are first converted into a Numpy array format using the img-to-array function. They are then preprocessed using the preprocess-input function from Keras Vgg16, making them suitable for loading into the pre-trained VGGNet16 model.

The key to this approach is the integration of the textual topics from LDA with the visual features extracted by VGGNet16. This integration facilitates training a modified version of the VGGNet16 model to correlate and predict themes for the provided images based on the combined textual and visual data.

For model evaluation, the true topics and the predicted topics were loaded and their accuracy was calculated. The model achieved high accuracy rates of 89% and 92% for the respective datasets utilized during the evaluation process. These results underscore the effectiveness of combining textual and visual analysis for topic modelling in images. This approach showcases the potential of multimodal analysis in understanding and categorizing complex datasets.

4.3.3. Evaluation metrics

To assess the effectiveness of the classification approach, the following metrics were utilized: Accuracy (Acc), Precision (P), Recall (R), and F1-score by class, which are commonly employed in Machine Learning and Information Retrieval. These metrics were illustrated through a confusion matrix. A confusion matrix, as shown in Table 4, records the count of occurrences between two raters, including the true/actual classification and the predicted classification [100].

Table 4 – Confusion matrix for binary classification

Confusion Matrix		Predicted	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

The confusion matrix of a binary classification has two rows and two columns. The first row shows how many negative samples were predicted as negative or positive and the second row shows how many positive samples were predicted as negative or positive. The following terms are defined based on this matrix:

- True Negative (TN): Refers to the number of negative samples correctly labelled as negative.
- True Positive (TP): Refers to the number of positive samples correctly labelled as positive.
- False Negative (FN): Refers to the number of positive samples incorrectly labelled as negative.
- False Positive (FP): Refers to the number of negative samples incorrectly labelled as positive.

The evaluation metrics can be presented using the previously defined terms. Accuracy is the ratio of samples correctly predicted among the total number of samples, expressing the number of both positive and negative samples correctly classified. The following formula is used to compute accuracy:

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the ratio of positive samples correctly predicted among the total number of samples predicted as positive. It expresses the number of correct positive predictions that were made. The following formula is used to compute the precision:

$$Precision(P) = \frac{TP}{TP + FP}$$

Recall is the ratio of positive samples correctly predicted among the total number of positive samples. It expresses the number of actual positive samples correctly classified. The following formula is used to compute the recall:

$$Recall(R) = \frac{TP}{TP + FN}$$

F1-score aggregates the precision and the recall into a single measure using the harmonic mean. It's the weighted average between precision and recall. The F1-score reaches its highest score at 1 and its lowest at 0. The following formula is used to compute the F1-score:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

4.3.4. Multimodal detector

4.3.4.1. **Baselines:** To evaluate the performance of VilBERT on the false information detection task, a comparison was made against other models, including single-modality and multimodal models.

(1) **single-modality models**:

(a) **Text only**: The text-based false information detection model was evaluated by fine-tuning the $BERT_{BASE}$ model. This model takes only the text as input and is fed to the pre-trained $BERT_{BASE}$. To assess the importance of the text within the image, the model for the text of the news only, $BERT_T$, was used. In this case, the input of the model is the concatenation of the text with news and the text within the news, denoted as $BERT_{T+IT}$. This was performed to evaluate the performance of the multimodal models as well. Additionally, it is important to note that for the Weibo dataset, bert-base-chinese was used because it is trained on cased Chinese simplified and traditional text.

(b) **Image only**: Here, the investigation focuses solely on the images. To accomplish this, VGG-19 and ResNet-34 are utilized.

(2) **multimodal models**: For the evaluation of the multimodal model, a fusion model is defined by concatenating the features from $BERT_T$ and ResNet-34. Subsequently, a Multilayer Perceptron (MLP) is trained on top of this fusion. Additionally, other existing multimodal models such as SpotFake, AMFB, FND-SCTI, HMCAN, and BDANN are included in the comparison. These models were chosen for comparison because they were trained on the same datasets as the ones used, namely Twitter and Weibo., namely Twitter and Weibo.

A fair comparison was then made based on four evaluation metrics as presented in Table 5, namely the classification accuracy, precision, recall and F1-score metrics stated by the corresponding authors. These evaluation metrics are commonly employed for false information detection.

The results as shown in Table 5 demonstrate that VilBERT outperforms the baseline models described above in terms of accuracy.

The results, as shown in Table 5, demonstrate that VilBERT not only outperforms the baseline models in accuracy, but also excels in precision, recall, and F1-score evaluations. Specifically, on the Twitter dataset, VilBERT surpasses the baseline models in accuracy and F1-score for detecting fake news. On the Weibo dataset, it outperforms the baselines across all metrics except recall for fake news detection, and all but the F1-score for real news detection. This comprehensive performance underscores VilBERT's robustness in identifying authenticity, confirming its efficacy across various types of content within social media networks.

In this study, each dataset was divided into two parts: 80% was assigned to training and 20% to testing. Although VilBERT was originally designed for various vision and language challenges, recent research has indicated that learning visiolinguistic feature representations may be transferred across tasks. As a result, ViLBERT is fine-tuned across datasets by passing the element-wise product of the final image and text representations into a learned classification layer.

4.3.5. Multimodal explainer

For the explanation part, LIME (Local Interpretable Model-agnostic Explanations) has been incorporated to elucidate the decision-making process of the multimodal detector, which relies on VilBERT, for both text and image data. Figure 6 presents an example of a tweet classified as fake news by this detector, demonstrating the functionality of LIME.

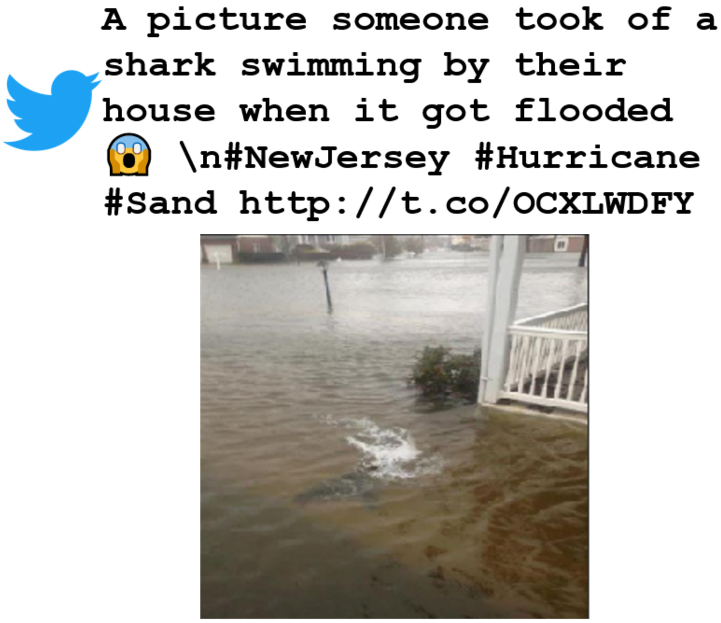
The explanation mechanism for images, as shown in Figure 7, begins by generating a dataset of image perturbations, or variations, centred around the instance being analyzed.

Table 5 – Results

Dataset	**Model**		**Accuracy**	**Fake News**			**Real News**		
				Precision	Recall	F1-score	Precision	Recall	F1-score
Twitter	Text only	$BERT_T$	0.572	0.602	0.586	0.597	0.543	0.553	0.544
		$BERT_{T+IT}$	0.577	0.612	0.574	0.598	0.551	0.564	0.556
	Image only	ResNet-34	0.624	0.712	0.567	0.6	0.558	0.72	0.62
		VGG-19	0.596	0.698	0.522	0.593	0.531	0.698	0.597
	Multi-modal	Fusion	0.7695	0.820	0.726	0.779	0.719	0.798	0.748
		SpotFake [263]	0.7777	0.751	0.900	0.82	0.832	0.606	0.701
		AMFB [149]	0.883	0.89	**0.95**	0.92	**0.87**	0.76	0.741
		HMCAN [218]	0.897	**0.971**	0.801	0.878	0.853	**0.979**	**0.912**
		BDANN [327]	0.830	0.810	0.630	0.710	0.830	0.930	0.880
		VilBERT	**0.898**	0.934	0.92	**0.926**	0.859	0.88	0.869
Weibo	Text only	$BERT_T$	0.680	0.731	0.715	0.709	0.667	0.676	0.669
		$BERT_{T+IT}$	0.682	0.739	0.72	0.71	0.672	0.684	0.673
	Image only	ResNet-34	0.694	0.701	0.634	0.698	0.698	0.711	0.699
		VGG-19	0.633	0.640	0.635	0.637	0.637	0.641	0.639
	Multi-modal	Fusion	0.8152	0.865	0.734	0.88	0.764	0.889	0.74
		SpotFake [263]	0.8923	0.902	**0.964**	0.932	0.847	0.656	0.739
		AMFB [149]	0.832	0.82	0.86	0.84	0.85	0.81	0.83
		FND-SCTI [324]	0.834	0.863	0.780	0.824	0.815	0.892	0.835
		HMCAN [218]	0.885	0.920	0.845	0.881	0.856	0.926	**0.890**
		BDANN [327]	0.842	0.830	0.870	0.850	0.850	0.820	0.830
		VilBERT	**0.9204**	**0.946**	0.948	**0.946**	**0.879**	**0.893**	0.885

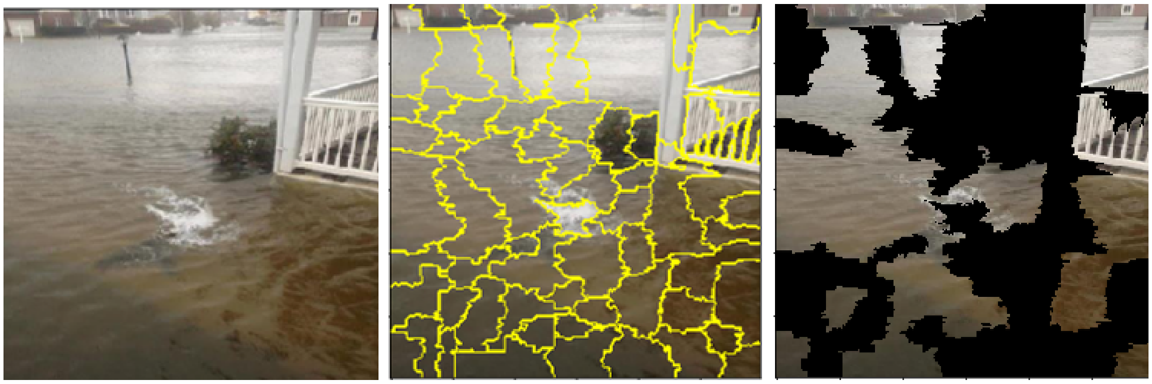
The model predicts the classification for each perturbation, and the significance of these perturbations is determined by assessing their similarity to the original image, with distances transformed into weights via a kernel function. These data perturbations, their predicted classifications, and corresponding weights are used to train a simpler, interpretable linear model. The coefficients of this linear model, each corresponding to a superpixel in the image, provide insights into the importance of specific image areas concerning the predicted classification. These coefficients are then used to generate explanations that help in understanding which areas of the image were most influential in the model's decision-making process.

The process starts with the creation of image perturbations, achieved by selectively activating or deactivating superpixels through the quickshift segmentation algorithm. The model then classifies these newly generated images. The cosine similarity between the original and perturbed images is computed to assign weights to the perturbations. Subsequently, a weighted linear regression model is fitted using these perturbations, their classifications, and weights. In this linear model, each coefficient correlates to a superpixel, indicating its importance in the class prediction. This strategy was originally proposed in the study by Ribeiro et al. [222].



A picture someone took of a shark swimming by their house when it got flooded 😱 \n#NewJersey #Hurricane #Sand http://t.co/OCXLWDFY

Figure 6 – Input tweet example



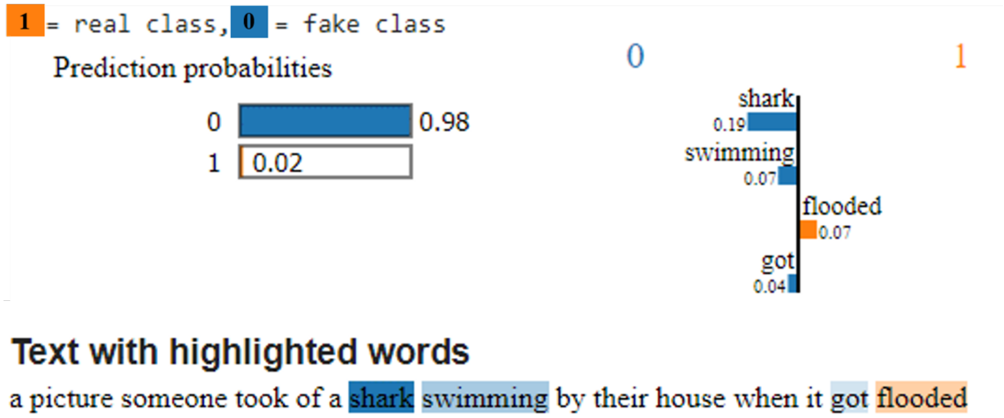
(a) (b) (c)

Figure 7 – (a) presents the original image (b) shows the superpixels that are generated using the quickshift segmentation algorithm (c) shows the area of the image that produced the prediction of the class (fake, in this case)

For textual data, the LIME Text Explainer is employed. This involves manipulating the original text by systematically omitting random words to create alternate versions. These versions are then classified into different categories (i.e., fake or real), enabling the assessment of how the absence or presence of certain words impacts the classification. The original publication by Ribeiro et al. [222] also detailed this method.

The output of LIME consists of a set of explanations that delineate the contribution of each word or image feature to the model's prediction (i.e., fake, real) for a given data sample.

The resulting analysis, as illustrated in Figure 8, aids in identifying which words or image elements are most influential in the model's classification decision.



Text with highlighted words

a picture someone took of a shark swimming by their house when it got flooded

Figure 8 – LIME explanations for textual data

4.3.6. Discussion

This chapter highlights the efficacy of using topic representation to discern between fake and real news by evaluating the topic similarity in both text and image content compared to known fake and real news corpus. The challenge of integrating topic modelling for textual and visual content in false information detection is significant, yet the approach has shown promising results in predicting news veracity through the coherence of captured topics.

To test the topic modelling component, its two sub-models (i.e., topic modelling for textual data and topic modelling for image data) are executed on a subset of 1000 samples from a labelled dataset. The hypothesis was that if the topics in the text and image of an online post are divergent, the post is likely fake. Out of these samples, 722 news items exhibited differing themes in text and images, and notably, 496 of these (i.e., 68%) were initially classified as fake.

Additionally, distinct performance patterns were observed in the experimental results, with the models showing better results with the Weibo dataset as compared to the Twitter dataset. This discrepancy could be attributed to the richer visual content in Weibo and the linguistic complexities post-segmentation in the Chinese dataset, which offered more informative content than the relatively shorter sentences in the Twitter dataset.

When assessing single-modality models, it was found that the image-only model was less effective than the text-only model. This suggests that textual content is more critical than visual information in identifying fake news. Although BERT shows competent performance in both single-modality and multimodal frameworks, it is outperformed by multimodal approaches that integrate textual and visual features.

These results support the idea that a combination of image and text analysis enhances performance compared to using either modality in isolation. Notably, the pre-trained ViLBERT model surpassed other baseline models, underscoring the transferability of learning the semantic relationship between visual and linguistic elements across different tasks. The pre-trained multi-task model demonstrated exceptional competence in correlating image and text signals. However, it is important to note that ViLBERT did not consistently achieve the highest scores in recall, precision, and F1-score metrics, particularly showing a stronger performance on the Weibo dataset than on the Twitter dataset, which was more imbalanced.

4.4. Conclusion

In this chapter, EXMULF, an integral component of the FACTS-ON framework was thoroughly explored. This system, designed to address the detection of false information in Online Social Networks (OSN), highlights the significance of not only identifying fake news but also ensuring that OSN users understand the rationale behind such identification. EXMULF, as an explainable multimodal content-based fake news detection system, has been shown to effectively process both textual and visual content, determining their authenticity and providing users with clear and interpretable explanations.

Recognizing the vital role of content in early detection owing to its immediate availability as opposed to auxiliary data such as social engagement, user response, and propagation patterns, which can only be obtained after the news has spread. In the development of EXMULF, advanced tools such as VilBERT (Vision-and-Language BERT) were utilized for aligning text with associated images, and LIME (Local Interpretable Model-agnostic Explanations) was integrated to offer transparent justifications for the system's decisions. Through extensive experimentation using multimodal datasets like Twitter and Weibo, EXMULF has demonstrated its efficiency in detecting false information, surpassing ten existing state-of-the-art models. This achievement marks a significant advancement in the field of information authenticity on social media platforms, being the first study to my knowledge to provide a fully explainable multimodal content-based fake news detection system employing both VilBERT and LIME models.

As the discussion on EXMULF concludes, the focus now shifts to the next chapter, which introduces MythXpose "Multimodal Content and Social Context-based System for Explainable False Information Detection with Personality Prediction" Building upon the foundational work of EXMULF, MythXpose represents an advancement in the FACTS-ON framework, combining content-based analysis with the innovative social context-based module, PERSONA, for assessing OSN users' personality traits. The next chapter will detail how MythXpose, extending beyond the capabilities of EXMULF, enriches the detection of deceptive content on social media by incorporating explainability mechanisms and personality trait analysis, thus contributing to a more nuanced and user-centric methodology in the battle against false information in OSN.

Chapter 5

Multimodal Content and Social Context-based System for Explainable False Information Detection with Personality Prediction: MythXpose

5.1. Introduction

Building on the foundational work presented in Chapter 4, this chapter delves deeper into the MythXpose system "Multimodal Content and Social Context-based System for Explainable False Information Detection with Personality Prediction", an advanced system designed to tackle the pervasive issue of false information on online social networks based on the content and the social context information (i.e., OSN user's personality traits). MythXpose, advancing beyond the capabilities of EXMULF introduced in Chapter 4, represents a fusion of content-based analysis (i.e., EXMULF) and the innovative social context-based module, PERSONA for Personality-Based Evaluation for Reliable Social Online News Analysis. This fusion of modules within MythXpose not only enriches the understanding of deceptive content on social media but also marks a significant stride towards more sophisticated, context-aware, and user-centric methodologies in false information detection.

The PERSONA module, specifically tailored to assess the personality traits of OSN users, brings a unique dimension to my approach to combating fake news. By analyzing user behaviours and tendencies, PERSONA offers critical insights that enrich the system's capability for explainable and accurate detection of false information.

MythXpose is dedicated to ensuring transparency and user understanding in the battle against false information on OSN. The system incorporates sophisticated mechanisms for explainability, ensuring that its approach to false information detection is both effective and transparent to users. This chapter will delve into the detailed functionalities and empirical validations of MythXpose's components, highlighting their individual and collective roles in achieving reliable and explainable false information detection on OSN. This integration of advanced modules, as illustrated in Figure 1, positions MythXpose as the second critical component in the FACTS-ON framework, enhancing its overall strategy for tackling false information in digital spaces.

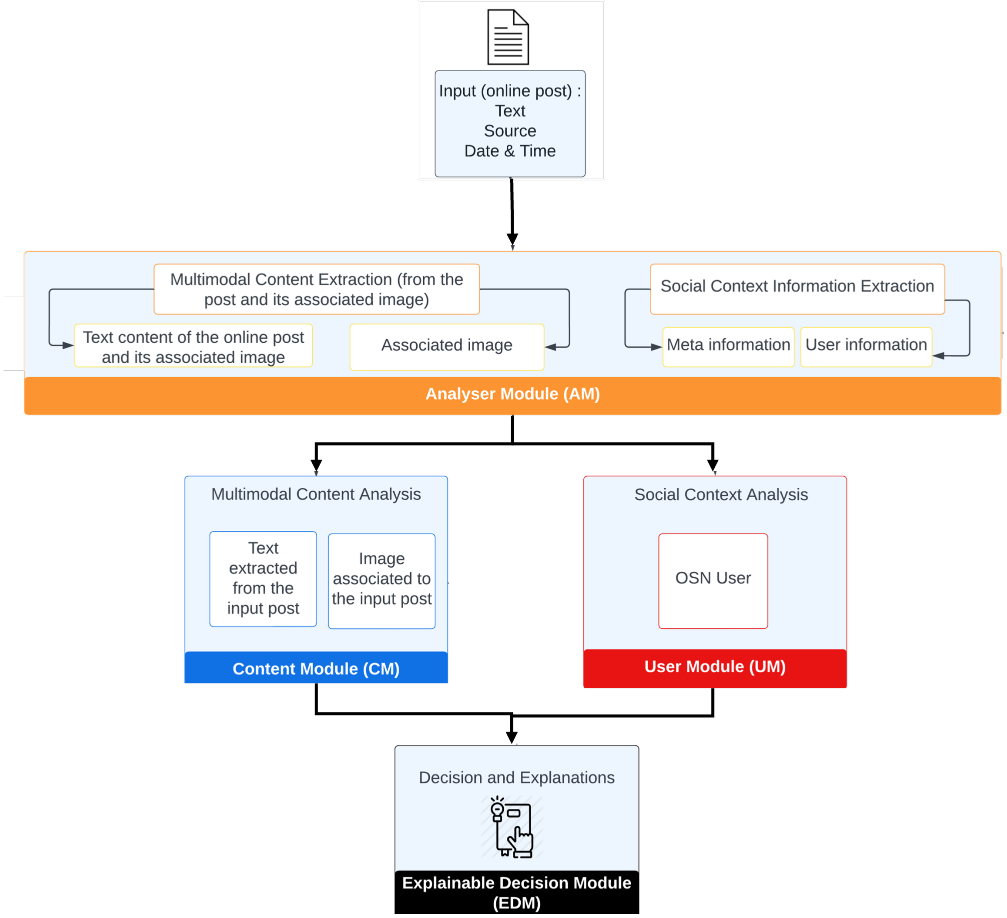

Input (online post) :
Text
Source
Date & Time
Multimodal Content Extraction (from the post and its associated image)
Text content of the online post and its associated image
Associated image
Social Context Information Extraction
Meta information
User information
Analyser Module (AM)
Multimodal Content Analysis
Text extracted from the input post
Image associated to the input post
Content Module (CM)
Social Context Analysis
OSN User
User Module (UM)
Decision and Explanations
Explainable Decision Module (EDM)

Figure 1 – FACTS-ON Modules Mapped to MythXpose. Diagram showing how MythXpose incorporates key FACTS-ON modules (i.e., Analyzer, Content, User, and Explainable Decision Modules).

5.2. MythXpose system overview

The MythXpose system is designed for the detection and explanation of false information within online social networks (OSN). It achieves this through the integration of personality prediction and multimodal content analysis. Comprising three distinct modules, namely EXMULF (Explainable Multimodal Content-based Fake News Detection System) [16], PERSONA (Personality-Based Evaluation for Reliable Social Online News Analysis), and MEXDM (Multimodal Explainable Decision Making), MythXpose offers a multifaceted approach to combat the proliferation of false information and enhance online content reliability. The architecture of MythXpose is illustrated in Figure 2.

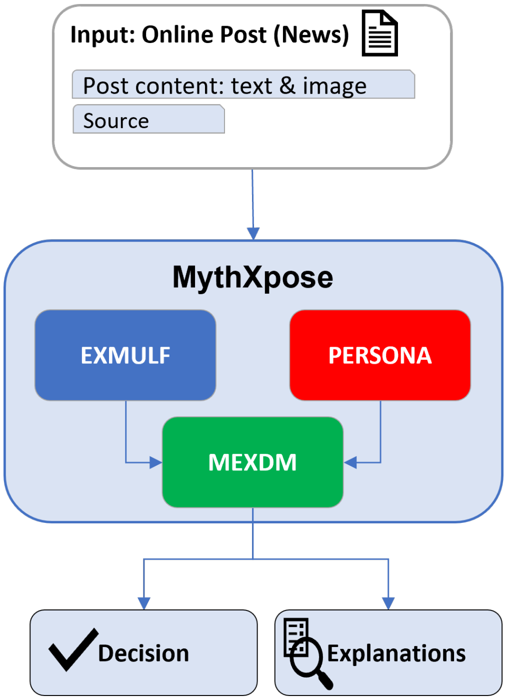


Figure 2 – The architecture of MythXpose

5.2.1. EXMULF: multimodal content analysis

At the core of the MythXpose system is EXMULF (Explainable Multimodal Content-based False Information Detection System), which is introduced and defined in Chapter 4. Rooted in the VILBERT (Visual-Linguistic BERT) model, EXMULF performs in-depth multimodal content analysis, evaluating both textual and visual components of online content. By leveraging VILBERT, it enhances its understanding of content authenticity, effectively discerning between genuine and false information.

5.2.2. PERSONA: personality prediction

In the intricate task of identifying and mitigating the spread of false information on online social networks (OSN), the PERSONA (Personality-Based Evaluation for Reliable Social Online News Analysis) module of MythXpose adopts a pioneering approach by analyzing personality traits. This approach is premised on the observation that susceptibility to fake news, misinformation, and disinformation (i.e., false information) varies among individuals [49], influenced by a complex interplay of cognitive abilities, emotional responses, and personal beliefs.

PERSONA's focus on personality traits as predictors for susceptibility to online false information is substantiated by extensive research. For instance, Li et al. [156] highlight how emotional responses and analytic thinking impact susceptibility to false information during crises like the COVID-19 outbreak. Similarly, studies by Roozenbeek et al. [225], Borukhson et al.[40], and Bronstein et al. [46] emphasize the role of analytic thinking and cognitive biases in shaping individuals' responses to misinformation.

Further, Saltor et al [231] delve into thinking dispositions and styles as crucial factors in discerning fake news, while Van Der Linden [284] and Nan et al. [192] offer insights into general susceptibility to misinformation and health misinformation, respectively. Insight into problem-solving abilities and their correlation with reduced susceptibility to fake news is presented by Salvi et al. [232].

Building on these insights, researchers investigate the connection between personality traits and susceptibility to false information. For example, Tulin et al. [281] discuss how personality influences social capital, a factor that can be crucial in understanding how individuals interact with information on OSN. Additionally, Fatke's research on personality traits and political ideology [85] provides a framework for understanding the political dimensions of misinformation susceptibility.

By considering these diverse perspectives, PERSONA offers a comprehensive approach to detecting and understanding the spread of fake news. It not only identifies potential false information but also provides insights into the psychological profiles of individuals who are most likely to be influenced by or share false information. This approach recognizes the diversity in human behaviour and psychology as a critical element in combating the spread of fake news on OSN.

5.2.3. MEXDM: multimodal explainable decision making

MEXDM (Multimodal Explainable Decision Making) serves as the decision-making and explainability module within MythXpose. Integrating the outcomes of EXMULF and PERSONA, MEXDM orchestrates the fusion of multimodal content analysis and personality-based evaluation. This combination enhances the accuracy of false information detection and offers valuable insights into decision-making processes. Empirical results further substantiate these findings.

5.3. Methodology

The MythXpose system's methodology, as illustrated in Figure 3, demonstrates the integration of various models within each module. This structure is key to the system's capability in detecting and analyzing false information on online social networks.

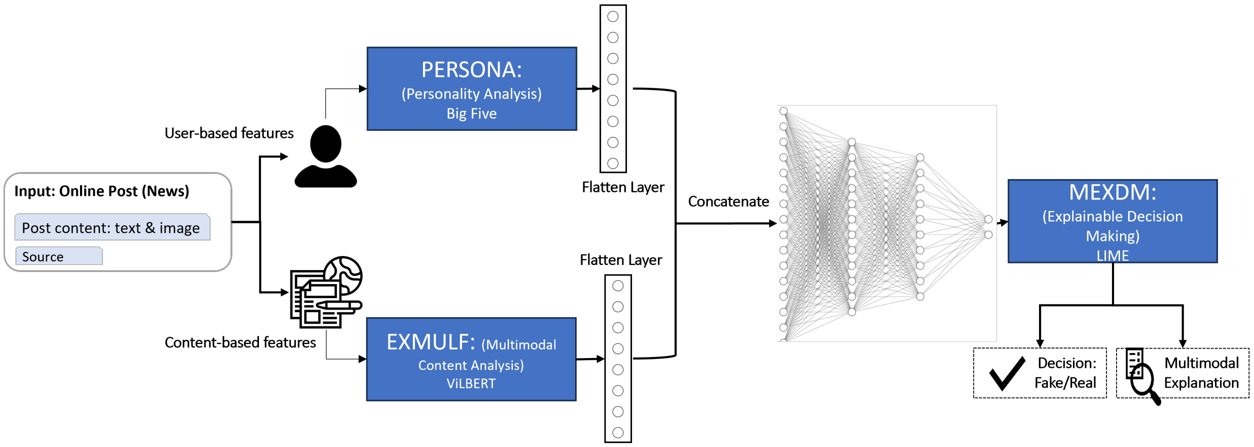


Figure 3 – MythXpose methodology overview

5.3.1. Personality prediction (PERSONA)

This section delves into the PERSONA module within MythXpose, starting with its data collection from the mypersonality dataset. It discusses the variety of machine learning algorithms applied for personality prediction.

The PERSONA module, integral to MythXpose, is dedicated to assessing the reliability of online news sources by analyzing the Big Five personality traits of the individuals behind these sources. The Big Five personality traits, also known as the Five-Factor Model, comprise Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These traits provide a comprehensive framework for understanding human personality, making them highly relevant for evaluating the personalities of individuals associated with news sources.

Leveraging the extensive mypersonality dataset, a foundational resource in text-based personality prediction, PERSONA's development involved a wide array of machine learning algorithms chosen for their specific strengths in predictive modelling.

To ensure the robustness and accuracy of personality predictions, various techniques were implemented. These included K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression (LR), Multinomial Naive Bayes (MNB), and advanced Transformer-based models such as Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) and Bidirectional Encoder Representations from Transformers (BERT). DistilBERT is a smaller, faster, cheaper, and lighter version of BERT, designed to provide most of the performance benefits of BERT with significantly reduced size and complexity, making it more efficient for deployment in practical applications [1]. Each model's performance was rigorously evaluated through a process of training, hyperparameter tuning, and cross-validation to fine-tune the predictive capabilities. This rigorous process played a pivotal role in creating a reliable module capable of providing insightful and nuanced personality assessments.

[1] `https://huggingface.co/docs/transformers/model_doc/distilbert`

5.3.2. Multimodal explainable decision making (MEXDM)

This subsection delves into the critical function of MEXDM within the MythXpose system, highlighting its integral role in both decision-making processes and explainability aspects as depicted in Figure 4. MEXDM stands as a cornerstone in the system, adeptly synthesizing and interpreting insights derived from the EXMULF (Explainable Multimodal Fusion) and PERSONA modules.

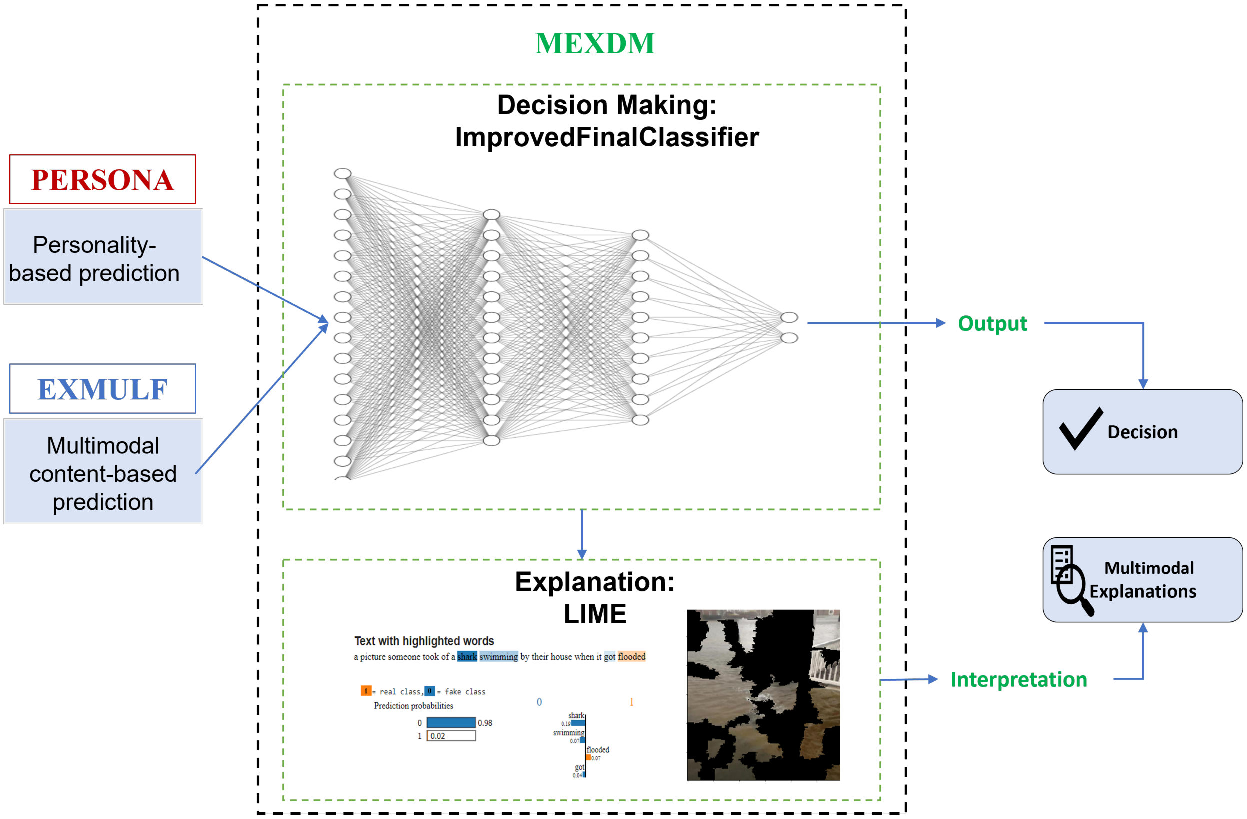


Figure 4 – MEXDM overview

A focal point of this discussion is the detailed architecture of the "Improved Final Classifier", a neural network that forms the backbone of MEXDM. This classifier is intricately designed with multiple layers, including fully connected layers (fc1, fc2, fc3), Rectified Linear Units (ReLU) activations (relu1, relu2), and a sigmoid function at the final stage. Each layer plays a pivotal role: the fully connected layers are responsible for processing and integrating features, while the ReLU activations introduce non-linearity, enhancing the network's ability to capture complex patterns. The sigmoid layer at the end serves to output precise, probabilistic predictions, which are crucial for decision-making.

The architecture's ingenious design allows for the seamless integration of features from both the EXMULF and PERSONA modules. EXMULF contributes by providing a rich, multimodal understanding of data, while PERSONA adds a layer of personalized context, ensuring that decisions are tailored and relevant. The "Improved Final Classifier" efficiently consolidates these diverse inputs, ensuring that the predictions are not only accurate but also comprehensible.

This section further explores how the MEXDM's architecture enables it to perform dual functions effectively. Firstly, in decision-making, it leverages the consolidated information to make informed, context-aware decisions. Secondly, in explainability, it ensures that these decisions are transparent and interpretable, aligning with the ethos of the MythXpose system to provide understandable and trustworthy AI solutions. The interplay between these two functions and their reliance on the sophisticated architecture of the MEXDM forms a core part of this discussion, underlining the system's commitment to both performance and transparency.

5.4. Experimental results

In this section, the experimental results of MythXpose are presented, consisting of three key components: EXMULF, PERSONA, and MEXDM. Each component is evaluated separately to assess its performance.

5.4.1. PERSONA results

The PERSONA component is responsible for predicting the personality traits and behavioural patterns of individuals who are likely to spread false information. Comprehensive experiments were conducted to develop and evaluate the predictive capabilities of the PERSONA model.

5.4.1.1. **Dataset:** The myPersonality dataset[2], a widely used resource in personality research, provides a rich compilation of Facebook users' status updates along with their corresponding Big Five personality traits scores across five dimensions: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness. This dataset was utilized in the development of the personality prediction model PERSONA and it served as the foundation for the creation of a robust and accurate predictive model for personality traits. An overview of this dataset, detailing its essential components, is depicted in Table 1.

5.4.1.2. **Personality prediction models:** Various machine learning and transformer-based models, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression (LR), Multinomial Naive Bayes (MNB), DistilBERT, and BERT, were explored. Each model underwent rigorous training and fine-tuning to maximize its predictive performance. The selection of these models was strategic, as it enabled the exploration of various approaches to personality prediction, ranging from traditional machine learning

[2] https://github.com/jcl132/personality-prediction-from-text/blob/master/data/myPersonality/mypersonality_final.csv

Table 1 – Overview of the myPersonality Dataset

Column	**Description**
AUTHID	Unique hashed user ID
STATUS	Facebook status update text
sEXT, sNEU, sAGR, sCON, sOPN	Personality scores (Extraversion, Neuroticism, Agreeableness, Conscientiousness, Openness)
cEXT, cNEU, cAGR, cCON, cOPN	Binary classification of personality traits
DATE	Date and time of the status update
NETWORKSIZE	Size of the user's network
BETWEENNESS	Betweenness centrality in the network
NBETWEENNESS	Normalized betweenness centrality
DENSITY	Density of the user's network
BROKERAGE	Brokerage in the network
NBROKERAGE	Normalized brokerage
TRANSITIVITY	Transitivity of the network

Table 2 – PERSONA results

Model		Accuracy	Precision	Recall	F1 score
Machine Learning	KNN	75	37	50	42
	SVM	62	58	50	51
	LR	61	57	51	52
	MNB	62	60	47	47
Transformers	BERT	97	96	98	97
	DistilBERT	96	95	98	98

methods like KNN, SVM, and LR to more advanced neural network-based models such as DistilBERT and BERT.

Table 2 summarizes the performance of these personality prediction models, showcasing their accuracy, precision, recall, and F1 scores.

Among the machine learning models, KNN exhibited good accuracy but struggled with precision. SVM and LR showed balanced performance, while MNB excelled in precision but had lower recall. In contrast, transformer models, BERT and DistilBERT, demonstrated exceptional performance, with BERT slightly outperforming DistilBERT.

5.4.2. MEXDM results

MEXDM, the Multimodal Explainable Decision-Making module, integrates the outputs of EXMULF and PERSONA to enhance false information detection. This section provides insights into the combined performance of these two components and evaluates the effectiveness of early intervention based on personality prediction.

Table 3 – MEXDM results

Model	Accuracy	Fake news			Real news		
		Precision	Recall	F1 score	Precision	Recall	F1 score
EXMULF	89.9	93.4	92	92.6	85.9	88	86.9
PERSONA	57.1	84	33	47	32	83	47
MEXDM	92.6	93	98	94	87	90.2	89

5.4.2.1. **Dataset:** For this evaluation, the Twitter dataset [3], which consists of both text and image data, was utilized. This dataset facilitates the assessment of the capabilities of EXMULF and PERSONA in a real-world scenario.

5.4.2.2. **Combined model performance:** The personality prediction model was integrated with the EXMULF model, and the features generated by both models were merged. The combined features were subsequently processed through a classifier known as the "Improved Final Classifier" which consisted of neural network layers. The process began with the initial layer (fc1), which transformed the feature representation from 4 to 128 dimensions, followed by a Rectified Linear Unit (ReLU) activation (relu1) to introduce non-linearity. The subsequent layer (fc2) further reduced the dimensionality to 64 features, followed by another ReLU activation (relu2). The penultimate layer (fc3) condensed the representation to a single output, and finally, the sigmoid activation function (sigmoid) was applied to this output.

Table 3 summarizes the performance metrics of three different models in classifying online posts as fake or real. The "EXMULF" model, the "PERSONA" model, and the combined models in "MEXDM" are compared based on accuracy, precision, recall, and F1 score.

The EXMULF model exhibited exceptional performance, achieving an impressive accuracy of 89.9%. It demonstrated robust precision and recall rates for both fake and real news, resulting in a remarkable F1 score of 92.6% for fake news and 86.9% for real news.

In contrast, the Personality-only model presented a lower accuracy of 57.1%. It grappled with a delicate balance between precision (84% for fake news and 32% for real news) and recall (33% for fake news and 83% for real news), leading to F1 scores of 47% for both categories.

However, the MEXDM model outshone the others by achieving an accuracy of 92.6% and maintaining well-balanced precision and recall rates (93% for fake news and 98% for real news). Consequently, it achieved strong F1 scores of 94% for fake news and 89% for real news.

In summary, the MEXDM model clearly demonstrated superior performance in accurately categorizing both fake and real news articles.

The results demonstrate that the integration of PERSONA with EXMULF significantly enhances the overall performance of fake content detection. The MEXDM model achieves

[3] `https://github.com/MKLab-ITI/image-verification-corpus`

the highest accuracy and F1 score, showcasing the effectiveness of early intervention based on personality traits in identifying and mitigating the spread of fake content (Myths).

In conclusion, the comprehensive evaluation of the MythXpose components (EXMULF, PERSONA, and MEXDM) highlights the importance of multimodal detection, personality prediction, and their combined application in mitigating the impact of fake content dissemination.

5.5. Discussion

The experimental results previously presented provide valuable insights into the effectiveness of the PERSONA, and MEXDM models in explainable false information detection and personality prediction. In this discussion, valuable insights into the implications of these results, their significance, and potential avenues for further research.

5.5.1. PERSONA results

The PERSONA model, designed for personality prediction, exhibited a different set of outcomes. While it is not directly involved in false information detection, its performance is pivotal in understanding potential spreaders of false information. The model achieved an accuracy of 57.1%, which, although lower than EXMULF, provides valuable insights into the personalities of individuals likely to engage in the dissemination of false information.

A noteworthy aspect of the PERSONA model's results is the trade-off between precision and recall. It excelled in precision for fake news (84%) but struggled with recall (33%). Conversely, it demonstrated high recall for real news (83%) but at the expense of precision (32%). This suggests that the PERSONA model is adept at identifying specific individuals with a propensity for spreading fake news, but it may generate false positives for real news articles and false negatives for fake news articles.

5.5.2. MEXDM results

The MEXDM model represents the culmination of the EXMULF and PERSONA models, combining their strengths to enhance false information detection. This integration not only resulted in improved accuracy (92.6%) but also achieved a harmonious balance between precision (93%) and recall (98%) for fake news, along with strong performance in real news classification (87% precision and 90.2% recall).

The MEXDM model's remarkable F1 scores of 94% for fake news and 89% for real news indicate its superior capability in accurately categorizing news articles. By leveraging both personality prediction and content analysis, it effectively addresses the source's personality traits while scrutinizing the news content, significantly mitigating the spread and impact of false information.

5.6. Conclusion

In this chapter, I have introduced MythXpose, a sophisticated component of the FACTS-ON framework, designed to enhance the detection and explanation of false information in online social networks (OSN). MythXpose represents an evolution beyond the capabilities of EXMULF, previously discussed, by integrating not only multimodal content analysis but also social context-based insights, including the use of personality prediction based on the Big Five personality traits.

By analyzing both the content and the social context, specifically through the lens of user personality traits, MythXpose offers a unique perspective in evaluating the reliability of sources in OSN. This integration not only improves the accuracy of detecting false information but also facilitates proactive intervention. Proactive measures enabled by MythXpose include early identification of likely misinformation spreaders or preemptive verification of information from sources with profiles associated with previous dissemination of falsehoods, effectively preventing the proliferation of false information before it gains traction.

The application of personality prediction as a key element of social context analysis in MythXpose provides several critical advantages. It enhances the understanding of potential false information spreaders, improves early detection capabilities, and reduces the spread of false information by addressing personality traits that may influence the sharing of false content.

Furthermore, the integration of the MEXDM (Multimodal Explainable Decision Module) ensures that the decision-making process of MythXpose is transparent and comprehensible to users, fostering trust and enhancing the user experience.

As the discussion transitions from MythXpose, the next chapter introduces the ExFake system "Explainable False Information Detection system based on Content, Context and External Evidence". ExFake expands upon the FACTS-ON framework by combining content analysis, social context exploration including user historical sharing behaviour, and external evidence assessment, all supported by explainable AI. The subsequent chapter will explore how ExFake integrates these diverse components to provide a more holistic solution for combating false information in OSN.

Chapter 6

Explainable False Information Detection based on Content, Context and External Evidence: ExFake

6.1. Introduction

In an era dominated by digital information dissemination, the challenge of identifying and combating false information, commonly referred to as "fake news", has grown increasingly complex. Traditional approaches to false information detection often rely solely on the content of online posts. While this provides valuable insights, it paints an incomplete picture. False information can be strategically crafted to mimic truth, making content-based methods susceptible to manipulation. Additionally, these approaches largely disregard the rich landscape of social context that surrounds online posts. Therefore, the exploration of auxiliary information is deemed crucial for the effective detection of false information. Context-based approaches explore the surrounding data outside of news content, which can be an effective direction and have some advantages in areas where the content approaches based on text classification can run into issues.

Recognizing the limitations of traditional content-based detection methods, which often fail to fully capture the intricacies of fake news, the *ExFake* system introduces a more comprehensive approach. ExFake, standing for **Ex**plainable **Fake** news detection, is designed as an advanced component within the FACTS-ON framework, incorporating a broader range of detection mechanisms.

ExFake is constructed with an abstract view of the FACTS-ON modules, integrating an *Analyzer Module* for initial data processing, a *Content Module* for detailed analysis of online post *content*, a *User Module* for assessing *social context* information, a *Search Module* for *external evidence* evaluation, and an *Explainable Decision Module* for providing clear, understandable *explanations*. These modules, though *named differently* within ExFake, align with the overarching structure of FACTS-ON, embodying its *multifaceted* approach to false information detection. This comprehensive approach, as illustrated in Figure 1, positions ExFake as a pioneering system in the FACTS-ON framework, enhancing the capabilities to detect and understand false information in the digital age.

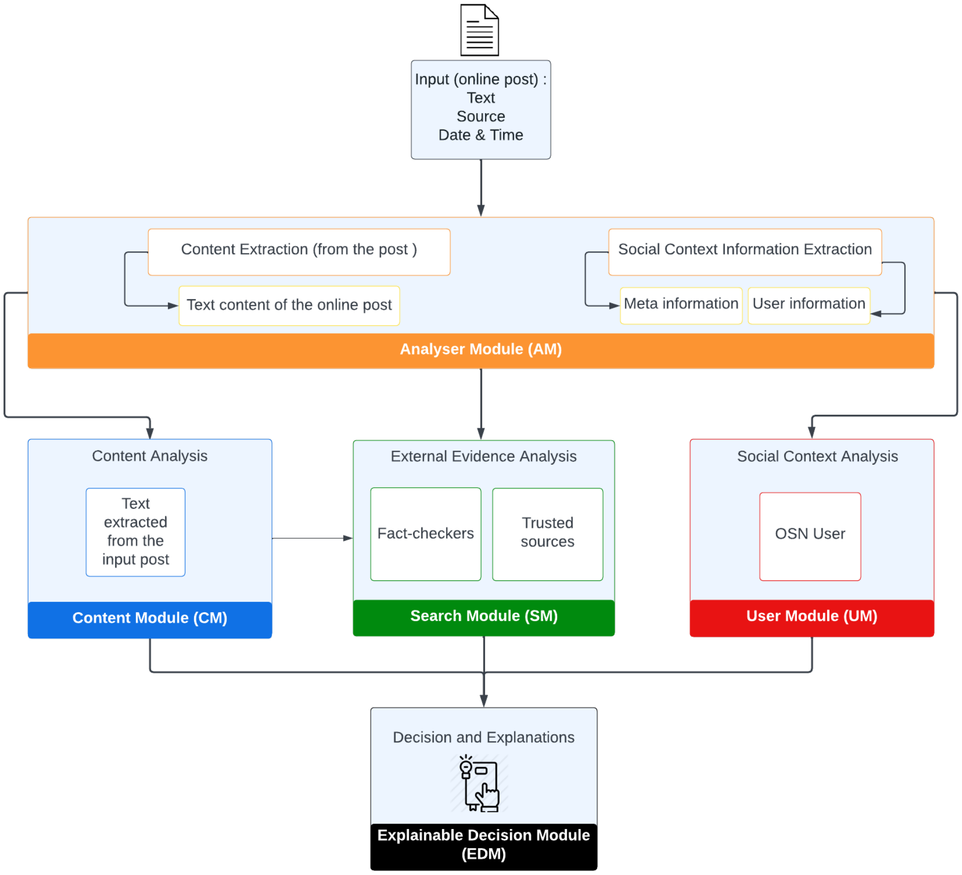


Figure 1 – FACTS-ON Modules Mapped to ExFake. Diagram showing how ExFake integrates key FACTS-ON modules (i.e., Analyzer, Content, User, Search, and Explainable Decision Modules).

However, it is important to recognize that not all *social context* information is equally accessible promptly. Some aspects, such as historical sharing behaviour for OSN users and the publication date of an online post, are available before false content spreads. These early indicators can significantly expedite the detection and identification of fake content. In contrast, other social context information, like social engagement, user response, and propagation patterns, can only be obtained after false content has already spread widely, limiting its usefulness for early detection.

Furthermore, ExFake does not stop at social context information, it extends to *external evidence*. Unlike existing methods that often overlook the valuable insights offered by techniques such as *web search* and *fact-checking*, ExFake leverages these *external sources*.

This *proactive* integration of *external evidence* equips ExFake to identify and address false content *early* in its dissemination, potentially saving valuable time and resources.

Additionally, the lack of *explainability* in false information detection models poses significant challenges. Users are often presented with black-box decisions, leaving them without a clear understanding of why a particular piece of information is classified as false. This opacity hinders users' ability to make informed decisions about the credibility of online content.

The *ExFake* system, which stands for **Ex**plainable **Fake** news detection, introduced in this chapter, addresses these multifaceted challenges. ExFake, as discussed in the in the Discussion Section of Chapter 2, is novel in its comprehensive integration of **content** analysis, **social context** evaluation, and **external evidence** assessment. Additionally, ExFake navigates the intricate landscape of online information, providing users with transparent and interpretable **explanations** for its decisions.

This chapter investigates the fusion of *content* analysis, *social context* exploration, and *external evidence* from *trusted sources*, thereby ushering in a new era of false information detection. ExFake leverages data from reputable *fact-checking organizations* and *named entities* (i.e., *official Twitter accounts*), continually updating its knowledge to adapt to evolving information landscapes. By assessing *source credibility* and employing *similarity analysis*, ExFake uncovers hidden patterns and relationships within large data volumes.

Moreover, ExFake introduces a *Bayesian average* technique for source credibility analysis. This technique is similar to the *five-star rating systems* familiar in *e-commerce*. This method assesses users' likelihood to share false information based on their historical sharing behaviour.

The core of ExFake is its *decision-making module*, driven by *neural networks*, which calculates an output score for the input online post. This output score referred to as the "confidence percentage", quantifies the credibility of the post, indicating the likelihood of it being fake or genuine. However, ExFake goes beyond being a mere black-box classifier. Its explanation module, built on Explainable AI (XAI) and Natural Language Inference (NLI) techniques, provides users with a deeper insight into its decision-making process. This module bridges the gap between the automated model's output and user understanding, encouraging critical thinking and enabling informed decision-making.

To validate the efficacy of ExFake, it was implemented and evaluated using the publicly available FakeNewsNet dataset. This dataset was enhanced with additional features, including the "*legitimacy score*", which assesses the credibility of online users based on their historical sharing and posting behaviour, particularly focusing on whether they tend to share more genuine or fake content. Additionally, the dataset incorporates data retrieved from *named entities*, enriching the analysis. This chapter presents a pioneering exploration of false information detection in ExFake, which integrates content analysis, social context examination, and external evidence evaluation with explainability.

6.2. ExFake yystem overview: explainable fake news Ddetection system

In this section, details about the system, ExFake are provided. ExFake operates by continuously collecting data from specific sources. The system consists of two main components: one for processing data and assessing sources, which includes three modules (i.e., **Ex-Fact**, **Ex-Source**, and **Ex-Entity**), and another single module, **Ex-Decision**, which calculates the confidence percentage of input data while providing an explanation. The architecture of ExFake is illustrated in Figure 2.

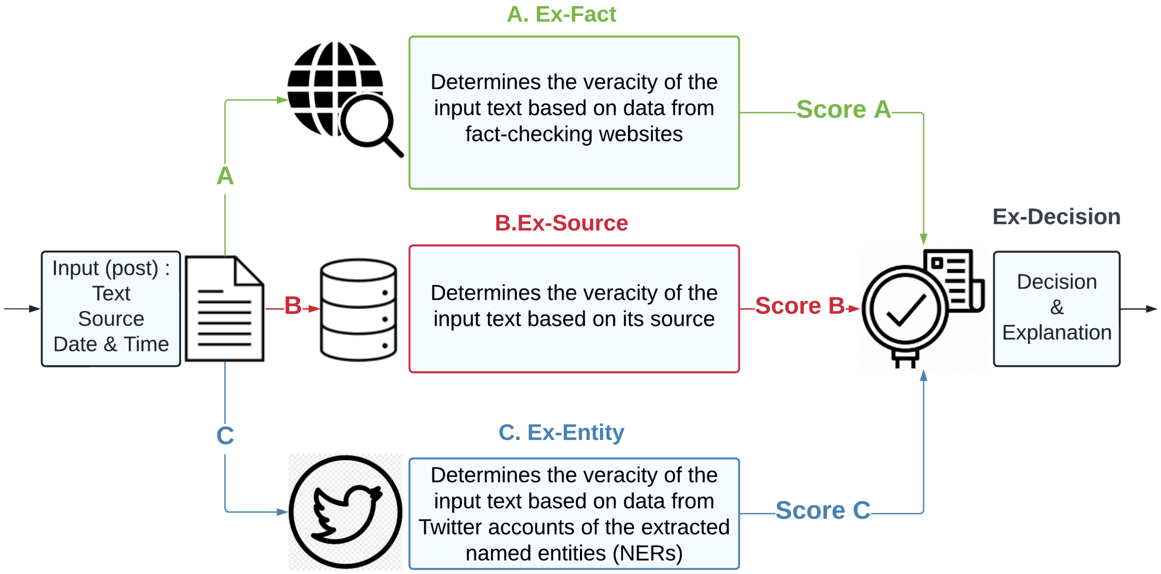

A. Ex-Fact
Determines the veracity of the input text based on data from fact-checking websites
Score A
A
B.Ex-Source
Input (post) : Text Source Date & Time
B
Determines the veracity of the input text based on its source
Score B
Ex-Decision
Decision & Explanation
C
C. Ex-Entity
Determines the veracity of the input text based on data from Twitter accounts of the extracted named entities (NERs)
Score C

Figure 2 – ExFake architecture

The Ex-Fake algorithm, presented in Algorithm 6.2.1, explains how ExFake works when given an online post as input. The system expects an input in the form of a Twitter post and starts by extracting and preprocessing data, including the publication date and time of the input tweet.

Algorithm 6.2.1. Ex-Fake Algorithm

```
data preprocessing()
$t = 0$;
For $i \in \{0, \ldots, 2\}$ :
    Do in parallel :
        Ex-Fact();
        Ex-Source();
        Ex-Entity();
    End do in parallel
    sleep(t);       %freezes the execution of the loop
    Ex-Decision(t); %computes the final percentage and generates the
                    explanation at the timestamp t
    $t = t + 1$;
End for
update();           %updates the dataset and the legitimacy score of all
                    users
Return Decision: (Ex-Decision.percentage, Ex-Decision.explanation);
```

The ExFake system provides three percentages at different timesteps, as indicated by the for-loop. While a percentage is returned at each timestep, ExFake continues to receive data (from fact-checking websites and named entities extracted from the input text) during the freeze time caused by the sleep() function. Once all the scores are computed, the Ex-Decision module calculates the final confidence percentage of the input post and generates an explanation. Finally, the dataset and legitimacy scores of all users are updated, and the final decision is made.

6.2.1. Ex-Fact

The Ex-Fact module, depicted in Figure 3, is responsible for computing a score based on metadata extracted from the input post, including text, date, and time. It also utilizes data obtained from trusted fact-checking websites. Although, for simplicity reasons, the dataset currently includes only PolitiFact.com. This module's primary task is to identify articles closely related to the input post among all past and upcoming articles on the fact-checking website. Using these highly similar articles, the module calculates a score based on the inferred relationship with the input post. The limitation of using a single fact-checking organization, and the solutions to address this issue, will be further explored in the upcoming chapter on the *AFCC system* (i.e., Chapter 8).

The initial output score for text similarity in the Ex-Fact module ranges from -1 to 1. This score is then normalized to a range of 0 to 1 for clarity and consistency. In this normalized scale, if the score of an article is 0.8 or higher, it means the article is considered similar to the input post. Specifically, a score of 0.8 or above indicates a strong resemblance in content or subject matter between the input post and the article from the fact-checking website. Such a high score suggests that the article may be discussing the same topic, presenting similar facts or assertions as those found in the input post.

Therefore, an article is deemed closely related to the input post if its normalized score is 0.8 or greater. This threshold of 0.8 is set to ensure that only articles with a high degree of similarity to the input post are taken into account for further analysis.

The final score in Ex-Fact is the average of points obtained based on these similarity assessments. To compute this final score, the following rules are applied to articles that meet this predefined threshold of similarity:

— An article that entails the input post is worth 100 points. In the context of NLI, "entailment" refers to a situation where the truth of one statement (i.e., the article in this case) logically guarantees the truth of another statement (i.e., the input post). Essentially, if an article from a fact-checking website supports or verifies the content of the input post, it is given the highest score of 100 points.

— A neutral article is worth 50 points. In NLI, "neutrality" indicates that the article neither explicitly supports nor contradicts the input post. It may provide related information but does not directly affirm or refute the post's content. Hence, it receives a mid-range score of 50 points.

— An article that contradicts the input post is worth 0 points. In NLI, "contradiction" occurs when the truth of one statement (i.e., the article) implies the falsehood of the other (i.e., the input post). An article that fact-checks and disproves the input post's content will thus receive the lowest score of 0 points.

6.2.2. Ex-Source

The Ex-Source module, as illustrated in Figure 4, computes and returns the legitimacy score of the online user (i.e., source) who posted the input post, based on their past online posts (i.e., posting history). For *new users* who do not have a posting history, this module assigns them an initial legitimacy score of *50%*. This starting score serves as a neutral baseline, reflecting an average level of trustworthiness. As the user continues to post content, their legitimacy score will be dynamically adjusted. The adjustments are based on the credibility percentages of their posts and their comparison to other users on the platform. The legitimacy score is calculated using the Bayesian average method, which combines the credibility of individual posts with the average credibility across all users, ensuring a balanced

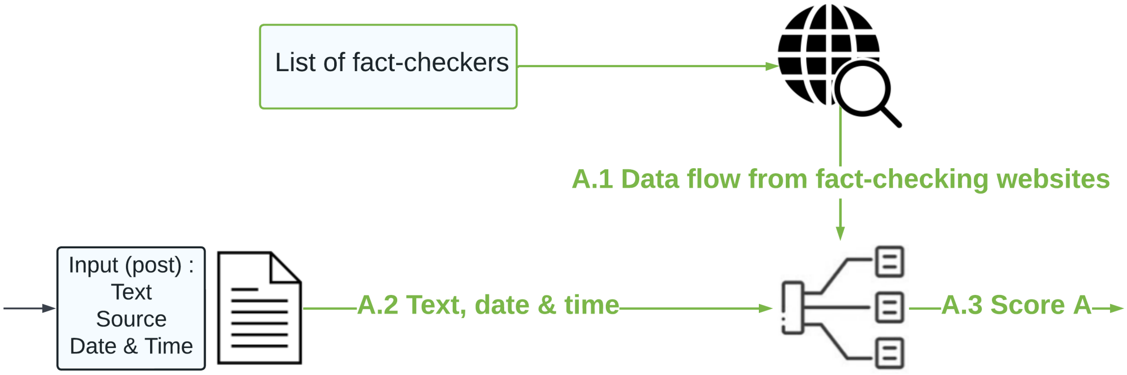


Figure 3 – Ex-Fact module

and statistically sound measure of each user's reliability. The Ex-Source module maintains a record of the credibility percentages of each user's past posts, and after each complete system execution in response to a request (i.e., an input post), updates the legitimacy scores for all users. This dynamic approach ensures that the legitimacy score is continually updated to reflect the most recent online behaviour and content dissemination patterns of the users.

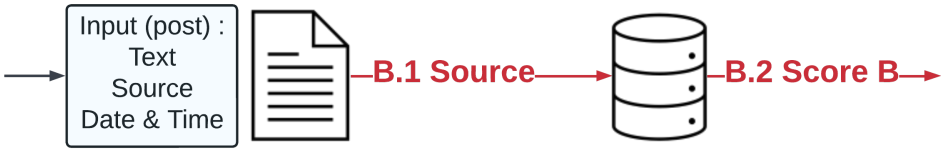


Figure 4 – Ex-Source module

The legitimacy score in the Ex-Source module follows the same logic of the five-star rating system commonly used in e-commerce websites. It employs the Bayesian average technique[1] to provide a more balanced and fair evaluation, especially for users with fewer past posts (i.e., those below a certain threshold). This technique ensures that users with limited posting history exert a lesser influence on their final score compared to users with a more substantial posting history.

Consequently, a user with a long history of posting truthful content will have a higher legitimacy score than a new user or one with fewer posts. For example, a user who has over one hundred genuine posts will be deemed more credible than another user with only two posts, even if both are genuine. The legitimacy score (LS) of a user/source is calculated to produce a value between 0 and 100 (i.e., percentage), as defined in Eq. (6.2.1). A higher

[1] `https://www.algolia.com/doc/guides/managing-results/must-do/custom-ranking/how-to/bayesian-average/`, last accessed on February 17, 2023.

score represents greater trustworthiness for posts made by this user/source, indicating a more reliable contributor to the online social network.

$$LS = \frac{M \times N + D \times E}{N + E} \tag{6.2.1}$$

Where M represents the mean of the percentage of all the received posts from this user/source, N denotes the number of posts received from this user/source, D stands for the mean of all percentages across the entire database, and E indicates the minimum number of received posts to be listed. It is important to note that E was set to 1.

The decision to use *post history* instead of assigning a score to a user based on features such as their location, job title, number of followers, etc., was made due to the limited research that utilized post history on Twitter. Additionally, unlike in other research, the legitimacy score in ExFake, is assigned to a user based on their previous tweets and the previous tweets of all users using the Bayesian average. Experimental results have indicated that when used alone, the legitimacy score outperforms most state-of-the-art baselines.

6.2.3. Ex-Entity

The Ex-Entity module, as illustrated in Figure 5, operates similarly to the Ex-Fact module, with the main difference being the source of external data. Ex-Entity receives data from legitimate entities, specifically the named entities mentioned in the input post. The module's first step is to extract all the named entities from the input post using named entity recognition (NER). The second step involves retrieving data from these named entities. Subsequently, Ex-Entity identifies similar posts that have been posted by the accounts of the extracted named entities and computes a score based on the inferred relationship with the input.

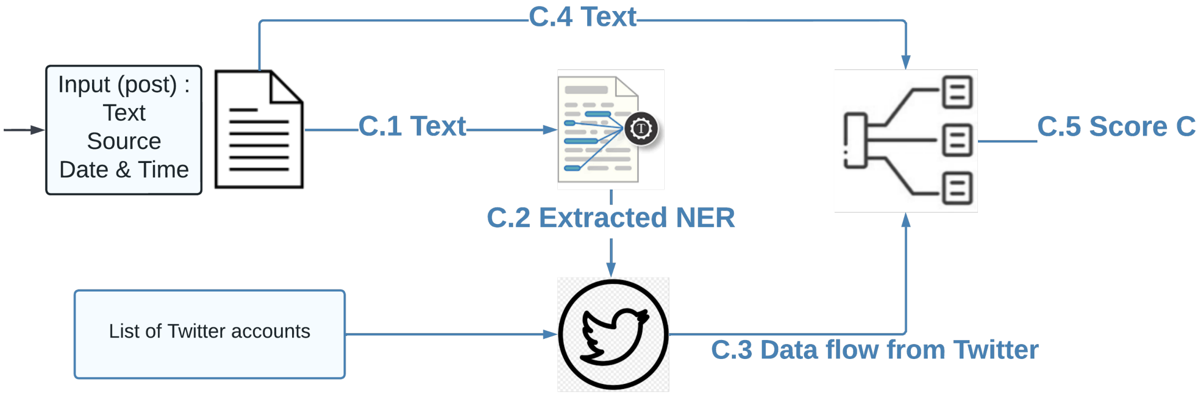


Figure 5 – Ex-Entity module

For example, For example, considering the following input tweet: "Clinton: Trump called women dogs: At the first presidential debate at Hofstra University, Hillary… #Skibabs", the module identifies *named entities* such as Trump, Hofstra University, and Hillary. ExFake maintains a database of the official certified Twitter accounts of frequently mentioned named entities. Ex-Entity initiates data retrieval from the official social accounts of individuals like Donald Trump and Hillary Clinton. It then searches for similar tweets from named entities that can be linked to the input tweet. If the input post lacks any named entities or if relevant named entities meeting the threshold cannot be retrieved, the module assigns an initial score of 50%. This initial score serves as a neutral baseline, indicating a lack of substantial evidence to either confirm or refute the veracity of the input post.

6.2.4. Ex-Decision

The Ex-Decision module, described in Figure 6, has two key tasks: decision-making and providing explanations. Firstly, it computes the final confidence percentage of the input post based on scores from the three preceding modules, namely Ex-Fact, Ex-Source, and Ex-Entity. Secondly, it furnishes explanations to system users. These explanations are derived from two of the previous modules, Ex-Fact and Ex-Entity.

For Ex-Fact and Ex-Entity modules, explanations are generated based on the final Natural Language Inference (NLI) task of each of these modules. This NLI-based explanation offers insights into the background of the confidence percentage assigned to users' input posts. This enables OSN users to assess the trustworthiness of the input posts before sharing them on social networks.

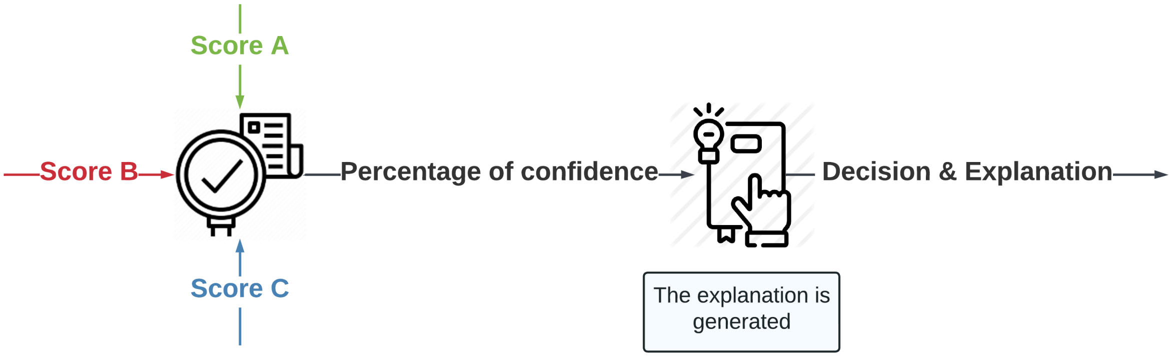


Figure 6 – Ex-Decision module

Ex-Decision calculates the confidence percentage using a neural network model, as depicted in Figure 7. The scores from Ex-Fact (score A), Ex-Source (score B), and Ex-Entity (score C) serve as inputs X_i. These inputs are weighted by neuron weights W_i, followed by the addition of a bias b. The sum Z of these results is then passed through a sigmoid activation function

$f(z)$, as defined in Eq. (6.2.2). The output of this activation function ranges between 0 and 1 and is converted into a percentage, representing the final confidence of the input post (i.e., tweet).

$$f(z) = \frac{1}{1 + e^{-x}} \tag{6.2.2}$$

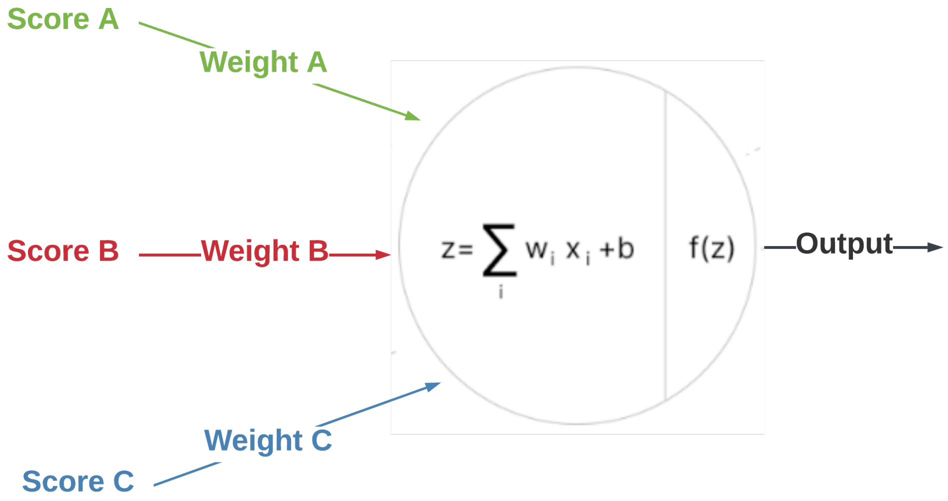


Figure 7 – Ex-Decision Neural Networks

Unlike typical systems that merely find and highlight important words within the input text, the explainer in ExFake is designed to cultivate specific user behaviour when confronted with information on online social networks. Ex-Decision returns the text from trusted organizations and legitimate entities, highlighting the words that most influenced the result within these trusted and legitimate texts. Trusted organization text is typically an article from a fact-checking website, while legitimate entity text is a post from a named entity mentioned in the input text. Additionally, this module provides information about the sources that published the post, including the published dates and times, and links to the articles or tweets. The aim is to demonstrate the simplicity of verifying certain information by visiting the mentioned named entity's official account or a reliable website.

6.3. Evaluation and discussion

This section describes the experimental evaluation of ExFake compared to benchmark methods on the FakeNewsNet data repository [252]. It begins with the detailing of the evaluation metrics, the benchmark used in the comparative analysis, and the data collection and processing. Subsequently, the empirical results on the FakeNewsNet dataset are presented, followed by a discussion of the results obtained.

6.3.1. Evaluation metrics

ExFake is a multi-class classification problem. Unlike binary-class classification problems, ExFake returns a percentage of confidence regarding the trustworthiness of the input post's content. This percentage is mapped to the rating labels defined in the *PolitiFact* dataset, as illustrated in Table 1. Therefore, to assess the performance of ExFake, the evaluation employed metrics such as *macro-accuracy*, *macro-precision*, *macro-recall*, and *macro-F1-score*, as outlined in Table 2.

Table 1 – Mapping scheme of the PolitiFact dataset labels to the percentage of confidence

Dataset label	***ExFake percentage of confidence***
True	[87% - 100%]
Mostly true	[70% - 86%]
Half true	[53% - 69%]
Barely true	[37% - 52%]
False	[21% - 36%]
Pants on fire	[0% - 20%]

Table 2 – Definition of classification evaluation metrics used in this study

Metric	**Formula**	**Description**
Macro-accuracy	$Macro_{accuracy} = \frac{\sum_{k=1}^{K} Accuracy_k}{K}$	Average accuracy across classes
$Accuracy_k$	$A_k = \frac{TP+TN}{TP+TN+FP+FN}$	Accuracy for class k
Macro-precision	$Macro_{precision} = \frac{\sum_{k=1}^{K} Precision_k}{K}$	Average precision across classes
$Precision_k$	$P_k = \frac{TP}{TP+FP}$	Precision for class k
Macro-recall	$Macro_{recall} = \frac{\sum_{k=1}^{K} Recall_k}{K}$	Average recall across classes
$Recall_k$	$R_k = \frac{TP}{TP+FN}$	Recall for class k
Macro-F1-score	$Macro_{F_1} = \frac{\sum_{k=1}^{K} F_{1_k}}{K}$	Average F1-score across classes
F_{1_k}	$F_{1_k} = 2 \times \frac{P_k \times R_k}{P_k + R_k}$	F1-score for class k

In the equations, listed in Table 2, K is the number of labelling classes. TP denotes $TruePositive$ signifying the correctly detected instances of fake news. TN stands for $TrueNegative$ and represents the number of negative samples accurately identified as non-fake news. FP refers to $FalsePositive$ indicating instances incorrectly labelled as fake news when they are not. Lastly, FN corresponds to $FalseNegative$ representing instances mistakenly identified as genuine when, in fact, they are fake news.

6.3.2. Data collection and processing

This subsection provides an overview of the dataset employed, the methodology for data gathering, and the steps involved in refining and preparing the data for subsequent analysis.

6.3.2.1. **Dataset:** One of the main issues faced by researchers is the scarcity of comprehensive and community-driven fake news datasets. Not only are existing datasets sparse, but they also lack a variety of aspects commonly required in research, such as news content, social context, and spatial data. The FakeNewsNet data repository [252] has been employed in ExFake to address this lack of quality datasets. To the best of my knowledge, FakeNewsNet is recognized as the sole dataset that incorporates a wide array of features, including detailed news content, comprehensive social context, and spatiotemporal information, as indicated in Table 3. This table presents statistical data collected from PolitiFact. Notably, FakeNewsNet not only includes news articles but also associated tweet URLs, enhancing its depth and suitability for research purposes.

PolitiFact is a dedicated fact-checking website that primarily focuses on political content. Its team of journalists and domain experts assesses political posts, often in the form of tweets, and assigns them one of several labels, including "pants on fire", "false", "barely true", "half true", "mostly true", and "true". In contrast, Gossip Cop specializes in fact-checking entertainment-related stories. GossipCop rates news stories on a scale from 0 to 10, with higher scores indicating a higher degree of truthfulness. In this context, a higher score implies that the news is more likely to be accurate.

The work undertaken enhances the diversity of the dataset by adding additional social context features. Initially, the system identifies and extracts named entities from the input text. The Twitter usernames of the most frequently mentioned named entities within the dataset are manually incorporated into a dedicated database. When a named entity is recognized, and its Twitter username is available in the database, ExFake establishes a data connection with this corresponding Twitter account, facilitating the receipt of relevant data. Additionally, each source whose input has been processed by ExFake is assigned a legitimacy score. Thus, posts from named entities mentioned in the input and the legitimacy score enrich the FakeNewsNet dataset with more social context features. Consequently, the FakeNewsNet dataset gains increased richness through the inclusion of posts from named entities mentioned

Table 3 – Statistics of the FakeNewsNet repository

	Category	***Features***	***PolitiFact***	
			Fake	***Real***
NewsContent	Linguistic	# News articles	432	624
		# News articles with text	420	528
	Visual	# News articles with images	336	447
SocialContext	User	# Users posting tweets	95,553	249,887
		# Users involved in likes	113,473	401,363
		# Users involved in retweets	106,195	346,459
		# Users involved in replies	40,585	18,6675
	Post	# Tweets posting news	164,892	399,237
	Response	# Tweets with replies	11,975	41,852
		# Tweets with likes	31,692	93,839
		# Tweets with retweets	23,489	67,035
	Network	# Followers	405,509,460	1,012,218,640
		# Followees Average	449,463,557	1,071,492,603
		# followers Average	1299.98	982.67
		# followees	1440.89	1040.21
Spatiotemporal Information	Spatial	# User profiles with locations	217,379	719,331
		# Tweets with locations	3,337	12,692
	Temporal	# Timestamps for news pieces	296	167
		# Timestamps for response	171,301	669,641

in the input post and the incorporation of legitimacy scores of the sources, which contribute to an augmented set of social context features.

To facilitate the experiments, a decision was made to work with a subset of the data. This choice was driven by the extensive number of experiments and the time required for their execution. Table 4 displays the distribution of data within the experiment dataset.

For this subset, a total of 15,000 tweets were randomly selected and divided into three subsets: training, validation, and test. The training set comprises 80% of the subset dataset, while both the validation and test sets each account for 10%. The same proportion was applied to a selection of 150 articles, ensuring the inclusion of relevant content on the same topics as the chosen tweets.

6.3.2.2. **Data preprocessing:** In the data preprocessing phase, the time complexity of training and fine-tuning the models had to be addressed. Consequently, the number of features from the FakeNewsNet dataset was limited. The complete dataset was exclusively utilized for performance comparisons between ExFake and the benchmark. Articles and tweets published in the year 2016 or content directly related to that specific year were the focus of attention. This selection criterion encompassed tweets from the beginning of 2017 or

Table 4 – Experimental-data distribution

Features	**Training subset**	**Validation subset**	**Test subset**
Content of articles	120	15	15
Articles author's name	112	13	14
Article published date and time	120	15	15
Content of tweet	12000	1500	1500
Tweet author's name (username)	9812	878	941
Tweet published date and time	11868	1282	1330

the end of 2015 but related to articles published in 2016. This temporal focus was motivated by the significant propagation of fake news during the 2016 US presidential election.

The preprocessing involved retaining all relevant tweet information, including information about the author (i.e., user details), labels (i.e., target categories or classifications), publication date, and time, which served as the input for ExFake. The results of this feature transformation are illustrated in Table 5. The preprocessing procedure began with the conversion of all text to lowercase, followed by the removal of stop words, punctuation marks, emojis, and symbols. Subsequently, the data was transformed into tokens, with the final step involving stemming the dataset.

Table 5 – Features of the transformed dataset

Aspect	***Source***	***Features***
Content-based	News article	The text content of the article
	Post (tweet)	The text content of the tweet
Context-based	News article	Author's name
		Published date and time
	Post (tweet)	Author's name
		Published date and time

6.3.3. Experiments

This section addresses the experiments carried out as part of this work. The first set of experiments is designed to elucidate the significance of each module within ExFake. The second set explores the performance of ExFake across different timesteps. These initial experiments employ the sub-dataset detailed in Section 6.3.2.1. Additionally, a comparative analysis of ExFake's results with other approaches utilizing the same dataset is provided. For this comparative analysis, the complete dataset was utilized.

The structure of this section unfolds as follows. First, an overview of the experimental configurations is presented. Subsequently, the findings from the timesteps experiments are explained. Then, experiments involving various configurations of ExFake's modules are discussed. Finally, an evaluation of ExFake's performance in comparison to other works using the same dataset is provided.

6.3.3.1. **Experimental settings:** In order to embark on the exploration of various experiments, it is imperative to establish consistent experimental settings, which were applied across the experiments detailed in the subsequent subsections (i.e., 6.3.3.2 and 6.3.3.3).

All experiments were conducted within the same computing environment - Google Colaboratory, utilizing a Colab Pro account. For a comprehensive understanding of this environment's specifications, please refer to Table 6.

Table 6 – Google Colab Pro environment specifications

Specification	**Value**
CPU model	Intel (R) Xeon (R)
CPU frequency	2.30 GHz
Number of CPU cores	2
RAM (Random Access Memory)	26.30 GB
Disk space	34 GB
GPU	NVIDIA Tesla P100 - PCIE
GPU memory	16 GB

6.3.3.2. **ExFake timesteps evaluation:** ExFake, operates in three distinct timesteps to provide percentage confidence regarding the input post. During these timesteps, the system continually ingests data from various sources. The objective is to gather additional relevant information between the timesteps to enhance the trustworthiness of the returned percentage, based on the most recent information acquired.

In this section, the focus lies on determining the optimal duration for each timestep, striking a balance between performance and speed. The aim is to ensure that the system delivers accurate results within a reasonable timeframe. Therefore, experiments are conducted to evaluate different timestep durations.

The experimentation was initiated with the employment of three different approaches. The first approach implements short-term timesteps, the second employs mid-term timesteps, and the last approach utilizes long-term timesteps. For each of these approaches, three distinct values were evaluated, which were arbitrarily selected.

Table 7 shows the different durations tested for each approach (i.e., short-term, mid-term, and long-term). The "time" column of "Timestep 1" represents the waiting time between the reception of a request (i.e., an online post) and the launch of the system. The "time" column of the other timesteps indicates the duration of the sleep() function as presented in the pseudocode of Algorithm 6.2.1) in the ExFake section (Section 6.2).

Table 7 – ExFake system Timesteps configuration and durations

Approach	Timestep 1	Timestep 2	Timestep 3
Short-term A	0 second	60 seconds	120 seconds
Short-term B	30 seconds	90 seconds	150 seconds
Short-term C	60 seconds	120 seconds	180 seconds
Mid-term A	1 minute	5 minutes	10 minutes
Mid-term B	2 minutes	8 minutes	16 minutes
Mid-term C	5 minutes	15 minutes	25 minutes
Long-term A	25 minutes	50 minutes	75 minutes
Long-term B	30 minutes	60 minutes	90 minutes
Long-term C	50 minutes	100 minutes	150 minutes

Inspired by Grandini et al. [101], the F1-score is used as a crucial metric to evaluate how well the algorithm performs on all classes and find the optimal timestep configuration.

Figure 8 depicts the results of the short-term approach experiments, where the time range is between 0 and 180 seconds after receiving a request (i.e., an input post). The macro F1-score ranges from a minimum of 58% to a maximum of 65%. Notably, these results indicate that attempting to provide a percentage shortly after a request does not produce satisfactory outcomes, and the score does not increase significantly over time.

The results of the mid-term approach experiments are presented in Figure 9. In this case, the time range considered is between 1 minute and 25 minutes. The recorded macro F1-score ranges from a minimum of 58% to a maximum of 85%. Notably, in this family of approaches, the highest score of 85% is achieved at the third timestep of the mid-term C configuration. This represents the highest performance observed with ExFake throughout all the conducted experiments and tests.

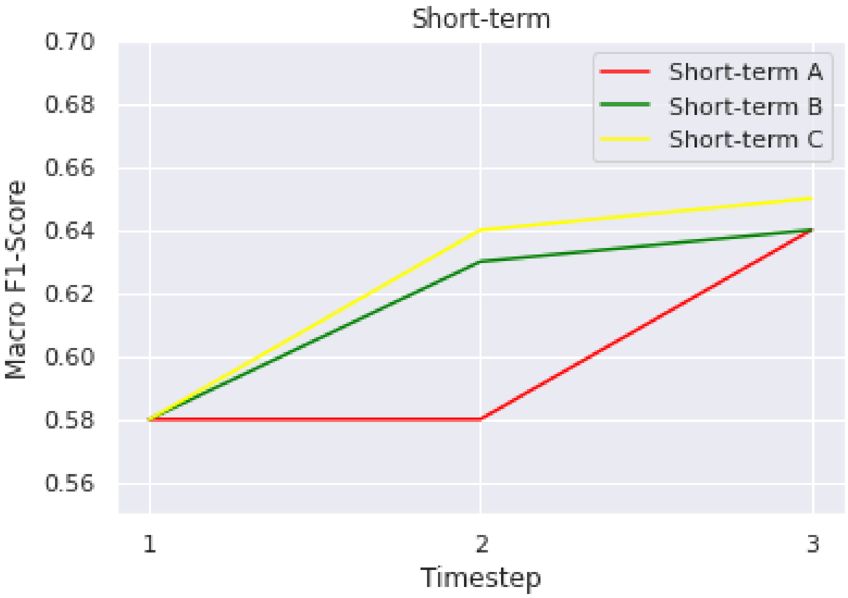


Figure 8 – Performance of Short-Term Timesteps

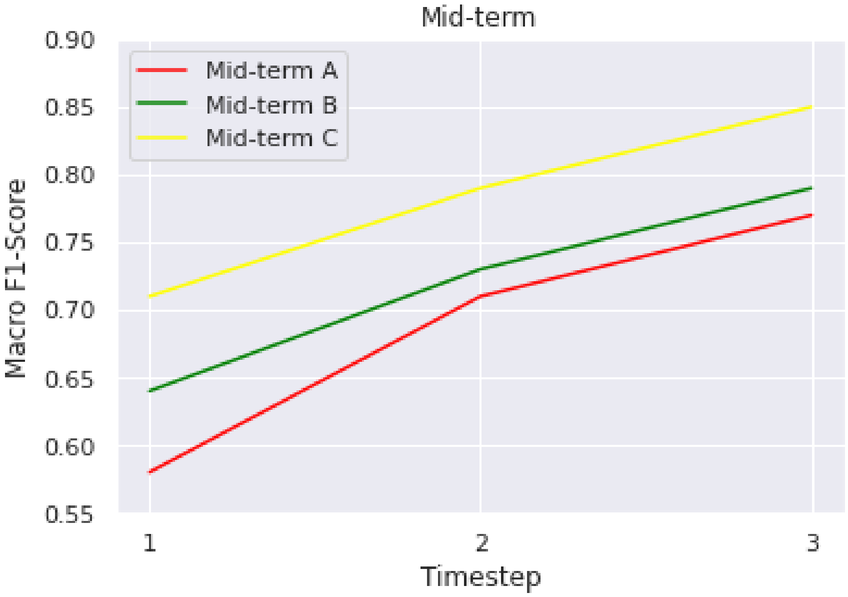


Figure 9 – Performance of Mid-Term Timesteps

The results of the long-term strategy experiment are depicted in Figure 10. The duration of the experiment spans from 25 to 75 minutes. It is noteworthy that the three lines in the graph closely overlap one another, resulting in a consistent score that does not vary across the different tested timesteps within this category. The macro F1-score consistently starts and remains at 85%. This phenomenon is attributed to the fact that between 25 and 75 minutes after the process initiation, ExFake did not encounter any significant posts or articles that met the required threshold to significantly influence the calculated confidence percentage of the input.

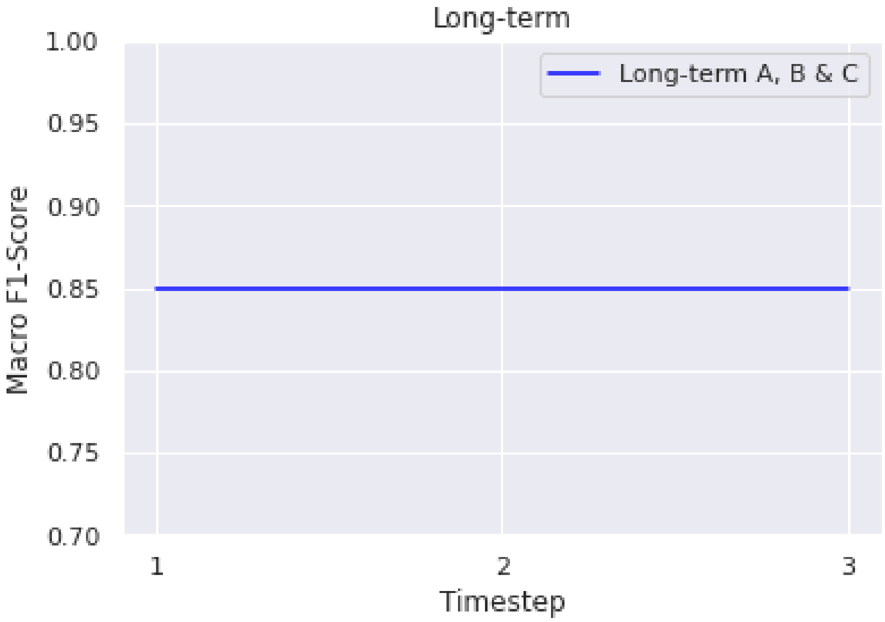


Figure 10 – Performance of Long-Term Timesteps

The foregoing observations prompted the exploration of a new approach that combines elements of both short-term and mid-term methodologies. The long-term approach was excluded from further consideration based on findings that showed no significant performance improvement beyond 25 minutes. This result is attributed to the construction of the dataset, which includes tweets associated with published news articles. Typically, these articles are fact-checked shortly before or after the corresponding tweets are posted. Therefore, further fine-tuning of ExFake beyond this time frame is unlikely to yield better results.

In the analysis of the short-term approach, it was observed that providing confidence percentages over a short period can be effective but does not consistently yield high performance. This behaviour can be explained by the fact that ExFake processes a substantial volume of data streams, requiring similarity checks and inferences for each retrieved article and post, which is a time-intensive process. Consequently, the short-term approach does not allow sufficient time for comprehensive computations, leading to limited improvements in results.

Yet, the mid-term approach allows for adequate time for calculations, leading to improved scores as the timesteps progress. However, achieving a balance between performance and speed is essential. Various combinations of the mid-term and short-term approaches were explored, resulting in the mix-term timesteps configuration. Table 8 depicts the selected mix-term timesteps for the system, determined after evaluating different configurations within this combined approach.

Figure 11 shows the macro F1-scores associated with the selected timesteps (i.e., Mix-Term Timesteps). It illustrates a nearly linear trend, commencing at a score of 64% and culminating at 85%.

Table 8 – Selected Timesteps for ExFake optimization

Approch	**Timestep**	**Duration**
Mix-term	Timestep 1	2 minutes
	Timestep 2	10 minutes
	Timestep 3	20 minutes

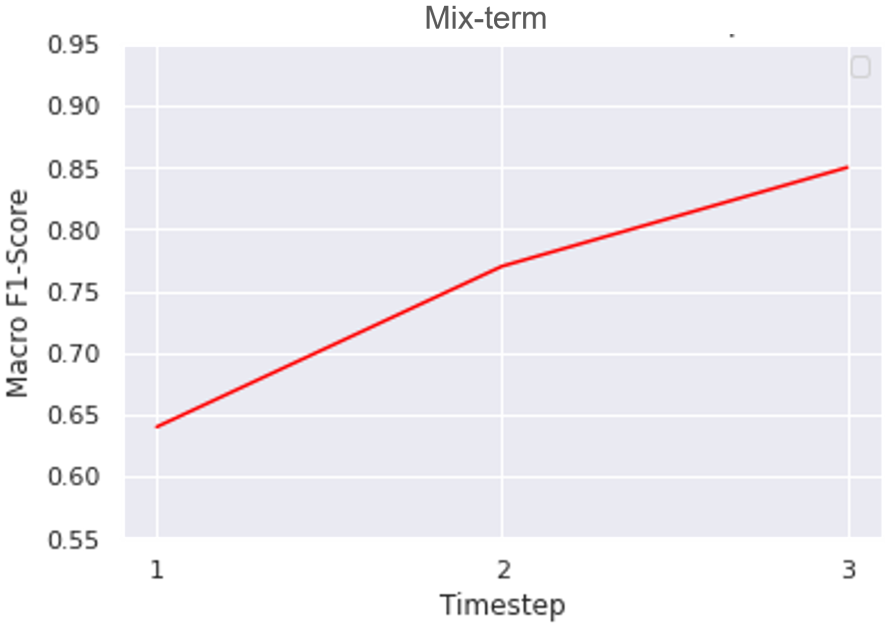


Figure 11 – Performance of Mix-Term Timesteps in achieving Macro F1-scores

6.3.3.3. **ExFake modules:** In the second experiment, the focus was on understanding the impact of each of the ExFake modules. The goal was to gain a deeper understanding of the functioning of each component within the system. To achieve this, tests were conducted with various combinations of these modules.

All experiments were performed using the data described in Subsection 6.3.2 and following the specific timesteps outlined in Subsection 6.3.3.2. A summary of the different module combinations can be found in Table 9.

It is important to note that these combinations exclude the component of the Ex-Decision module responsible for generating explanations. When ExDecision is integrated into the combinations, the focus is solely on the neural network's computation of the confidence percentage (i.e., the output score) returned by ExFake.

Table 9 – Module combinations in ExFake experiments

Experiment Combination	Ex-Fact	Ex-Source	Ex-Entity	Ex-Decision
Combination 1	X			
Combination 2		X		
Combination 3			X	
Combination 4	X			X
Combination 5		X		X
Combination 6			X	X
Combination 7	X	X		X
Combination 8		X	X	X
Combination 9	X		X	X

The Ex-Source module is primarily reliant on the historical data associated with each user. The legitimacy score produced by this module is independent of any computational process. Therefore, for experiments involving only Ex-Source or Ex-Source in conjunction with Ex-Decision, it is assumed that whenever a request is submitted by a particular user/source, Ex-Source would return the legitimacy score as it typically does. However, when it comes time to update the database, it is considered that ExFake has accurately determined the class, and the user's historical data is accordingly updated.

The initial combinations (i.e., Combinations 1, 2, and 3) consist of exploring the individual performance of the core modules in Ex-Fake, Ex-Fact, Ex-Source, and Ex-Entity, which are designed to work in conjunction. Testing the Ex-Decision module alone was not considered because it cannot operate without the support of at least one of the aforementioned modules. Figure 12 presents the macro F1-scores for Combinations 1, 2, and 3.

Combination 1, featuring only the Ex-Fact module, outperforms the first three combinations with a final score of 72% at timestep 3. Additionally, tests were conducted with varying timesteps using Ex-Fact alone. It was observed that the module's performance is directly proportional to the timestep duration. The score increases as the duration of each timestep increases, and vice versa.

Secondly, as anticipated, the score for Combination 2, which solely employs Ex-Source, remains consistent across different timesteps. The legitimacy score of a user/source only changes after the final result has been computed. During query execution, the legitimacy score remains unchanged.

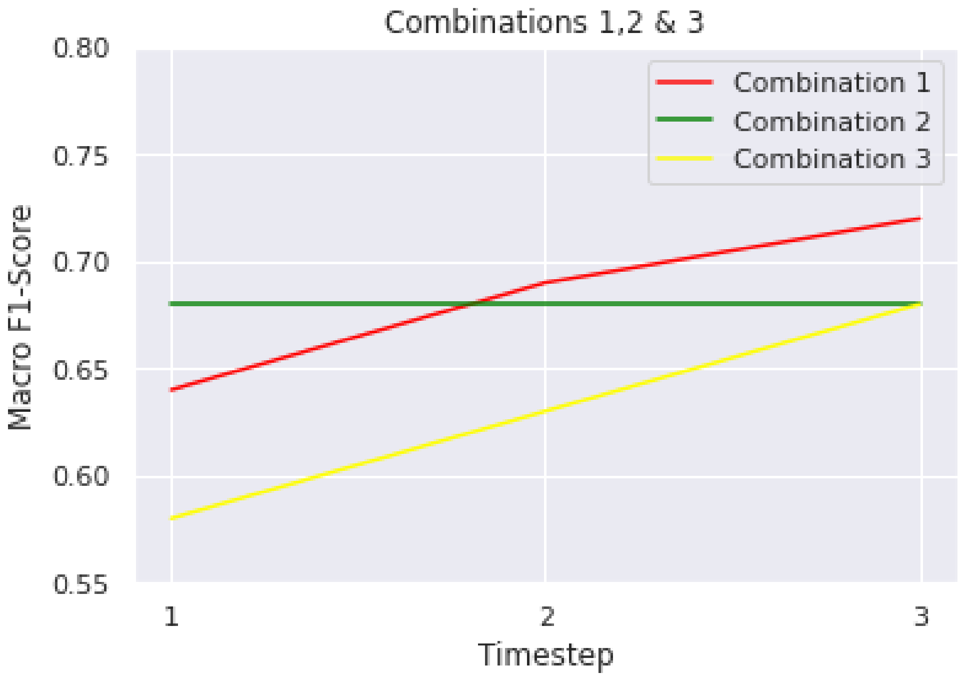


Figure 12 – The Macro F1-score of combinations 1, 2 and 3

However, expectations were not met when a score of 68% was achieved with Ex-Source operating in isolation. To explain this result, the frequency of posts received from individual user/sources was explored, as Ex-Source relies on the historical data of each user/source.

Figure 13 illustrates the percentage of users/sources with varying numbers of posts. Notably, only 1.3% of users/sources have more than 50 posts, while 21.3% of users/sources have between 11 and 50 posts. Consequently, the obtained result can be attributed to the fact that 35.3% of users/sources lack a substantial history as they have only made a single post.

The third combination, featuring only Ex-Entity, yielded the lowest score at timestep 1 and finished at timestep 3 with a score of 68%, on par with combination 2. Once again, the dataset was explored to comprehend this result.

The hypothesis revolved around the presence of named entities in the dataset because Ex-Entity operates on data streams received from named entities within input posts. It was discovered that 85.7% of tweets contain at least one named entity. Despite this relatively high percentage, the score remained notably distant from that of combination 1, which also extracts data streams. Further investigations were conducted to dissect this result. Consequently, it can be concluded that the result fell below expectations due to the realization that the database of official and certified named entity Twitter accounts is incomplete. It omits certain named entities, and among those listed in the database, several rarely tweet.

Combinations 4, 5, and 6 are quite similar to the first three combinations, with the addition of the Ex-Decision module at the end of each preceding module. Figure 14 presents

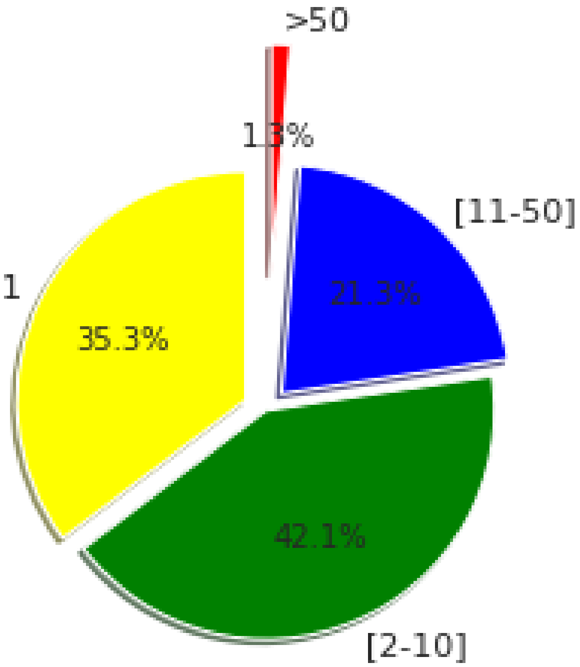


Figure 13 – The percentage of users/sources with a certain number of posts

the macro F1-score of these combinations. It is evident that the Ex-Decision module improved the performance of all the modules (Ex-Fact, Ex-Source, and Ex-Entity). While the increase is not huge, the inclusion of a neural network has a positive impact on the results.

The last three combinations are composed of a blend of two parallel-running modules and the Ex-Decision module. These mixtures could not be tested without the Ex-Decision module since it was required to compute the returned percentage. Each module, whether Ex-Fact, Ex-Source, or Ex-Entity, produces a score. Therefore, the neural network is essential to aggregate these scores into a single percentage. Figure 15 displays the macro F1-score of these combinations.

Combination 7, which combines the Ex-Fact and Ex-Source modules, attained the highest score at timestep 1 but finished second at timestep 3 with a score of 75%. Combination 8, merging the Ex-Source and Ex-Entity modules, scored the lowest across all timesteps. This outcome was anticipated as this combination amalgamates modules that underperformed when used independently or with Ex-Decision. Combination 9 outperformed all other presented combinations with a score of 78% at timestep 3. Once more, this was expected as it combines the Ex-Fact and Ex-Entity modules, which achieved the best results when used alone or with Ex-Decision.

This experimentation illuminated the impact of each module. It is evident that to achieve peak performance, all four modules of Ex-Fake (Ex-Fact, Ex-Source, Ex-Entity, and Ex-Decision) must be incorporated. As demonstrated earlier, particularly when amalgamating

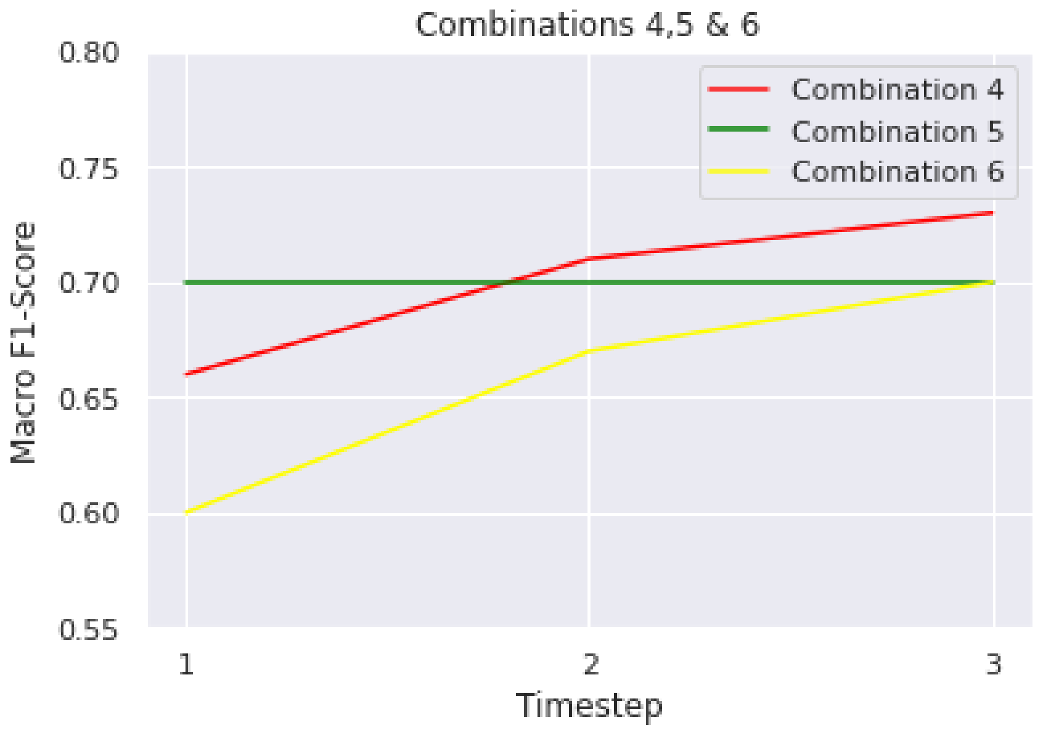


Figure 14 – The Macro F1-score of combinations 4, 5 and 6

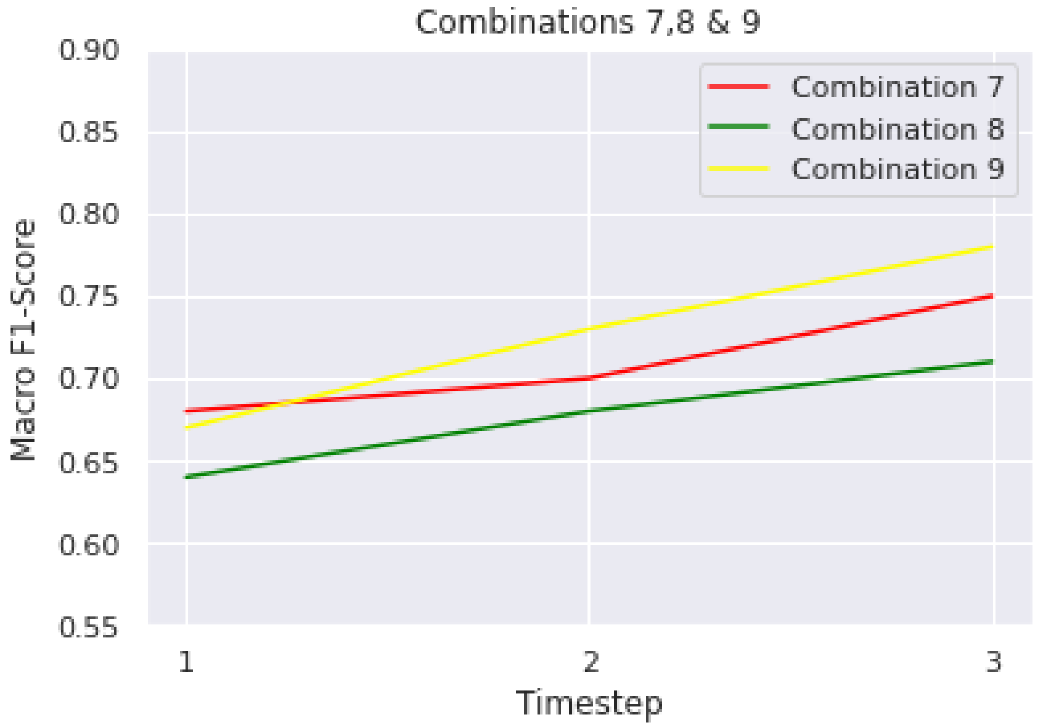


Figure 15 – The Macro F1-score of combinations 7, 8 and 9

modules, Ex-Decision is pivotal in consolidating different scores into a single percentage. Regarding the parallel-executing modules, while all are vital, with the dataset in use, it was observed that the Ex-Fact module had the most significant impact, followed by the Ex-Entity module. However, this order may vary depending on the dataset in use. Nevertheless, ExFake remains robust since each module compensates for the others. For instance, in a dataset lacking sufficient extractable named entities but featuring a comprehensive history of numerous users/sources, Ex-Source becomes more significant than Ex-Entity.

6.3.3.4. **Benchmark models:** In this subsection, the state-of-the-art models used to evaluate the performance of the ExFake system are presented. The performance of the ExFake system was benchmarked against seven learning models. These models align with those utilized by Shu et al. [252] in the FakeNewsNet framework, providing a foundation for a comprehensive performance comparison in the realm of fake news detection.

The models include four traditional machine learning approaches: Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), and Convolutional Neural Network (CNN). Additionally, three variations of the Social Article Fusion (SAF) models [250] were evaluated.

The SAF models [250] utilize an autoencoder with two-layer LSTM cells for encoding and decoding, along with a network of two-layer LSTM cells to capture the temporal patterns of user engagements. Figure 16 illustrates the architecture of the Social Article Fusion model (SAF). These SAF models are distinctive in their approach: SAF/S primarily targets the news content, SAF/A emphasizes the social context, particularly the temporal patterns of user engagements, and SAF combines both content-based and social context-based analyses for a more holistic approach. All three versions of the SAF model were used for the evaluation.

6.3.4. Results and discussion

In this section, the outcomes of the ExFake system are presented, composed of four key modules as outlined in Section 6.2. Sentence-BERT (SBERT) was utilized for text-similarity and natural language inference (NLI) tasks, while the Spacy tool was employed for named entity recognition (NER), and the Twitter web scraping API was integrated for data retrieval.

Specifically in ExFake, the SBERT (Sentence-BERT), a pre-trained BERT network utilizing Siamese and triplet network architectures to generate semantically relevant sentence embeddings, was used for the text-similarity task. The fine-tuning process was executed using the STS benchmark (Semantic Textual Similarity) [54], a benchmark renowned for assessing sentence similarity. Details of the hyperparameters used for SBERT training in the text similarity task are presented in Table 10. For the NLI task, SBERT underwent fine-tuning with the SNLI dataset [42]. The corresponding hyperparameters for this training are illustrated in Table 11.

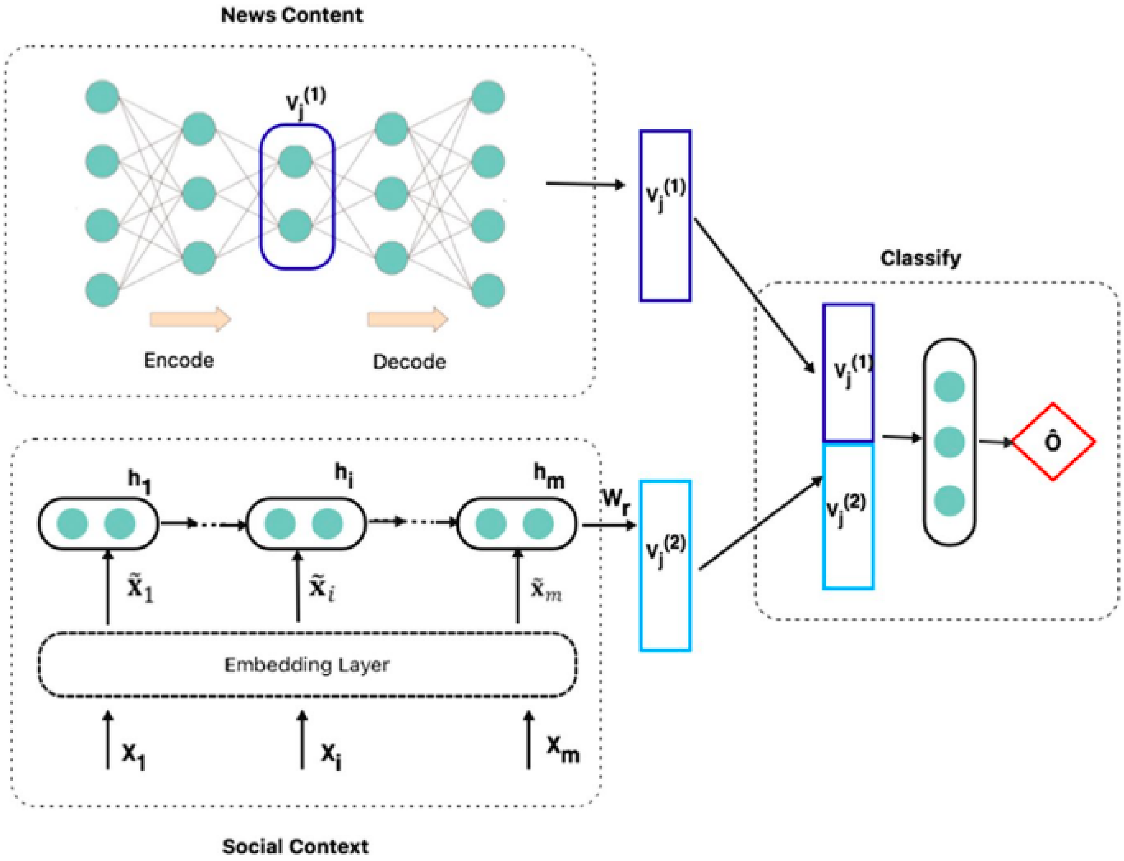


Figure 16 – The Architecture of the Social Article Fusion model (SAF) [250]

Table 10 – SBERT training hyper-parameters for the text similarity task

Hyper-Parameter	***Value***
Epochs number	4
Optimizer	Adam optimizer
Learning rate	2e-5
Weight decay	0.01
Regression loss	Mean squared-error loss
Batch-size	16
Random seeds	10

In the Named Entity Recognition (NER) task, the SpaCy tool was utilized to extract named entities from text. SpaCy recognizes various named entities and categorizes them into labels such as GPE (Geopolitical Entities: geographic entities like countries, cities, states, etc.), PER (persons: named individuals or families), ORG (organizations: companies,

Table 11 – SBERT training hyper-parameters for the NLI task

Hyper-Parameter	**Value**
Epochs number	5
Optimizer	Adam optimizer
Learning rate	1e-5
Weight decay	0.01
Regression loss	Sparse categorical cross-entropy loss
Batch-size	128

agencies, institutions, etc.), DATE (dates), LOC (locations), and MONEY (monetary values). In ExFake, the focus was on three labels: *ORG*, *PER*, and *GPE*. Subsequently, ExFake begins receiving data streams from the Twitter accounts of extracted named entities. For this process, ExFake utilizes a database containing Twitter accounts of these named entities. To establish this database, a manual search was conducted to identify and add the official and certified Twitter accounts of the top 100 named entities most frequently mentioned in the dataset. Table 12 presents the first 10 entries of this dataset.

Table 12 – List of Twitter accounts for the top 10 most mentioned named entities in the ExFake dataset

#	**Named Entity**	**Twitter Username**
1	Donald Trump	@realDonaldTrump
2	Hillary Clinton	@HillaryClinton
3	Barack Obama	@BarackObama
4	Bernie Sanders	@BernieSanders
5	The white house	@WhiteHouse
6	Ted Cruz	@tedcruz
7	FBI	@FBI
8	Mike Pence	@Mike_Pence
9	Marco Rubio	@marcorubio
10	Bill Clinton	@BillClinton

To address the explainability task, ExFake integrates an approach grounded in the NLI module, sharing similarities with the Lime (Local Interpretable Model-agnostic Explanations) method [222]. The primary goal is to identify the pivotal words that influence the model's predictions. This process involves systematically removing individual words from the hypothesis sentence and observing how this impacts the prediction score. The greater the variation in the score, the more confident it becomes that the removed word is indeed essential. This process culminates in providing explanations based on the list of important words.

Table 13 offers a comprehensive performance comparison between ExFake and seven other models, based on the performance evaluation metrics including macro-accuracy, macro-precision, macro-recall, and macro-F1-score.

Table 13 – Best performance comparison for fake news detection on FakeNewsNet

Model	**Metric**			
	Macro-accuracy	**Macro-precision**	**Macro-recall**	**Macro F1-score**
SVM	0.580	0.611	0.717	0.659
Logistic regression	0.642	0.757	0.543	0.633
Naive Bayes	0.617	0.674	0.630	0.651
CNN	0.629	0.807	0.456	0.583
Social Article Fusion /S	0.654	0.600	0.789	0.681
Social Article Fusion /A	0.667	0.667	0.579	0.619
Social Article Fusion	0.691	0.638	0.789	0.706
ExFake	**0.808**	**0.841**	**0.871**	**0.855**

The macro-accuracy of the baseline models hovers around 0.65, while ExFake excels with a macro-accuracy of 0.808. In terms of macro-precision, the baseline models exhibit a range of values from 0.600 (SAF/S) to 0.807 (CNN). ExFake surpasses this range with a macro-precision of 0.841. Among the baseline models, SAF achieves the best macro-recall, tied with SAF/S, at 0.789, while the CNN approach scores the lowest with 0.456. ExFake stands out with a macro-recall of 0.871. The most substantial performance improvement is observed in the macro F1-score, with ExFake achieving a score of 0.855. In comparison, the CNN model scores 0.583, and SAF leads the baseline models with 0.706. ExFake outperforms all state-of-the-art baseline models in the FakeNewsNet benchmark across all metrics.

The experiments not only highlight the enhanced performance of ExFake but also underscore the significance of two additional elements introduced to the fake news detection problem: the *legitimacy score* and the *combination of the two NLP tasks.* When used alone, the module that computes the legitimacy score based on Bayesian averaging yields an F1-score of 68%, outperforming five of the seven state-of-the-art baseline approaches. This behaviour is mirrored in the F1-score at the third timestep for the modules that encapsulate the text similarity and NLI tasks.

6.4. Conclusion

In this chapter, I have introduced the ExFake system, a novel approach within the FACTS-ON framework, designed for explainable fake news detection. ExFake distinguishes itself by integrating the content of the post, context-based auxiliary information, and data from external sources such as trusted fact-checking organizations and named entities. Comprising four modules—Ex-Fact, Ex-Source, Ex-Entity, and Ex-Decision, ExFake collaboratively assesses the credibility of an input post, providing users on online social networks (OSN) with clear explanations for enhanced discernment of false information.

The experiments conducted using real-world fake news datasets have underscored the efficacy of ExFake, demonstrating the importance of integrating various analytical components, namely text similarity, natural language inference, and data processing tasks for fake news detection.

During the development of ExFake, challenges were encountered in utilizing *multiple fact-checkers as external evidence*, primarily due to the difficulty in collecting data from various fact-checking organizations and the heterogeneity of their textual rating labels. Consequently, ExFake initially relied on only *one fact-checking organization*, underscoring the complexity posed by the diverse rating labels and decisions across different fact-checkers. This heterogeneity often resulted in different rating labels for the same news content, complicating the standardization of truthfulness ratings and the assessment of their credibility.

In response to these challenges, the next chapter introduces the Automated Fact-checkers Consensus and Credibility Assessment (AFCC) system, a significant development that emerged from the difficulties faced in the ExFake system. AFCC, another integral component of the FACTS-ON framework, specifically addresses the issues of heterogeneity in rating labels and decisions among fact-checkers. It focuses on building consensus and assessing the credibility of fact-checkers, thereby enhancing the trustworthiness of fake news detection. The system standardizes verdicts from various fact-checking organizations and evaluates their consensus, assigning credibility scores based on their historical accuracy and reliability. The AFCC system's development reflects my commitment to addressing the complexities of using multiple fact-checkers (i.e., fact-checking organizations) and ensuring that the conclusions drawn about the veracity of news are unified, unbiased, and credible.

The subsequent chapter on AFCC will explore how this system manages the challenges of diverse rating labels and decisions, automates the use of heterogeneous decisions from multiple fact-checking organizations, and evaluates the credibility of these organizations.

Integrating AFCC into the FACTS-ON framework significantly enhances its capabilities, enabling it to go beyond mere detection of false information. This integration empowers FACTS-ON to establish a trust-based detection approach, find consensus among various fact-checking organizations, and assess the credibility of these organizations based on that consensus. This added dimension of trust and credibility assessment reinforces FACTS-ON's effectiveness in combating false information in online social networks.

Chapter 7

Fact-checkers' Consensus Inference and Credibility Assessment for Trust-based Fake News Detection: AFCC

7.1. Introduction

Following the exploration of ExFake, this chapter delves into the Automated Fact-checkers Consensus and Credibility (AFCC) system, a critical component of the FACTS-ON framework. AFCC was developed in response to challenges encountered during the creation of ExFake, particularly in utilizing fact-checking organizations as external evidence. In ExFake, due to the complexities of collecting data from multiple fact-checkers and the heterogeneity of their textual rating labels, reliance was initially placed on a single fact-checker. This limitation highlighted the need for a system that could effectively manage and standardize the diverse rating labels and decisions from various fact-checking organizations.

Fact-checking-based systems fall into two broad categories: manual fact-checking, which involves human assessment, and automatic fact-checking, which leverages machine-based methods. In this chapter, the primary focus is on the former. Fact-checking, in essence, is the task of determining the accuracy of statements, typically carried out by trained professionals [276]. This task has led to the development of various fact-checking websites, some of which are signatories of the International Fact-Checking Network's (IFCN) code of principles, indicating their adherence to established best practices in the field [1]. More details about fact-checking organizations and the IFCN, including their role and importance, are provided in the Technical background Section of Chapter 3.

Fact-checking organizations evaluate the veracity of claims and news using a range of discrete textual rating labels. For instance, PolitiFact [2] uses the six rating labels, encompassing

[1] `https://ifcncodeofprinciples.poynter.org/signatories`, last access date: 30-12-2023.

[2] `https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/`, last access date: 30-12-2023.

"pants-fire", "false", "mostly false", "half true", "mostly true", and "true", while Snopes [3] employs sixteen different labels, including "true", "mostly true", "mixture", "mostly false", "false", "unproven", "outdated", "miscaptioned", "correct attribution", "misattributed", "scam", "legend", "labeled satire", "originated satire", "recall", and "lost legend".

Concurrently, scholars studying journalism, bias, misinformation, disinformation, and fake news have raised questions about trust in fact-checkers [10, 195]. They have explored when and how fact-checking influences beliefs [298] and how multiple fact-checked claims gain more attention than those that are not fact-checked [221]. These discussions have also delved into the challenges of achieving consensus among fact-checkers, especially when assessing claims that fall within ambiguous scoring ranges such as "half true" or "mostly false" [157].

Moreover, both experts from industry and academia have discussed various issues related to fact-checkers. They have questioned whether fact-checkers can agree on what is true [4], examined when fact-checkers disagree [5] and even questioned who fact-checks the fact-checkers and reveal that the objectivity of fact-checking websites has been subject to doubt [6]. Additionally, inquiries have been made regarding potential biases among fact-checkers [7]. It is essential to recognize that some analyses have highlighted potential biases in fact-checking organizations, emphasizing the importance of adhering to facts. A thorough analysis of Matt Shapiro's work on the Paradox project revealed that PolitiFact is biased in its fact-checking [8]. Therefore, some call on fact-checkers to stick to the facts since the power of fact-checkers is greatest when they dissect an outright lie or confirm a verifiable truth [9].

Hence, the rating scheme, comprising discrete textual labels employed by various fact-checkers, remains challenging to utilize as available evidence in automated false information detection systems for assessing the accuracy of news according to multiple fact-checkers and for measuring the credibility of both the news source and the fact-checking organization. These challenges arise due to the following factors:

— The heterogeneity of the rating labels. Each fact-checking organization uses different rating labels as well as a different number of rating labels. Thus, for a given news content different rating labels are assigned by several fact-checking organizations.

[3] https://www.snopes.com/fact-check-ratings/, last access date: 30-12-2023.

[4] https://www.poynter.org/fact-checking/2017/can-fact-checkers-agree-on-what-is-true-new-study-doesnt-point-to-the-answer/, last access date: 30-12-2023.

[5] https://www.politifact.com/article/2011/oct/21/when-fact-checkers-disagree/, last access date: 30-12-2023.

[6] https://www.acsh.org/news/2019/11/04/debunkers-debunked-who-fact-checks-fact-checkers-14378, last access date: 30-12-2023.

[7] https://www.allsides.com/blog/6-ways-fact-checkers-are-biased, last access date: 30-12-2023.

[8] https://theparadoxproject.org/2016/12/16/2016124mostly-false-politifact-and-bias/, last access date: 30-12-2022.

[9] https://www.bnnbloomberg.ca/fact-checkers-need-to-stick-to-the-facts-1.1350935, last access date: 30-12-2023.

- The heterogeneity of the rating decisions. There are many fact-checking organizations, each giving a different rating decision related to the truthfulness of the same news content (i.e., multiple ratings for a single content). This makes it difficult to standardize its truthfulness rating decisions and to assess its credibility. Thus, it is also difficult for people to assess its veracity without any additional information.

To address these issues, the main contributions of this chapter are then to:

- Manage the heterogeneity of the rating labels. For this, a classification system is proposed to map the rating labels from different fact-checking organizations to a unified rating scheme.
- Automate the use of the heterogeneous decisions made by multiple fact-checking organizations. First, textual rating labels are converted into numerical rating values. Then, the consensus among the different rating decisions is computed so that only one single decision value (i.e., numerical rating value) is obtained from the many initial decision values.
- Evaluate the credibility of fact-checking organizations to identify and mitigate the impact of disparate ratings. Indeed, it has been proven that credibility indicators can reduce the propensity to share fake news [318].

While prior work by Tchechmedjiev et al. [274] in the ClaimsKG framework established a foundation by developing knowledge graphs for fact-checked claims, it primarily focused on normalizing textual rating labels into four basic categories (i.e., true, false, mixture, other). This approach aimed to generate a public corpus of structured information about claims and related metadata. In contrast, the AFCC system represents a significant evolution in this field. Unlike ClaimsKG, which is restricted to only six fact-checking organizations (i.e., africacheck.org, factscan.ca, politifact.com, snopes.com, checkyourfact.com, truthorfiction.com), AFCC not only unifies and quantifies these diverse textual labels into a numerical format but also introduces a consensus-based approach for synthesizing these ratings. Moreover, AFCC extends the analysis to encompass a wider array of fact-checking organizations, thereby providing a more comprehensive perspective. Additionally, AFCC innovates by incorporating a novel module for assessing and adjusting the credibility of fact-checkers. This addition is pivotal as it addresses potential biases and inconsistencies in the fact-checking process, thereby enhancing the overall trustworthiness and utility of the system in the battle against false information. Collectively, these contributions in AFCC offer a more robust and comprehensive framework for evaluating fact-checked claims and enhancing the trustworthiness of fact-checking organizations.

7.2. Problem statement

7.2.1. Motivation scenarios

To illustrate the Automated Fact-checkers Consensus and Credibility problem (AFCC), some motivational scenarios are introduced to aid in understanding the motivations and potential applications of the approach. The following simplified examples serve as motivation scenarios, providing insights into the vision of the approach.

Heterogeneity of rating labels based on the same factual premises (i.e., news/claim): There are cases where fact-checkers may agree on the decision regarding a news/claim, but they use different organization-specific rating labels to express their judgment. The following example, also illustrated in Table 2, showcases this scenario. Different fact-checkers, while unanimously recognizing the falsehood of the claim, employed distinct rating labels. Notably, this claim was later validated as true by scientists. This situation underscores the complexity in standardizing and interpreting decisions of fact-checkers, underlining the necessity for a unified approach to evaluate news credibility.

— Example 1 [10]: In 2020 fact-checkers disputed the origin of COVID-19. While they dismissed the possibility that it was manipulated in a lab, PolitiFact rated the news/claim as "Pants on Fire!" while FactCheck.org rated it as "faulty". Although both fact-checkers agreed on their decisions on the given news/claim, they used heterogeneous textual rating labels.

Table 1 – Divergent rating labels for identical claim by Fact-Checkers

Claim	**Fact-checker**	**Rating**
COVID-19 originated in a laboratory	**PolitiFact**	Pants on Fire
	Factcheck.org	Faulty

Heterogeneity of rating decisions based on the same factual premises (i.e., news/claim): In other cases, fact-checkers may disagree with the decision on the same news/claim. The following examples, also outlined in Table 2, demonstrate disputes and disagreements between several fact-checkers based on the same factual premises (i.e., news/claim). The first and third examples illustrate that the agreement rate is much lower in the fuzzy areas and for statements in the more ambiguous scoring range (i.e., "Half true" or "Mostly false"). The second example highlights that fact-checkers can arrive at different conclusions even for claims in the assertive scoring range (i.e., "whopper" or "absurd").

[10] `https://www.realclearpolitics.com/articles/2021/06/03/how_fact-checkers_mishandled_the_covid-19_origin_debate.html`, last access date: 30-12-2023.

— Example 2[11]: Based on the same factual premises (i.e., news/claim), Washington Post's fact-checker judged ex-Florida Governor Jeb Bush's statement that his state "led the nation in job creation", "mostly false" and PolitiFact called it "half true".

— Example 3[12]: In 2011, Joe Biden has been making comments about connections between murders, rapes, cops and budget cuts. Factcheck.org called his words a "whopper". The Washington Post's Fact-checker said "absurd", and PolitiFact said "Mostly True".

— Example 4[13]: On the claim that Ben Carson said "illegal immigrants who get caught voting should be stripped of citizenship", Snopes[14] rated it "Mostly True", in contrast to PolitiFact's[15] rating of "Mostly False". This illustrates differences in interpretation and context assessment by fact-checkers.

Table 2 – Divergent rating decisions by Fact-Checkers on identical claims

Claim	**Fact-checker**	**Rating**
Jeb Bush's stated, "Florida led the nation in job creation".	**Factcheck.org**	Mostly false
	PolitiFact	Half true
Joe Biden's comments about connections between murders, rapes, cops and budget cuts.	**Factcheck.org**	Whopper
	Washington Post	Absurd
	PolitiFact	Mostly True
Ben Carson said "illegal immigrants who get caught voting should be stripped of citizenship".	**PolitiFact**	Mostly false
	Snopes	Mostly true

7.2.2. Problem formulation

This section introduces the fundamental concepts and formal mathematical notations used in subsequent sections, summarized in Table 3.

The Automated Fact-checkers Consensus and Credibility problem (AFCC) in online social networks for detecting fake news is defined as a 3-tuple: $AFCC = (F, L, w)$. Each component represents:

[11] `https://www.bnnbloomberg.ca/fact-checkers-need-to-stick-to-the-facts-1.1350935`, last access date: 30-12-2023.

[12] `https://www.politifact.com/article/2011/oct/21/when-fact-checkers-disagree/`, last access date: 30-12-2023.

[13] `https://misinforeview.hks.harvard.edu/article/fact-checking-fact-checkers-a-data-driven-approach/`, last access date: 30-12-2023.

[14] `https://www.snopes.com/fact-check/ben-carson-voter-fraud/`, last access date: 30-12-2023.

[15] `https://www.politifact.com/factchecks/2019/feb/08/facebook-posts/ben-carson-illegal-immigrants-should-be-stripped/`, last access date: 30-12-2023.

— $F = \{f_i \mid 1 \leq i \leq n\}$ represents a set of n fact-checking organizations (i.e., fact-checkers), signatories of the IFCN code of principles, and that verify (i.e., fact-check) an online news/claim w. Each fact-checker $f_i \in F$ is described by a quadruple $f_i = \langle name^i, L^i, Cr^i, Hr^i \rangle$, where $name^i$ is the name of the fact-checking organization, $L^i = \{l^i_j \mid 1 \leq j \leq p_i\}$ is its set of p_i textual rating labels (with p_i indicating the number of unique rating labels used by fact-checker f_i, reflecting the specificity and diversity of its rating system), Cr^i denotes its credibility, which will be the object of the Section 7.3.6, and Hr^i contains its rating history detailing the past ratings given by the fact-checker f_i to various news/claims.

— $L = \bigcup\limits_{i=1}^{n} L^i$ where $|L| = m$ and $m > n$, denotes the set of $m > n$ textual rating labels obtained by the union of all fact-checking organizations' label sets. This notation means that L is the union of all the sets of textual rating labels L^i from each fact-checking organization f_i, with n being the total number of fact-checking organizations. The cardinality of L, denoted as $|L|$, equals m, representing the total number of unique textual rating labels across all fact-checking organizations.

— $w = \langle metaData, RD \rangle$ represents the fact-checked news/claim that is defined by its metadata $metaData$ and the set of rating decisions RD made by the fact-checkers. $metaData$ contains the headline of the news, its textual content, its source, and the publication date and time. $RD = \{(f_i, l^i_x) \mid 1 \leq i \leq n, l^i_x \in L^i\}$, where l^i_x is f_i's attributed rating label for an x^{th} news w and $l^i_x = l^i_j$ for some $1 \leq j \leq p_i$.

The goal of AFCC is to process all rating labels in RD regarding a fact-checked news/claim w to compute a single numerical consensus decision D: $f_{consensus}(F, L, w) = D$, which will be further explored in Section 7.3.5.

7.3. AFCC: automated fact-checkers consensus and credibility

7.3.1. AFCC architecture

The AFCC system is designed to enhance trust in fake news detection by employing a consensus and credibility-based approach. Rather than the traditional majority voting method, which is prevalent in tasks where multiple annotations are reconciled to establish a unified truth, AFCC utilizes a "majority rating" concept. This term more accurately reflects the process of evaluating news veracity through collective assessments (i.e., "ratings") from various fact-checkers, as opposed to "votes".

Researchers from diverse fields are increasingly employing majority voting techniques to tackle a range of challenges. This approach finds applications not only in web services [166], but also in ensemble methods for intrusion detection systems [26], performance evaluation

Table 3 – Summary of core concepts and formal mathematical notation

Fact-checking organizations	The set of fact-checking organizations.	F
	International Fact-Checking Network.	IFCN
	The i'th fact-checker.	f_i
	The number of fact-checking organizations.	n
	The name of the fact-checking organization.	name
	A measure of the credibility of the fact-checker i.	Cr^i
Labels	The set of all textual rating labels.	L
	The set of textual rating labels of the i'th fact-checking organization.	L^i
	The j^{th} textual rating label of the fact-checking organization i (f_i).	l_j^i
	The textual rating label attributed to an x^th news/claim w by the fact-checking organization i (f_i).	l_x^i
	The number of textual rating labels.	m
Online news/claim	The fact-checked news/claim.	w
	The metadata of the news/claim contains the headline of the news, its textual content, its source, and the publication date and time.	metadata
	The set of rating decisions.	RD
Decision	The final numerical consensus decision.	D

within the domains of biology and medicine [31, 270], particularly in the context of COVID-related research, strategic decision-making in the realms of business and management [143], and the enhancement of safety protocols within aircraft systems [129]. Throughout this chapter, the approach is referred to as "majority rating" instead of "majority voting" to better capture the nature of "ratings" provided by various fact-checkers, which are distinct from mere "votes".

The AFCC system aims to facilitate the automated assessment of the accuracy of fact-checked news or claims (i.e., news truthfulness scoring) by measuring the consensus from various fact-checkers' ratings and by evaluating their credibility. Presently, the system is limited to analyzing content that is in English, encompassing news content and claims. The architecture of AFCC is composed of three main modules: the transformation of textual evaluations into a standardized numerical rating (i.e., unified text-to-numerical rating transformation), the derivation of consensus from these ratings (i.e., consensus inference), and the evaluation and recalibration of the credibility of the fact-checkers (i.e., fact-checker credibility assessment and update). These modules collectively form the integral AFCC system. The architecture of AFCC is depicted in Figure 1, providing an overview of the system's components and interactions.

In the design phase (i.e., steps 1 to 3 in Figure 1), the dataset is built based on the collection of fact-checking organizations F and their corresponding rating labels L (i.e., steps 1). Subsequently, a unified rating scheme is formulated through the clustering of textual rating labels, followed by their mapping to appropriate numerical clusters (i.e., step 2). The

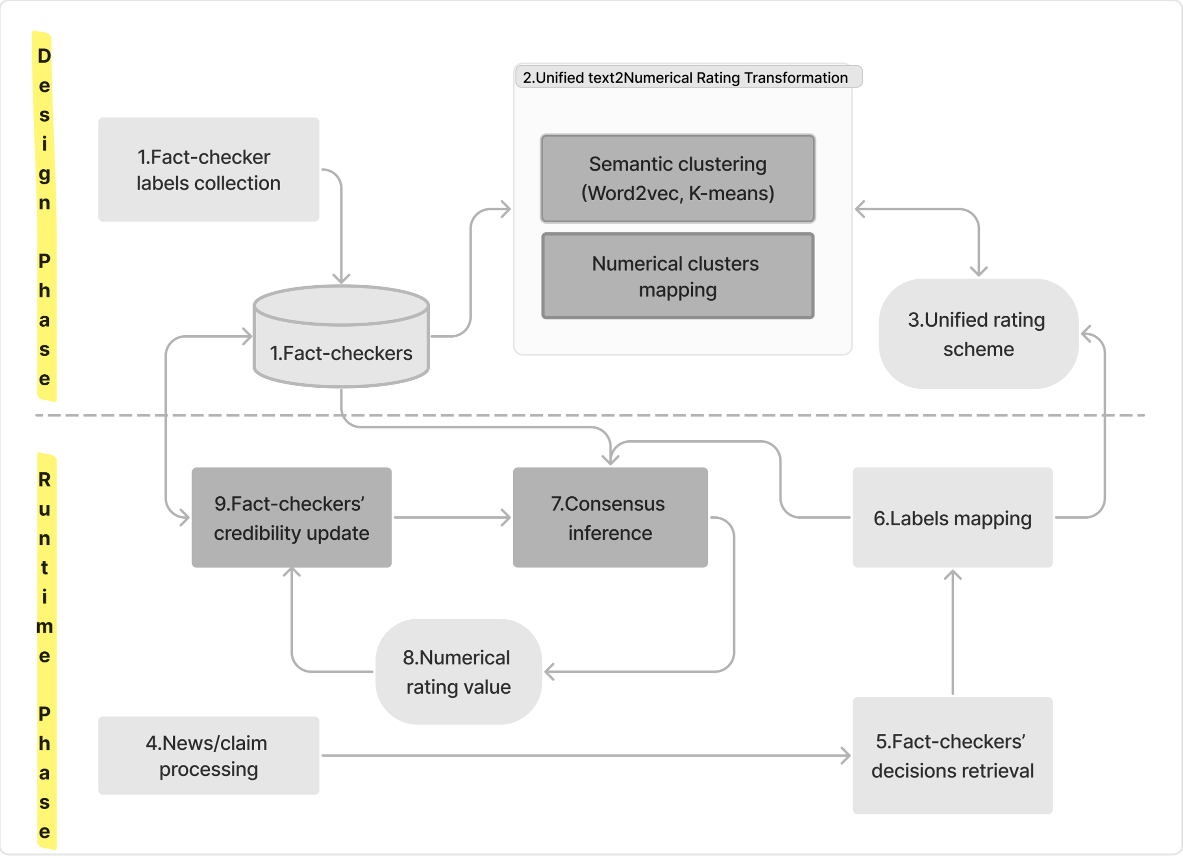

Design Phase
2.Unified text2Numerical Rating Transformation
1.Fact-checker labels collection
Semantic clustering (Word2vec, K-means)
Numerical clusters mapping
1.Fact-checkers
3.Unified rating scheme
Runtime Phase
9.Fact-checkers' credibility update
7.Consensus inference
6.Labels mapping
8.Numerical rating value
4.News/claim processing
5.Fact-checkers' decisions retrieval

Figure 1 – The architecture of AFCC

latter consists of quantifying the decisions of fact-checkers into numerical values. This unified scheme (i.e., step 3) plays a pivotal role in quantifying fact-checkers' decisions, facilitating the measurement of the veracity of the fact-checked news/claim w through consensus inference, and enabling the evaluation of fact-checker credibility (i.e., all of these are the steps in the runtime phase).

In the runtime phase, the various decisions RD made by all fact-checking organizations $\{f_i \in F\}$ regarding the veracity of a given fact-checked news/claim w are collected and mapped to the unified rating scheme (i.e., steps4 to 6). This mapping results in numerical rating values that undergo processing to infer consensus, yielding a single numerical rating value that serves as the ultimate decision regarding the truthfulness of the scrutinized content w (i.e., step 7 and 8). Furthermore, in this phase, fact-checkers' credibility is updated based on the inferred numerical consensus decision (i.e., step 9). All steps in each phase will be detailed in the subsequent sections.

7.3.2. Fact-checkers and labels collection

This module is dedicated to the semi-automated extraction of data, encompassing fact-checkers, their textual rating labels, and their corresponding definitions, from unstructured fact-checking organizations' data. It is worth noting that this data extraction is semi-automated due to restrictions imposed by some fact-checking websites that prohibit web scraping. The primary objective of this module is to build a dataset comprising fact-checkers and the discrete textual values they employ to rate the veracity of news content and claims. This dataset undergoes periodic updates and serves as a crucial component in the unified text-to-numerical rating transformer and consensus inference processes. The following steps outline the approach:

(1) Selection of Fact-checking Organizations: I begin by selecting a list of English-language fact-checking organizations among the ones listed by the International Fact-Checking Network (IFCN).

(2) Collection of Textual Rating Labels: Subsequently, the textual rating labels used by each fact-checking website are collected to provide a comprehensive overview of their rating systems.

(3) Periodic Updates: To ensure the dataset remains current and reflective of evolving fact-checking practices, steps 1 and 2 are routinely revisited and repeated.

7.3.3. Unified text-to-numerical rating transformer

As illustrated in the motivational scenarios, fact-checkers often use diverse and inconsistent textual rating labels when assessing the veracity of the same news or claims. To address this issue, this module aims to unify and quantify these discrete textual values of ratings across multiple fact-checkers into a single unified textual and numerical rating scheme. The unified numerical rating values can then be utilized to measure the veracity of fact-checked news or claims. Thus, shifting from discrete textual rating labels to unified textual and numerical rating values. This transformation is achieved through the following steps:

(1) Mapping Textual Rating Labels: the textual rating labels are mapped to a unified textual rating scheme using word2vec and k-means clustering algorithms. Word2vec is an unsupervised learning model that creates word embedding in the field of Natural Language Processing (NLP) with a deep learning model working in the back-end. The neural network model is used to learn word associations from a large corpus of text, such as the identification of synonyms and antonym words once a model has been trained, by measuring cosine similarity. In this Module, Word2vec is used to group textual rating labels based on their semantic meanings. To this end, semantically similar rating labels are grouped into the same class of labels (i.e., cluster). K-means

clustering, an unsupervised learning algorithm, further refines these clusters as it is used to generate the clusters of the textual rating labels. For instance, textual rating labels from PolitiFact, such as "pants-fire" and "false", may be grouped into the "very negative" cluster, while labels like "mostly false", "half true", and "mostly true" may belong to the "almost positive" cluster, and "true" is aligned with a "positive" cluster. However, the initial clusters were not always accurate due to the limited dataset, necessitating fine-tuning with the GoogleNews dataset[16] to enhance the accuracy of the clusters. Figure 2 provides a visualization of the clusters obtained from the application of Word2Vec and K-means clustering models. It is important to note that the axes in this figure represent the features used for clustering, and they are unrelated to the clustering algorithm itself. However, the resulting clusters from the use of k-means and Word2Vec were not perfect, some textual rating labels did not go to the appropriate clusters. This imperfection underscores the need for refining these clusters to ensure accurate representation and analysis. In response to this, Table 4 details the refined clusters. This refinement process involved mapping textual rating labels used by fact-checkers to the appropriate clusters. This table is crucial in understanding how textual ratings provided by fact-checkers are quantitatively represented, a process that will be explained in the subsequent step, and linked to specific levels of truthfulness, thereby offering a more structured approach to assessing the credibility of content.

(2) Quantifying Rating Labels: To facilitate the measurement of news or claim veracity (i.e., consensus inference) based on multiple fact-checker decisions, the unified textual rating labels were converted into discrete numerical rating values ranging from 1 to 5, as detailed in Table 5. For instance, for PolitiFact, "pants-fire" and "false" are quantified as "very negative = 1", while "mostly false", "half true", and "mostly true" are converted to "almost positive = 4". Similarly, for Africa Check[17], "incorrect" is mapped to "very negative = 1", "Misleading", "exaggerated", and "understated" are mapped to "negative = 2", "Unproven" is mapped to "neutral = 3", "Mostly correct" is mapped to "almost positive = 4", and "Correct" is mapped to "positive = 5". These numerical ratings allow for consistent evaluation and comparison of fact-checked content.

A set of unified rating classes is defined as $C = \{c_1, c_2,..., c_k| \ 1 \leq k \leq 5\}$, where each $c_i = \langle name_i, Nr_i, rl_i \rangle$. In this context, $name_i$ represents the label for the class within the cluster, $Nr_i \in NR$ denotes the numerical rating value mapped to the class, and rl_i denotes the set of rating labels belonging to the class c_i. Importantly, it is ensured that $rl_i \subset L$, and $\bigcap\limits_{i=1}^{n} rl_i = \emptyset$, indicating that each class c_i contains distinct rating labels, and that the rating labels within each class c_i are unique.

[16] `https://www.kaggle.com/datasets/leadbest/googlenewsvectorsnegative300`, last access date: 30-12-2023.

[17] `https://africacheck.org/`, last access date: 30-12-2023.

$NR = \{1, 2, 3, 4, 5\}$ is the set of numerical rating values defined in terms of positivity, negativity, and neutrality as shown in Table 5. The main objective behind using NR is to craft a numerical rating system built upon interval-level measurements, similar to those employed in eCommerce platforms. NR is structured to ensure symmetric scaling and equidistant quality of attributes. By maintaining these two proprieties in the rating system NR, additional valuable information can be obtained from these ratings, and proper interval-level analyses can be conducted.

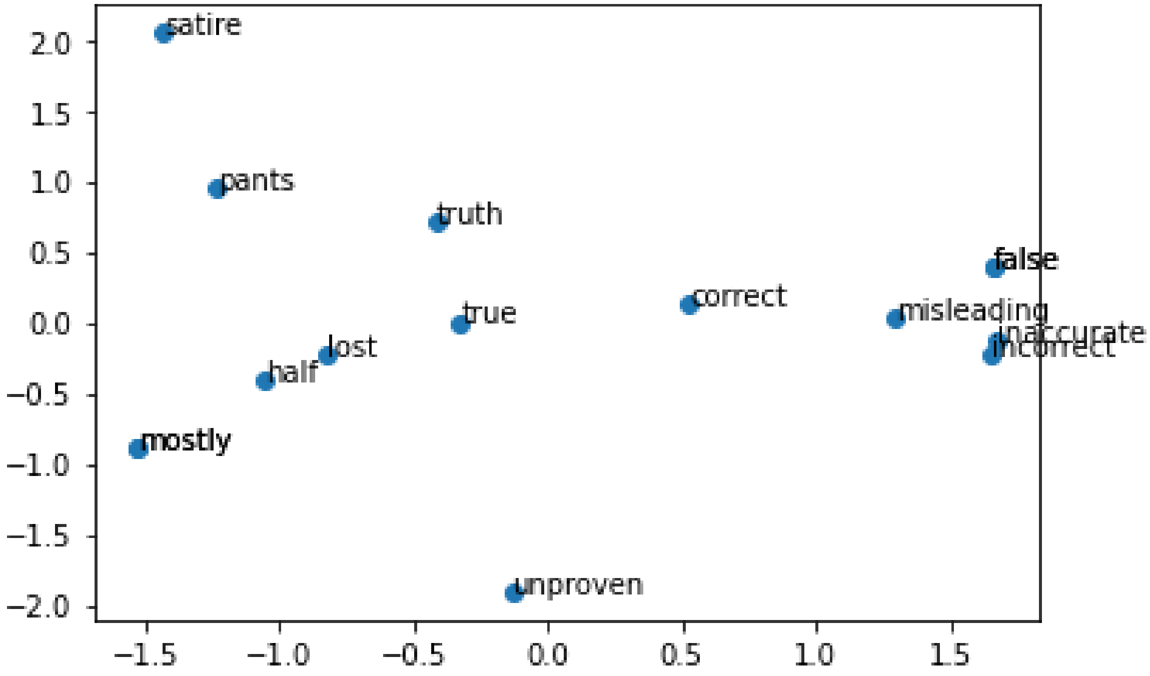


Figure 2 – Clustering result using word2vec and K-means pre-trained models

Table 4 – Distribution of textual rating labels retrieved from the list of fact-checking organizations

Degree of truthfulness	**Very negative**	**Negative**	**Neutral**	**Almost Positive**	**Positive**
Numerical rating value	**1**	**2**	**3**	**4**	**5**
Textual rating values used by fact-checkers	Incorrect False Fake Blatantly false Void Pants on fire Inaccurate Five ????? Marks: Totally false Four Pinocchios Bottomless Pinocchio An Upside-Down Pinocchio	Misleading Exaggerated Understated Just in case Satire FFS! (For Facts' Sake) Scam Labeled Satire Originated as Satire Originated as Satire One ? Mark: Exaggeration	Unproven No evidence Unverified Verify Unsupported Legend Lost legend Verdict Pending	Mostly correct Mostly true Partly true Half true Mixture Mostly false False headline Outdated Miscaptioned Misattributed Four ???? Marks: Mostly false Two ?? Marks: Misinterpretation Four ???? Marks: Mostly false One Pinocchio Two Pinocchios Three Pinocchios	Correct Investigation Analysis Explainer True Correct attribution Recall The Geppetto Checkmark

Table 5 – Distribution of numerical rating values and their degrees of truthfulness

Degree of truthfulness	Very negative	Negative	Neutral	Almost Positive	Positive
Numerical rating value	1	2	3	4	5

7.3.4. Fact-checker decisions retrieval and labels mapping

This module involves retrieving all rating decision labels RD made by multiple fact-checking organizations f_i regarding a fact-checked news/claim w and assigning the appropriate corresponding numerical rating value Nr_i associated with the retrieved rating label, as illustrated in Figure 3. To achieve this, Google Fact Check Tool APIs[18] are utilized for retrieving all decisions made by fact-checkers regarding a fact-checked claim. Label mapping is then performed using the unified rating scheme established during the design phase.

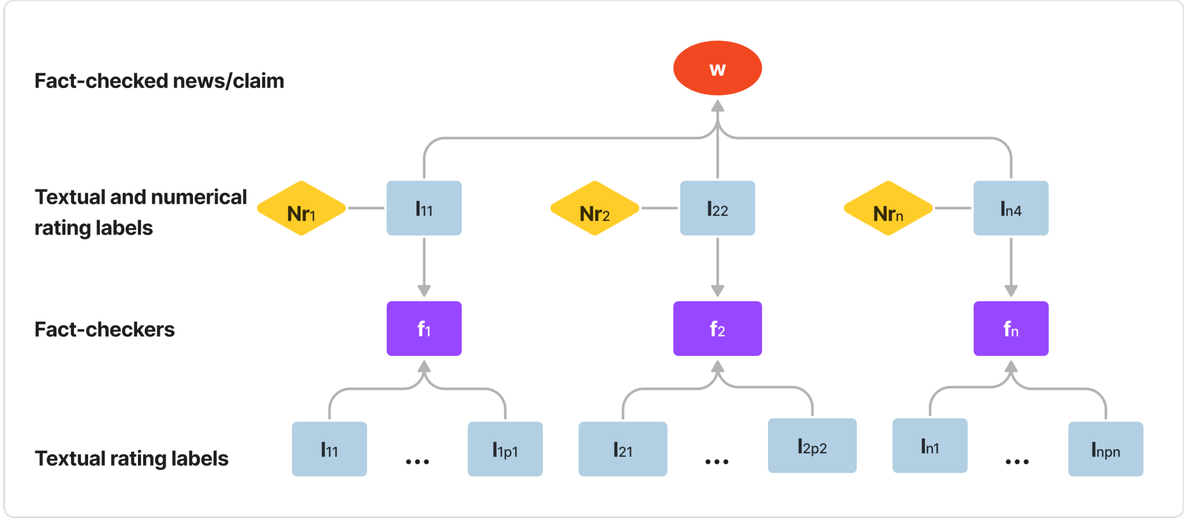


Figure 3 – Decisions retrieval of fact-checked news/claim w

7.3.5. Consensus inference

This module is responsible for computing the consensus among the decisions (i.e., rating values) of multiple fact-checking organizations regarding a fact-checked news or claim. It aims to consolidate these numerical rating values obtained during the label mapping of fact-checkers' retrieved decisions into a single numerical rating value that represents the collective decision about the truthfulness of the fact-checked content.

However, the consensus calculation involves several factors that need precise definition and measurement. First, it necessitates gathering the various decisions and assessments made by multiple fact-checking organizations concerning a given news or claim. Second,

[18] `https://developers.google.com/fact-check/tools/api`, last access date: 30-12-2023.

it entails evaluating the credibility of these fact-checkers to assign appropriate weights to their rating decisions. This step aims to establish trust and counter potential deceptive activities in credibility management, thereby aiding in identifying, preventing, and detecting malicious behaviour by entities or groups (i.e., colluding entities) posing as trusted fact-checker organizations (i.e., acting as trusted fact-checker organizations).

Consequently, in the designed model, the consensus on a given content (i.e., news or claim) is calculated as a weighted average, taking into account both fact-checkers' ratings and their corresponding credibility. Formally, the consensus is defined as:

$$consensus(w_x) = \frac{\sum_{i=1}^{k} (Nr_i^x \times Cr(f_i))}{\sum_{i=1}^{k} Cr(f_i)} \tag{7.3.1}$$

Where Nr_i^x represents the numerical rating value of the i'th fact-checking organization f_i of the x'th news w_x, k denotes the number of fact-checking organizations that fact-checked the news w_x (i.e., evaluated its veracity and rated its truthfulness), and $Cr(f_i)$ signifies the credibility of the fact-checking organization f_i which will be elaborated upon in Section 7.3.6.

To provide a more comprehensive understanding of the consensus inference process, Algorithm 1 is presented with a detailed pseudo-code that outlines the steps involved.

Algorithm 1 Consensus inference algorithm

```
Intput: $w, F, C$
Output: $Consensus$
Begin
  $SumCr \leftarrow 0$
  $SumRateCredibility \leftarrow 0$
for f in F do
    $RateLabel \leftarrow getFactRating(w)$  ▷ extract the rating label provided by fact-checker $f$ for the news or claim $w$ from their website
    if $RateLabel$ then
        $factRate \leftarrow mapping(C, RateLabel)$  ▷ map the rating label to its corresponding class in the set of classes $C$
        $credibility \leftarrow credibility\ assessment(f, w, \sigma)$  ▷ see Algorithm 2
        $SumRateCredibility \leftarrow SumRateCredibility + factRate * credibility$
        $SumCr \leftarrow SumCr + credibility$
        $update\_Crediblity(f, credibility)$
        $update\_ratingHistory(f, factRate)$
    end if
end for
$Consensus \leftarrow SumRateCredibility/SumCr$  ▷ Eq. 7.3.1
End
```

The code in Algorithm 1 outlines the consensus inference for a given news or claim (w) using the ratings provided by a set of fact-checkers (F) and the unified rating scheme classes (C). The output is the consensus score of the news or claim.

The variables ($SumCr$) and ($SumRateCredibility$) are the initialization variables. The rating labels of each fact-checker in (F) for the given news or claim (w) are extracted using the function ($getFactRating(w)$). If the rating label is valid, the algorithm maps this label to one of the classes in (C). Then, the credibility of the fact-checker (*credibility*) is calculated as defined in Algorithm 2.

The function (*credibility assessment*(f, w, σ)) at line 8 calculates the adjusted credibility of the fact-checker ($f \in F$) as defined in Eq. (7.3.2) which is mainly based on three parameters: the *news/claim* (w), the *ratings* provided by other fact-checkers for the same content (w), and the *historical behaviour of the fact-checker* (i.e., *rating history* $\overline{H}r$). The details of how to calculate the credibility of a fact-checker (f) are discussed in section 7.3.6 and defined in Algorithm 2.

The variables ($SumRateCredibility$) and ($SumCr$), as well as the credibility and rating history of each fact-checker, are updated based on the previous calculations. Finally, the consensus score of the news or claim (w) is calculated as defined in Eq. (7.3.1). The resulting score is then returned as the output of the algorithm.

7.3.6. Fact-checkers credibility

One of the main limitations of fact-checking rating systems is that they assume all ratings to be honest and unbiased. However, in reality, there is a distinction in the credibility given to different sources, with a tendency to favour and assign more weight to those regarded as more trustworthy [275]. This module focuses on evaluating the credibility of fact-checkers while detecting and mitigating the effects of disparate ratings based on the same factual premises (i.e., news/claim).

To identify and address unfair or inconsistent ratings, the credibility assessment task considers two key metrics. Firstly, unbiased (i.e., trusted/ honest) fact-checkers are distinguished from biased (i.e., dishonest) ones by determining the extent of their deviation from the majority ratings. Secondly, the past behaviour of each fact-checker is analyzed in terms of their previous rating decisions. Algorithm 2 outlines the steps for calculating the credibility of a particular fact-checker f_i.

A *majority rating* approach is used, wherein the "uniformity of ratings" reflects their accuracy. The basic idea is that ratings that significantly deviate from the majority are considered less reliable and should be given less weight when determining consensus. However, there may be cases in which a fact-checker exhibits bias [19]. In such instances, relying solely

[19] `https://theparadoxproject.org/2016/12/16/2016124mostly-false-politifact-and-bias/`, last access date: 30-12-2022.

on the majority rating approach may not suffice. To address this, a *credibility adjustment parameter* is introduced that can increase or decrease the credibility of fact-checkers, along with a penalty parameter based on their historical behaviour.

The credibility of a fact-checker, as defined in Eq. (7.3.2), determines the extent to which users of online social networks can trust the ratings provided by fact-checkers regarding the veracity of a given news or claim.

$$Cr_t(f_i) = (Cr_{t-1}(f_i) + \alpha \times \beta) \times \overline{H}r \tag{7.3.2}$$

Where t and $t-1$ are the times at which credibility is computed, denoting current credibility and past credibility, respectively. α serves as the credibility adjustment normalizing parameter, defined in Eq. (7.3.3). β signifies the extent of credibility change resulting from agreement or disagreement with the majority decisions, as specified in Eq. (7.3.5). Finally, $\overline{H}r$ represents the penalty attributed due to the historical behaviour of fact-checker f_i, as defined in Eq. (7.3.6).

$$\alpha = Cr(f_i) \times (1 - \frac{|Nr_i - Mj|}{max(NR)}) \tag{7.3.3}$$

Where Nr_i is the reported numerical rating of the fact-checker f_i, Mj is the majority rating as defined in Eq. (7.3.4) and $max(NR)$ is the maximum value in NR which is used to scale the value of α into the range [0-1]. Equation (7.3.3) states that the value of the normalizing factor α depends on the credibility of the fact-checker $Cr(f_i)$ and the absolute difference between the fact-checkers current feedback Nr_i and the majority rating Mj. Multiplying by the credibility of the fact-checker allows the ratings of the more credible fact-checkers to have a bigger impact on consensus inference value than those of the less credible ones.

To screen the ratings based on their deviations from the majority ratings and to dilute the effects of unfair or inconsistent ratings, a majority rating system is adopted. The majority rating is defined using a data clustering technique. The k-means clustering algorithm is used on all current reported ratings to create the clusters. The most densely populated cluster is then labelled as the "majority cluster" and the centroid of the majority cluster denoted Mj, is taken as the majority rating.

$$Mj = centroid(argmax(C_k)) \qquad \forall k \tag{7.3.4}$$

Where k is the total number of clusters, $argmax(C_k)$ gives the largest cluster, and $centroid(x)$ gives the geometric center of the cluster x.

Fact-checkers credibility is then adjusted based on the agreement/ disagreement with the majority decisions β as defined in Eq. (7.3.5) and based on their past behaviour Hr defined in Eq. (7.3.6). β is made up of the Euclidean distance between the majority rating (Mj) and the reported numerical rating (Nr_i), and the standard deviation in all the reported ratings.

β takes a negative value if the reported rating is far from the majority rating. The change in credibility due to majority rating, denoted by β is defined as:

$$\beta = \begin{cases} 1 - \frac{\sqrt{(Mj-Nr)^2}}{\sigma} & \text{if } \sqrt{(Mj-Nr)^2} < \sigma \\ -(1 - \frac{\sigma}{\sqrt{(Mj-Nr)^2}}) & \text{otherwise} \end{cases} \tag{7.3.5}$$

Where σ represents the standard deviation in all the reported ratings. The closer a reported rating is to the majority rating, the more the credibility of the fact-checker organization is increased through the credibility adjustment parameter β. Otherwise, a negative value of β indicates a decrease in the credibility of an outlier fact-checker.

However, there may be cases in which the majority of fact-checkers collude to manipulate their ratings and deliver a biased judgment on a particular news item or claim. Consequently, relying solely on a majority rating system is inadequate for accurately evaluating the credibility of fact-checkers. This limitation is addressed by complementing the majority rating approach with an adjustment based on a fact-checker's past behaviour, denoted as Hr. This concept draws inspiration from the logic underpinning the five-star rating systems often used in e-commerce platforms [234].

To mitigate attempts of rating manipulation, the Bayesian average method [20] is employed instead of a simple arithmetic mean. The simplest way to explain this is that although all fact-checkers ratings are considered, not all ratings have the same impact (or "weight") on the consensus inference (i.e., the final numerical rating decision). For instance, consider two fact-checkers: f_1, who has assessed only one news item or claim, and f_2, who has evaluated ten. It is evident that they should not have an equal impact on credibility assessment. If a simple mean were used to calculate their past behaviour, both fact-checkers would be assigned a legitimacy score of 100%. However, f_2, by virtue of examining more news items and claims, naturally commands greater trust and expertise than f_1.

However, by utilizing the Bayesian average method to compute the past behaviour of fact-checkers (as defined in Eq. (7.3.6)), it is ensured that the impact of fact-checker f_2 is weighted more heavily than fact-checker f_1 in credibility assessment and consequently, in consensus inference. In essence, the Bayesian average provides a more equitable basis for comparing fact-checkers by relativizing ratings in this manner. It guarantees that fact-checkers with fewer ratings and less accurate historical ratings (i.e., ratings that significantly diverge from the majority) exert less influence in credibility adjustment and consensus inference due to the higher penalty parameter associated with their historical behaviour.

$$\overline{Hr}_{bayesAvg}(f_i) = \frac{W^i_{avg} \times W^i_{count} + W_{mean} \times W_{min}}{W^i_{count} + W_{min}} \tag{7.3.6}$$

[20] https://www.algolia.com/doc/guides/managing-results/must-do/custom-ranking/how-to/bayesian-average/, last access date: 30-12-2023.

Where W^i_{avg} represents the arithmetic average of the news or claims that fact-checker f_i has accurately evaluated, as defined in Eq. (7.3.7), W_{mean} signifies the mean of all the news or claims fact-checked across the entire dataset. W^i_{count} is the count of news or claims that fact-checker f_i has evaluated accurately, and W_{min} sets the minimum number of fact-checked news or claims required for a fact-checker f_i to be included in the assessment. W_{min} was set to 1, considering all fact-checkers involved in the fact-checking process.

$$W^i_{avg}(f_i) = n\frac{\sum_{i=1}^{k} h_i}{\sum_{i=1}^{n}\sum_{j=1}^{k} h_{ij}} \tag{7.3.7}$$

Where h_i, as defined in Eq. (7.3.8), denotes a submission where fact-checker f_i provided a rating deemed "accurate" in their previous ratings. k represents the total number of news or claims fact-checked by f_i, and n stands for the number of fact-checkers.

$$h_i = \begin{cases} 1 & \text{if } Nr_{(t-1)} \simeq Consensus_{(t-1)} \\ 0 & \text{otherwise} \end{cases} \tag{7.3.8}$$

This parameter h_i was chosen to account for the previous accurate ratings made by fact-checkers. Its purpose is to ensure that fact-checkers with lower accurate ratings carry less weight in the credibility adjustment due to the high penalty parameter associated with their historical behaviour.

Algorithm 2 details the process of calculating the credibility of a fact-checker (f) by utilizing both a piece of news or claim (w) and the fact-checker's historical ratings, considering the standard deviation (σ). This algorithm takes as input a news or claim (w) and a fact-checker (f) and returns the credibility score of the fact-checker.

The algorithm starts by clustering the ratings $(w.RD)$ of the news/claim made by various fact-checkers using the k-means clustering method. The largest cluster is selected, and its centroid is computed as (Mj).

Next, the algorithm calculates the distance (d) between the fact-checker (f)'s rating and the centroid of the largest cluster, denoted as Mj. If the distance (d) is less than the threshold (σ), the β value is calculated as $1 - \frac{d}{\sigma}$; otherwise, it is calculated as $-(1 - \frac{\sigma}{d})$. This step quantifies the change in credibility due to agreement or disagreement with the majority decisions, as per Eq. (7.3.5).

The algorithm proceeds to calculate the penalty parameter (Hr) by applying a Bayesian average to the fact-checker's historical ratings, as defined in Eq. (7.3.6). The penalty parameter (Hr) represents the change in credibility based on the past behaviour of the fact-checker (f).

The current credibility of the fact-checker $(currentCredibility)$ is obtained, and the α value is computed. The calculation of the α value takes into account the disparity between the fact-checker's rating for the news or claim and the centroid of the largest cluster, as well as their previous credibility score.

Finally, the new credibility score for the fact-checker (*newCredibility*) is calculated by combining the current credibility (*currentCredibility*), α, and β values, and then multiplying them by the penalty parameter (Hr). The resulting score is returned as the new credibility score.

This algorithm effectively assesses a fact-checker's credibility while considering their historical ratings and the standard deviation (σ) for consensus inference, providing a comprehensive measure of their trustworthiness in evaluating the veracity of news or claims.

Algorithm 2 Credibility assessment (w, f, σ)

```
Intput: $w, f$
Output: $Credibility$
Begin
$clusters \leftarrow kMeans(w.RD)$    ▷ $w.RD$ contains the set of ratings of the news/claim $w$ made by fact-checkers.
$cluster \leftarrow largestCluster(clusters)$
$Mj \leftarrow centroid(cluster)$
$d \leftarrow \sqrt{(Mj - Nr)^2}$
if $d < \sigma$ then    ▷ Eq. 7.3.5
    $\beta \leftarrow 1 - \frac{d}{\sigma}$
else
    $\beta \leftarrow -(1 - \frac{\sigma}{d})$
end if
$Hr \leftarrow bayesianAverage(f.Hr)$    ▷ Eq. 7.3.6
$currentCredibility \leftarrow getCredibility(f)$    ▷ get current credibility of f
$\alpha \leftarrow oldCredibility * (1 - \frac{|getFactRate(f,w) - Mj|}{max(getAllFactRate(f,w))})$    ▷ The function $getFactRate(f, w)$ returns the rating made by fact-checker $f$ for the news $w$, and the function $getAllFactRate(f, w)$ returns all the ratings made by distinct fact-checkers for the same news.
$newCredibility \leftarrow (currentCredibility + \alpha * \beta) * Hr$
return newCredibility
End
```

7.4. Evaluation and discussion

No existing datasets are available to evaluate AFCC as proposed, to the best of my knowledge, this work is among the pioneering efforts in addressing the challenge of fact-checker consensus inference. In this section, the approach for evaluating the correctness and effectiveness of the proposed solution is outlined. The evaluation strategy includes both analytical and empirical methods.

The theoretical evaluation leverages analytical models, which are mathematical models with closed-form solutions. These models provide a high-level and generalized perspective, enabling predictions of system behaviour. This approach differs from experimental evaluation, which delves into detailed system behaviours [312]. Due to the absence of suitable datasets for empirical analysis, simulation has been selected as the method for empirical evaluation. Simulations are increasingly recognized as valuable for assessing system performance in recent research.

A comprehensive evaluation, incorporating both analytical and empirical analyses of the system, is presented in this section.

7.4.1. Analytical method

To accurately predict the behaviour of the Automated Fact-Checker Consensus Inference System (AFCC), a general mathematical analytical function, as outlined in Eq. (7.3.1), is employed. Analysis of Equations (7.3.2), (7.3.3), and (7.3.5) reveals that the consensus function is primarily dependent on the credibility function, which can be simplified to Eq. (7.4.1).

$$Cr_t(f_i) = \begin{cases} Cr_{t-1}(f_i) \times (1 + \gamma \times (1 - \frac{|Nr_i - Mj|}{max(NR)})) \times \overline{H}r & \text{if } f_i \text{ unbiased} \\ Cr_{t-1}(f_i) \times (1 - \varphi \times (1 - \frac{|Nr_i - Mj|}{max(NR)})) \times \overline{H}r & \text{otherwise} \end{cases} \tag{7.4.1}$$

For analytical purposes, γ and φ replace the factors from Eq. (7.3.3) and Eq. (7.3.5), representing the change amount in credibility. Similarly, $\overline{H}r$ measures the number of bad decisions in the past and is used as a replacement for Eq. (7.3.6). Constant values are assigned to these parameters in the analytical model.

For the purpose of easy comparison and visualization of result quality, max-normalization is employed to normalize the consensus values obtained from both analytical and simulation methods. Through this normalization, the maximum rating value (5) is mapped to (1) and the rest of the values fall within the interval (0,1]. The consensus max-normalization Ω is defined in Eq. (7.4.2):

$$\Omega = \frac{consensus}{max(NR)} \tag{7.4.2}$$

7.4.2. Simulation method

Since the model is not complicated and only a few results are required, the simulation is kept as simple as possible.

7.4.2.1. **Scenarios:** The following scenarios are provided to assess the simulation and analytical method results in different cases separately. These scenarios are designed based on the proportions of unbiased and biased fact-checkers to simulate real-world behaviour. The idea behind this is to check the behaviour of fact-checkers and the impact of the adjustment method. For example, when the number of biased fact-checkers exceeds the number of unbiased ones, the consensus model must adjust ratings to reflect the actual situation after a given number of rated news/claims. In this experiment, a fact-checker is considered unbiased if their rating consistently aligns with the real evaluation value ($realEval$), in other words, the difference between r and $realEval$ is less than a defined threshold ϵ, as per Eq. (7.4.3). In this case, ϵ is set to 1. Conversely, a fact-checker is considered biased if they rate news/claims incorrectly for the majority of cases.

$$|r - realEval| < \epsilon \tag{7.4.3}$$

In the real world, three main distribution groups of fact-checkers can be observed: where the number of unbiased and biased fact-checkers is equal, where the number of unbiased fact-checkers is greater than the number of biased ones, and where the number of biased ones is greater than the number of unbiased ones. Table 6 outlines the different scenarios and the corresponding percentages of unbiased and biased individuals (i.e., fact-checkers) in each scenario.

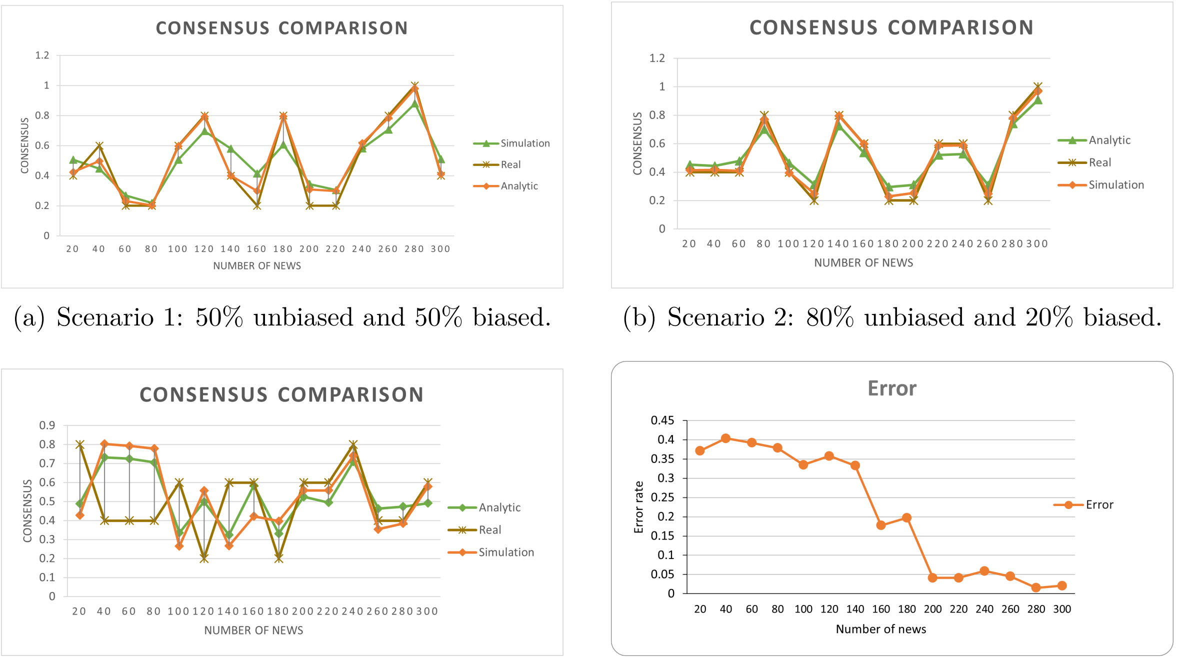
Table 6 – Population scenarios: distribution of unbiased and biased individuals

Scenario	**Unbiased**	**Biased**
Scenario 1	50%	50%
Scenario 2	80%	20%
Scenario 3	20%	80%

7.4.3. Results

In all three scenarios, the fact-checkers' rating values are randomly generated using a Gaussian distribution. This allows for the generation of values between 1 and 5 with finite variance. The evaluation is conducted using a dataset of 100 fact-checkers. The change amount in credibility parameters, represented by γ and φ, is set to 0.5. To simulate real-world data and account for the lack of datasets, the real rating value ($realEval$) of a given news or claim is randomly generated to facilitate the creation of both unbiased and biased fact-checkers.

The execution results of the simulations are depicted in Figure 4. In Scenario 1, where the percentages of unbiased and biased fact-checkers are equal (i.e., 50% each), both simulation and analytical results closely match the real values curve, with a negligible deviation.



(a) Scenario 1: 50% unbiased and 50% biased.

(b) Scenario 2: 80% unbiased and 20% biased.

(c) Scenario 3: 20% unbiased and 80% biased.

(d) Error.

Figure 4 – Execution scenarios

In Scenario 2, given 80% unbiased fact-checkers and 20% biased fact-checkers, the simulation and analytical results closely align with each other and exhibit no significant deviation compared to real values. In other words, the unbiased fact-checkers consistently remain unbiased. This is attributed to the fact that most fact-checkers are unbiased and their rating decisions closely resemble the real evaluations of news/claims. The simulation results curve outperforms the analytical model as it is closer to a real-world scenario of news/claim rating, which explains the good results of the proposed solution. Therefore, the efficiency of the proposed approach can be concluded, given that both the simulation and analytical evaluation results exhibit similar behaviour.

However, in Scenario 3, where only 20% of fact-checkers are unbiased, and 80% are biased, two significant observations can be made from the results shown in subfigure 4(c). Firstly, there is a substantial disparity between the values obtained from the analytical model, the simulation, and the real values. This discrepancy is primarily attributed to the higher proportion of biased fact-checkers (i.e., 80%) compared to unbiased ones (i.e., 20%). Secondly, as the number of news increases, both the analytical and simulation results progressively converge toward the real values. This convergence can be attributed to the credibility adjustment process.

The error curve depicted in the subfigure 4(d) underscores the significance of the adjustment process. In particular, the difference between the real values and the values of the simulation method decreases over time. Starting from the point where the number of news reaches 200, the curve levels off towards a value of 0. This phenomenon is explained by the adjustment mechanism, which relies on the majority rating, placing significant weight on ratings closely aligning with actual values, and incorporating a penalty for ratings that deviate from real values.

7.5. Conclusion

In this chapter, I have introduced the AFCC (Automated Fact-Checkers Consensus and Credibility) system, a key part of my FACTS-ON framework, which enhances trust in fake news detection on online social networks. The AFCC system innovatively merges the e-commerce five-star rating system with majority voting models to synthesize insights from multiple fact-checkers. This approach is critical in assessing the credibility of fact-checking organizations, establishing trust in their ratings, and achieving consensus on news items or claims based on collective assessments. By transforming multiple, diverse ratings into a singular, coherent decision, the AFCC model significantly strengthens the trustworthiness of information on online social networks.

In this chapter, I detailed a unique process for transforming textual ratings into numerical values using unsupervised learning techniques such as Word2vec and K-means clustering. Additionally, I proposed two algorithms for consensus inference and credibility assessment of fact-checkers, inspired by majority rating principles and the e-commerce five-star rating system. These algorithms were evaluated using both analytical and simulation methods on a large scale, demonstrating the AFCC model's effectiveness in performing consensus inference and credibility assessment.

As this chapter concludes, I will next present the final chapter of this dissertation, which focuses on the conclusion and future directions of my research. This concluding chapter will recap the main findings and contributions of the FACTS-ON framework, encompassing systems like EXMULF, ExFake, MythXpose, and AFCC. It will also explore potential avenues for future research, aiming to further enhance these systems, investigate new methodologies, and address emerging challenges in the evolving field of fake news detection in online social networks.

Chapter 8

Conclusion and future work

8.1. Contributions

In this dissertation, I have made advances in combating fake news, misinformation, and disinformation on online social networks. My contributions, validated through rigorous empirical experimentation, are innovative and impactful, addressing critical aspects of detection and understanding of online false information (i.e., fake news, misinformation and disinformation). These contributions, closely aligned with the respective chapters, represent a substantial advancement in the field of fake news, misinformation and disinformation studies. By adopting an interdisciplinary approach and employing innovative methodologies, my work sets new benchmarks for research and opens up new avenues for further exploration and development in this critically important area. My key contributions, aligned with the respective chapters, are as follows:

8.1.1. Foundational review and understanding

In **Chapter 1**, my work involved conducting a thorough review of the concepts and challenges related to fake news, misinformation, and disinformation (i.e., false information) in social media. This foundational review was essential to establish a clear understanding of the false information landscape, differentiating between various types of deceptive content such as content-based fake news, and intent-based fake news. The comprehensive nature of this review set the stage for the more targeted research that followed.

In **Chapter 2**, I extended this groundwork by critically examining various detection methods and techniques used in the field. This involved an in-depth analysis of existing tools and methodologies, assessing their strengths and weaknesses in detecting fake news. This chapter not only provided a comprehensive survey of current techniques but also critically analyzed the state-of-the-art. Furthermore, it clearly positioned my contributions in relation to the existing state-of-the-art, highlighting the advancements and unique aspects of my work within the field.

8.1.2. FACTS-ON framework

Chapter 3 is dedicated to the FACTS-ON framework, a cornerstone of my dissertation. This framework represents a novel and comprehensive approach to tackling false information challenges. I designed FACTS-ON to address the multifaceted nature of fake news, incorporating various modules and functionalities that work together to detect and analyze deceptive content.

8.1.3. Explainable multimodal content-based fake news detection

In **Chapter 4**, I focused on the development EXMULF, a system for detecting fake news using multimodal content analysis. A key aspect of EXMULF is its explainability, which ensures that the detection process is transparent and understandable. By integrating various content types, such as text and images, and providing clear explanations for its detection decisions, EXMULF represents a significant step forward in making fake news detection systems more user-friendly and trustworthy.

8.1.4. Multimodal content and social context-based system for explainable false information detection with personality prediction

In **Chapter 5**, I introduced "MythXpose", a system that combines content analysis with social context information to detect false information on online social networks. MythXpose builds upon the EXMULF system, introduced in **Chapter 4**, and incorporates the PERSONA module, which assesses the personality traits of Online Social Network (OSN) users. This fusion enhances the ability to detect deceptive content while maintaining transparency through explainability mechanisms. MythXpose represents a significant step toward more user-centric and context-aware false information detection.

8.1.5. Explainable false information detection based on content, context, and external evidence

In **Chapter 6**, I introduced and explored the ExFake system, which notably integrates content analysis with context and external evidence to detect false information. Central to this system is the incorporation of explainability features, which are crucial in providing clear insights into why certain content is identified as fake. This involves not just analyzing the content itself, but also understanding the surrounding context and utilizing external evidence to provide a comprehensive and transparent rationale for each detection decision. My approach ensures that the detection process is not only effective but also transparent and understandable for users, addressing one of the key challenges in the field of fake news detection, the need for systems to be as interpretable as they are accurate. The effectiveness of this holistic and explainable approach was demonstrated through comprehensive empirical experiments.

8.1.6. Fact-Checkers' consensus inference and credibility assessment for trust-based fake news detection

Chapter 7 introduces the AFCC system, which is a novel approach to leveraging the collective insights of fact-checkers for trust-based fake news detection. This system is designed to aggregate and analyze the judgments of various fact-checking sources, thereby enhancing the credibility assessment of news items and claims. The effectiveness of the AFCC system was evaluated using analytical and simulation methods due to the absence of comprehensive datasets in this novel area of research. As one of the pioneering efforts to address the challenge of fact-checker consensus inference, this evaluation underscores the system's potential in offering a more nuanced and reliable assessment of news authenticity, particularly important in today's rapid information-sharing environment. This step marks a key advancement in strengthening the trustworthiness and accuracy of fake news detection processes.

8.2. Future perspectives

In the dynamic landscape of online social networks, the ongoing battle against false information represents a continuously evolving challenge. In my dissertation, while I have made significant progress in combating fake news, misinformation, and disinformation on online social networks, there are inherent limitations in my current solutions that I aim to address in my future research. Recognizing these limitations is crucial for the ongoing development of more effective and comprehensive systems. The key limitations and the areas for future enhancements include:

8.2.1. Combining systems for comprehensive detection and explanation

In future work, I plan to combine all the systems developed in my research (EXMULF, MythXpose, ExFake, and AFCC) to operate collectively in detecting and predicting false information. This integrated approach aims to leverage the strengths of each system to provide comprehensive explanations and a more robust detection mechanism. By unifying these systems, the potential to detect, analyze, and explain false information on a broader and more effective scale is significantly increased.

8.2.2. Real-time detection and learning

In my dissertation, I have predominantly utilized historical data for detecting fake news and misinformation, which poses limitations in the dynamic and rapidly evolving landscape of online social networks. Recognizing this, my future research will pivot towards developing real-time detection systems. These systems will be crucial for promptly detecting, predicting, and mitigating the spread of false information as it emerges. By incorporating *data stream mining* techniques, I aim to evolve my models to process and analyze real-time data, moving beyond the constraints of historical analysis. This advancement is critical for improving responsiveness and timeliness, addressing a key limitation in my current approach and significantly enhancing the effectiveness of false information mitigation strategies.

8.2.3. Complexity in multimodal data integration

While my research has incorporated multimodal data, the integration of these diverse data types (i.e., text, audio, video) poses significant complexities. The current methodologies may not fully exploit the potential of multimodal data, especially in terms of correlating different types of content for more accurate detection. Enhancing the sophistication of multimodal integration will be a key focus, aiming to leverage the full spectrum of available data more effectively.

8.2.4. Bot detection and analysis

The current approach may not sufficiently distinguish between human users and non-human entities like bots and cyborgs, which are increasingly sophisticated in mimicking human behaviour. Future research will concentrate on refining bot detection techniques, using advanced machine learning models to better identify and differentiate between genuine and artificial behaviours in the network.

8.2.5. Enhancement of datasets

The datasets used in my current research may not fully represent the diverse and complex nature of fake news across different platforms and contexts. There is a need for more comprehensive and varied datasets that encompass a broader range of fake news instances. Future work will involve the creation and utilization of such datasets to enhance the accuracy and generalizability of detection models.

8.2.6. Explainability and user comprehension

While I have integrated explainability into my detection systems, the current level of explainability may not be sufficient for all users to fully understand the rationale behind certain detections. Enhancing the clarity and accessibility of explanations provided by these systems will be crucial in making them more user-friendly and trustworthy.

A key aspect of this, future direction involves considering users' evaluations of the explanations. This means actively seeking feedback from users regarding the effectiveness and understandability of the explanations offered by the system. By assessing how well users comprehend the rationale behind certain detections, modifications and improvements can be made to tailor the explanations to better meet their needs.

This user-centric approach to refining explainability not only aims to improve the overall user experience but also to ensure that the systems are more trustworthy and reliable. Understanding how users interact with and perceive the explanations will provide valuable insights into how these systems can be optimized for greater effectiveness and user satisfaction. This step is crucial in bridging the gap between sophisticated detection technologies and the diverse comprehension levels of users, ultimately leading to more accessible and effective tools for combating false information in online social networks.

8.2.7. Ethical and privacy considerations in personality-based predictions

Future research will also focus on the ethical implications and privacy concerns arising from personality-based predictions used in identifying potential spreaders of fake news. This includes establishing clear guidelines and protocols to ensure that such predictive models do not infringe on individual privacy rights or lead to unintended discrimination or stigmatization. There's a need to develop mechanisms to make these predictions more transparent and accountable.

8.2.8. Mitigating data bias and improving generalizability

Another critical area of focus will be to address potential biases in the datasets used for training the models. Ensuring that the datasets are diverse and representative of various demographics is vital for preventing biased predictions. Additionally, future research will aim to test and refine these models in real-world scenarios to confirm their effectiveness outside controlled experimental settings, thus improving their generalizability and reliability.

8.2.9. Incorporation of advanced AI tools like ChatGPT and Bing

Incorporating advanced AI tools such as ChatGPT and Bing, both of which are Large Language Models (LLMs), into the FACTS-ON framework presents an exciting frontier. These tools, with their extensive knowledge bases and sophisticated natural language processing capabilities, offer promising avenues for enhancing fact-checking and false information detection. However, current challenges such as maintaining up-to-date information, ensuring unbiased content, continual updates to keep up with the evolving nature of language and false information strategies, and navigating privacy concerns need to be addressed. Exploring the integration of these tools will involve careful consideration of these limitations to fully harness their potential in the battle against fake news on online social networks.

8.2.10. Adapting to evolving social media dynamics

Lastly, recognizing the dynamic nature of online social networks, future research will explore adaptive algorithms that can evolve with changing user behaviours, trends, and the emergence of new social media platforms. This will involve developing models that are not only responsive to the current landscape but are also flexible enough to adjust to future changes in social media use and fake news tactics.

8.3. Conclusion

In conclusion, this dissertation has delved into addressing the pervasive challenge of combating false information (i.e., fake news, misinformation and disinformation) spread on online social networks. Through an in-depth exploration of multiple aspects (i.e., multimodal content, context and external evidence together with explainability), this research has contributed to the existing body of knowledge on counterfeit truths in the realm of online social networks. The insights gained from this dissertation shed light on combating the intricate web of fake news, misinformation, and disinformation.

The significance of this dissertation extends beyond its academic contributions, resonating with the evolving dynamics of our digital society. By unveiling the mechanisms that underpin the spread of counterfeit truths, individuals and communities are empowered to make more informed decisions within the online landscape. Furthermore, the synthesis of knowledge, insights, and practical applications derived from this dissertation paves the way for a more robust and trustworthy virtual environment.

References

[1] Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter et Mohammad Rezwanul Huq : Detecting fake news using machine learning and deep learning algorithms. *In 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE, 2019.

[2] Abdullah-All-Tanvir, Ehesas Mia Mahir, S M Asiful Huda et Shuvo Barua : A hybrid approach for identifying authentic news using deep learning methods on popular Twitter threads. *In International Conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–6. IEEE, 2020.

[3] Omar Abu Arqoub, Adeola Abdulateef Elega, Bahire Efe Özad, Hanadi Dwikat et Felix Adedamola Oloyede : Mapping the scholarship of fake news research: A systematic review. *Journalism Practice*, 16(1):56–86, 2022.

[4] Sajjad Ahmed, Knut Hinkelmann et Flavio Corradini : Development of fake news model using machine learning through natural language processing. *International Journal of Computer and Information Engineering*, 14(12):454–460, 2020.

[5] Esma Aïmeur, Gilles Brassard et Jonathan Rioux : Data privacy: An end-user perspective. *International Journal of Computer networks and Communications Security*, 1(6):237–250, 2013.

[6] Esma Aïmeur, Hicham Hage et Sabrine Amri : The scourge of online deception in social networks. *In 2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1266–1271. IEEE, 2018.

[7] Alberto Alemanno : How to counter fake news? A taxonomy of anti-fake news approaches. *European Journal of Risk Regulation*, 9(1):1–5, 2018.

[8] Raed Alharbi, Minh N Vu et My T Thai : Evaluating fake news detection models from explainable machine learning perspectives. *In ICC 2021-IEEE International Conference on Communications*, pages 1–6. IEEE, 2021.

[9] Hunt Allcott et Matthew Gentzkow : Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[10] Jennifer Allen, Antonio A Arechar, Gordon Pennycook et David G Rand : Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393,

2021.

[11] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild et Duncan J Watts : Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14):eaay3539, 2020.

[12] Daniel Allington, Bobby Duffy, Simon Wessely, Nayana Dhavan et James Rubin : Health-protective behaviour, social media usage and conspiracy belief during the Covid-19 public health emergency. *Psychological Medicine*, pages 1–7, 2020.

[13] Patricia Alonso-Galbán et Claudia Alemañy-Castilla : Curbing misinformation and disinformation in the covid-19 era: a view from cuba. *MEDICC review*, 22:45–46, 2022.

[14] Sacha Altay, Anne-Sophie Hacquin et Hugo Mercier : Why do so few people share fake news? it hurts their reputation. *New Media & Society*, 24(6):1303–1324, 2022.

[15] Sabrine Amri, Henri-Cedric Mputu Boleilanga et Esma Aïmeur : Exfake: Towards an explainable fake news detection based on content and social context information. *In International Conference on Security & Management (SAM'23)*, 2023. Accepted.

[16] Sabrine Amri, Dorsaf Sallami et Esma Aïmeur : Exmulf: An explainable multimodal content-based fake news detection system. *In International Symposium on Foundations and Practice of Security*, pages 177–187. Springer, 2022.

[17] Jack Andersen et Sille Obelitz Søe : Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news–the case of Facebook. *European Journal of Communication*, 35(2):126–139, 2020.

[18] Oberiri Destiny Apuke et Bahiyah Omar : Fake news and Covid-19: Modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, 56:101475, 2021.

[19] Oberiri Destiny Apuke, Bahiyah Omar, Elif Asude Tunca et Celestine Verlumun Gever : The effect of visual multimedia instructions against fake news spread: A quasi-experimental study with nigerian students. *Journal of Librarianship and Information Science*, page 09610006221096477, 2022.

[20] Reema Aswani, SP Ghrera, Arpan Kumar Kar et Satish Chandra : Identifying buzz in social media: A hybrid approach using artificial bee colony and k-nearest neighbors for outlier detection. *Social Network Analysis and Mining*, 7(1):1–10, 2017.

[21] Mihai Avram, Nicholas Micallef, Sameer Patil et Filippo Menczer : Exposure to social engagement metrics increases vulnerability to misinformation. *arXiv preprint arXiv:2005.04682*, 2020.

[22] Adam Badawy, Kristina Lerman et Emilio Ferrara : Who falls for online political manipulation? *In Companion Proceedings of The 2019 World Wide Web Conference*, pages 162–168, 2019.

[23] Pritika BAHAD, Preeti SAXENA et Raj KAMAL : Fake news detection using bi-directional LSTM-recurrent neural network. *Procedia Computer Science*, 165:74–82, 2019.

[24] Jonathan BAKDASH, Char SAMPLE, Monica RANKIN, Murat KANTARCIOGLU, Jennifer HOLMES, Sue KASE, Erin ZAROUKIAN et Boleslaw SZYMANSKI : The future of deception: Machine-generated and manipulated images, video, and audio? *In 2018 International Workshop on Social Sensing (SocialSens)*, pages 2–2. IEEE, 2018.

[25] Meital BALMAS : When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research*, 41(3):430–454, 2014.

[26] Alwi M BAMHDI, Iram ABRAR et Faheem MASOODI : An ensemble based approach for effective intrusion detection using majority voting. *Telkomnika (Telecommunication Computing Electronics and Control)*, 19(2):664–671, 2021.

[27] João Pedro BAPTISTA et Anabela GRADIM : Understanding fake news consumption: A review. *Social Sciences*, 9(10), 2020.

[28] João Pedro BAPTISTA et Anabela GRADIM : A working definition of fake news. *Encyclopedia*, 2(1):632–645, 2022.

[29] Zach BASTICK : Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in Human Behavior*, 116:106633, 2021.

[30] Cédric BATAILLER, Skylar M BRANNON, Paul E TEAS et Bertram GAWRONSKI : A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*, 17(1):78–98, 2022.

[31] Marwa BEN JABRA, Anis KOUBAA, Bilel BENJDIRA, Adel AMMAR et Habib HAMAM : Covid-19 diagnosis in chest x-rays using deep learning and majority voting. *Applied Sciences*, 11(6):2884, 2021.

[32] Alessandro BESSI et Emilio FERRARA : Social bots distort the 2016 US presidential election online discussion. *First Monday*, 21(11-7), 2016.

[33] Amrita BHATTACHARJEE, Kai SHU, Min GAO et Huan LIU : Disinformation in the online information ecosystem: Detection, mitigation and challenges. *arXiv preprint arXiv:2010.09113*, 2020.

[34] Bimal BHATTARAI, Ole-Christoffer GRANMO et Lei JIAO : Explainable tsetlin machine framework for fake news detection with credibility score assessment. *arXiv preprint arXiv:2105.09114*, 2021.

[35] Md Momen BHUIYAN, Amy X ZHANG, Connie Moon SEHAT et Tanushree MITRA : Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[36] David M Blei, Andrew Y Ng et Michael I Jordan : Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[37] Leticia Bode et Emily K Vraga : In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638, 2015.

[38] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou et Yiannis Kompatsiaris : Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

[39] Alessandro Bondielli et Francesco Marcelloni : A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55, 2019.

[40] David Borukhson, Philipp Lorenz-Spreen et Marco Ragni : When does an individual accept misinformation? an extended investigation through cognitive modeling. *Computational brain & behavior*, 5(2):244–260, 2022.

[41] Alexandre Bovet et Hernán A Makse : Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1):1–14, 2019.

[42] Samuel R Bowman, Gabor Angeli, Christopher Potts et Christopher D Manning : A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[43] Nadia M Brashier, Gordon Pennycook, Adam J Berinsky et David G Rand : Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5), 2021.

[44] Paul R Brewer, Dannagal Goldthwaite Young et Michelle Morreale : The impact of real news about “fake news”: Intertextual processes and political satire. *International Journal of Public Opinion Research*, 25(3):323–343, 2013.

[45] Rex P Bringula, Annaliza E Catacutan-Bangit, Manuel B Garcia, John Paul S Gonzales et Arlene Mae C Valderama : “who is gullible to political disinformation?”: predicting susceptibility of university students to fake news. *Journal of Information Technology & Politics*, 19(2):165–179, 2022.

[46] Michael V Bronstein, Gordon Pennycook, Lydia Buonomano et Tyrone D Cannon : Belief in fake news, responsiveness to cognitive conflict, and analytic reasoning engagement. *Thinking & Reasoning*, 27(4):510–535, 2021.

[47] Francesco Buccafurri, Gianluca Lax, Serena Nicolazzo et Antonino Nocera : Tweetchain: An alternative to blockchain for crowd-based applications. *In International Conference on Web Engineering*, pages 386–393. Springer, 2017.

[48] Sheldon Burshtein : The true story on fake news. *Intellectual Property Journal*, 29(3), 2017.

[49] Dustin P Calvillo, Alex León et Abraham M Rutchick : Personality and misinformation. *Current Opinion in Psychology*, page 101752, 2023.

[50] Matteo Cardaioli, Stefano Cecconello, Mauro Conti, Luca Pajola et Federico Turrin : Fake news spreaders profiling through behavioural analysis. *In CLEF (Working Notes)*, 2020.

[51] Fernando Cardoso Durier da Silva, Rafael Vieira et Ana Cristina Garcia : Can machines learn to detect fake news? A survey focused on social media. *In Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[52] Elinor Carmi, Simeon J Yates, Eleanor Lockley et Alicja Pawluczuk : Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, 9(2):1–22, 2020.

[53] Marlie Celliers et Marie Hattingh : A systematic review on fake news themes reported in literature. *In Conference on e-Business, e-Services and e-Society*, pages 223–234. Springer, 2020.

[54] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio et Lucia Specia : Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

[55] Yize Chen, Quanlai Li et Hao Wang : Towards trusted social networks with blockchain technology. *arXiv preprint arXiv:1801.02796*, 2018.

[56] Lu Cheng, Ruocheng Guo, Kai Shu et Huan Liu : Towards causal understanding of fake news dissemination. *arXiv preprint arXiv:2010.10580*, 2020.

[57] Ming Ming Chiu et Yu Won Oh : How fake news differs from personal lies. *American Behavioral Scientist*, 65(2):243–258, 2021.

[58] Myojung Chung et Nuri Kim : When I learn the news is false: How fact-checking information stems the spread of fake news via third-person perception. *Human Communication Research*, 47(1):1–24, 2021.

[59] Jonathan Clarke, Hailiang Chen, Ding Du et Yu Jeffrey Hu : Fake news, investor attention, and market reaction. *Information Systems Research*, 2020.

[60] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan *et al.* : Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4):1073–1095, 2020.

[61] Botambu Collins, Dinh Tuyen Hoang, Ngoc Thanh Nguyen et Dosam Hwang : Fake news types and detection models on social media a state-of-the-art survey. *In Asian Conference on Intelligent Information and Database Systems*, pages 562–573. Springer, 2020.

[62] Nadia K Conroy, Victoria L Rubin et Yimin Chen : Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[63] Nicole A Cooke : Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The library Quarterly*, 87(3):211–221, 2017.

[64] Michele Coscia et Luca Rossi : Distortions of political bias in crowdsourced misinformation flagging. *Journal of the Royal Society Interface*, 17(167):20200020, 2020.

[65] Theodora Dame Adjin-Tettey : Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. *Cogent Arts & Humanities*, 9(1):2037229, 2022.

[66] Madeleine de Cock Buning : *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation.* Publications Office of the European Union, 2018.

[67] S Deepak et Bhadrachalam Chitturi : Deep neural approach to fake-news identification. *Procedia Computer Science*, 167:2236–2243, 2020.

[68] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala et Fabiana Zollo : Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):1–22, 2019.

[69] John Demuyakor et Edward Martey Opata : Fake news on social media: Predicting which media format influences fake news most on facebook. *Journal of Intelligent Communication*, 2(1), 2022.

[70] Ronald Denaux et Jose Manuel Gomez-Perez : Linked credibility reviews for explainable misinformation detection. *In International Semantic Web Conference*, pages 147–163. Springer, 2020.

[71] Hossein Derakhshan et Claire Wardle : Information disorder: Definitions. *In Understanding and Addressing the Disinformation Ecosystem*, pages 5–12, 2017.

[72] Angel N Desai, Diandra Ruidera, Julie M Steinbrink, Bruno Granwehr et Dong Heun Lee : Misinformation and disinformation: The potential disadvantages of social media in infectious disease and how to combat them. *Clinical Infectious Diseases*, 74(Supplement_3):e34–e39, 2022.

[73] Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova : Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[74] Lusiana Citra Dewi, Alvin Chandra *et al.* : Social media web scraping using social media developers api and regex. *Procedia Computer Science*, 157:444–449, 2019.

[75] Giandomenico Di Domenico, Jason Sit, Alessio Ishizaka et Daniel Nunan : Fake news, social media and marketing: A systematic review. *Journal of Business Research*,

124:329–341, 2021.

[76] Nicholas Dias, Gordon Pennycook et David G Rand : Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, 1(1), 2020.

[77] Karen Watts DiCicco et Nitin Agarwal : Blockchain technology-based solutions to fight misinformation: A survey. *In Disinformation, Misinformation, and Fake News in Social Media*, pages 267–281. Springer, 2020.

[78] Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang et Farzin Deravi : Understanding conspiracy theories. *Political Psychology*, 40:3–35, 2019.

[79] Stephanie Edgerly, Rachel R Mourão, Esther Thorson et Samuel M Tham : When do audiences verify? How perceptions about message and source influence audience verification of news headlines. *Journalism & Mass Communication Quarterly*, 97(1):52–71, 2020.

[80] Jana Laura Egelhofer et Sophie Lecheler : Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, 43(2):97–116, 2019.

[81] Mohamed K Elhadad, Kin Fun Li et Fayez Gebali : A novel approach for selecting hybrid features from online news textual metadata for fake news detection. *In International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 914–925. Springer, 2019.

[82] ERGA : *'FAKE NEWS' AND THE INFORMATION DISORDER*. European Broadcasting Union (EBU), 2018.

[83] ERGA : *Notions of disinformation and related concepts*. European Regulators Group for Audiovisual Media Services (ERGA), 2021.

[84] Álex Escolà-Gascón : New techniques to measure lie detection using Covid-19 fake news and the Multivariable Multiaxial Suggestibility Inventory-2 (MMSI-2). *Computers in Human Behavior Reports*, 3:100049, 2021.

[85] Matthias Fatke : Personality traits and political ideology: A first global assessment. *Political Psychology*, 38(5):881–899, 2017.

[86] Lisa Fazio : Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, 1(2), 2020.

[87] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer et Alessandro Flammini : The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[88] DJ Flynn, Brendan Nyhan et Jason Reifler : The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150, 2017.

[89] Paula Fraga-Lamas et Tiago M Fernández-Caramés : Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality. *IT Professional*, 22(2):53–59, 2020.

[90] Daniel Freeman, Felicity Waite, Laina Rosebrock, Ariane Petit, Chiara Causier, Anna East, Lucy Jenner, Ashley-Louise Teale, Lydia Carr, Sophie Mulhall *et al.* : Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological Medicine*, pages 1–13, 2020.

[91] Adrien Friggeri, Lada Adamic, Dean Eckles et Justin Cheng : Rumor cascades. *In Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

[92] Santiago Alonso García, Gerardo Gómez García, Mariano Sanz Prieto, Antonio José Moreno Guerrero et Carmen Rodríguez Jiménez : The impact of term fake news on the scientific community. Scientific performance and mapping in web of science. *Social Sciences*, 9(5), 2020.

[93] R Kelly Garrett et Robert M Bond : Conservatives' susceptibility to political misperceptions. *Science Advances*, 7(23):eabf1234, 2021.

[94] Anastasia Giachanou, Esteban A Ríssola, Bilal Ghanem, Fabio Crestani et Paolo Rosso : The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers. *In International Conference on Applications of Natural Language to Information Systems*, pages 181–192. Springer, 2020.

[95] Anastasia Giachanou, Guobiao Zhang et Paolo Rosso : Multimodal fake news detection with textual, visual and semantic information. *In International Conference on Text, Speech, and Dialogue*, pages 30–38. Springer, 2020.

[96] Anastasia Giachanou, Guobiao Zhang et Paolo Rosso : Multimodal multi-image fake news detection. *In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654. IEEE, 2020.

[97] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B Everett *et al.* : Fake news vs satire: A dataset and analysis. *In Proceedings of the 10th ACM Conference on Web Science*, pages 17–21, 2018.

[98] Mohammad Hadi Goldani, Saeedeh Momtazi et Reza Safabakhsh : Detecting fake news with capsule neural networks. *Applied Soft Computing*, 101:106991, 2021.

[99] Itay Goldstein et Liyan Yang : Good disclosure, bad disclosure. *Journal of Financial Economics*, 131(1):118–138, 2019.

[100] M. Grandini, E. Bagli et G. Visani : Metrics for multi-class classification: an overview. *CoRR*, abs/2008.05756, 2020.

[101] Margherita Grandini, Enrico Bagli et Giorgio Visani : Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

[102] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson et David Lazer : Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425):374–378, 2019.

[103] Rosanna E Guadagno et Karen Guttieri : Fake news and information warfare: An examination of the political and psychological processes from the digital sphere to the real world. *In Research Anthology on Fake News, Political Warfare, and Combatting the Spread of Misinformation*, pages 218–242. IGI Global, 2021.

[104] Andrew Guess, Jonathan Nagler et Joshua Tucker : Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1):eaau4586, 2019.

[105] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang et Zhiwen Yu : The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36, 2020.

[106] Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu et Miao Yu : Exploiting emotions for fake news detection on social media. *arXiv preprint arXiv:1903.01728*, 2019.

[107] Ashish Gupta, Han Li, Alireza Farnoush et Wenting Jiang : Understanding patterns of covid infodemic: A systematic and pragmatic approach to curb fake news. *Journal of business research*, 140:670–683, 2022.

[108] Louisa Ha, Loarre Andreu Perez et Rik Ray : Mapping recent development in scholarship on fake news and misinformation, 2008 to 2017: Disciplinary contribution, topics, and impact. *American Behavioral Scientist*, 65(2):290–315, 2021.

[109] Ammara Habib, Muhammad Zubair Asghar, Adil Khan, Anam Habib et Aurangzeb Khan : False information detection in online content and its role in decision making: A systematic literature review. *Social Network Analysis and Mining*, 9(1):1–20, 2019.

[110] Hicham Hage, Esma Aïmeur et Amel Guedidi : Understanding the landscape of online deception. *In Research Anthology on Fake News, Political Warfare, and Combatting the Spread of Misinformation*, pages 39–66. IGI Global, 2021.

[111] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta et Wazir Zada Khan : An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 117:47–58, 2021.

[112] Tarek Hamdi, Hamda Slimi, Ibrahim Bounhas et Yahya Slimani : A hybrid approach for fake news detection in Twitter based on user features and graph embedding. *In International Conference on Distributed Computing and Internet Technology*, pages 266–280. Springer, 2020.

[113] Michael Hameleers : Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the us and netherlands. *Information, Communication & Society*, 25(1):110–126, 2022.

[114] Michael Hameleers, Anna Brosius et Claes H de Vreese : Whom to trust? media exposure patterns of citizens with perceptions of misinformation and disinformation related to the news media. *European Journal of Communication*, page 02673231211072667, 2022.

[115] Michael Hameleers, Thomas E Powell, Toni GLA Van Der Meer et Lieke Bos : A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2):281–301, 2020.

[116] Kris Hartley et Minh Khuong Vu : Fighting fake news in the Covid-19 era: Policy insights from an equilibrium model. *Policy Sciences*, 53(4):735–758, 2020.

[117] Haya R Hasan et Khaled Salah : Combating deepfake videos using blockchain and smart contracts. *Ieee Access*, 7:41596–41606, 2019.

[118] Kaiming He, Georgia Gkioxari, Piotr Dollár et Ross Girshick : Mask r-cnn. *In Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[119] Srinidhi Hiriyannaiah, AMD Srinivas, Gagan K Shetty, GM Siddesh et KG Srinivasa : A computationally intelligent agent for detecting fake news using generative adversarial networks. *Hybrid Computational Intelligence: Challenges and Applications*, pages 69–96, 2020.

[120] Seyedmehdi Hosseinimotlagh et Evangelos E Papalexakis : Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. *In Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*, 2018.

[121] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan et Ming Zhou : Compare to the knowledge: Graph neural fake news detection with external knowledge. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763, 2021.

[122] Steve Huckle et Martin White : Fake news: A technological approach to proving the origins of content, using blockchains. *Big Data*, 5(4):356–371, 2017.

[123] Jordan S Huffaker, Jonathan K Kummerfeld, Walter S Lasecki et Mark S Ackerman : Crowdsourced detection of emotionally manipulative language. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[124] Cherilyn Ireton et Julie Posetti : *Journalism, fake news & disinformation: Handbook for journalism education and training.* Unesco Publishing, 2018.

[125] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang et Guandong Xu : Deep learning for misinformation detection on online social networks: A survey and new perspectives.

Social Network Analysis and Mining, 10(1):1–20, 2020.

[126] Max Ismailov, Michail Tsikerdekis et Sherali Zeadally : Vulnerabilities to online social network identity deception detection research and recommendations for mitigation. *Future Internet*, 12(9):148, 2020.

[127] Maurice Jakesch, Moran Koren, Anna Evtushenko et Mor Naaman : The role of source and expressive responding in political news evaluation. *In Computation and Journalism Symposium*, 2019.

[128] Kathleen Hall Jamieson : *Cyberwar: How Russian hackers and trolls helped elect a president: What we don't, can't, and do know*. Oxford University Press, 2020.

[129] Fengcheng Ji, Dongping Ming, Beichen Zeng, Jiawei Yu, Yuanzhao Qing, Tongyao Du et Xinyi Zhang : Aircraft detection in high spatial resolution remote sensing images combining multi-angle features driven and majority voting cnn. *Remote Sensing*, 13(11):2207, 2021.

[130] Shengyi Jiang, Xiaoting Chen, Liming Zhang, Sutong Chen et Haonan Liu : User-characteristic enhanced model for fake news detection in social media. *In CCF International Conference on Natural Language Processing and Chinese Computing*, pages 634–646. Springer, 2019.

[131] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang et Jiebo Luo : Multimodal fusion with recurrent neural networks for rumor detection on microblogs. *In Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816, 2017.

[132] Zhiwei Jin, Juan Cao, Yongdong Zhang et Jiebo Luo : News verification by exploiting conflicting social viewpoints in microblogs. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

[133] Tee Wee Jing et Raja Kumar Murugesan : A theoretical framework to build trust and prevent fake news in social media using blockchain. *In International Conference of Reliable Information and Communication Technology*, pages 955–962. Springer, 2018.

[134] S Mo Jones-Jang, Tara Mortensen et Jingjing Liu : Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist*, 65(2):371–388, 2021.

[135] Andreas Jungherr et Ralph Schroeder : Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media+ Society*, 7(1):2056305121988928, 2021.

[136] Rohit Kumar Kaliyar : Fake news detection using a deep neural network. *In 2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–7. IEEE, 2018.

[137] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang et Soumendu Sinha : Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44, 2020.

[138] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis et Vassilios Peristeras : A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, 23(5):1301–1326, 2021.

[139] Jozef Kapusta, L'ubomír Benko et Michal Munk : Fake news identification based on sentiment and frequency analysis. *In International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning*, pages 400–409. Springer, 2019.

[140] Sawinder Kaur, Parteek Kumar et Ponnurangam Kumaraguru : Automating fake news detection system using multi-level voting model. *Soft Computing*, 24(12):9049–9069, 2020.

[141] Sayeed Ahsan Khan, Mohammed Hazim Alkawaz et Hewa Majeed Zangana : The use and abuse of social media for spreading fake news. *In 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pages 145–148. IEEE, 2019.

[142] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf et Manuel Gomez-Rodriguez : Leveraging the crowd to detect and reduce the spread of fake news and misinformation. *In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 324–332, 2018.

[143] Soo Young Kim et Arun Upneja : Majority voting ensemble with a decision trees for business failure prediction during economic downturns. *Journal of Innovation & Knowledge*, 6(2):112–123, 2021.

[144] David Klein et Joshua Wueller : Fake news: A legal perspective. *Journal of Internet Law (Apr. 2017)*, 20(10):5–13, 2017.

[145] Shimon Kogan, Tobias J Moskowitz et Marina Niessner : Fake news: Evidence from financial markets. *Available at SSRN 3237763*, 2019.

[146] James H Kuklinski, Paul J Quirk, Jennifer Jerit, David Schwieder et Robert F Rich : Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3):790–816, 2000.

[147] Srijan Kumar et Neil Shah : False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.

[148] Srijan Kumar, Robert West et Jure Leskovec : Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes. *In Proceedings of the 25th International Conference on World Wide Web*, pages 591–602, 2016.

[149] Rina Kumari et Asif Ekbal : Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications*, 184: 115412, 2021.

[150] Lukas Kurasinski et Radu-Casian Mihailescu : Towards machine learning explainability in text classification for fake news detection. *In 2020 19th IEEE International*

Conference on Machine Learning and Applications (ICMLA), pages 775–781. IEEE, 2020.

[151] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro et Damiano Spina : Crowdsourcing truthfulness: The impact of judgment scale and assessor bias. *In European Conference on Information Retrieval*, pages 207–214. Springer, 2020.

[152] Candice Lanius, Ryan Weber et William I MacKenzie : Use of bot and content flags to limit the spread of misinformation among social networks: A behavior and attitude survey. *Social Network Analysis and Mining*, 11(1):1–15, 2021.

[153] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild *et al.* : The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[154] Thai Le, Kai Shu, Maria D Molina, Dongwon Lee, S Shyam Sundar et Huan Liu : 5 sources of clickbaits you should know! Using synthetic clickbaits to improve prediction and distinguish between bot-generated and human-written headlines. *In 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 33–40. IEEE, 2019.

[155] Stephan Lewandowsky : Climate change, disinformation, and how to combat it. *Annual Review of Public Health*, 42, 2020.

[156] Ming-Hui Li, Zhiqin Chen et Li-Lin Rao : Emotion, analytic thinking and susceptibility to misinformation during the covid-19 outbreak. *Computers in Human Behavior*, 133:107295, 2022.

[157] Chloe Lim : Checking how fact-checkers check. *Research & Politics*, 5(3):2053168018786848, 2018.

[158] Yang Liu et Yi-Fang Wu : Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. *In Proceedings of the AAAI Conference on Artificial Intelligence*, pages 354–361, 2018.

[159] Jiasen Lu, Dhruv Batra, Devi Parikh et Stefan Lee : Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[160] Yi-Ju Lu et Cheng-Te Li : Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*, 2020.

[161] Mufan Luo, Jeffrey T Hancock et David M Markowitz : Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research*, 49(2):171–195, 2022.

[162] Lauren Lutzke, Caitlin Drummond, Paul Slovic et Joseph Árvai : Priming critical thinking: Simple interventions limit the influence of fake news about climate change on

Facebook. *Global environmental change*, 58:101964, 2019.

[163] Rakoen Maertens, Frederik Anseel et Sander van der Linden : Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology*, 70:101455, 2020.

[164] Atik Mahabub : A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers. *SN Applied Sciences*, 2(4):1–9, 2020.

[165] Syed Mahbub, Eric Pardede, ASM Kayes et Wenny Rahayu : Controlling astroturfing on the internet: A survey on detection techniques and research challenges. *International Journal of Web and Grid Services*, 15(2):139–158, 2019.

[166] Zaki Malik et Athman Bouguettaya : Rater credibility assessment in web services interactions. *World Wide Web*, 12(1):3–25, 2009.

[167] Deepak Mangal et Dilip Kumar Sharma : Fake news detection with integration of embedded text cues and image features. *In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 68–72. IEEE, 2020.

[168] Chris Marsden, Trisha Meyer et Ian Brown : Platform values and democratic elections: How can the law regulate digital disinformation? *Computer Law & Security Review*, 36:105373, 2020.

[169] Elio Masciari, Vincenzo Moscato, Antonio Picariello et Giancarlo Sperlí : Detecting fake news by image analysis. *In Proceedings of the 24th Symposium on International Database Engineering & Applications*, pages 1–5, 2020.

[170] Valeria Mazzeo et Andrea Rapisarda : Investigating fake and reliable news sources using complex networks analysis. *Frontiers in Physics*, 10:886544, 2022.

[171] Valeria Mazzeo, Andrea Rapisarda et Giovanni Giuffrida : Detection of fake news on covid-19 on web search engines. *Frontiers in physics*, 9:685730, 2021.

[172] Sarah McGrew : Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, 145:103711, 2020.

[173] Sarah McGrew, Joel Breakstone, Teresa Ortega, Mark Smith et Sam Wineburg : Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory & Research in Social Education*, 46(2):165–193, 2018.

[174] Priyanka Meel et Dinesh Kumar Vishwakarma : Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986, 2020.

[175] Priyanka Meel et Dinesh Kumar Vishwakarma : Han, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences*, 567:23–41, 2021.

[176] James Meese, Jordan Frith et Rowan Wilken : Covid-19, 5G conspiracies and infrastructural futures. *Media International Australia*, 177(1):30–46, 2020.

[177] Miriam J Metzger, Ethan H Hartsell et Andrew J Flanagin : Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research*, 47(1):3–28, 2020.

[178] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad et Nasir Memon : The role of the crowd in countering misinformation: A case study of the Covid-19 infodemic. *arXiv preprint arXiv:2011.05773*, 2020.

[179] Dimitrios Michail, Nikos Kanakaris et Iraklis Varlamis : Detection of fake news campaigns using graph convolutional networks. *International Journal of Information Management Data Insights*, 2(2):100104, 2022.

[180] Paul Mihailidis et Samantha Viotty : Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American Behavioral Scientist*, 61(4):441–454, 2017.

[181] Mahdi Mirhoseini, Spencer Early, Nour El Shamy et Khaled Hassanein : Actively open-minded thinking is key to combating fake news: A multimethod study. *Information & Management*, 60(3):103761, 2023.

[182] Mahdi Mirhoseini, Spencer Early et Khaled Hassanein : All eyes on misinformation and social media consumption: A pupil dilation study. *In NeuroIS Retreat*, pages 73–80. Springer, 2022.

[183] Mahdieh Mirzabeigi, Mahsa Torabi et Tahereh Jowkar : The role of personality traits and the ability to detect fake news in predicting information avoidance during the covid-19 pandemic. *Library Hi Tech*, 2023.

[184] Rahul Mishra : Fake news detection using higher-order user to user mutual-attention progression in propagation paths. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 652–653, 2020.

[185] Shubha Mishra, Piyush Shukla et Ratish Agarwal : Analyzing machine learning enabled fake news detection techniques for diversified datasets. *Wireless Communications and Mobile Computing*, 2022, 2022.

[186] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji et Eric Ragan : Machine learning explanations to prevent overtrust in fake news detection. *arXiv preprint arXiv:2007.12358*, 2020.

[187] Maria D Molina, S Shyam Sundar, Thai Le et Dongwon Lee : "fake news" is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist*, 65(2):180–212, 2021.

[188] Christian Moro et James R Birt : Review bombing is a dirty practice, but research shows games do benefit from online feedback. *The Conversation*, 2022.

[189] Eni Mustafaraj et Panagiotis Takis Metaxas : The fake news spreading plague: Was it preventable? *In Proceedings of the 2017 ACM on Web Science Conference*, pages 235–239, 2017.

[190] Tyler WS Nagel : Measuring fake news acumen using a news media literacy instrument. *Journal of Media Literacy Education*, 14(1):29–42, 2022.

[191] Preslav Nakov : Can we spot the "fake news" before it was even written? *arXiv preprint arXiv:2008.04374*, 2020.

[192] Xiaoli Nan, Yuan Wang et Kathryn Thier : Why people believe health misinformation and who are at risk? a systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine*, page 115398, 2022.

[193] Elmie Nekmat : Nudge effect of fact-check alerts: Source influence and media skepticism on sharing of news misinformation in social media. *Social Media+ Society*, 6(1):2056305119897322, 2020.

[194] Thomas Nygren, Fredrik Brounéus et Göran Svensson : Diversity and credibility in young people's news feeds: A foundation for teaching and learning citizenship in a digital era. *Journal of Social Science Education*, 18(2):87–109, 2019.

[195] Brendan Nyhan, Ethan Porter, Jason Reifler et Thomas J Wood : Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, 42(3):939–960, 2020.

[196] Brendan Nyhan et Jason Reifler : Displacing misinformation about events: An experimental test of causal corrections. *Journal of Experimental Political Science*, 2(1):81–93, 2015.

[197] Ning Xin Nyow et Hui Na Chua : Detecting fake news with tweets' properties. *In 2019 IEEE Conference on Application, Information and Network Security (AINS)*, pages 24–29. IEEE, 2019.

[198] Iago Sestrem Ochoa, Gabriel de Mello, Luis A Silva, Abel JP Gomes, Anita MR Fernandes et Valderi Reis Quietinho Leithardt : Fakechain: A blockchain architecture to ensure trust in social media networks. *In International Conference on the Quality of Information and Communications Technology*, pages 105–118. Springer, 2019.

[199] Ali Orhan : Fake news detection on social media: the predictive role of university students' critical thinking dispositions and new media literacy. *Smart Learning Environments*, 10(1):1–14, 2023.

[200] Feyza Altunbey Ozbay et Bilal Alatas : Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications*, 540:123174, 2020.

[201] Pinar Ozturk, Huaye Li et Yasuaki Sakamoto : Combating rumor spread on social media: The effectiveness of refutation and warning. *In 2015 48th Hawaii International Conference on System Sciences*, pages 2406–2414. IEEE, 2015.

[202] Shivam B Parikh et Pradeep K Atrey : Media-rich fake news detection: A survey. *In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 436–441. IEEE, 2018.

[203] Kevin Parrish : Deep learning & machine learning: What's the difference? Online: `https://parsers.me/deep-learning-machine-learning-whats-the-difference/`, 2018. Accessed: 20-05-2020.

[204] Jeannette Paschen : Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management*, 29(2):223–233, 2019.

[205] Archita Pathak et Rohini K Srihari : Breaking! Presenting fake news corpus for automated fact checking. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 357–362, 2019.

[206] Jian Peng, Sam Detchon, Kim-Kwang Raymond Choo et Helen Ashman : Astroturfing detection in social media: A binary n-gram–based approach. *Concurrency and Computation: Practice and Experience*, 29(17):e4013, 2017.

[207] Gordon Pennycook, Adam Bear, Evan T Collins et David G Rand : The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11):4944–4957, 2020.

[208] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu et David G Rand : Fighting Covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780, 2020.

[209] Gordon Pennycook et David G Rand : Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.

[210] Gordon Pennycook et David G Rand : Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 88(2):185–200, 2020.

[211] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates et Gerhard Weikum : Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*, 2018.

[212] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff et Benno Stein : A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[213] Marialaura Previti, Victor Rodriguez-Fernandez, David Camacho, Vincenza Carchiolo et Michele Malgeri : Fake news detection using time series and user features classification. *In International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pages 339–353. Springer, 2020.

[214] Piotr Przybyla : Capturing the style of fake news. *In Proceedings of the AAAI Conference on Artificial Intelligence*, pages 490–497, 2020.

[215] Piotr Przybyła et Axel J Soto : When classification accuracy is not enough: Explaining news credibility assessment. *Information Processing & Management*, 58(5): 102653, 2021.

[216] Adnan Qayyum, Junaid Qadir, Muhammad Umar Janjua et Falak Sher : Using blockchain to rein in the new post-truth world and check the spread of fake news. *IT Professional*, 21(4):16–24, 2019.

[217] Feng Qian, Chengyue Gong, Karishma Sharma et Yan Liu : Neural user response generator: Fake news detection with collective user intelligence. *In IJCAI*, volume 18, pages 3834–3840, 2018.

[218] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang et Changsheng Xu : Hierarchical multi-modal contextual attention network for fake news detection. *In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 153–162, 2021.

[219] Shaina Raza et Chen Ding : Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362, 2022.

[220] Julio CS Reis, André Correia, Fabricio Murai, Adriano Veloso et Fabrício Benevenuto : Explainable machine learning for fake news detection. *In Proceedings of the 10th ACM conference on web science*, pages 17–26, 2019.

[221] Manoel Horta Ribeiro, Savvas Zannettou, Oana Goga, Fabrício Benevenuto et Robert West : Can online attention signals help fact-checkers to fact-check. *In 16th International Conference on Web and Social Media*, 2022.

[222] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin : " why should i trust you?" explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[223] Julie Ricard et Juliano Medeiros : Using misinformation as a political weapon: Covid-19 and Bolsonaro in Brazil. *Harvard Kennedy School Misinformation Review*, 1(3), 2020.

[224] Esteban Andres Rissola, Seyed Ali Bahrainian et Fabio Crestani : Personality recognition in conversations using capsule neural networks. *In IEEE/WIC/ACM International Conference on Web Intelligence*, pages 180–187, 2019.

[225] Jon Roozenbeek, Rakoen Maertens, Stefan M Herzog, Michael Geers, Ralf Kurvers, Mubashir Sultan et Sander van der Linden : Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, 17(3):547–573, 2022.

[226] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles et Sander Van Der Linden : Susceptibility to misinformation about Covid-19 around the world. *Royal Society Open Science*, 7(10):201199, 2020.

[227] Jon Roozenbeek et Sander van der Linden : Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1):1–10, 2019.

[228] Jon Roozenbeek, Sander van der Linden et Thomas Nygren : Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, 2020.

[229] Victoria L Rubin, Niall Conroy, Yimin Chen et Sarah Cornwell : Fake news or truth? Using satirical cues to detect potentially misleading news. *In Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, 2016.

[230] Natali Ruchansky, Sungyong Seo et Yan Liu : Csi: A hybrid deep model for fake news detection. *In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

[231] Joan Saltor, Itxaso Barberia et Javier Rodríguez-Ferreiro : Thinking disposition, thinking style, and susceptibility to causal illusion predict fake news discriminability. *Applied Cognitive Psychology*, 37(2):360–368, 2023.

[232] Carola Salvi, Nathaniel Barr, Joseph E Dunsmoor et Jordan Grafman : Insight problem solving ability predicts reduced susceptibility to fake news, bullshit, and overclaiming. *Thinking & Reasoning*, pages 1–25, 2022.

[233] Brinda Sampat et Sahil Raj : Fake or real news? understanding the gratifications and personality traits of individuals sharing fake news on social media platforms. *Aslib Journal of Information Management*, 74(5):840–876, 2022.

[234] Atiquer Rahman Sarkar et Shamim Ahmad : A new approach to expert reviewer detection and product rating derivation from online experiential product reviews. *Heliyon*, 7(7):e07409, 2021.

[235] Andrew J Schuyler : Regulating facts: A procedural framework for identifying, excluding, and deterring the intentional or knowing proliferation of fake news online. *U. Ill. JL Tech. & Pol'y*, 2019(1):211–240, 2019.

[236] Zonyin Shae et Jeffrey Tsai : AI blockchain platform for trusting news. *In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1610–1619. IEEE, 2019.

[237] Priyanshi Shah et Ziad Kobti : Multimodal fake news detection using a cultural algorithm with situational and normative knowledge. *In 2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–7. IEEE, 2020.

[238] Wenqian Shang, Mengyu Liu, Weiguo Lin et Minzheng Jia : Tracing the source of news based on blockchain. *In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 377–381. IEEE, 2018.

[239] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini et Filippo Menczer : Hoaxy: A platform for tracking online misinformation. *In Proceedings of the 25th International Conference Companion on World Wide Web*, pages 745–750, 2016.

[240] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini et Filippo Menczer : The spread of low-credibility content by social bots. *Nature Communications*, 9(1):1–9, 2018.

[241] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer et Giovanni Luca Ciampaglia : Anatomy of an online misinformation network. *PloS One*, 13(4):e0196087, 2018.

[242] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang et Yan Liu : Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[243] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla et Yan Liu : Covid-19 on social media: Analyzing misinformation in Twitter conversations. *arXiv preprint arXiv:2003.12309*, 2020.

[244] Cuihua Shen, Mona Kasra, Wenjing Pan, Grace A Bassett, Yining Malloch et James F O'Brien : Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media & Society*, 21(2):438–463, 2019.

[245] Mark P Shephard, David J Robertson, Narisong Huhe et Anthony Anderson : Everyday non-partisan fake news: Sharing behavior, platform specificity, and detection. *Frontiers in Psychology*, 14:1118407, 2023.

[246] Imani N Sherman, Elissa M Redmiles et Jack W Stokes : Designing indicators to combat fake media. *arXiv preprint arXiv:2010.00544*, 2020.

[247] Peining Shi, Zhiyong Zhang et Kim-Kwang Raymond Choo : Detecting malicious social bots based on clickstream sequences. *IEEE Access*, 7:28855–28862, 2019.

[248] Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H Nazer, Kaize Ding, Mansooreh Karami et Huan Liu : Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1385, 2020.

[249] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee et Huan Liu : defend: Explainable fake news detection. *In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[250] Kai Shu, Deepak Mahudeswaran et Huan Liu : Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, 25:60–71, 2019.

[251] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee et Huan Liu : Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[252] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee et Huan Liu : Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

[253] Kai Shu, Deepak Mahudeswaran, Suhang Wang et Huan Liu : Hierarchical propagation networks for fake news detection: Investigation and exploitation. *In Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637. AAAI press, 2020.

[254] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang et Huan Liu : Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[255] Kai Shu, Suhang Wang, Dongwon Lee et Huan Liu : Mining disinformation and fake news: Concepts, methods, and recent advancements. *In Disinformation, Misinformation, and Fake News in Social Media*, pages 1–19. Springer, 2020.

[256] Kai Shu, Suhang Wang et Huan Liu : Understanding user profiles on social media for fake news detection. *In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435. IEEE, 2018.

[257] Kai Shu, Suhang Wang et Huan Liu : Beyond news contents: The role of social context for fake news detection. *In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 312–320, 2019.

[258] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston et Huan Liu : Early detection of fake news with multi-source weak social supervision. *In ECML/PKDD (3)*, pages 650–666, 2020.

[259] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani et Huan Liu : The role of user profiles for fake news detection. *In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439, 2019.

[260] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera et Christopher Leckie : Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5):102618, 2021.

[261] Aditya Simha et K Praveen Parboteeah : The big 5 personality traits and willingness to justify unethical behavior—a cross-national examination. *Journal of Business Ethics*, 167:451–471, 2020.

[262] Vivek K Singh, Isha Ghosh et Darshan Sonagara : Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, 72(1):3–17, 2021.

[263] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru et Shin'ichi Satoh : Spotfake: A multi-modal framework for fake news detection. *In 2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE, 2019.

[264] Stavros Sintos, Pankaj K. Agarwal et Jun Yang : Selecting data to clean for fact checking: Minimizing uncertainty vs. maximizing surprise. *Proceedings of the VLDB Endowment*, 12(13):2408–2421, 2019.

[265] Jackie Snow : Can AI win the war against fake news? *MIT Technology Review*, December 2017. Online: `https://www.technologyreview.com/s/609717/can-ai-win-the-war-against-fake-news/`. Accessed: 3-10-2020.

[266] Gyuwon Song, Suhyun Kim, Haejin Hwang et Kwanhoon Lee : Blockchain-based notarization for social media. *In 2019 IEEE International Conference on Consumer Clectronics (ICCE)*, pages 1–2. IEEE, 2019.

[267] Kate Starbird, Ahmer Arif et Tom Wilson : Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[268] David Sterret, Dan Malato, Jennifer Benz, Liz Kantor, Trevor Tompson, Tom Rosenstiel, Jeff Sonderman, Kevin Loker et Emily Swanson : Who shared it? How Americans decide what news to trust on social media. Rapport technique, Norc Working Paper Series, WP-2018-001, 1–24, 2018.

[269] Robbie M Sutton et Karen M Douglas : Conspiracy theories and the conspiracy mindset: Implications for political ideology. *Current Opinion in Behavioral Sciences*, 34:118–122, 2020.

[270] Gopal S Tandel, Ashish Tiwari et OG Kakde : Performance optimisation of deep learning models using majority voting algorithm for brain tumour classification. *Computers in Biology and Medicine*, 135:104564, 2021.

[271] Edson C Tandoc Jr, Ryan J Thomas et Lauren Bishop : What is (fake) news? Analyzing news values (and more) in fake stories. *Media and Communication*, 9(1):110–119, 2021.

[272] Yahya Tashtoush, Balqis Alrababah, Omar Darwish, Majdi Maabreh et Nasser Alsaedi : A deep learning framework for detection of covid-19 fake news on social media platforms. *Data*, 7(5):65, 2022.

[273] Franklin Tchakounté, Ahmadou Faissal, Marcellin Atemkeng et Achille Ntyam : A reliable weighting scheme for the aggregation of crowd intelligence to detect fake news. *Information*, 11(6):319, 2020.

[274] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze et Konstantin Todorov : Claimskg: A knowledge graph of fact-checked claims. *In International Semantic Web Conference*, pages 309–324. Springer, 2019.

[275] Moshe Tennenholtz : Reputation systems: An axiomatic approach. *arXiv preprint arXiv:1207.4163*, 2012.

[276] James Thorne et Andreas Vlachos : Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*, 2018.

[277] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos et Arpit Mittal : The fact extraction and verification (fever) shared task. *arXiv preprint arXiv:1811.10971*, 2018.

[278] Kathie M d'I Treen, Hywel TP Williams et Saffron J O'Neill : Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):e665, 2020.

[279] Stephanie Jean Tsang : Motivated fake news perception: The impact of news sources and policy support on audiences' assessment of news fakeness. *Journalism & Mass Communication Quarterly*, page 1077699020952129, 2020.

[280] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant et Andreas Krause : Fake news detection in social networks via crowd signals. *In Companion Proceedings of the The Web Conference 2018*, pages 517–524, 2018.

[281] Marina Tulin, Bram Lancee et Beate Volker : Personality and social capital. *Social psychology quarterly*, 81(4):295–318, 2018.

[282] Santosh Kumar Uppada, K Manasa, B Vidhathri, R Harini et B Sivaselvan : Novel approaches to fake news and fake account detection in osns: user social engagement and visual content centric model. *Social Network Analysis and Mining*, 12(1):1–19, 2022.

[283] Sebastián Valenzuela, Carlos Muñiz et Marcelo Santos : Social media and belief in misinformation in mexico: A case of maximal panic, minimal effects? *The International Journal of Press/Politics*, page 19401612221088988, 2022.

[284] Sander Van Der Linden : Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28(3):460–467, 2022.

[285] Sander van der Linden, Costas Panagopoulos et Jon Roozenbeek : You are fake news: Political bias in perceptions of fake news. *Media, Culture & Society*, 42(3):460–470, 2020.

[286] Sander Van der Linden et Jon Roozenbeek : Psychological inoculation against fake news. *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation.*, 2020.

[287] Norman Vasu, Benjamin Ang, Terri-Anne Teo, Shashi Jayakumar, Muhammad Raizal et Juhi Ahuja : *Fake news: National security in the post-truth era.* RSIS, 2018.

[288] Alina Vereshchaka, Seth Cosimini et Wen Dong : Analyzing and distinguishing fake and real news to mitigate the problem of disinformation. *Computational and Mathematical Organization Theory*, pages 1–15, 2020.

[289] Mark Verstraete, Derek E Bambauer et Jane R Bambauer : Identifying and countering fake news. *Arizona Legal Studies Discussion Paper*, 73(17-15), 2017.

[290] JBJ Vilmer, Alexandre Escorcia, Marine Guillaume et Janaina Herrera : *Information manipulation: A challenge for our democracies.* In Report by the Policy Planning Staff (CAPS) of the Ministry for Europe and Foreign Affairs, and the Institute for Strategic Research (RSEM) of the Ministry for the Armed Forces., 2018.

[291] Dinesh Kumar Vishwakarma, Deepika Varshney et Ashima Yadav : Detection and veracity analysis of fake news via scrapping and authenticating the web search. *Cognitive Systems Research*, 58:217–229, 2019.

[292] Andreas Vlachos et Sebastian Riedel : Fact checking: Task definition and dataset construction. *In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, 2014.

[293] Christian von der Weth, Ashraf Abdul, Shaojing Fan et Mohan Kankanhalli : Helping users tackle algorithmic threats on social media: A multimedia research agenda. *In Proceedings of the 28th ACM International Conference on Multimedia*, pages 4425–4434, 2020.

[294] Soroush Vosoughi, Deb Roy et Sinan Aral : The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[295] Emily K Vraga et Leticia Bode : Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5):621–645, 2017.

[296] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan et Hannaneh Hajishirzi : Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.

[297] Ari Ezra Waldman : The marketplace of fake news. *U. Pa. J. Const. L.*, 20:845, 2017.

[298] Nathan Walter, Jonathan Cohen, R Lance Holbert et Yasmin Morag : Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3):350–375, 2020.

[299] Liqiang Wang, Yafang Wang, Gerard de Melo et Gerhard Weikum : Understanding archetypes of fake news via fine-grained classification. *Social Network Analysis and Mining*, 9(1):1–17, 2019.

[300] William Yang Wang : "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[301] Yangqian Wang, Hao Han, Ye Ding, Xuan Wang et Qing Liao : Learning contextual features with multi-head self-attention for fake news detection. *In International Conference on Cognitive Computing*, pages 132–142. Springer, 2019.

[302] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su et Jing Gao : Eann: Event adversarial neural networks for multi-modal fake news detection. *In Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[303] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng et Jing Gao : Weak supervision for fake news detection via reinforcement learning. *In Proceedings of the AAAI Conference on Artificial Intelligence*, pages 516–523, 2020.

[304] Yuxi Wang, Martin McKee, Aleksandra Torbica et David Stuckler : Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, 240:112552, 2019.

[305] Claire Wardle : Fake news. It's complicated. Online: `https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79`, 2017. Accessed: 3-10-2020.

[306] Claire Wardle : The need for smarter definitions and practical, timely empirical research on information disorder. *Digital Journalism*, 6(8):951–963, 2018.

[307] Claire Wardle et Hossein Derakhshan : Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, 27:1–107, 2017.

[308] Andrew P Weiss, Ahmed Alwan, Eric P Garcia et Julieta Garcia : Surveying fake news: Assessing university faculty's fragmented definition of fake news and its impact on teaching critical thinking. *International Journal for Educational Integrity*, 16(1):1–30, 2020.

[309] Liang Wu et Huan Liu : Tracing fake-news footprints: Characterizing social media messages by how they propagate. *In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 637–645, 2018.

[310] Liang Wu, Fred Morstatter, Kathleen M Carley et Huan Liu : Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90, 2019.

[311] Lianwei Wu et Yuan Rao : Adaptive interaction fusion networks for fake news detection. *arXiv preprint arXiv:2004.10009*, 2020.

[312] Yongwei Wu, Likun Liu, Jiayin Mao, Guangwen Yang et Weimin Zheng : An analytical model for performance evaluation in a computational grid. *In Proceedings of the 2007 Asian Technology Information Program's (ATIP's) 3rd Workshop on High Performance Computing in China: Solution Approaches to Impediments for High Performance Computing*, CHINA HPC '07, page 145–151, New York, NY, USA, 2007. Association for Computing Machinery.

[313] Yuanyuan Wu, Eric WT Ngai, Pengkun Wu et Chong Wu : Fake news on the internet: a literature review, synthesis and directions for future research. *Internet Research*, 2022.

[314] Kuai Xu, Feng Wang, Haiyan Wang et Bo Yang : Detecting fake news over online social media via domain reputations and content understanding. *Tsinghua Science and Technology*, 25(1):20–27, 2019.

[315] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi et Lin Wei : Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610, 2021.

[316] Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji et Xia Hu : Xfake: explainable fake news detector with visualizations. *In The World Wide Web Conference*, pages 3600–3604, 2019.

[317] Xin Yang, Yuezun Li et Siwei Lyu : Exposing deep fakes using inconsistent head poses. *In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[318] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon et Sameer Patil : Effects of credibility indicators on social media news sharing intent. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[319] Arefeh Yavary, Hedieh Sajedi et Mohammad Saniee Abadeh : Information verification in social networks based on user feedback and news agencies. *Social Network Analysis and Mining*, 10(1):1–8, 2020.

[320] Kasra Majbouri Yazdi, Adel Majbouri Yazdi, Saeid Khodayi, Jingyu Hou, Wanlei Zhou et Saeed Saedy : Improving fake news detection using k-means and support vector machine approaches. *International Journal of Electronics and Communication Engineering*, 14(2):38–42, 2020.

[321] Hua Yuan, Jie Zheng, Qiongwei Ye, Yu Qian et Yan Zhang : Improving fake news detection with domain-adversarial and graph-attention neural network. *Decision Support Systems*, page 113633, 2021.

[322] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn et Nicolas Kourtellis : The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–37, 2019.

[323] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner et Yejin Choi : Defending against neural fake news. *arXiv preprint arXiv:1905.12616*, 2019.

[324] Jiangfeng Zeng, Yin Zhang et Xiao Ma : Fake news detection for epidemic emergencies via deep correlations between text and images. *Sustainable Cities and Society*, 66:102652,

2021.

[325] Jiawei Zhang, Bowen Dong et S Yu Philip : Fakedetector: Effective fake news detection with deep diffusive neural network. *In 2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1826–1829. IEEE, 2020.

[326] Qiang Zhang, Aldo Lipani, Shangsong Liang et Emine Yilmaz : Reply-aided detection of misinformation via Bayesian deep learning. *In The World Wide Web Conference*, pages 2333–2343, 2019.

[327] Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao et Lizhen Cui : Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. *In 2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

[328] Xichen Zhang et Ali A Ghorbani : An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020.

[329] Xu Zhang, Svebor Karaman et Shih-Fu Chang : Detecting and simulating artifacts in GAN fake images. *In 2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.

[330] Xinyi Zhou, Atishay Jain, Vir V Phoha et Reza Zafarani : Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25, 2020.

[331] Xinyi Zhou, Jindi Wu et Reza Zafarani : Safe: Similarity-aware multi-modal fake news detection. *Advances in Knowledge Discovery and Data Mining*, 12085:354, 2020.

[332] Xinyi Zhou et Reza Zafarani : A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[333] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata et Rob Procter : Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.