

Université de Montréal

**Low-resource Suicide Ideation and Depression
Detection with Multitask Learning and Large
Language Models**

par

Pierre-William Breau

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Informatique

31 août 2023

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

**Low-resource Suicide Ideation and Depression Detection
with Multitask Learning and Large Language Models**

présenté par

Pierre-William Breau

a été évalué par un jury composé des personnes suivantes :

Philippe Langlais

(président-rapporteur)

Bang Liu

(directeur de recherche)

Jian-Yun Nie

(codirecteur)

Claude Frasson

(membre du jury)

Résumé

Nous évaluons des méthodes de traitement automatique du langage naturel (TALN) pour la détection d'idées suicidaires, de la dépression et de l'anxiété à partir de publications sur les médias sociaux. Comme les ensembles de données relatifs à la santé mentale sont rares et généralement de petite taille, les méthodes classiques d'apprentissage automatique ont traditionnellement été utilisées dans ce domaine. Nous évaluons l'effet de l'apprentissage multi-tâche sur la détection d'idées suicidaires en utilisant comme tâches auxiliaires des ensembles de données disponibles publiquement pour la détection de la dépression et de l'anxiété, ainsi que la classification d'émotions et du stress. Nous constatons une hausse de la performance de classification pour les tâches de détection d'idées suicidaires, de la dépression et de l'anxiété lorsqu'elles sont entraînées ensemble en raison de similitudes entre les troubles de santé mentale à l'étude. Nous observons que l'utilisation d'ensembles de données publiquement accessibles pour des tâches connexes peut bénéficier à la détection de problèmes de santé mentale. Nous évaluons enfin la performance des modèles ChatGPT et GPT-4 dans des scénarios d'apprentissage *zero-shot* et *few-shot*. GPT-4 surpasse toutes les autres méthodes testées pour la détection d'idées suicidaires. De plus, nous observons que ChatGPT bénéficie davantage de l'apprentissage *few-shot*, car le modèle fournit un haut taux de réponses non concluantes si aucun exemple n'est présenté. Enfin, une analyse des faux négatifs produits par GPT-4 pour la détection d'idées suicidaires conclut qu'ils sont dus à des erreurs d'étiquetage plutôt qu'à des lacunes du modèle.

Mots-clés: modèles de langage, idées suicidaires, classification de textes, apprentissage multitâche.

Abstract

In this work we explore natural language processing (NLP) methods to suicide ideation, depression, and anxiety detection in social media posts. Since annotated mental health data is scarce and difficult to come by, classical machine learning methods have traditionally been employed on this type of task due to the small size of the datasets. We evaluate the effect of multi-task learning on suicide ideation detection using publicly-available datasets for depression, anxiety, emotion and stress classification as auxiliary tasks. We find that classification performance of suicide ideation, depression, and anxiety is improved when trained together because of the proximity between the mental disorders. We observe that publicly-available datasets for closely-related tasks can benefit the detection of certain mental health conditions. We then perform classification experiments using ChatGPT and GPT-4 using zero-shot and few-shot learning, and find that GPT-4 obtains the best performance of all methods tested for suicide ideation detection. We further observe that ChatGPT benefits the most from few-shot learning as it struggles to give conclusive answers when no examples are provided. Finally, an analysis of false negative results for suicide ideation output by GPT-4 concludes that they are due to labeling errors rather than mistakes from the model.

Keywords: language models, suicide ideation, text classification, multitask learning.

Contents

Résumé	5
Abstract	7
List of tables	13
List of figures	15
List of acronyms and abbreviations	17
Acknowledgements	19
Chapter 1. Introduction	21
Chapter 2. Natural language processing for mental health applications ...	23
2.1. Types of data.....	23
2.2. Acquiring and labeling social media data.....	23
2.3. Tasks and methods	25
2.4. Suicide ideation	25
2.4.1. Definition.....	25
2.4.2. Related work	26
2.4.3. Preliminary work on the suicide ideation dataset	27
2.5. Ethical Concerns.....	27
2.6. Datasets used in this work	27
2.6.1. Suicide Ideation.....	27
2.6.2. Publicly available mental health related datasets	28
2.6.3. DATD	28
2.6.4. SAD	29
2.6.5. CrowdFlower	29
Chapter 3. Preliminary Analysis	31

3.1. Analysis of Annotations.....	31
3.1.1. Common Ambiguities.....	31
3.1.2. Analysis Labels.....	32
3.1.3. Results of Manual Evaluation.....	32
3.1.4. Duplicates.....	33
3.1.5. Updated dataset.....	34
3.2. Experiments with classical methods.....	34
3.2.1. Random Forest.....	34
3.2.2. SVM.....	35
3.2.3. Preprocessing.....	35
3.2.4. Features.....	35
3.2.5. Metrics.....	36
3.2.6. Results.....	36
3.2.7. Discussion.....	37
Chapter 4. Multi-Task Learning.....	39
4.1. Objectives.....	39
4.2. Language Models.....	39
4.2.1. BERT.....	40
4.3. Hard parameter sharing.....	40
4.4. Related work.....	40
4.5. Architecture.....	41
4.6. Tests.....	42
4.7. Results.....	42
4.7.1. Overall performance.....	43
4.7.2. Suicide ideation classification.....	44
4.8. Discussion.....	45
Chapter 5. ChatGPT and GPT-4.....	47
5.1. Related work.....	47
5.2. Overview of the models.....	48
5.3. Prompt design.....	48

5.4. Tests	50
5.5. Results	51
5.5.1. Classification results	51
5.5.2. Ablation study	52
5.5.3. Confusion matrix	53
5.5.4. False negatives	54
5.5.5. Non-answers	54
5.6. Discussion	55
Chapter 6. Conclusion	57
Chapter 7. Future Work	59
References	61
Appendix A. Prompts	67
A.1. Few-shot examples for the DATD task	67

List of tables

2.1	Suicide ideation dataset.....	28
3.1	Analysis of tweets labeled as "Suspected suicide ideation"	33
3.2	Analysis of tweets labeled as "Related to suicide".....	33
3.3	Analysis of tweets labeled as "Suicide other".....	33
3.4	Duplicates confusion matrix.....	34
3.5	Baseline classification results.....	36
4.1	Classification results using BERT as the base model.....	43
4.2	Classification results using MentalBERT as the base model.....	44
4.3	Per-class suicide ideation classification results using MentalBERT	44
4.4	Confusion matrix for suicide ideation	45
5.1	Classification results for ChatGPT and GPT-4.....	51
5.2	Per-class suicide ideation classification results for ChatGPT and GPT-4.....	53
5.3	GPT-4 four-shot confusion matrix for suicide ideation	53
5.4	Suicide ideation false negatives from the GPT-4 four-shot experiment	54

List of figures

4.1	Hard parameter sharing [10]	41
5.1	Example of context for ChatGPT prompt	51
5.2	Ablation study of few-shot learning on the suicide ideation dataset	52

List of acronyms and abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
CNN	Convolutional Neural Network
DATD	Depression and Anxiety in Twitter Dataset
GPT	Generative Pre-trained Transformer
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Model
LM	Language Model
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
MTL	Multi-Task Learning

NB	Naive Bayes
NLP	Natural Language Processing
RF	Random Forest
RNN	Recurrent Neural Network
SAD	Stress Annotated Dataset
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency

Acknowledgements

I would like to thank Professors Bang Liu and Jian-Yun Nie for advising me throughout this project.

Special thanks to Kyle, Suyuchen, Sifan, and everyone in the lab for their useful insights and collaboration.

I have benefited from the help and support of a great number of people without whom I could not have seen this project through. Thank you.

Chapter 1

Introduction

Every day, around 200 people in Canada make an attempt on their life¹. The topic of mental health has gained a lot of public attention in recent years, and especially since the pandemic when the global instability, confinement and public safety measures highlighted or exacerbated people's mental health conditions. We are witnessing a societal effort to destigmatize mental health, with causes such as Bell Let's Talk encouraging people to talk about their struggles and reach out for help. Coincidentally, large social media platforms like Twitter, Reddit and Facebook are making it easy for people to share their daily lives, interests and concerns with a broad audience. In this context, we are seeing a significant number of people sharing intimate details about their mental health on internet platforms. This information has proven to be of interest for researchers and public health agencies, since by monitoring social media posts, we could now get a sense of a population's general mental health by tracking what people willingly post about their own conditions. It could also be possible to provide help to individuals who show signs of distress, for example by expressing suicide ideation. But to do this at scale, such detection needs some sort of automation, hence the need for machine learning, more specifically Natural Language Processing (NLP) which focuses on textual data. Even though social media data is relatively easy to come by, it can prove tedious and costly to have the posts expertly labeled. For this reason, accurately labeled social media posts relating to mental health have historically been scarce, forcing researchers to resort to keyword searches, hashtags or regex queries to automate the annotation process [26]. Moreover, some mental health conditions can be hard to define objectively, making the definitions of the classification tasks themselves difficult—for example, it can be hard to come up with an objective definition for what constitutes suicide ideation within a single tweet, and researchers have constructed datasets with the number of labels ranging from 2 to 7 for this single task [1]. This inherent ambiguity can also lead to some inconsistencies in the data labeling.

¹<https://www.canada.ca/en/public-health/services/suicide-prevention/suicide-canada.html>

In this research, we focus on the task of detecting suicide ideation, and apply the same methods on anxiety and depression detection. The main research objectives are:

- Evaluate the quality of annotations for the suicide ideation dataset that we are working with.
- Obtain a baseline classification performance using a non-deep learning method.
- Evaluate using multi-task learning to palliate to the small suicide ideation dataset size.
- Examine the relationship between various mental health related tasks.
- Evaluate and compare the performance of modern GPT models on the two main tasks, i.e. suicide ideation, and depression and anxiety detection.

Thesis structure

- **Chapter 2** presents the various challenges relating to NLP for mental health applications including data acquisition and annotation. We make a brief overview of the most commonly-used methods then describe the suicide ideation detection task, and provide details about the datasets used in this work.
- In **Chapter 3**, we perform a thorough analysis of the suicide ideation dataset’s annotations. We then obtain baseline results for the suicide ideation and DATD (depression and anxiety) datasets using the Random Forest and SVM algorithms.
- **Chapter 4** details the multi-task experiments performed on different combinations of the datasets described in **Chapter 2**. We compare BERT and MentalBERT as base models.
- In **Chapter 5**, we test the ChatGPT and GPT-4 LLMs on the suicide ideation and DATD datasets using different prompting strategies. We perform an ablation study for the suicide ideation detection task by comparing few-shot to zero-shot learning using both models.
- In **Chapters 6 & 7**, we conclude then present potential avenues for future research in NLP for mental health and suicide ideation detection.

Chapter 2

Natural language processing for mental health applications

We present a brief overview of the challenges encountered when applying NLP to classify mental-health related social media data, including acquisition and labeling. We then present some of the most commonly used methods, specificities of the suicide ideation task, and ethical considerations to be mindful of when doing research in this field.

2.1. Types of data

The three predominant data sources for NLP research on mental health are [54]:

- Social media data (81%)

Usually obtained from Twitter or mental-health related subreddits, social media datasets are the most commonly used in this type of research.

- Interviews (7%)

Clinical interviews conducted to detect mental disorders in participants. The transcribed text can then be analyzed using NLP.

- Electronic health records (6%)

Data collected from patient files including clinical notes, diagnosis history, medication, etc.

2.2. Acquiring and labeling social media data

For the purposes of this research project, we focus on social media data as it is the most easily accessible. Labeling this data can prove to be a challenge, and researchers have found various methods to do so, with varying costs, scalability, and reliability. We mention some of these methods in this section.

- Expert annotations

The most valuable datasets are the ones that have been annotated by experts in the

field of mental health. Given the high cost of employing healthcare professionals, these datasets are usually very small. Data augmentation techniques can be used to have enough data to train models on, but this somewhat dilutes the value of the data. The suicide ideation dataset that we use in this work, collected by Trenholm and Mavinkurve [49], was assembled by downloading tweets of interest via keyword search, then labeled by annotators with appropriate training in psychology.

- Crowdsourced labels

Crowdsourcing via platforms such as Amazon Mechanical Turk¹ or CrowdFlower can help label large amounts of data at a relatively low cost. However, since this technique employs untrained workers, and the labels obtained using this technique can be unreliable. The CrowdFlower dataset² for emotion detection in tweets is labeled in this manner.

- Surveys

Surveys involve asking social media users to answer a questionnaire, and using the provided answers to label the participating users' social media posts accordingly. De Choudhury et al. [15] ask crowd workers to answer a clinical depression survey and provide a history of their social media posts. The posts are then labeled based on their answers.

- Subforums

A lot of research using Reddit data uses mental health related subreddits—r/Anxiety, r/Depression, r/PTSD, r/SuicideWatch, etc.—as labels to classify the posts [45, 22, 29]. Shen and Rudzicz [45] collected a binary anxiety detection dataset by taking posts from r/anxiety and other anxiety-related subreddits, along with a control set of posts from subreddits not associated with mental health.

- Hashtags

Doan et al. [17] obtained a dataset of tweets containing stress- and relaxation-related hashtags (e.g. "#stress" and "#relax"), then used the hashtags to label the tweets as "stress" or "relaxation".

- Regular expressions

Coppersmith et al. [11, 14] make use of regular expressions to filter through a large twitter dataset and detect posts where a user presents a self-diagnosis. By searching for tweets matching e.g. "I was diagnosed with X.", they are able to identify users with various mental health conditions, then construct a dataset by collecting all available tweets posted by these users.

- Volunteered data

Some datasets are constructed from volunteered social media data by people who

¹<https://www.mturk.com/>

²<https://data.world/crowdfLOWER/sentiment-analysis-in-text>

suffer from a known mental health condition. Coppersmith et al. [14] collect a dataset from OurDataHelps³ of social media posts provided by people who have committed a suicide attempt, or their loved ones.

2.3. Tasks and methods

NLP has been used to identify a wide range of mental disorders found in social media posts. Classifiers have been successfully built to detect depression [11, 43, 47], anxiety [45], post-traumatic stress disorder (PTSD) [11], obsessive-compulsive disorder (OCD) [12], and suicide ideation, among others. A lot of the methods used rely on feature extraction and feature engineering. Commonly used features include Linguistic Inquiry and Word Count (LIWC) [39]—a tool used to detect mental disorders based on linguistic patterns—, part-of-speech tagging, bag-of-words, tf-idf, and n-grams.

Due to the scarce availability of annotated data, classical classification methods such as support vector machine (SVM) and random forest have historically performed well against their deep learning counterparts such as convolutional neural networks (CNN) or recurrent neural networks (RNN) that need large amounts of data to be trained. In general, we notice that SVMs obtain better results than other classical methods. More recently, approaches making use of multi-task learning and large language models (LLMs) have attempted to make up for this lack of data by using auxiliary datasets or models already pretrained on large quantities of unstructured text. We look at some of those methods in more detail in sections 4.4 and 5.1.

2.4. Suicide ideation

2.4.1. Definition

According to Beck et al. [4], suicide ideation is the planning or wishing to commit suicide by individuals who have not recently made a suicide attempt. The task of detecting suicide ideation is to identify within a given textual document such plans or wishes. Most classification tasks relating to suicide ideation are formulated as binary tasks [1], although some propose a scale, e.g. "Strongly concerning", "Possibly concerning", and "Safe to ignore" [36]. Others also include suicide-related topics in their label definitions, e.g. suicide campaigning or prevention [8].

Although someone who presents suicidal behaviour can also show signs of depression (e.g. depressed mood, loss of interest or pleasure [48]) or anxiety (e.g. fear, worry, feeling overwhelmed [32]), detecting suicide ideation goes beyond merely identifying symptoms of

³<https://ourdatahelps.org/>

a given condition. It aims to determine whether the author of a social media post or other piece of text *intends* to make an attempt on their life.

2.4.2. Related work

As with other mental health related tasks, most suicide ideation detection methods rely heavily on feature extraction. We summarize some of them here.

Matero et al. [30] tested three separate logistic regression models on the CLPsych-2019 Shared Task dataset [56], a dataset composed of reddit posts from 621 users who have posted on r/SuicideWatch. The posts are annotated by taking account four families of suicide risk factors: *Thoughts*, *Feelings*, *Logistics*, and *Context*. There are four different classes of posts: *No risk*, *Low risk*, *Moderate risk*, and *Severe risk*. Their approaches use a combination of user-level and message-level features, namely BERT embeddings, which are aggregated using an RNN model with attention. They find that using dual-context modeling, i.e. modeling language related to suicide separately from posts not originating from r/SuicideWatch, helps determining the level of risk in two ways: features from the suicide context help determine whether or not the post presents a risk at all, while features from the normal context such as personality factors help distinguish between low, moderate, and severe risk.

Ramírez-Cifuentes et al. [42] collected a dataset of 1200 Spanish-language tweets and employed clinicians to label them as control (885/1200), suicidal ideation risk (115/1200), and doubtful (200/1200). Based on these results, they then collected and annotated more tweets from the users presenting a risk, to obtain a total of 513 tweets found to present suicidal ideation risk. The authors then formulated a binary classification task for which they tested Random Forest, Logistic Regression, Multi-Layer Perceptron (MLP), SVM, and CNN. They used LIWC, bag-of-words (BoW), n-gram, social networks and psychological (SNPSY), and image features in a multimodal approach. They found that SVM performs best on this task, and found that the most relevant features were self-references, the number of tweets posted, and the elapsed time between each tweet.

Tadesse et al. [47] formulated a binary classification task using a dataset of Reddit posts collected by Ji et al. [20] composed of 3549 "suicide-indicative" posts from r/SuicideWatch and 3652 "non-suicidal" posts from other subreddits. The authors proposed an architecture combining a long short-term memory (LSTM) model with a CNN, and using word2vec embeddings. The LSTM-CNN architecture was found to outperform classical methods such as SVM, NB, RF, and Extreme Gradient Boosting using TF-IDF, BoW, and statistical features.

2.4.3. Preliminary work on the suicide ideation dataset

Preliminary experiments performed on the suicide ideation dataset [49] described in Section 2.6.1 compared naive Bayes, random forest and BERT classifiers for this task, and found BERT to be the best performing of the three. However, due to some discrepancies in the dataset (Section 3.1), we have to make some modifications to it such as duplicate removal, and thus obtain new baselines for this updated dataset.

2.5. Ethical Concerns

As with any personal health information, data relating to mental health must be kept confidential. Even though all data used in this research has been sanitized of any identifiers, for privacy reasons, Twitter does not allow downloaded tweets to be published [54]. In line with Benton et al. [5] and Coppersmith et al. [14], all examples presented in this work are thus paraphrased to protect the users' identities.

2.6. Datasets used in this work

2.6.1. Suicide Ideation

The main dataset used for this research was collected using the Twitter API by Trenholm and Mavinkurve [49] using a keyword search consisting of 92 rules. The rules include words or phrases associated with suicide ideation, depressive feelings, depression symptoms, fear, hurt, loneliness, drug abuse, or psychological disorders. 6322 of the collected tweets were annotated and separated into four classes:

- (1) "**Suspect suicide ideation, self-harm or suicidal behavior:** this class contains tweets expressing thinking about suicide, planning suicide. It does not include references to previous thoughts or attempts, unless related as still existing;"
- (2) "**Related to suicide:** this class includes tweets talking about suicide, news, friends-relatives, support or hot lines, suicide awareness campaigns, opinion about suicide, reporting/talking about a previous attempt(s) or thoughts;"
- (3) "**Non-related to suicide:** this class is about everything else, just depression-anxiety, another mental health, mental suffering, insomnia, etc.);"
- (4) "**Suicide other:** this class includes all other uses of the suicide word, suicide in the figurative way (political/strategical suicide, "this annotation job is so boring that makes me want to kill myself", etc.)."

Given our objective to analyze the suicide ideation task and the limitations of this dataset, we deem it necessary to include the full, original class definitions. After the labeling, the

tweets are distributed as described in Table 2.1. To balance the dataset, we trim the "Non-related to suicide" class in a similar fashion to [49], and keep a random selection of 200 tweets from that class.

Table 2.1. Suicide ideation dataset

Class	Number of samples
Suspected suicide ideation	106
Related to suicide	245
Non-related to suicide	5853
Suicide other	118

2.6.2. Publicly available mental health related datasets

Since annotated datasets are not easily accessible, we obtain public datasets related in some way to mental health. Through the use of multi-task learning, we hope to answer the following questions:

- What are the relationships between these tasks and suicide ideation?
- Can we use these datasets as auxiliary tasks in a multi-task learning setting to improve classification performance for suicide ideation?

Due to the presumed proximity of depression and anxiety to suicide ideation, we further test our methods on the DATD dataset.

2.6.3. DATD

The Depression and Anxiety in Twitter Dataset (DATD) [38] is comprised of 1050 tweets annotated by non-expert humans. They are first obtained using a keyword search on Twitter’s API, then split into two categories:

- **Positive instances:** tweets where the author appears to be suffering from anxiety or depression (548 samples)
- **Negative instances:** everything else (502 samples)

To validate the three annotators’ labeling, 150 tweets are labeled by each, with an average agreement of 80%. Tweets that are found to contain depression or anxiety are labeled as "MENTAL_HEALTH", and the others, as "OTHER". Here are two example tweets and their associated labels:

Summertime depression is hitting hard right now. **mental_health**
I get anxiety looking at my bank account. LOL. **other**

2.6.4. SAD

The Stress Annotated Dataset (SAD) [31] contains 6850 short sentences written by crowd workers answering the question, "What is stressing you out right now?". The worker is then asked to label the description according to nine stressor categories: **work, school, financial problem, emotional turmoil, social relationships, family issues, health, fatigue, or physical pain, everyday decision making, other**. The labels are then validated by a second group of workers. Example sentences and their labels:

Exams start this week and I'm scared. **school**

I'm worried about the tension in my office. **work**

2.6.5. CrowdFlower

The CrowdFlower dataset⁴ contains 39 740 tweets labeled with 13 different emotions: **empty, sadness, enthusiasm, neutral, worry, love, fun, happiness, hate, surprise, relief, boredom, anger**. The tweets are labeled by crowd workers. As others have pointed out, this dataset is noisy [40] and classifiers generally perform poorly on this task [35]. Some example tweets from this dataset:

This is my last day working at my job today, sad times. **sadness**

I can't sleep yet I have to wake up super early tomorrow. **worry**

⁴<https://data.world/crowdfLOWER/sentiment-analysis-in-text>

Chapter 3

Preliminary Analysis

3.1. Analysis of Annotations

Preliminary tests and a summary assessment of the suicide ideation dataset collected by Trenholm and Mavinkurve [49] suggested that there may be some inconsistencies in the labeling of the tweets. It was thus deemed relevant to perform a manual review of the dataset. Given the disproportionately large number of tweets labeled as "Non-related to suicide", and the relative simplicity of its labeling compared to the other classes, only "positive" samples were analysed, i.e. classes "Suspected suicide ideation", "Related to suicide", and "Suicide other".

3.1.1. Common Ambiguities

The most commonly found sources of ambiguity in the labeling are:

- Sadness/pain labeled as suicide ideation or related to suicide;
- Figurative use of suicide-related words labeled as suicide ideation;
- Reference to death/dying without explicitly alluding to suicide;
- Ambiguous terms or turns of phrases labeled as suicide ideation without clear context to support it;
- Reference to prior attempts or someone else's suicidal thoughts labeled as suicide ideation;
- Tweets labeled as "Suicide other" with seemingly no link to suicide.

Based on this observation, we define a set of analysis labels for each class.

3.1.2. Analysis Labels

Tweets from each of the three positive classes are labeled according to the most common ambiguities or inaccuracies for that class. The "analysis" labels are as follows:

Suspected suicide ideation:

- (1) The tweet clearly refers to suicide ideation in the past but not at the moment.
e.g. *I found pills in my bag. I probably would have succeeded had I had them when I tried to kill myself.*
- (2) The tweet provides some context suggesting it is not suicide ideation.
e.g. *I have such a stomach ache I wish I could die.*
- (3) The tweet is ambiguous. For example, it may refer to one's death without any explicit mention of suicide. The tweet, "I want to go to the Met when I die." falls into this category, as this person could very well be a fan of art, and refer to the Metropolitan Museum as heaven-like.
e.g. *Why is life just suffering?*
- (4) The label appears to be correct.

Related to suicide:

- (1) The tweet is not obviously related to suicide
e.g. *@user Need help?*
- (2) The label appears to be correct.

Suicide other:

- (1) The tweet is related to suicide.
e.g. *"Our differences make us equal." A mom grieves for her daughter after her suicide.*
- (2) The tweet is not related to suicide.
e.g. *And love laughs at the barren twilights of death.*
- (3) The label appears to be correct.

3.1.3. Results of Manual Evaluation

All tweets in the "positive" classes were labeled according to the definitions presented above. By quantifying some of the apparent labeling issues in the dataset, we hope to gain a better understanding of the classification results later on.

It is not a surprising result that expertly-labeled tweets could appear as ambiguous to an untrained observer, and a similar comment could be made of the 10 samples found not to express suicide ideation in Table 3.1. However, in the case of the 5 tweets referring to past ideation, since the condition can be clearly verified, we can confidently consider them as being mislabeled.

Table 3.1. Analysis of tweets labeled as "Suspected suicide ideation"

Label	Number of samples
Past suicide ideation	5
Not suicide ideation	10
Ambiguous	19
Correctly labeled	72
Total	106

Table 3.2. Analysis of tweets labeled as "Related to suicide"

Label	Number of samples
Not related to suicide	11
Correctly labeled	234
Total	245

A review of the "Related to suicide" class (Table 3.2) found the labeling to be consistent, with 234/245 judged as correctly labeled.

Table 3.3. Analysis of tweets labeled as "Suicide other"

Label	Number of samples
Related to suicide	5
Not related to suicide	46
Correctly labeled	67
Total	118

The "Suicide other" class was found to be the least consistent, with 46 tweets being labeled as "Not related to suicide" compared to the 67 that were correctly labeled. This tells us that there is a lot of ambiguity around this class, as a significant portion of the tweets with this label were not found to obviously satisfy its definition.

3.1.4. Duplicates

Upon inspection of the data, it was discovered that some duplicates were present in the dataset. Further analysis showed that of the 530 duplicates, not all pairs had matching labels, indicating that the duplication had occurred prior to the labeling, and giving us the opportunity to observe the labeling consistency on this part of the dataset.

To do so, we define *confusion pairs* as the unordered sets of labels given to two identical tweets. We then count each set and present the results in Table 3.4. Triples are counted as 1.5 pairs if the labeling is consistent.

We find that 7/12 pairs containing at least one tweet labeled as suicide ideation are mismatched. Interestingly, the confusion occurs mostly with the "Non-related to suicide"

Table 3.4. Duplicates confusion matrix

Label	Ideation	Related	Non-related	Other
Ideation	5			
Related	0	3.5		
Non-related	6	1	238.5	
Other	1	0	9	1

Ideation: Suspected suicide ideation; Related: Related to suicide;
 Non-related: Non-related to suicide; Other: Suicide other

class. The "Related to suicide" sample size is very small but fairly consistent, whereas the "Non-related to suicide" tweets are the most consistently labeled of all. Surprisingly, 10/11 of the pairs containing at least one "Suicide other" are mismatched, further highlighting the ambiguity presented in Table 3.3. Note that since there is no ground truth in this context, the resulting confusion matrix is of triangular shape.

3.1.5. Updated dataset

Upon removal of the duplicates and some empty tweets, the new dataset is composed of 91 "Suspected suicide ideation", 223 "Related to suicide", 200 "Non-related to suicide", and 101 "Suicide other".

3.2. Experiments with classical methods

Trenholm and Mavinkurve [49] concluded that BERT outperformed classical classification methods such as Random Forests and Naive Bayes on the suicide ideation task. To validate this result, we start by obtaining baseline results using Random Forests and SVM since it is often found to perform well on this type of task. We test these methods on the two main tasks: suicide ideation and DATD.

3.2.1. Random Forest

Random Forest is an ensemble learning method based on decision trees, where each tree is trained on a random subset of the training data. The randomization is done to reduce overfitting which commonly occurs in decision trees. Classification is done via a voting mechanism, and the prediction corresponds to the class voted by the greater number of trees.

We implement this method using the `RandomForestClassifier` class of the `sklearn.ensemble` Python module. A grid search is performed to optimize the number of estimators (trees) and maximum depth hyperparameters, with 5-fold cross validation.

For suicide ideation, we use 50 estimators with a max depth of 100, and for DATD, 30 estimators and a max depth of 50.

3.2.2. SVM

Support vector machine (SVM) [33] is a machine learning method which can be used for both regression and classification tasks. It works by calculating a hyperplane that optimally separates the classes. It does so by maximizing the distance, or margin, between the hyperplane and the nearest datapoints. The data samples nearest to the hyperplane are called support vectors, giving the algorithm its name.

Similarly to Random Forest, we implement this method using the SVC class of the `sklearn.svm` Python module. We use the default radial basis function kernel and optimize the C and $gamma$ hyperparameters using grid search and cross validation. C acts as a regularization parameter. A larger value of C will allow for smaller margins if it improves classification performance, while a smaller C will incur larger margins and a simpler decision function¹. The $gamma$ parameter determines how much importance is given to any single sample in the training set. A larger $gamma$ value can lead to overfitting since the algorithm will attempt to account for each sample instead of extracting some generalization over the training data. After grid search and cross-validation, both tasks are found to perform best with $C = 100000$ and $gamma = 0.00001$.

3.2.3. Preprocessing

Before training the classifiers, it is important to perform pre-processing of the textual data to reduce noise in the features. To do so, we remove user mentions from tweets before performing punctuation removal, stopword removal (e.g. "the", "is", "and"), and word lemmatization (e.g. "scared" -> "scare").

3.2.4. Features

We use two features: text length based on the hypothesis that emotional state may impact the length of utterances, and TF-IDF. TF-IDF, or term frequency-inverse document frequency, is a commonly used feature in NLP. It gives the relative importance of a given term based on how many times it occurs in a document compared to how many documents it appears in. Basically, the more it appears in a single document while being rare in others, the more it is important. Conversely, if it is not frequently used in a document but is commonly found otherwise, it means that the word is not important in this context.

TF-IDF formula:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

¹https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

Where:

$$\text{TF}(t, d) = \frac{\text{count of term } t \text{ in document } d}{\text{number of terms in document } d}$$

$$\text{IDF}(t, D) = \log \left(\frac{\text{number of documents in collection } D}{\text{number of documents containing term } t} \right)$$

3.2.5. Metrics

The metrics we use to evaluate our classifiers' performance are:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{total samples}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

We further consider two other metrics: weighted- and macro-averaged F1 score. The weighted-averaged F1 score metric, as its name implies, is the average of F1 scores across all classes, weighted proportionally to the number of samples per class. The macro-averaged F1 score is simply the arithmetic mean of all F1 scores. Since the suicide ideation dataset is imbalanced, i.e. some classes have more samples than others, good accuracy results could be achieved by performing well on the most voluminous classes, at the expense of the less-represented ones. Looking at these metrics can thus give us a better understanding of how the classifier performs across all classes.

3.2.6. Results

Table 3.5. Baseline classification results

Task	Method	Accuracy	Precision	Recall	WF1	MF1
Suicide ideation	Random Forest	0.65	0.63	0.65	0.59	0.51
	SVM	0.65	0.63	0.65	0.62	0.55
DATD	Random Forest	0.66	0.69	0.66	0.64	0.64
	SVM	0.69	0.69	0.69	0.68	0.68

WF1: weighted-averaged F1 score;
MF1: macro-averaged F1 score

Notice in Table 3.5 that the macro-averaged F1 scores are significantly lower than the weighted-averaged F1 scores for the suicide ideation task, indicating uneven performance of the classifiers on the four different classes.

In terms of accuracy, the two methods obtain equal results for the suicide ideation task, while the SVM classifier outperforms Random Forest on the DATD task.

3.2.7. Discussion

Random Forest and SVMs can be trained efficiently using a small number of data points, which can be valuable when working on tasks with limited resources, where there is not enough data to train, for example, neural networks which require large amounts of training data. However, they do have some limitations that motivate us to find better methods. Namely, Random Forest can sometimes be prone to overfitting, meaning that what it learns when training doesn't generalize well to unseen data. It can also be sensitive to class imbalance, which can be an issue for suicide ideation detection since the dataset is not well balanced. SVMs are sensitive to noise and will perform poorly when there is too much overlap between the classes, because it is difficult to define hyperplanes to separate them.

But most importantly, since detecting suicide ideation is a fairly complex task, these methods lack the language understanding capabilities to make classification decisions based on the context in which certain words are used. The same suicide or depression related words can also be used figuratively, and the difference between true ideation and a joke can often be subtle, even for humans. For this reason, we want to use NLP methods that better capture meaning in text. In the following chapters we experiment with multi-task learning using BERT and MentalBERT, and conclude with classification experiments using GPT-4 and ChatGPT.

Chapter 4

Multi-Task Learning

4.1. Objectives

We apply a multi-task learning approach to the suicide ideation classification task in order to gain a better understanding of the problem. More specifically, we aim to:

- (1) Evaluate if publicly available datasets related to mental health can help palliate to the lack of data in our expert-labeled dataset.
- (2) Examine the relationship, if any, between the suicide ideation task and the others, particularly DATD. The hypothesis is that if two tasks are closely related, multi-task learning could have an effect similar to data augmentation and improve classification performance.
- (3) Evaluate if a model pre-trained on unstructured text related to mental health can help improve classification performance.

We start with a brief review of language models and multi-task learning, then look at related research in the field of mental health, and finally describe our experimental setup and results.

4.2. Language Models

Language models are probabilistic models of language. They allow us perform various tasks including translation, text generation and classification. For example, simple models such as n-grams predict the most probable next word given a string of n words. In this work, we make use of large language models (LLMs) which have been trained on large amounts of unstructured text and are able to capture the context in which they make predictions by relying on the self-attention mechanism of transformers.

4.2.1. BERT

In this section we perform experiments using the BERT (Bidirectional Encoder Representations from Transformers) [16] family of language models, which are made up of Transformer [50] encoder layers. Transformers-based models use self-attention and positional encodings to produce deeply contextualized representations of words. By processing inputs in parallel, they are able to capture the relationships of each word in a given input with every other word. BERT-based models are pre-trained on large quantities of unstructured text on two different tasks:

- **Masked Language Modeling**, where some tokens in the input are masked, and the model is asked to predict the masked tokens.
- **Next Sentence Prediction**, where given two sentences A and B, the objective is to answer whether B is the actual sentence that follows A.

After pre-training, BERT can be fine-tuned to a specific task. Since the pre-trained model already possesses some language understanding, fine-tuning can be done with much a much smaller amount of data than if one were to train a task-specific model from scratch.

The two models we use are BERT_{BASE}, a 110M-parameter model pre-trained on 800M words from the BooksCorpus [55] and 2,500M words from English Wikipedia, and MentalBERT [21], also with 110M parameters, pretrained on a mental health-related subreddit corpus of 13,671,785 sentences. We are interested in learning whether pre-training specifically on text pertaining to mental health can improve classification performance on suicide ideation, depression and anxiety detection.

4.3. Hard parameter sharing

Multi-task learning (MTL) [9] is the concept of training a single model on multiple tasks to achieve better generalization or obtain complementary knowledge on a given main task from closely-related auxiliary tasks. While there exist a number of different architectures to perform MTL [10], the most common ones use hard parameter sharing, where all tasks share a common set of parameters. Figure 4.1 shows an example of hard parameter sharing in a tree-like architecture, where all tasks share representations h_s and have their own task-specific representations h_t being trained in parallel.

4.4. Related work

Previous research has looked into the possibility of using multi-task learning for mental health applications. We summarize a few of the approaches and their key findings:

Kirinde Gamaarachchige and Inkpen [23] present a multi-task, multi-channel convolutional neural network (CNN) model to detect depression and PTSD from the CLPsych 2015

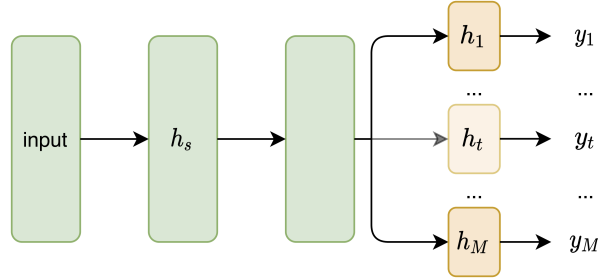


Fig. 4.1. Hard parameter sharing [10]

Twitter dataset [13]. The authors also train a separate multi-task CNN model for emotion classification (anger,fear,joy,sadness). The emotion classifier is then used to predict emotions found in the main dataset, and the output emotion prediction is concatenated with manually engineered features before being fed as inputs to the model. This approach beats the baseline results from a multi-class SVM classifier, and achieves higher F1-score and AUC than a standard multi-class CNN, suggesting that there are benefits to multi-task learning on closely-related mental health tasks such as depression and PTSD detection.

Sarkar et al. [44] propose AMMNet, a multi-task model where shared document-level features are learned in parallel with task-specific features before being concatenated and fed into task-specific final dense layers for classification. The shared feature learning module is based on a CNN architecture with BERT-base word embeddings, and the task-specific feature learner performs topic modeling to learn the topic of a given document. This method is trained on Reddit posts collected from mental-health related subreddits and labeled as 'Anxiety', 'Depression', and 'Others'. This method is found to outperform classical machine learning methods such as SVM and Random Forest. model: tasks: datasets: key findings:

Benton et al. [6] compare single-task learning using a multilayer perceptron to multi-task learning via a 2-layer feedforward network where the first hidden layer is shared among all tasks and the second is replicated for each mental health condition of the CLPsych 2015 Twitter dataset [13]. The model's shared representation aims to take advantage of the high level of comorbidity between mental health disorders. The authors find that the multi-task learning approach using character n-gram features outperforms single-task learning in all cases, and that adding an extra gender detection task can further improve classification performance.

4.5. Architecture

Parallel, hard parameter sharing MTL as seen in Figure 4.1 is chosen for its simplicity and interpretability as a comparison tool. Since parameters are shared between all tasks, we can infer that if we see a performance increase when training multiple tasks together, it means

that they share some common characteristics that are being represented by these parameters. We base our experiments on the multi-task-NLP project¹. In this hard parameter sharing architecture, all tasks share a same language model, and have their own task-specific layers for classification. Distinction between main and auxiliary tasks is made by varying the task-specific loss weight parameters, which are coefficients applied to each task’s loss to determine the importance of each task when updating parameters. This architecture also allows us to compare BERT and MentalBERT as shared models.

4.6. Tests

Given that there are 30 possible task-model combinations to choose from, it is important to select the best experiments to answer our research objectives:

- (1) First we train all tasks individually to obtain a baseline result.
- (2) We then train all tasks in parallel to answer our first question and evaluate the benefits of the auxiliary tasks on the overall classification performance.
- (3) Because we suspect the CrowdFlower emotion task to have a negative effect on our results, we repeat the experiment without this task.
- (4) To analyze the relationship between suicide ideation and DATD, we train the two tasks together. To validate our results with MentalBERT, we run the same experiment a second time with DATD as the main task.
- (5) Finally, to compare the overall performance of BERT and MentalBERT, we replicate the experiments for the two models.

4.7. Results

Since there is a significant number of task combinations and experiments, hyper-parameters such as learning rate, batch size, and number of epochs were not fully optimized for each test. We have found that training in this setup took longer than typical BERT fine-tuning (6-8 epochs), and that a learning rate of 1e-5 with a batch size of 32 works best in most configurations. Loss weight for the main task is set to 1, and 0.2 or 0.4 for auxiliary tasks.

Because the aim is to gain a better understanding of the suicide ideation and depression detection tasks, the results are mostly meant to be compared to each other to see what test setup or base model favors performance on those tasks. The multi-task learning setup is used as a comparison tool to analyze interactions between the tasks.

¹<https://github.com/hellohaptik/multi-task-NLP>

4.7.1. Overall performance

To evaluate the classification performance of our models on the four different tasks, we consider accuracy as well as weighted- and macro-averaged F1 scores. Given the uneven class distributions of some of the datasets, we look at macro-averaged F1 score as a good indicator of overall classification performance.

Table 4.1 shows results for the first round of experiments using BERT as the base model. We first note that results for single-task suicide ideation and SAD improve upon the random forest and SVM baselines in Table 3.5. Although not all tasks improve their performance in the "All tasks" setting, each task has at least one setting in which macro-averaged F1 score is improved, and CrowdFlower emotion is the only task whose accuracy is highest when trained individually.

Table 4.1. Classification results using BERT as the base model

Tasks	Suicide ideation			SAD			DATD			CF Emotion		
	Acc.	WF1	MF1	Acc.	WF1	MF1	Acc.	WF1	MF1	Acc.	WF1	MF1
SUI	<u>0.70</u>	0.67	0.60									
SAD				<u>0.78</u>	<u>0.77</u>	<u>0.78</u>						
DATD							0.70	0.70	0.70			
CF										0.40	0.35	<u>0.19</u>
SUI*+DATD	0.69	0.64	0.57				<u>0.73</u>	<u>0.73</u>	<u>0.73</u>			
SUI*+SAD+DATD	0.71	0.70	0.65	0.80	0.80	0.80	0.72	0.71	0.71			
All tasks	<u>0.70</u>	<u>0.68</u>	<u>0.64</u>	0.77	<u>0.77</u>	0.77	0.74	0.74	0.74	<u>0.37</u>	0.35	0.22

Acc.: Accuracy; WF1: weighted-averaged F1 score; MF1: macro-averaged F1 score
 SUI: Suicide ideation; CF: CrowdFlower; * denotes a main task

In Table 4.2, we first observe that using MentalBERT boosts performance on all tasks, confirming that the pretraining on mental-health related text is beneficial. The enhanced performance in SUI+DATD settings also highlight a closeness between the suicide ideation and DATD tasks. The SAD task also appears to help performance for suicide ideation, since the SUI*+SAD+DATD setting gives the highest macro-averaged F1 score, whereas adding CrowdFlower Emotion decreases all performance metrics on the other three tasks.

A performance increase for suicide ideation detection when trained alongside stressor classification using the SAD dataset is consistent with Wilburn and Smith [52] who found that stressful life events and low self-esteem were significantly related to the development of suicidal thoughts. We also know that preexisting mental health disorders are a consistent risk factor for suicide ideation [34], that depression and anxiety have a high rate of comorbidity, and that they frequently coexist with other psychiatric disorders [2]. This could potentially

explain the improved results obtained when suicide ideation, depression, and anxiety detection are trained together. The only dataset whose addition doesn't help detecting suicide ideation is the CrowdFlower dataset, which can likely be explained by the poor quality of the labeling as discussed in Section 2.6.

Table 4.2. Classification results using MentalBERT as the base model

Tasks	Suicide			SAD			DATD			CF Emotion		
	Acc.	WF1	MF1	Acc.	WF1	MF1	Acc.	WF1	MF1	Acc.	WF1	MF1
SUI	0.73	0.70	0.63									
SAD				0.80	0.80	0.80						
DATD							<u>0.78</u>	<u>0.78</u>	<u>0.78</u>			
CF										0.41	0.36	<u>0.20</u>
SUI*+DATD	0.75	<u>0.71</u>	<u>0.66</u>				<u>0.78</u>	0.77	0.77			
SUI+DATD*	0.67	0.62	0.55				0.82	0.82	0.82			
SUI*+SAD+DATD	<u>0.74</u>	0.72	0.67	<u>0.79</u>	<u>0.79</u>	<u>0.79</u>	0.73	0.73	0.73			
All tasks	0.73	<u>0.71</u>	<u>0.66</u>	0.78	0.78	0.78	0.71	0.71	0.71	<u>0.36</u>	<u>0.34</u>	0.21

Acc.: Accuracy; WF1: weighted-averaged F1 score; MF1: macro-averaged F1 score

SUI: Suicide ideation; CF: CrowdFlower; * denotes a main task

4.7.2. Suicide ideation classification

Not all classes in the suicide ideation task have the same degree of importance. Namely, it is much more important to classify the tweets labeled as "Suspected suicide ideation" correctly than it is for "Suicide other". For this reason, we look into more detail at how the MentalBERT model performs on each class for three different settings: suicide ideation only, all tasks, and suicide ideation combined with DATD.

Table 4.3. Per-class suicide ideation classification results using MentalBERT

Tasks	Ideation			Related			Non-related			Other		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SUI	0.71	0.67	0.69	0.79	0.86	0.83	0.69	0.90	0.78	0.60	0.15	0.24
All tasks	0.65	0.72	0.68	0.76	0.84	0.80	0.76	0.85	0.80	0.60	0.25	0.36
SUI*+DATD	0.75	0.83	0.79	0.82	0.84	0.83	0.68	0.90	0.77	0.75	0.15	0.25

P: Precision; R: Recall; F1: F1 score

SUI: Suicide ideation; * denotes a main task

While the SUI*+DATD task combination has lower performance on the "Non-related to suicide" and "Suicide other" classes, it obtains significantly better results for "Suspected suicide ideation", the class we are most interested in. This could be due to depression and

anxiety detection being more closely related to the detection of suicide ideation than that of figurative uses of suicide-related words.

Table 4.4. Confusion matrix for suicide ideation

Label\Predicted	Ideation	Related	Non-related	Other
Ideation	15	2	1	0
Related	2	37	5	0
Non-related	0	3	36	1
Other	3	3	11	3

Ideation: Suspected suicide ideation; Related: Related to suicide;

Non-related: Non-related to suicide; Other: Suicide other

Note: This matrix is based on results from the Suicide ideation + DATD MentalBERT model.

The confusion matrix obtained for the MentalBERT SUI*+DATD model (Table 4.4) shows that suspected suicide ideation tweets are most commonly mislabeled as being related to suicide. "Suicide other" tweets are most commonly confused with "Non-related to suicide". The latter observation appears consistent with our analysis of duplicate tweets in Table 3.4.

4.8. Discussion

We first confirm the preliminary findings by Trenholm and Mavinkurve [49] that the BERT base model obtains better classification results than classical methods on the suicide ideation task. BERT obtains a macro-averaged F1 score of 0.60 compared to 0.55 and 0.51 for SVM and Random Forest.

MentalBERT is shown to improve performance on all tasks. DATD benefits the most from this model, with a 0.08 improvement on its F1 score compared to 0.03 for suicide ideation.

Multi-task learning with the acquired public datasets helps improve performance on suicide ideation, indicating a similarity between the tasks.

We find that the macro-averaged F1 score doesn't tell the whole story, and that it is important to analyze the per-class metrics for suicide ideation detection. In doing so, we are able to observe that training for depression and anxiety detection in parallel with suicide has the most impact on detecting suicide ideation, even though performance drops for "Suicide other".

There appears to be a close proximity suicide ideation and depression/anxiety detection, since both improve in classification when the other one is its auxiliary task.

Finally, the confusion matrix (Table 4.4) doesn't have all the same issues found in the duplicates analysis from Table 3.4—namely confusion between suicide ideation and non-related to suicide—, except for "Suicide other" for which we see a very similar pattern of being confused for "Non-related to suicide".

Chapter 5

ChatGPT and GPT-4

OpenAI’s ChatGPT and GPT-4 [37] models have seen a rapid adoption in a variety of contexts ranging from customer support chatbots to virtual assistants embedded in productivity tools such as PowerBI or code generation tools like GitHub Copilot¹. While some experiments have already been made using these models on mental health tasks, research in this area is still in its infancy. In this chapter, we evaluate their performance on the suicide ideation and DATD tasks.

5.1. Related work

Ghanadian et al. [18] performed experiments on the University of Maryland Reddit suicidality dataset [46, 56] using ChatGPT in zero- and few-shot settings. The zero-shot experiment outperformed few-shot learning on this task, and both DistilBERT and AIBERT obtained higher F1 scores on this task than ChatGPT in either setting. It was found that ChatGPT gives less inconclusive results when using few-shot learning, and that increasing the model’s temperature setting also reduces the inconclusiveness rate but at the cost of poorer classification performance.

Lamichhane [24] tested ChatGPT (gpt-3.5-turbo) on three annotated Reddit datasets for stress, depression, and suicidality detection. The author found that although the model was outperformed by a fine-tuned BERT model, the results were promising given that the tests were performed in a zero-shot setting. The model had its worst performance on the suicidality detection task, with a 0.37 F1 score compared to 0.73 and 0.86 for stress and depression.

Yang et al. [53] compared GPT-3 and ChatGPT on emotion recognition in conversations (ERC), causal emotion entailment (CEE), depression, stress and suicide tasks. While ChatGPT does not beat BERT or MentalBERT in terms of classification performance, a human evaluation found that it generates reliable explanations that significantly outperform those

¹<https://github.com/features/copilot>

provided by GPT-3. The authors further tested "emotion-enhanced prompts" prompting the model to consider the emotion expressed in a given post before performing classification in a chain-of-thought (CoT) manner. This method provided better results than zero-shot or regular CoT. **Emotion-enhanced prompt:** *Post: "[Post]". Consider the emotions expressed from this post to answer the question: Is the poster likely to suffer from very severe [Condition]? Only return Yes or No, then explain your reasoning step by step.*

Amin et al. [3] tested ChatGPT on personality prediction, sentiment analysis as well as suicide and depression detection tasks. The authors formulated zero-shot prompts starting with "What is your guess[...]" for each task, e.g. "What is your guess for the sentiment of the text "text", answer positive, neutral, or negative? it does not have to be correct.", and found the classification performance to be similar to bag-of-words models and significantly lower than a fine-tuned RoBERTa model.

5.2. Overview of the models

GPT (Generative Pre-trained Transformer) models are multi-layer decoder-only transformers [28] pre-trained using the generative pre-training objective [41] which consists of predicting the next word in a sequence. GPT models are particularly good at text generation, but have found applications in a variety of NLP tasks such as translation or text classification, as classification tasks can be framed as text generation where the output tokens correspond to the predicted labels.

The experiments in this chapter are conducted using the following models:

- **ChatGPT**² (gpt-3.5-turbo) is a model fine-tuned for chat applications from the 175M-parameter GPT-3 model.
- **GPT-4** [37] is OpenAI's most recent and most capable model, which "has been trained on an unknown but extremely vast corpus of web-text data?" [7] Although its training details are not publicly known, it has shown state-of-the-art performance on various tasks, notably scoring a 90th percentile on the Uniform Bar Exam³.

5.3. Prompt design

The chosen prompts for the ChatGPT and GPT-4 tests on suicide ideation and DATD are inspired by Huang et al. [19]. The prompts follow the same basic structure: *Given post: 'tweet'. [Question]? Explain why in one or two sentences.* For a prompt to be effective in a zero-shot setting, it must accurately and concisely describe each class. For binary classification tasks, this is straightforward as the task can usually be presented as a yes/no

²<https://openai.com/blog/chatgpt>

³<https://openai.com/research/gpt-4>

question. For the suicide ideation task, we formulate the following prompt to differentiate between the four classes:

Given Post: 'sample tweet'. Answer whether the post expresses suicide ideation, is related to suicide, is not related to suicide, or uses suicide-related words in a figurative way. Explain why in one or two sentences.

Based on this template, we can elaborate a few-shot context with examples highlighting some of the ambiguities presented in Section 3.1.1. We include the full few-shot context for the suicide ideation task.

Few-shot examples for the suicide ideation tests

User: Given Post: 'As far as I can see in the future, the outlook is bleak. As I get older, the thought of suicide crosses my mind as I don't have any money saved for retirement I spent years unemployed or doing part-time work and I am left with nothing.'. Answer whether the post expresses suicide ideation, is related to suicide, is not related to suicide, or uses suicide-related words in a figurative way. Explain why in one or two sentences.

Assistant: The post expresses suicide ideation. It directly mentions thoughts of suicide due to a bleak financial outlook, indicating a connection to suicidal thoughts and the difficult situation.

User: Given Post: 'Quality of life depends on your mental health and the quality of your thoughts. The woman in this picture contemplated suicide and suffered through depression and anxiety, but today is happy and healthy. Mental illness doesn't have to be for life.'. Answer whether the post expresses suicide ideation, is related to suicide, is not related to suicide, or uses suicide-related words in a figurative way. Explain why in one or two sentences.

Assistant: The post is related to suicide as it mentions thoughts of suicide. However, it does not constitute suicide ideation because it is referring to someone else, and the contemplation of suicide occurred in the past.

User: Given Post: 'Could anyone drop off some Taco Bell? I'm virtually on my death bed.'. Answer whether the post expresses suicide ideation, is related to suicide, is not related to suicide, or uses suicide-related words in a figurative way. Explain why in one or two sentences.

Assistant: The post is not related to suicide. The user describes being on their death bed, but there is no mention of suicide or suicide ideation.

User: Given Post: 'Standing in line behind two crypto bros in bomber jackets discussing whether they'd date Cardi B and I want to die.'. Answer whether the post

expresses suicide ideation, is related to suicide, is not related to suicide, or uses suicide-related words in a figurative way. Explain why in one or two sentences.

Assistant: The post uses suicide-related words in a figurative way. The phrase "I want to die" is used to express the user's strong discomfort or annoyance with the situation, not an actual desire to commit suicide.

The benefits of asking the model for an explanation are twofold:

- the provided explanations can help us interpret the model's answer;
- in-context explanations have been shown to help classification performance [25].

We follow the same pattern for the DATD prompt, and use the examples given in [38] to formulate few-shot examples. Since the annotators were given the instruction to label ambiguous posts as positives, we compose two versions of the prompt, one without this instruction, and another with the instruction appended to the question:

Prompt A:

Given Post: 'tweet'. Answer yes or no for whether the user appears to be suffering from depression or anxiety.

Prompt B:

Given Post: 'tweet'. Answer yes or no for whether the user appears to be suffering from depression or anxiety. Answer yes if the post is ambiguous or does not provide enough information.

The full few-shot examples for the DATD task can be found in Appendix A.1.

5.4. Tests

ChatGPT and GPT-4 experiments are run on the two main tasks for this project. In both cases, we want to compare the different models' performance by running the same tests with ChatGPT and GPT-4. For the suicide ideation task, we run an ablation study to analyze the impact of each example on overall and per-class performance.

The experiments are run through OpenAI's API. The model used for the ChatGPT experiments is *gpt-3.5-turbo*. The few-shot examples are provided in the "messages" JSON array as presented in Figure 5.1. For each sample we append a different question to this context, since the last message is the actual question being asked.

Since the models output textual statements, they have to be parsed in order to obtain classification results. For the DATD task, all answers start with either "Yes." or "No.", with the former corresponding to "depression or anxiety". There are more cases to consider for the suicide ideation task because of its four classes, but the process is mostly the same.

```

1  [
2    {
3      "role": "system",
4      "content": "You are a helpful assistant."
5    },
6    {
7      "role": "user",
8      "content": "Given Post: 'I feel sick to my stomach, I hate
↪ having such bad anxiety'. Answer yes or no for whether the user
↪ appears to be suffering from depression or anxiety."
9    },
10   {
11     "role": "assistant",
12     "content": "Yes. The user states that they have anxiety."
13   }
14 ]

```

Fig. 5.1. Example of context for ChatGPT prompt

5.5. Results

5.5.1. Classification results

Table 5.1. Classification results for ChatGPT and GPT-4

Task	Method	Accuracy	Precision	Recall	WF1	MF1
Suicide ideation	ChatGPT zero-shot	0.54	0.62	0.54	0.48	0.41
	ChatGPT few-shot	0.69	0.69	0.69	0.67	0.61
	GPT-4 zero-shot	0.71	0.73	0.71	0.70	0.65
	GPT-4 few-shot	0.82	0.83	0.82	0.79	0.75
DATD	ChatGPT (A)	0.77	0.81	0.77	0.76	0.76
	ChatGPT (B)	0.74	0.79	0.74	0.73	0.73
	GPT-4 (A)	0.78	0.79	0.78	0.78	0.78
	GPT-4 (B)	0.71	0.82	0.71	0.68	0.68

WF1: weighted-averaged F1 score;
MF1: macro-averaged F1 score

In Table 5.1, we observe that GPT-4 significantly outperforms ChatGPT on the suicide ideation task in both zero- and few-shot (with 4 examples) settings. Both models have an increased performance when given few-shot examples, but this improvement is much more pronounced in the case of ChatGPT, whose macro-averaged F1 score jumps by 0.20 compared to 0.10 for GPT-4. Another remarkable result is that GPT-4 obtains better classification performance across all metrics when compared to ChatGPT with few-shot examples.

For the DATD task, we see that both models perform worse when prompted to answer "Yes" whenever the post is ambiguous, even though that same instruction was given to the human annotators. The instruction was given to humans in an attempt to lower false negatives, but in our experiments this results in too many false positives output by the models. Interestingly, GPT-4 suffers a bigger drop in performance compared to ChatGPT, presumably because it considers more posts to be ambiguous. Decreased performance for both models could mean that detecting ambiguity is a more difficult task than detecting depression and anxiety, as recent work by Liu et al. [27] has found that GPT-4 performs poorly on certain disambiguation tasks.

5.5.2. Ablation study

We perform an ablation study on the suicide ideation task to measure the impact of the added prompt context on both models.

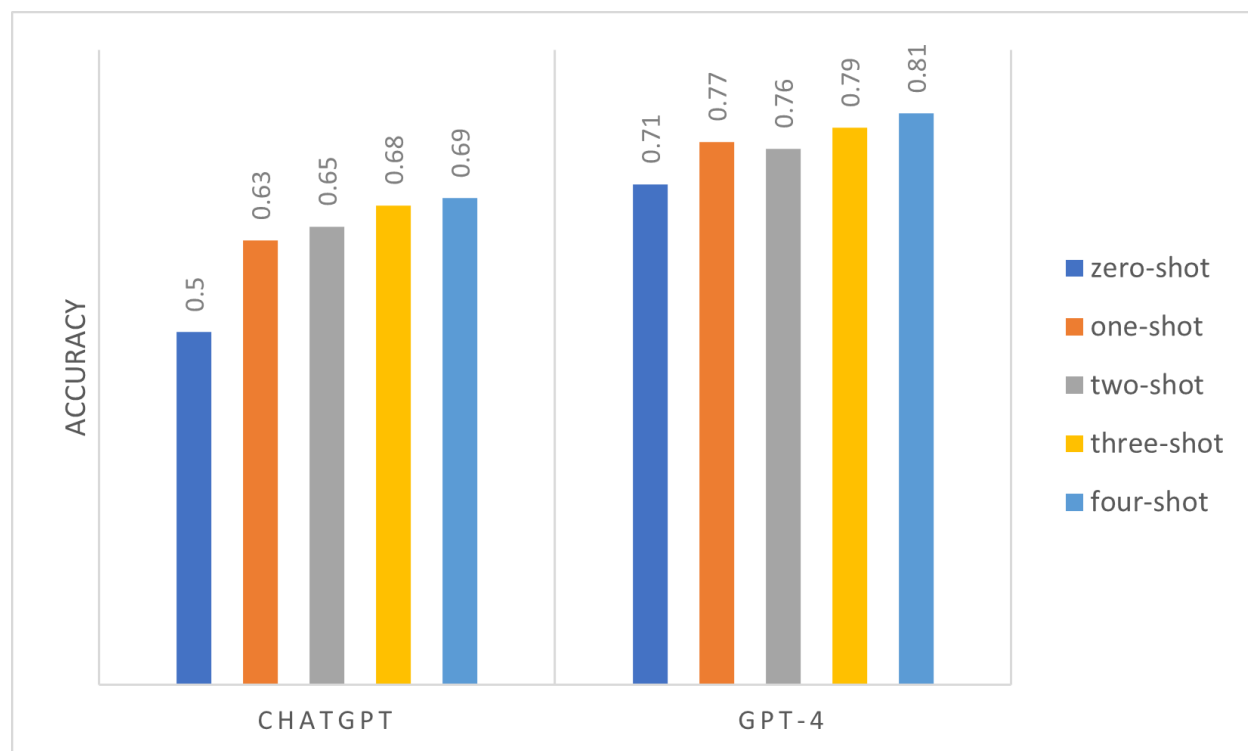


Fig. 5.2. Ablation study of few-shot learning on the suicide ideation dataset

Of the two models, ChatGPT benefits the most from the in-context examples, with a gain of 0.19 in accuracy compared to 0.10 for GPT-4. A lot of it is due to ChatGPT giving non-answers in the zero-shot setting, hence the jump in performance from zero- to one-shot.

Table 5.2. Per-class suicide ideation classification results for ChatGPT and GPT-4

Model and context	Ideation			Related			Non-related			Other		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ChatGPT zero-shot	0.32	0.94	0.48	0.89	0.18	0.30	0.77	0.90	0.83	0.0	0.0	0.0
ChatGPT one-shot	0.53	0.94	0.68	1.0	0.59	0.74	0.73	0.82	0.78	0.08	0.05	0.06
ChatGPT two-shot	0.49	0.94	0.64	0.96	0.59	0.73	0.81	0.88	0.84	0.08	0.05	0.06
ChatGPT three-shot	0.55	0.94	0.69	0.93	0.64	0.76	0.73	0.90	0.81	0.18	0.10	0.13
ChatGPT four-shot	0.53	0.94	0.68	0.88	0.66	0.75	0.76	0.88	0.81	0.30	0.15	0.20
GPT-4 zero-shot	0.67	0.89	0.76	1.0	0.64	0.78	0.68	0.97	0.80	0.31	0.20	0.24
GPT-4 one-shot	0.73	0.89	0.80	0.90	0.86	0.88	0.73	0.93	0.81	0.50	0.15	0.23
GPT-4 two-shot	0.70	0.89	0.78	0.95	0.80	0.86	0.72	0.97	0.83	0.43	0.15	0.22
GPT-4 three-shot	0.79	0.83	0.81	0.91	0.91	0.91	0.72	0.95	0.82	0.50	0.15	0.23
GPT-4 four-shot	0.83	0.83	0.83	0.91	0.93	0.92	0.73	0.95	0.83	0.83	0.25	0.38

P: Precision; R: Recall; F1: F1 score

Results from the multi-task learning experiments showed the importance of looking at the classification results for each class in the suicide ideation task. Table 5.2 shows that GPT-4 performs best at detecting suicide ideation in the four-shot setting, while ChatGPT has its highest F1 score when given three examples. Overall, the added context gives a significant improvement on both models.

The other main observation from this table is that GPT-4 zero-shot obtains higher F1 scores than ChatGPT four-shot across all classes except for "Non-related to suicide".

Since the progression from zero- to four-shot is made in the same order as presented in the table, i.e. the first example is of suspected suicide ideation, the second is related to suicide, etc., we can observe the impact of each added example to its own class. We find that only "Suspected suicide ideation" and "Suicide other" improve when an example from that class is added. For both models, adding an example from "Related to suicide" or "Non-related to suicide" decreases the F1 score for these classes.

5.5.3. Confusion matrix

Table 5.3. GPT-4 four-shot confusion matrix for suicide ideation

Label\Predicted	Ideation	Related	Non-related	Other
Ideation	15	1	1	0
Related	2	41	1	0
Non-related	0	1	38	1
Other	1	2	12	5

Ideation: Suspected suicide ideation; Related: Related to suicide;
Non-related: Non-related to suicide; Other: Suicide other

The confusion matrix for GPT-4 four-shot results follows a similar pattern to what was obtained by the multi-task learning setup. Notice that there is one less element in the Ideation row due to a non-answer from the model. Even though the model has multi-task learning setup for each class, we see that confusion between "Suicide other" and "Non-related to suicide" is even more pronounced than what was found previously, further reaffirming our results from the analysis of annotations.

5.5.4. False negatives

Given the nature of the task, and especially since tweets expressing suicide ideation are rare, a good classifier should limit false negatives as much as possible. Table 5.3 reveals that GPT-4 with four-shot context made two such mistakes for the "Suspected suicide ideation" class. We analyze them in Table 5.4 to understand the problem.

Table 5.4. Suicide ideation false negatives from the GPT-4 four-shot experiment

Tweet	Prediction	Explanation
This could be my last post here. This year. If you had a heart attack and went into a depression thinking you wouldn't hear from me again, I got you.	Non-related to suicide	It uses dramatic language to suggest the end of the user's posts for the year, but there is no mention or suggestion of suicide or suicide ideation.
Suicidal thoughts suck. I was 14 years old at the time, my dad announced we couldn't afford the football equipment. Grandpa owned the golf course life is confusing at times.	Related to suicide	It mentions suicidal thoughts in the context of past experiences, indicating a connection to the topic of suicide, but it does not express current suicide ideation.

GPT-4 actually appears to be right in these two cases. If we refer to the analysis labels from Section 3.1.2, the first false negative from Table 5.4 was labeled as "Not suicide ideation", while the second one was found to be ambiguous. The GPT-4 explanations further support these predictions, as in the first case, the user is making a joke, while in the second post, they are referring to the past.

5.5.5. Non-answers

Responses that do not address the question are considered as non-answers and affect classification performance negatively. Examples of non-answers: *The post expresses suicide-related thoughts and emotions*, *The post does not express suicide ideation*. ChatGPT gave

10 such non-answers in the zero-shot setting compared to 0 for GPT-4, but only 1 in the four-shot experiment.

Since the prompt asked a simpler yes/no question for the binary DATD task, there was no such problem in that case.

5.6. Discussion

Interestingly, while GPT-4 far outperforms ChatGPT and all other methods on the suicide ideation task, we observe that the use of ChatGPT and GPT-4 does not provide better results for DATD compared to the multi-task learning experiments. This could be due to some confusion induced by the few-shot examples, since contrarily to the suicide ideation examples which were taken from the dataset, the examples for DATD were adapted directly from the instructions given to the annotators, and special care was not taken to show examples of common ambiguities as was done for suicide ideation.

We find that ChatGPT results are consistent with [24] as the model is outperformed by the BERT baseline from Table 4.1.

The ablation study on the suicide ideation task showed that for both ChatGPT and GPT-4, four-shot learning performs best, and most of the performance boost is obtained from the first example, i.e. from zero- to one-shot learning. ChatGPT benefits the most from the examples, namely because it produces a lot more non-answers than GPT-4 in the zero-shot setting.

While the explanations provided by the GPT models do not constitute interpretability of how the model produced an answer, they can give us cues to help form an opinion on the quality of the predicted label. In Table 5.4, the explanations provided add relevant context to the prediction decisions.

Analysis of the confusion matrix for GPT-4 on the suicide ideation task highlights a recurring theme of "Suicide other" tweets being confused with "Non-related to suicide". Since this pattern was also found in Tables 3.4 and 4.4, it further supports the hypothesis that there is an inherent issue with this class definition and/or how it was labeled.

Since producing a high number of false negatives is arguably the worst result for a suicide ideation classifier, careful inspection of the two false negatives for the class "Suspected suicide ideation" concludes that they were due to labeling errors. The explanations provided by GPT-4 correctly point out that one of the tweet is not related to suicide ideation at all, while the other one does not constitute suspected ideation since it refers to the past.

In line with [18], we observe less inconclusive results, or "non-answers", when using few-shot prompting compared to zero-shot.

Chapter 6

Conclusion

One of the main challenges when developing NLP solutions for suicide ideation and other mental health tasks is data scarcity. In this research, we explored ways to make up for this lack of data using multi-task learning and large language models. Focusing our efforts on suicide ideation detection and evaluating our methods on anxiety and depression detection, we fulfilled our research objectives:

- We performed a manual review of the suicide ideation dataset to highlight discrepancies and ambiguities in the labeling. We also analyzed the labels of duplicate tweets and found that some classes, namely "Suicide other", were seemingly inconsistently labeled.
- We obtained baseline results using classical methods on the suicide ideation and DATD datasets. SVM was found to outperform Random Forest on both tasks.
- We evaluated the use of multi-task learning with publicly-available datasets relating to mental health as auxiliary tasks, and obtained improved classification results on the suicide ideation task.
- By comparing classification results for different task combinations in a multi-task learning setting, we observed a closeness between suicide ideation, depression and anxiety (DATD), since training them together improved classification for both tasks.
- Finally, we tested ChatGPT and GPT-4 on suicide ideation and DATD. We performed an ablation study to measure the impact of few-shot examples on suicide ideation detection, and tested different prompt formulations for DATD. GPT-4 with few-shot examples proved to be the best performing classifier for the suicide ideation task.

Overall, we have gained a better understanding of the suicide ideation task by investigating issues and ambiguities relating to its annotations, analyzing its relationship to other mental health related tasks, and obtaining good classification results using modern LLMs.

Chapter 7

Future Work

A variety of research directions could be explored for both the multi-task learning and the ChatGPT/GPT-4 experiments.

Further multi-task learning experiments could be made using auxiliary classification tasks for non-mental health related risk factors of suicide ideation such as gender, age, education level and marital status [34].

Since risk factors have a potential causal relationship with suicide ideation, a hierarchical multi-task learning architecture to model the causal relations.

Since ChatGPT is much cheaper than GPT-4 to use, there is a good incentive to find ways to boost its performance. Deeper analysis of its most common mistakes could be made in order to craft a better prompt template. The task itself could also be simplified by removing the "Suicide other" class which brings an unnecessary distraction to the task of accurately detecting suicide ideation.

Another way to reduce cost and make the GPT-4 solution a more viable option would be to find a way to compress the prompt context. The few-shot method that we employed introduces a lot of redundancy since the question is repeated for each example. A prompt context that would allow for multiple examples shown to the model without repeating the question could reduce the number of tokens used.

Finally, the concept of multi-task learning could be introduced in a chain-of-thought [51] fashion to incite the model to answer intermediate questions pertaining to common causes of suicide ideation to formulate an assessment of the various risk factors before answering.

References

- [1] Asma Abdulsalam and Areej Alhothali. Suicidal ideation detection on social media: A review of machine learning methods. *arXiv preprint arXiv:2201.10515*, 2022.
- [2] Yemi Aina and Jeffrey L Susman. Understanding comorbidity with depression and anxiety disorders. *The Journal of the American Osteopathic Association*, 106(5 Suppl 2):S9—14, May 2006. ISSN 0098-6151. URL <http://europepmc.org/abstract/MED/16738013>.
- [3] Mostafa M Amin, Erik Cambria, and Björn W Schuller. Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *arXiv preprint arXiv:2303.03186*, 2023.
- [4] Aaron T Beck, Maria Kovacs, and Arlene Weissman. Assessment of suicidal intention: the scale for suicide ideation. *Journal of consulting and clinical psychology*, 47(2):343, 1979.
- [5] Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102, 2017.
- [6] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multi-task learning for mental health using social media text, 2017.
- [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [8] Pete Burnap, Walter Colombo, and Jonathan Scourfield. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 75–84, 2015.
- [9] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.
- [10] Shijie Chen, Yu Zhang, and Qiang Yang. Multi-task learning in natural language processing: An overview. *arXiv preprint arXiv:2109.09138*, 2021.
- [11] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical*

- psychology: From linguistic signal to clinical reality*, pages 51–60, 2014.
- [12] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10, 2015.
- [13] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39, 2015.
- [14] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860, 2018.
- [15] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137, 2013.
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [17] Son Doan, Amanda Ritchart, Nicholas Perry, Juan D Chaparro, Mike Conway, et al. How do you# relax when you’re# stressed? a content analysis and infodemiology study of stress-related tweets. *JMIR Public Health and Surveillance*, 3(2):e5939, 2017.
- [18] Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. Chatgpt for suicide risk assessment on social media: Quantitative evaluation of model performance, potentials and limitations. *arXiv preprint arXiv:2306.09390*, 2023.
- [19] Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*, 2023.
- [20] Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018, 2018.
- [21] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*, 2021.
- [22] Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):11846, 2020.
- [23] Prasadith Kirinde Gamaarachchige and Diana Inkpen. Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI*

- 2019), pages 54–64, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6208. URL <https://aclanthology.org/D19-6208>.
- [24] Bishal Lamichhane. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*, 2023.
- [25] Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, 2022.
- [26] Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708, 2021.
- [27] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. We’re afraid language models aren’t modeling ambiguity. *arXiv preprint arXiv:2304.14399*, 2023.
- [28] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018.
- [29] Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635, 2020.
- [30] Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 39–44, 2019.
- [31] Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7, 2021.
- [32] Sadaf Munir and Veronica Takov. Generalized anxiety disorder. In *StatPearls [Internet]*. StatPearls Publishing, 2022. <https://www.ncbi.nlm.nih.gov/books/NBK441870/>.
- [33] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.
- [34] Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Jordi Alonso, Matthias Angermeyer, Annette Beautrais, Ronny Bruffaerts, Wai Tat Chiu, Giovanni De Girolamo, Semyon Gluzman, et al. Cross-national prevalence and risk factors for suicidal ideation,

- plans and attempts. *The British journal of psychiatry*, 192(2):98–105, 2008.
- [35] Laura Ana Maria Oberländer and Roman Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th international conference on computational linguistics*, pages 2104–2119, 2018.
- [36] Bridianne O’dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. Detecting suicidality on twitter. *Internet Interventions*, 2(2): 183–188, 2015.
- [37] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [38] David Owen, Jose Camacho Collados, and Luis Espinosa-Anke. Towards preemptive detection of depression and anxiety in twitter. *arXiv preprint arXiv:2011.05249*, 2020.
- [39] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.
- [40] Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. *arXiv preprint arXiv:2109.10255*, 2021.
- [41] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.
- [42] Diana Ramírez-Cifuentes, Ana Freire, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Alejandro Velazquez, Josep Maria Gonfaus, and Jordi González. Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. *Journal of medical internet research*, 22(7):e17758, 2020.
- [43] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 99–107, 2015.
- [44] Shailik Sarkar, Abdulaziz Alhamadani, Lulwah Alkulaib, and Chang-Tien Lu. Predicting depression and anxiety on reddit: a multi-task learning approach. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 427–435, 2022. doi: 10.1109/ASONAM55673.2022.10068655.
- [45] Judy Hanwen Shen and Frank Rudzicz. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65, 2017.
- [46] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, 2018.

- [47] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. *Ieee Access*, 7:44883–44893, 2019.
- [48] Julio C Tolentino and Sergio L Schmidt. Dsm-5 criteria and depression severity: implications for clinical practice. *Frontiers in psychiatry*, 9:450, 2018.
- [49] Wally Trenholm and Mai Mavinkurve. Artificial intelligence (ai) pilot project for insight into suicide-related behaviours on social media platforms. Technical report, Sightline Innovation, University of Montreal and McGill University, 2020.
- [50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [52] Victor R Wilburn and Delores E Smith. Stress, self-esteem, and suicidal ideation in late adolescents. *Adolescence*, 40(157), 2005.
- [53] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyang Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with chatgpt. *arXiv preprint arXiv:2304.03347*, 2023.
- [54] Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46, 2022.
- [55] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [56] Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33, 2019.

Appendix A

Prompts

A.1. Few-shot examples for the DATD task

User: Given Post: 'I feel sick to my stomach, I hate having such bad anxiety'. Answer yes or no for whether the user appears to be suffering from depression or anxiety.

Assistant: Yes. The user states that they have anxiety.

User: Given Post: 'Counselling fixed my depression and anxiety'. Answer yes or no for whether the user appears to be suffering from depression or anxiety.

Assistant: No. The user states that they have had depression or anxiety in the past.

User: Given Post: '@user I wish you all the best in beating your anxiety'. Answer yes or no for whether the user appears to be suffering from depression or anxiety.

Assistant: No. The user is referring to a fellow user who may have depression or anxiety.

User: Given Post: 'Nothing gives me anxiety more than the tills at Aldi'. Answer yes or no for whether the user appears to be suffering from depression or anxiety.

Assistant: No. The user is temporarily anxious due to a short or superfluous event.

User: Given Post: 'Depression is not taken serious enough'. Answer yes or no for whether the user appears to be suffering from depression or anxiety.

Assistant: Yes. The post is ambiguous.