Université de Montréal

Classification automatique de commentaires synchrones dans les vidéos de danmaku

par

Youyang Peng

Département de linguistique et de traduction
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès arts (M.A.)
en linguistique

Janvier 2024

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Classification automatique de commentaires synchrones dans les vidéos de danmaku

présenté par

Youyang Peng

a été évalué par un jury composé des personnes suivantes :

François Lareau
(président-rapporteur)

Antoine Venant
(directeur de recherche)

Patrick Drouin
(membre du jury)

Résumé

Le danmaku désigne les commentaires synchronisés qui s'affichent et défilent directement en surimpression sur des vidéos au fil du visionnement. Bien que les danmakus proposent à l'audience une manière originale de partager leur sentiments, connaissances, compréhensions et prédictions sur l'histoire d'une série, etc., et d'interagir entre eux, la façon dont les commentaires s'affichent peut nuire à l'expérience de visionnement, lorsqu'une densité excessive de commentaires dissimule complètement les images de la vidéo ou distrait l'audience.

Actuellement, les sites de vidéo chinois emploient principalement des méthodes par mots-clefs s'appuyant sur des expressions régulières pour éliminer les commentaires non désirés. Ces approches risquent fortement de surgénéraliser en supprimant involontairement des commentaires intéressants contenant certains mots-clés ou, au contraire, de sous-généraliser en étant incapables de détecter ces mots lorsqu'ils sont camouflés sous forme d'homophones. Par ailleurs, les recherches existantes sur la classification automatique du danmaku se consacrent principalement à la reconnaissance de la polarité des sentiments exprimés dans les commentaires.

Ainsi, nous avons cherché à regrouper les commentaires par classes fonctionnelles, à évaluer la robustesse d'une telle classification et la possibilité de l'automatiser dans la perspective de développer de meilleurs systèmes de filtrage des commentaires.

Nous avons proposé une nouvelle taxonomie pour catégoriser les commentaires en nous appuyant sur la théorie des actes de parole et la théorie des gratifications dans l'usage des médias, que nous avons utilisées pour produire un corpus annoté. Un fragment de ce corpus a été co-annoté pour estimer un accord inter-annotateur sur la classification manuelle.

Enfin, nous avons réalisé plusieurs expériences de classification automatique. Celles-ci comportent trois étapes : 1) des expériences de classification binaire où l'on examine si la machine est capable de faire la distinction entre la classe majoritaire et les classes minoritaires,

2) des expériences de classification multiclasses à granularité grosse cherchant à classifier les commentaires selon les catégories principales de notre taxonomie, et 3) des expériences de classification à granularité fine sur certaines sous-catégories. Nous avons expérimenté avec des méthodes d'apprentissage automatique supervisé et semi-supervisé avec différents traits.

Mots clés : Danmaku ; annotation du corpus ; taxonomie ; traitement automatique des langues ; classification automatique de textes ; apprentissage automatique

Abstract

Danmaku denotes synchronized comments which are displayed and scroll directly on top of videos as they unfold. Although danmaku offers an innovative way to share their sentiments, knowledge, predictions on the plot of a series, etc., as well as to interact with each other, the way comments display can have a negative impact on the watching experience, when the number of comments displayed in a given timespan is so high that they completely hide the pictures, or distract audience.

Currently, Chinese video websites mainly ressort to keyword approaches based on regular expressions to filter undesired comments. These approaches are at high risk to overgeneralize, thus deleting interesting comments coincidentally containing some keywords, or, to the contrary, undergeneralize due to their incapacity to detect occurrences of these keywords disguised as homophones. On another note, existing research focus essentially on recognizing the polarity of sentiments expressed within comments.

Hence, we have sought to regroup comments into functional classes, evaluate the robustness of such a classification and the feasibility of its automation, under an objective of developping better comments filtering systems.

Building on the theory of speech acts and the theory of gratification in media usage, we have proposed a new taxonomy of danmaku comments, and applied it to produce an annotated corpus. A fragment of the corpus has been co-annotated to estimate an inter-annotator agreement for human classification.

Finally, we performed several automatic classification experiments. These involved three steps: 1) binary classification experiments evaluating whether the machine can distinguish the most frequent class from all others, 2) coarse-grained multi-class classification experiments aiming at classifying comments within the main categories of our taxonomy, and

3) fine-grained multi-class classification experiments on specific subcategories. We experimented both with supervised and semi-supervised learning algorithms with diffrent features.

Keywords: Danmaku; corpus annotation; taxonomy; natural language processing; automatic text classification; machine learning

Table des matières

Liste des tableaux

Table des figures

Liste des sigles et des abréviations

ACP	analyse en composantes principales.
BERT	Représentations bidirectionnelles d'encodeurs à partir de transformateurs (en anglais: *bidirectional Encoder Representations from Transformers).*
CBOW	sac de mots continu (en anglais: *continuous bag of words*).
FBR	fonction de base radiale.
LSTM	Réseaux de Mémoire à Long Terme et de Court Terme (en anglais : Long Short-Term Memory (LSTM) networks).
Onehot	encodage un parmi n.
SDM	sac de mots.
SDM Fréquence	Sac-de-mots de Fréquence.
SDM TF-IDF	Sac-de-mots de TF-IDF.
SDM Word2Vec	Sac-de-mots Word2Vec.
SVM	Séparateur à vaste marge.
TF-IDF	fréquence de terme-fréquence inverse de document.

W2V Word2vec.

Remerciements

Ce travail de mémoire n'aurait pas pu être accompli sans le soutien des personnes suivantes :

Je tiens à remercier ma famille, en particulier ma mère et ma sœur, pour leur soutien et leur compréhension inconditionnels dans tous les aspects de ma vie depuis toujours.

Je suis surtout reconnaissante à mon directeur, Dr Antoine Venant, qui m'a donné beaucoup de courage tout au long de mes études de maîtrise, m'a accordé de précieux conseils sur mon mémoire, et a généreusement consacré son temps à répondre à mes questions ainsi qu'à examiner mon travail. En tant que directeur de recherche, il est toujours optimiste et patient tout en restant rigoureux dans le travail et la recherche.

J'aimerais aussi exprimer ma profonde gratitude envers Dr François Lareau, qui m'a offert l'opportunité de poursuivre mes études à l'Université de Montréal. Sans lui, je n'aurais pas pu réaliser mon rêve d'enfance. Je lui suis également reconnaissante d'avoir établi des relations très amicales entre les professeurs et étudiants. Je n'oublierai jamais le temps que j'ai passé à l'OLST et les moments de repas partagés entre amis et professeurs.

De plus, je suis redevable à Dr Patrick Drouin pour sa volonté de faire partie de mon jury de mémoire et d'investir son temps à la révision de mon mémoire.

Enfin, je ne peux pas passer sous silence l'accompagnement de mes amis. Ils constituent tous ma source de bonheur. Merci à Li Liu, ma chère colocataire et *camarade*. Merci à Jingyun Song avec qui j'ai partagé de nombreux moments émotionnels. Merci à Shiyu Li qui est un ami toujours très serviable et amusant. Merci à tous les amis de l'OLST qui m'ont donné beaucoup d'assistance dans mes études et les amis que j'ai rencontrés à Montréal pour être toujours là avec moi.

Chapitre 1

Introduction

Avec l'expansion des médias sociaux au cours de ces dernières années, l'audience des médias est devenue plus active que jamais. Contrairement aux médias traditionnels dont la relation entre le contenu et l'audience est unidirectionnelle (l'audience ne peut que recevoir les informations), les utilisateurs des médias sociaux disposent de différents moyens pour s'exprimer : sur des sites de vidéos, par exemple, les utilisateurs ont la possibilité de commenter, cliquer sur « aimer » ou « ne pas aimer », partager, s'abonner, et ainsi de suite. Le danmaku, une forme particulière de commentaire proposée par certains sites japonais et chinois, a largement contribué à accroître leur audience.

Le danmaku réfère à un format de commentaires défilant de manière superposée aux images des vidéos. L'application de ce genre de commentaire est omniprésente, notamment sur les sites de vidéos japonais et chinois. Depuis sa naissance, les particularités du danmaku ont été visées par plusieurs études.

《中国网络评论发展报告（2019）》(Rapport annuel sur le développement des commentaires en ligne en Chine (2019)) (Zhao, 2019) indique que la consultation de commentaires est devenue une partie importante des activités en ligne du fait que le nombre des commentaires est une indication de la popularité et de la qualité du contenu de l'information (normalement, plus le contenu est prisé et bien accueilli par l'audience, plus de commentaires il reçoit) mais que le contenu des commentaires peut aussi être significatif dans le sens où il peut tout à fait exister beaucoup de commentaires défavorables ou futiles pour un contenu de média.

Généralement, l'audience traite le contenu des médias de deux manières : la consommation et la participation (Khan, 2017). La consommation consiste en la lecture du contenu textuel ou audiovisuel (blog, vidéo, commentaire des autres utilisateurs, etc.). La participation consiste en l'interaction entre les utilisateurs et le contenu ainsi que l'interaction entre les utilisateurs : la publication et le partage du contenu ainsi que l'expression de ses opinions sur le contenu. Il faut remarquer que le commentaire, souvent considéré uniquement comme une façon d'exprimer l'opinion, est dorénavant également désigné comme le *contenu généré par les utilisateurs* (en anglais : *user-generated content*), faisant l'objet de la navigation des autres spectateurs. Les gens consomment les informations quand ils lisent un commentaire. Quand ils publient un commentaire, ils participent à l'interaction contenu-utilisateur et/ou utilisateur-utilisateur. La notion d'audience active, qui ne concernait que la sélection et l'interprétation proactive du contenu (Livingstone, 2003), fait référence désormais à une audience autonome, qui participe activement à l'interaction. La théorie des usages et gratifications (Katz & Blumler, 1974) suppose que l'audience opte consciemment pour (via consommation ou participation) le média qui pourrait répondre à ses besoins et qu'elle a conscience des raisons de ses choix médiatiques.

Nous nous demandons quelles sont les motivations derrière les danmakus sur les sites de vidéos. Dans le cadre de la théorie des usages et gratifications, Katz et al. (1973) ont établi une classification en cinq catégories pour les besoins psychologiques et sociaux relatifs à l'usage des médias. Ces derniers se déclinent comme suit :

(1) Besoins cognitifs : besoins liés au renforcement de l'information, de la connaissance et de la compréhension. Par exemple, les gens regardent les nouvelles pour obtenir des informations factuelles.

(2) Besoins affectifs : besoins liés au renforcement de l'expérience esthétique, plaisante et émotionnelle. À titre d'exemple, les gens regardent des comédies afin de s'octroyer une source d'amusement.

(3) Besoins personnels-intégratifs : besoins liés au renforcement de la crédibilité, de la confiance, de la stabilité et du statut. Cela concerne la validation de l'identité personnelle. Par exemple, les documentaires historiques suscitent souvent la fierté du peuple d'un pays spécifique en renforçant leur identité nationale.

(4) Besoins sociaux-intégratifs : besoins liés au renforcement du contact avec la famille, les amis et le monde, pour illustrer, les gens regardent une série ou un film pour pouvoir en discuter avec les autres.

(5) Besoins de détente : besoins liés à l'évasion ou à la libération de la tension. Par exemple, les gens ont recours au divertissement léger, tel que la musique populaire ou les émissions de variétés sans nécessairement attendre de bénéfices intellectuels ou émotionnels.

La publication de danmakus sur les sites de vidéos permet donc vraisemblablement de satisfaire, d'une part, les besoins de l'auteur des commentaires, et d'autre part, ceux des lecteurs.

Par ailleurs, même si le danmaku constitue un bon mécanisme de partage d'émotion, d'interaction et d'intégration, il est aussi critiqué pour disperser l'attention des spectateurs : ces derniers risquent de manquer du contenu intéressant en lisant les danmakus. Parfois, les commentaires sont si nombreux qu'ils cachent complètement les images derrière, nuisant ainsi considérablement à l'expérience de visionnement. Par conséquent, il serait bénéfique que les commentaires indésirables puissent être filtrés pour ne garder que les commentaires qui répondent aux besoins spécifiques des spectateurs. La méthode la plus répandue pour enlever les commentaires non désirés est l'utilisation d'expressions régulières pour masquer les commentaires contenant les termes définis par un utilisateur. Par exemple, bilibili [1] (un site de partage de vidéo) permet à ses utilisateurs de masquer les commentaires reconnus par une expression régulière dans la configuration de danmaku au-dessous de la vidéo. Pourtant, cette méthode a des limites manifestes : (1) il y a une forte probabilité d'enlever par erreur les commentaires intéressants comprenant un certain mot-clé ; (2) les gens peuvent tout à fait utiliser des homophones quand ils rédigent un commentaire pour échapper à la capture.

1. https ://www.bilibili.com/

Ainsi, nous essayons de proposer un alternative qui traitent les commentaires indésirables d'une façon plus subtile. L'effet idéal serait que les utilisateurs soient capables de masquer les commentaires qui ne les intéressent pas ou qui ne sont pas en rapport avec leur propre commentaire.

Pour ce faire, il nous faut établir une taxonomie explicite des motivations et des fonctions des danmakus. Mais les recherches existantes sur la classification automatique de danmaku se concentrent majoritairement sur la reconnaissance de la polarité des sentiments exprimés dans les commentaires, faisant l'hypothèse implicite que tous les commentaires ont pour fonction d'exprimer un sentiment ou une opinion. Dans le cadre de ce projet, nous proposons, pour la première fois, une taxonomie plus générale visant à catégoriser les danmakus selon leur fonction, en nous inspirant de la théorie des gratifications dans l'usage des médias et de la théorie des actes de parole. Notre contribution réside également dans la création d'un corpus annoté fondé sur les fonctions des danmakus à des fin de classification, ainsi que dans l'entraînement et l'évaluation de plusieurs modèles de référence (machines à vecteurs de supports, combinés avec des traits discrets ou des plongements lexicaux) pour différentes granularités de la tâche de classification.

Dans un premier temps, nous avons essayé de relever les caractéristiques importantes des danmakus et les regrouper par classes de similarité. Sur la base de ce travail et de certains concepts de la théorie des actes de parole, nous avons proposé une taxonomie pour catégoriser les danmakus. Par la suite, une partie de notre corpus a été annotée manuellement afin de vérifier la solidité de la taxonomie et, de construire un corpus annoté pour la classification automatique. En vue d'examiner la fiabilité de l'annotation, une co-annotation a été faite sur une partie des annotations initiales, et plusieurs scores d'accord d'inter-annotateurs ont été calculés. L'accord obtenu est assez bon, au moins pour certains niveaux de granularité dans les annotations. Les données annotées ont été utilisées dans des expériences de classifications automatiques, qui se déclinent en trois parties : des expériences de classification binaire, des expériences de classification multiclasses à granularité grosse , et des expériences de multiclasses à granularité fine. Pour chaque partie, nous avons appliqué des méthodes

d'apprentissage supervisé et semi-supervisé sur des vecteurs de mots formés de diverses manières allant de l'encodage Onehot aux plongements lexicaux, en passant par la pondération TF-IDF.

Le reste du mémoire est organisé comme suit : d'abord, nous introduisons le terme de danmaku, en mettant en évidence leurs caractéristiques distinctives. Et puis, nous présentons notre corpus, en détaillant sa constitution et le processus de pré-traitement auquel il a été soumis. Ensuite, nous développons notre taxonomie, et nous examinons dans quelle mesure les êtres humains sont capables de distinguer les différentes catégories de commentaires, et si cette catégorisation peut également être réalisée de manière automatique. Enfin, nous résumons les résultats obtenus, leurs limites et entamons une discussion sur les pistes de recherche futures.

1.1. Le danmaku

1.1.1. Définition du danmaku

Le danmaku, littéralement « rideau de balles », est un type de commentaire synchronisé qui défile de droite à gauche sur l'écran comme réaction à une scène spécifique ou interagir avec d'autres utilisateurs (cf. **Figure** 1.1). Dans ce mémoire, le terme « danmaku » est utilisé pour désigner le média, le système de sous-titre à part entière, alors que sa forme plurielle « danmakus », fait référence aux commentaires individuels.

Il s'agit originellement un terme militaire signifiant une frappe intense de tirs concentrés sur une zone spécifique pour former un barrage (« japandict.com », s. d.). Ce terme est ensuite introduit dans les jeux vidéos de tir. En 2006, le site de partage de vidéos japonais Niconico a utilisé pour la première fois le terme *danmaku* pour désigner l'effet de « mitraillage » de commentaires sur l'écran (comme dans les jeux de tirs), puis, petit à petit, le terme a commencé à être utilisé pour faire référence aux commentaires eux-mêmes. Depuis 2008, la fonction de commentaires danmaku a été importée en Chine par les sites de streaming vidéo tels que Bilibili, Acfun, Youku, mais aussi par différentes autres plateformes comme celles de vidéos courtes ou de vidéos diffusées en direct.

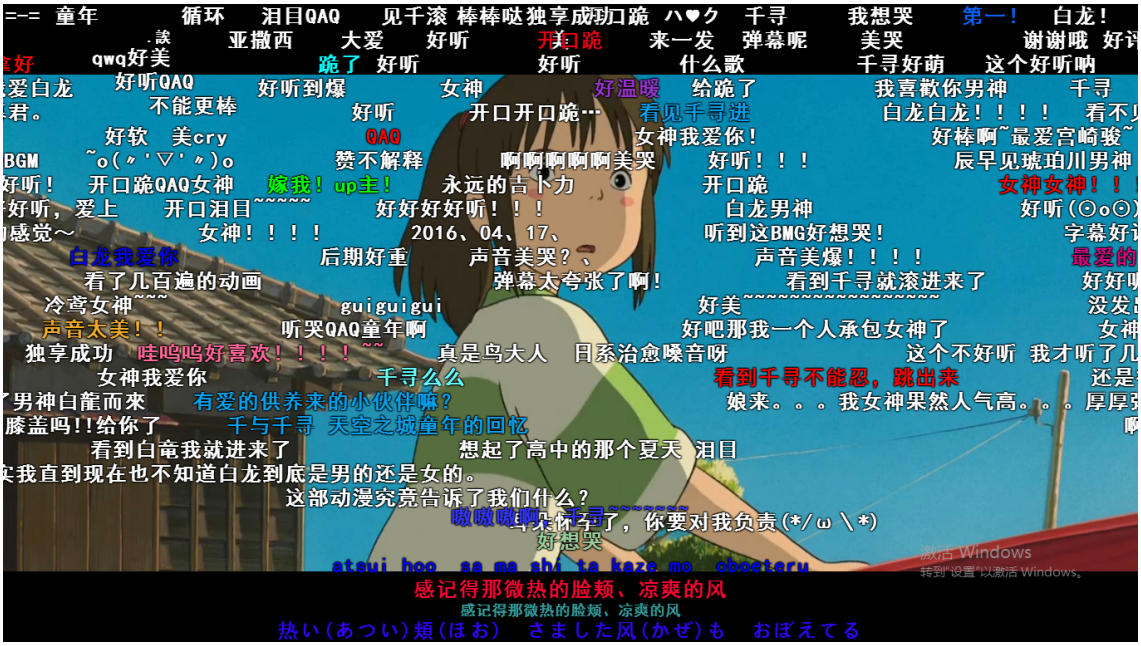

Figure 1.1 – Une capture d'écran de vidéo avec commentaire de danmaku en Mandarin superposé sur une image de vidéo

1.1.2. Caractéristiques du danmaku

Plusieurs caractéristiques distinguent le danmaku des commentaires traditionnels :

(1) Interactivité

Le danmaku, en tant que forme de contenu médiatique, est intrinsèquement interactif, qu'il s'agisse d'une interaction explicite ou implicite. L'interaction explicite se traduit souvent par des questions, des réponses, des salutations, des ordres, et ainsi de suite, alors que l'interaction implicite s'exprime sous forme de monologue. Par exemple, lorsqu'un spectateur émet un jugement sur une scène, tel que « Ça m'écœure. », bien que l'interlocuteur ne soit pas explicitement convoqué dans la phrase, l'auteur a vraisemblablement l'intention implicite de susciter une résonance émotionnelle chez les autres.

(2) Simultanéité

À la différence des commentaires traditionnels qui se publient souvent après le visionnement de la vidéo et se trouvent dans une zone fixe (normalement sous la vidéo), les danmakus sont affichés pendant la lecture des vidéos et se superposent avec les images de la vidéo. Dès leur publication, les danmakus défilent de droite à gauche jusqu'à ce qu'ils disparaissent de l'écran (ceux qui sont plus longs se déplacent plus vite). La simultanéité se résume en deux points : 1) l'affichage d'un commentaire au même secteur temporel que la scène qu'il commente (cf. **Figure** 1.2) ; 2) la proximité temporelle des commentaires reliés. La raison est que l'interprétation d'un danmaku dépend de son contexte, qui se compose des commentaires qui sont autour, et de la position temporelle du commentaire (dans la vidéo). Si un commentaire est affiché trop en retard, il y a de fortes chances qu'il devienne inintelligible et qu'il ne reçoive pas la réaction attendue. Il est à noter que le temps d'affichage des commentaires est indépendant du moment auquel ces commentaires sont rédigés et publiés dans la réalité. Par exemple, un commentaire A qui critique une scène aux dix premières secondes d'une vidéo peut apparaître approximativement en même temps qu'un commentaire B, qui concerne cette même scène, même si le commentaire A est publié un mois avant le commentaire B (toutefois, certains vieux danmakus sont parfois éliminés par la plateforme pour économiser des ressources). Il en résulte une expérience de co-visionnement pseudo-synchronisé.

(3) Anonymat

Contrairement aux commentaires traditionnels, les danmakus sont anonymes : d'un côté, l'audience est ignorante de l'identité des auteurs des commentaires (leur nom d'utilisateur ou toute autre information), de l'autre côté, quand ils publient un commentaire, leur identité est aussi cachée. De ce fait, les danmakus sont dépourvus de fonction de mention (l'usage du symbole « @ » suivi du nom de l'utilisateur auquel l'on veut s'adresser, ce qui est très commun sur les médias sociaux tels que X (Twitter) ou Youtube). De ce fait, la détection des interactions explicites est plus délicate.

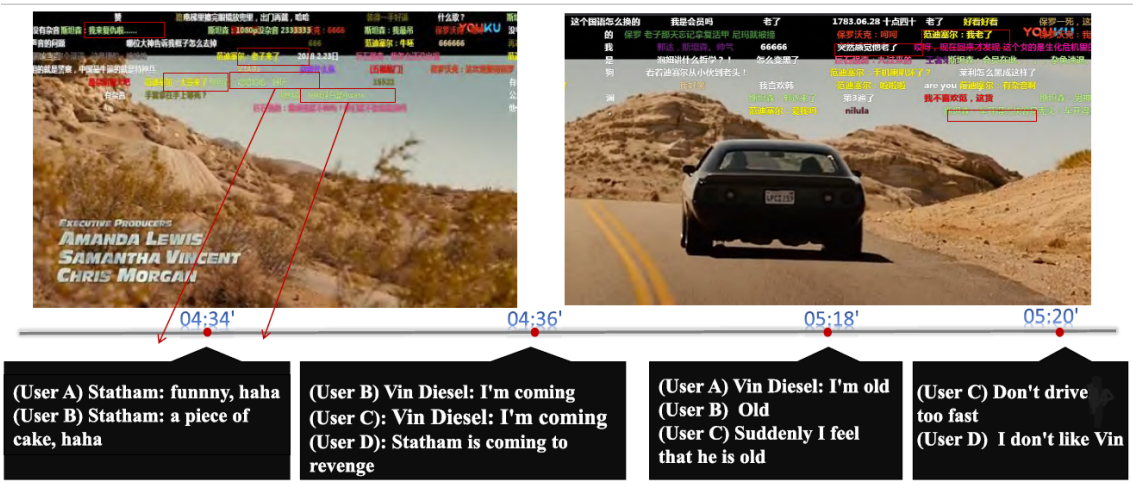

Figure 1.2 – Des danmakus publiés aux points temporels différents dans une vidéo (Bai et al., 2019).

(4) Spontanéité

En raison du temps d'affichage limité des commentaires dont nous avons parlé plus haut, et du contenu restreint sur lesquels ils portent (typiquement, le lecteur ou l'auteur d'un commentaire n'ont vu qu'une section au lieu de l'ensemble de la vidéo, contrairement à ce qui passe avec les commentaires traditionnels qui sont le plus souvent rédigés et lus après que nous aovns visionné l'intégralité de la vidéo), les danmakus véhiculent généralement des messages courts et spontanés plutôt qu'une vision globale. De plus, le caractère anonyme des commentaires encourage cette spontanéité : les utilisateurs peuvent s'exprimer sans beaucoup de réflexion, car la perception de leur commentaire ne rejaillit pas directement sur leur image.

Ces caractéristiques globales de danmaku les distinguent des commentaires traditionnels et nous poussent à explorer davantage ce que les utilisateurs accomplissent en publiant ce genre de commentaires particuliers.

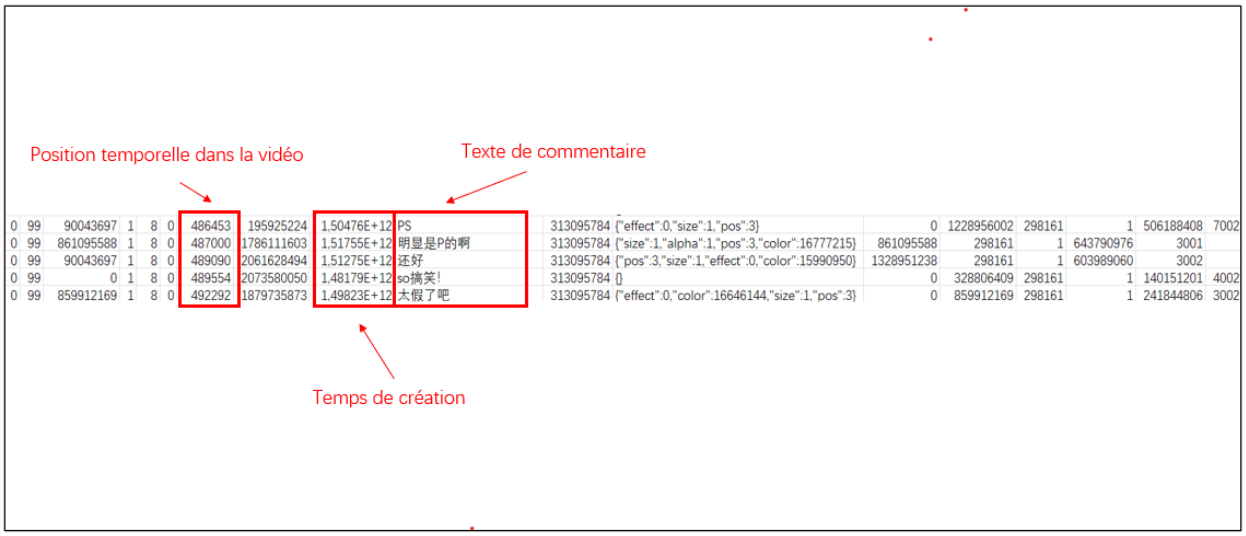

Figure 1.3 – Un extrait du corpus original et les trois colonnes gardées pour la recherche

1.2. Corpus et annotation

1.2.1. Description du corpus

Le corpus utilisé dans cette recherche a été partagé avec nous par les auteurs de l'article « Stories That Big Danmaku Data Can Tell as a New Media » (Bai et al., 2019). Ces données ont été aspirées depuis le site de vidéos chinois Youku, l'un des plus importants sites de vidéos en Chine, entre janvier et avril 2018. Ce corpus chinois comprend 8 156 fichiers au format .tsv incluant 57 176 457 commentaires et 1 383 319 tokens, émis par 6 259 558 utilisateurs. Un fichier de vidéo englobe en moyenne 7 010 danmakus dont la longueur moyenne est de 9 caractères. Il s'agit du plus grand jeu de données de danmaku existant au moment de la rédaction de ce mémoire. Chaque fichier contient des lignes de textes de commentaires et leurs méta-données issues d'un épisode. Nous n'avons conservé que trois colonnes pertinentes pour la recherche, soit la date de création du commentaire (heure Unix : temps écoulé depuis le 1er janvier 1970 00:00:00 UTC, en milisecondes), le texte du commentaire, et la position temporelle relative du commentaire dans la vidéo (en millisecondes) (voir la **Figure** 1.3).

Les genres de séries représentés sont très variés dans ce corpus : les séries dramatiques sont les plus nombreuses, suivies par les séries romantiques.

Les commentaires sont courts. Ils sont majoritairement d'une longueur de quatre caractères (il s'agit principalement de mots ou de petites locutions). Cela est dû, d'une part, à la politique de plateforme de Youku (une longueur maximale de 40 caractères était autorisée. Les commentaires dépassant cette limite peuvent être publiés mais seuls les 40 premiers caractères seront affichés à l'écran), d'autre part, comme nous l'avons mentionné précédemment, à la spontanéité des commentaires.

À l'instar des langages sur d'autres plateformes d'Internet, les danmakus se caractérisent par des occurrences fréquentes d'expressions informelles, comme des émoticônes, des émojis, des termes en vogue, des jeux de mots homophoniques, des mélanges de langues, etc..

Bien que la position temporelle des commentaires et la position relative des commentaires reliés soient essentielles pour la compréhension des commentaires, nous n'avons pas libre accès aux ressources originales sur Youku. D'un côté, beaucoup de séries sur Youku sont inaccessibles au Canada en raison des droits d'auteur. D'un autre, pour raison de lisibilité des images de vidéo, Youku supprime régulièrement des danmakus trop anciens pour que le nombre total de danmakus dans une vidéo ne dépasse pas une certaine limite (qui peut varier selon la longueur d'une vidéo). De ce fait, certains danmakus sont susceptibles d'avoir été enlevés de la vidéo depuis l'aspiration du corpus, s'il y a trop de commentaires dans cette série.

1.2.2. Statistiques et biais

La taille des fichiers dans le corpus peut varier dramatiquement, allant de 1 jusqu'à 30822 commentaires. Cela peut s'expliquer normalement par la popularité d'une vidéo : une vidéo plus connue est susceptible d'avoir une audience plus large et de déclencher plus de réactions de spectateurs et d'interactions.

Afin d'avoir un aperçu global de la distribution des mots, nous avons généré un nuage de mots (cf. **Figure** 1.4) à partir de l'ensemble du corpus, avec la classe `WordCloud` de la bibliothèque de Python `wordcloud`, les mots vides (une liste combinant les listes de mots

vides d'HIT (« HIT Stopwords », 2019) et de NLTK (Bird et al., 2009)) ont été exclus avant cette procédure. Plus un mot est fréquent dans le corpus, plus sa taille est grande dans la représentation. On peut constater que les mots les plus fréquents sont les noms de personnes (de gens dans la vie réelle ou de personnages dans une série). Ces mots proviennent très probablement de pratiques de "flood" (la publication massive de messages pour faire en sorte qu'ils inondent l'écran pour des raisons diverses, ce qui est souvent ennuyant voire agaçant pour les autres spectateurs). Il s'agit typiquement de messages que veut esquiver l'audience en regardant une vidéo car ils suscitent rarement la sympathie chez les spectateurs. Par ailleurs, le corpus contient de nombreux *hapax* (les mots qui ne se présentent qu'une seule fois dans le corpus), il s'agit souvent de fautes typographiques, d'abréviations, de kaomojis (terme japonais qui se traduit littéralement comme caractère de visage, désignant des constructions qui se servent de caractères ASCII pour exprimer des émotions) et de néologismes.

Figure 1.4 – Nuage de mots sur l'ensemble du corpus de danmaku dans lequel le mot le plus fréquent est « 武盼婷 », un nom de fille commun dans la vie réelle.

1.2.3. Pré-traitement

Le corpus original contenait des commentaires en double dans chaque fichier, donc la première étape a consisté à enlever ces répétitions. Comme le corpus est divisé en différents fichiers, et qu'il n'y a pas d'indexation globale des commentaires dans le corpus original, nous avons également ajouté dans tous les fichiers une colonne d'index. L'index de chaque commentaire est une paire, dont le premier élément est le numéro de fichier qui le contient (chaque fichier/épisode a un numéro spécifique) et le second élément est la position du

commentaire dans le fichier. Les deux informations sont séparées par le symbole « # ».

Pour la segmentation des textes de commentaire, on s'est servi de l'implémentation Python de `Jieba` (Sun, 2019), un outil dédié à la segmentation de textes chinois. Nous avons appliqué le mode précis dans `Jieba` pour faire des segmentations plus exactes. Le principe de l'algorithme de segmentation est de générer un graphe acyclique orienté en utilisant une méthode de balayage de graphe de mots basée sur un dictionnaire de préfixes chinois. Un algorithme de programmation dynamique est ensuite appliqué pour trouver la segmentation la plus probable en fonction de la fréquence des mots. Pour gérer les mots non répertoriés, le système modélise la distribution des séquences de caractères en fonction des choix de segmentation avec un modèle de Markov caché et utilise l'algorithme de Viterbi pour trouver la segmentation maximisant la vraisemblance de la séquence de caractères observée. Nous étions globalement satisfaits du résultat de la segmentation après une observation préliminaire. Cependant, nous avons quand même repéré des erreurs évidentes, surtout lorsqu'il s'agit de mots ou expressions relativement rares. Par exemple, les mots rares décrivant les armes tels que : « 穿刺力差 » :

chuān cì lì chà
穿刺 力 差。
perforation force mauvais
'La force de perforation est mauvaise.'

Celui-ci est segmenté à tort comme « 穿刺/力差 ». Ce genre de problème se produit souvent dans le cas d'une séquence de mots où les premiers caractères ont une forte probabilité a priori de former un mot. Dans ce cas-là, les caractères derrière vont former un mot inconnu erroné : dans notre exemple, « 穿刺力差 », les deux premiers caractères sont susceptibles de former le mot « 穿刺 » (la perforation), ce qui mène à la combinaison insignifiante de caractères « 力差 », alors que la bonne segmentation est « 穿刺力/差 ».

En observant les textes des commentaires, nous pouvons remarquer qu'il existe des émojis, des émoticônes et des séquences de chiffres constituant des mots d'argot internet comme « 66666 » (ce qui signifie « bien fait », peu importe le nombre de « 6 » dans l'expression, par

exemple, ces variations—« 666 » et « 666666666666 » ont le même sens). Ces séquences sont bien segmentées comme des unités textuelles, de même que la ponctuation et les nombres. Ces données ont été conservées car elles sont pertinentes pour notre tâche de classification. De plus, nous n'avons pas supprimé les mots vides et la ponctuation au cours du pré-traitement car nous avons pré-entraîné des modèles de plongements lexicaux que le retrait des mots vides pourraient influencer.

Pour les besoins de notre recherche, nous avons extrait du corpus original trois colonnes qui étaient jugées utiles : la date de création du commentaire, le texte du commentaire, ainsi que la position temporelle relative du commentaire dans la vidéo. Cela nous a permis de rendre compte du contexte des commentaires, c'est-à-dire des commentaires temporellement proches du commentaire en question, et de la position temporelle de chaque commentaire dans la vidéo.

1.3. Catégorisation de danmaku

En observant les commentaires, nous pouvons remarquer que leur contenu est très varié. À première vue, nous pouvons discriminer les messages subjectifs (souvent sous forme de critique) des messages informatifs (qui ont pour objet de donner des informations sans exprimer un sentiment). Étant donné le caractère *interactif* et *social* des danmakus, il n'est pas difficile de déceler que chaque commentaire est doté plus ou moins d'une certaine intention du locuteur : ce dernier veut *faire* quelque chose à travers ses mots. Autrement dit, chaque commentaire est un acte de parole. Ainsi, nous pouvons baser une taxonomie sur les actes de paroles des commentaires.

1.3.1. Les danmakus et leur acte illocutoire

Les commentaires peuvent exister sous différentes formes, soit déclarative, interrogative ou bien impérative. Dans le cadre de cette recherche, la catégorisation est fondée sur l'acte *illocutoire* (Austin, 1962) des commentaires. Austin (1962) définit la réalisation d'un acte illocutoire comme « la réalisation d'acte *en* disant quelque chose » et il réfère aux différents types de fonction du langage comme *forces illocutoires*. Prenons un exemple dans le corpus

de danmaku : en disant « Quelle est la musique d'ouverture ? », le locuteur demande à son interlocuteur le nom de la musique d'ouverture de la série, ce qui constitue un acte illocutoire.

Cerner l'acte illocutoire d'un commentaire consiste à comprendre la *véritable* intention communicative de son auteur. La détection de l'acte illocutoire est relativement simple quand celui-ci est direct : Yule (1996) indique que lorsqu'il existe une relation directe entre la forme de l'énoncé (par exemple, trois types de phrase élémentaires telles que déclarative, interrogative et impérative) et sa fonction (assertion pour la structure déclarative, question pour l'interrogative, et commande pour l'impérative), il s'agit là d'un *acte de parole direct*. En suivant cette définition, l'exemple « Quelle est la musique d'ouverture ? » donné plus haut est un acte de parole direct, car l'énnoncé revêt une forme interrogative, et l'acte illocutoire correspondant est de poser une question. Mais, toujours d'après Yule (1996), si aucune relation directe ne se trouve entre la structure et la fonction, nous avons alors un *acte de parole indirect*. Autrement dit, la structure de ces énoncés ne traduit pas directement leur force illocutoire. Dans les termes de Chierchia (1990), *la force de phrase*, force signalée par la forme de la phrase, se distingue de la force illocutoire.

Ainsi, les questions rhétoriques, qui n'attendent pas de réponse, sont considérées comme des assertions ayant une polarité négative (Han, 2002). Prenons comme exemple la phrase suivante :

yī kǒu qì pǎo shí lǐ
一 口 气 跑 十 里？
un bouchée souffle courir dix li
'Courir dix lis (500 mètres) d'un seul souffle ?'

Bien que la phrase prenne une forme interrogative (typiquement associée à une force illocutoire d'interrogation), le locuteur veut en réalité *affirmer* qu'il est hors de question pour le personnage de courir 500 mètres d'un souffle. L'utilisation d'une question rhétorique n'est qu'un renforcement de soupçon ou de négation.

Toutefois, cette définition de l'acte de parole indirect a des limitations car les forces de phrase ne sont pas toutes universelles à travers les langues d'une part, et ne sont pas aussi nombreuses que les actes de parole d'autre part. Alternativement, Searle (1979) a proposé

une autre définition de l'acte de parole indirect (primaire) selon laquelle ce dernier est réalisé au moyen de l'exécution d'un autre acte illocutoire (secondaire).

Searle (1979) distingue l'illocution—première intention réelle du locuteur en énonçant une phrase donnée, que l'on peut rapprocher du *sens locuteur* chez Grice (ce que le locuteur *veut dire*), de l'illocution secondaire (littérale), qui est déterminée par le sens phrastique (sens littéral) de l'énoncé.

La complexité réside toujours dans l'identification de l'acte de parole indirect. Et pour ce faire, Searle (1979) a suggéré quelques outils nécessaires incorporant la théorie de l'acte de parole, certains principes généraux de conversation coopérative (Grice, 1975), ainsi que l'information contextuelle partagée entre le locuteur et l'interlocuteur et en dernier, l'aptitude à l'inférence de l'interlocuteur.

En effet, les actes de langage indirects présentent d'importants défis dans le cadre de notre tâche d'annotation. Et nous allons illustrer amplement ce point à l'aide de davantage d'exemples concrets dans la section 2.4 du chapitre suivant.

1.3.2. Critères de classification

Nous avons assigné une étiquette *None (Nul)* aux commentaires dépourvus de sens défini, ce qui est dû notamment à l'incomplétude de la phrase, ou aux erreurs typographiques. Par contre, aucune étiquette n'est donnée aux commentaires qui sont sémantiquement explicites, mais difficile à catégoriser dû au manque de contexte spécifique. Par exemple, « 任天堂 » (*Nintendo*). Dans cet exemple, nous sommes bien conscients que le locuteur parle d'une compagnie japonaise de jeu vidéo. Pourtant, nous ne parvenons pas à déterminer avec certitude la force illocutoire de cet énoncé en nous appuyant seulement sur les commentaires qui sont autour. Pour les autres commentaires, il se peut qu'un commentaire soit associé à plusieurs classes, sachant qu'un énoncé est susceptible d'effectuer plus d'un acte illocutoire, ou d'être ambigu.

Pour la classification, nous définissons en premier lieu les catégories principales selon trois axes : **la subjectivité, la pertinence et la temporalité**.

La **subjectivité** est un concept difficile à cerner, car tous les mots choisis par un locuteur sont indicatifs de son attitude et de sa stratégie rhétorique. Cet aspect de la subjectivité est désigné dans Smet et Verstraete (2006) comme la « subjectivité pragmatique ». Une autre subjectivité concerne les caractéristiques sémantiques d'une expression, à savoir la « subjectivité sémantique » selon Smet et Verstraete (2006). Par exemple, dans le cadre de la linguistique cognitive, la subjectivité constitue la référence au locuteur dans l'interprétation d'une construction (Traugott, 1989) ou la mention implicite du locuteur dans un élément d'énoncé (Langacker, 1985). Dans ces cas-là, l'insertion de la perspective du locuteur influence les traits sémantiques des expressions (subjectification). Par exemple, dans la phrase « The earthquake is going to destroy the city » (Le tremblement de terre va détruire la ville), le sens de la construction « is going to » (va) évolue d'une signification purement spatiale vers une signification impliquant le temps. Apparemment, le tremblement de terre, qui est inanimé, n'est pas capable de se déplacer. La construction « be going to » reflète plutôt *le chemin mental* du locuteur, c'est-à-dire, le *balayage mental* d'après Langacker (1990). Il existe également cette forme de subjectivité dans la langue chinoise. He (2014) en a donné un exemple avec le mot « 所以 ». Originellement, ce mot était employé comme une locution prépositionnelle signifiant « l'endroit pour faire quelque chose », comme dans la phrase suivante :

zhōng xìn suǒ yǐ jìn dé yě
忠 信, 所 以 进 德 也。

fidélité sincérité endroit pour.faire.quelque.chose perfectionner vertu PTCL

'La formation de la fidélité et de la sincérité est là où l'on perfectionne la vertu.'

Plus tard, ce mot est devenu une conjonction causale. Aujourd'hui, en chinois moderne, « 所以 » est utilisé afin d'introduire une inférence qui engage le jugement subjectif du locuteur : il n'est plus nécessaire que les propositions conjointes par « 所以 » soient normalement corrélées pour que la phrase soit acceptable. « 所以 » est donc devenu un marqueur discursif exprimant l'attitude du locuteur. Et selon Traugott et Dasher (2002), les marqueurs discursifs sont précisément les expressions subissant le processus de la subjectification. Par exemple :

wǒ dōu bù gǎn xiāng xìn zhè me duō rén shì lái huānyíng wǒ men de
我 都 不 敢 相信 这么 多 人 是 来 欢迎 我们 的。

je même NEG oser croire tellement nombreux personne être venir accueillir nous PTCL

'Je n'ose pas croire qu'il y ait autant de gens venus pour nous accueillir.'

suǒ yǐ shuō hái shì huí jiā hǎo a
所以 说 还 是 回 家 好 啊。

DSC dire plutôt être rentrer maison bien PTCL

'Il est préférable de rentrer chez soi.'

Dans cette phrase, « 所以 » doit être perçu comme un marqueur discursif portant son sens procédural au lieu de son sens conceptuel (Fraser, 1999). C'est-à-dire, le mot ne sert ici qu'à introduire de la cohésion dans la phrase.

Dans cette recherche, nous adoptons une conception de la subjectivité plus spécifique, qui se concentre essentiellement sur les cas qui sont susceptibles d'engendrer un *désaccord sans faute* (Kölbel, 2003), c'est-à-dire, « une situation où il y a un penseur A, un penseur B, et une proposition p, tel que A croit (juge) que p et B croit que $\neg p$, sans que ni A ni B n'aient tort ». Selon Kölbel (2003), ce type d'énoncé implique souvent les sujets esthétique, culinaire, moral, probabilité, justification des croyances, goût etc. Par exemple, le commentaire « 这剧太烂了。 » (La série est vraiment nulle.) est subjectif selon notre définition, car un autre commentateur pourrait dire « La série est exceptionnelle. » sans nécessairement que l'un des deux ait tort. D'une certaine manière, nous pouvons aussi considérer les cas de désaccord sans faute comme l'intervention de la perspective du locuteur.

La définition de la **pertinence** est plus directe : un commentaire est considéré comme pertinent quand son contenu est directement en lien avec le contenu médiatique : qu'il s'agisse du scénario, de la musique de fond, du cast de la série, et ainsi de suite.

Enfin, la **temporalité** se définit comme une forme de (quasi-) simultanéité entre le commentaire et ce qu'il commente. Il est à noter que la pertinence n'équivaut pas la temporalité et vice versa : les commentateurs peuvent très bien parler à la fin d'une série d'une scène qui s'affiche au début de la série, dans ce cas là, les commentaires sont pertinents mais non temporels ; un commentaire peut aussi concerner un sujet provoqué par la scène actuelle,

sans que le contenu ne se rapporte avec la série. De cette manière, le commentaire est non pertinent mais temporel.

Le contenu des commentaires explicitement interactifs (qui attendent manifestement la réponse des autres utilisateurs ou qui réagissent à un autre ou plusieurs commentaires) est diversifié (les commentaires des autres catégories admettent également des réactions d'autres spectateurs, mais ils n'ont pas d'acte de parole explicite d'interaction). Ces commentaires ne se différencient pas par les trois axes indiqués ci-haut, car ils peuvent être subjectifs ou objectifs, temporels ou non temporels, pertinents ou non pertinents.

Alors que la subjectivité et la pertinence permettent de définir les trois catégories principales, les sous-catégories se distinguent selon la nature détaillée de leur acte illocutoire, et la temporalité. Nous allons préciser leur définition dans le guide d'annotation.

Il est possible d'attribuer plusieurs étiquettes à chaque commentaire s'il est pourvu de plus d'une force illocutoire (Searle, 1969) (un commentaire peut être composé d'une phrase ayant plusieurs forces illocutoires ou de plusieurs phrases ayant des forces illocutoires différentes). Il n'y a donc pas d'exigence quant au nombre maximal d'étiquettes. Pour les commentaires explicitement interactifs, nous avons non seulement annoté leur(s) classe(s), mais aussi leur(s) référence(s) que nous nommons *pivot* —— ce sont le(s) commentaire(s) avec lesquels le commentaire considéré interagit.

Chapitre 2

Annotation

2.1. Guide d'annotation

En nous appuyant sur les critères explicités dans le chapitre précédent, nous avons élaboré un guide d'annotation représentant une hiérarchie de catégories, avec les catégories principales subdivisées en plusieurs sous-catégories. Le but de cette section est de le décrire.

Lors d'un premier examen de commentaires choisis aléatoirement dans le corpus, nous avons identifié des fonctions très variées pour les commentaires étudiés. Nous avons d'abord regroupé les commentaires en classes fonctionnelles assez spécifiques (selon leurs actes illocutoires) avant de chercher à regrouper différentes fonctions au sein de classes plus génériques.

Pertinent	**Subjectif**	**Catégorie principale**
1	1	1
	0	3
0	1	4
	0	

Tableau 2.1 – Trois catégories principales établies selon deux axes : la subjectivité et la pertinence. Dans les colonnes *Pertinent* et *Subjectif*, *1* dénote la valeur booléenne *Vrai* (présence d'un trait) ; *0* le cas inverse.

Quatre catégories principales ont ainsi été établies, parmi lesquelles trois, soit les classes 1, 3 et 4, ont été définies selon les deux axes spécifiés dans le chapitre précédent section 1.3.2, à savoir la pertinence et la subjectivité. D'abord, la pertinence distingue les classes 1 et 3 de la classe 4 : les classes 1 et 3 sont pertinentes, alors que la classe 4 est non pertinente. Ensuite, les classes 1 et 3 se discriminent par la subjectivité : la classe 1 est non subjective, tandis que la classe 3 est subjective. Par contre, la subjectivité ne permet pas de différencier les commentaires dans la classe 4, car ils sont susceptibles d'être subjectifs ou non. En revanche, les commentaires de la classe 2, qui se caractérisent par une interactivité explicite, ne se définissent pas par la subjectivité ni la pertinence. Ces commentaires peuvent être subjectifs ou non subjectifs, pertinents ou non pertinents (le **Tableau** 2.1 montre la distribution des catégories principales 1, 3, 4 par rapport aux deux axes. La classe 2 est exclue de ce tableau, car cette catégorie n'est pas distinguée des autres par ces deux axes). Finalement, la temporalité ne sert donc pas à discriminer les catégories principales. À la place, elle est utilisée pour différencier les sous-catégories. Le **Tableau** 2.2 est une extension du Tableau 2.1. Il récapitule la distribution des catégories par rapport aux trois axes (la pertinence, la subjectivité, et la temporalité).

Les critères et les exemples des catégories principales et de leurs sous-catégories sont définis comme suit :

(1) **Commentaires explicatifs/instructifs**. Ces commentaires manifestent l'intention de leur auteur de fournir des informations factuelles affectant (favorablement, la plupart du temps) l'expérience de visionnement des autres utilisateurs. Ces informations peuvent être de différentes natures. Il peut s'agir de commentaires liés à une scène spécifique ou d'informations générales sur la série :

1.1) Des sous-titres expliquant le sens d'un dialogue.

1.2) Des avertissements portant sur le contenu d'une scène à venir pouvant choquer, ou au contraire, sur un contenu ou un détail à ne pas manquer, par exemple :

i. 注意：前方高能。

‘Attention : haute énergie à venir (pour avertir qu’une scène effrayante va suivre).’

ii. 注意：前方高萌。

‘Attention : haut niveau de « mignonnerie » à venir (pour prévenir de l’arrivée de quelque chose de mignon à l’écran.’

1.3) Des explications relatives à l’intrigue ou à d’autres informations générales relatives à la série (musique de fond, acteur, lieu de tournage, etc.) :

i. 这是金鸡湖。

‘C’est le Lac de Jinji (lieu de tournage de la série).’

ii. 逆战中的反派肖。

‘C’est le vilain Xiao dans le film *Contre-attaque* (information sur l’expérience de l’acteur).’

(2) **Interactions sociales**. Ces commentaires manifestent l’intention de leur auteur d’initier ou de poursuivre une interaction linguistique avec d’autres utilisateurs. Il peut s’agir, par exemple, de poser une question, donner un ordre, contester une information, ou encore répondre à une question. Lorsqu’il s’agit d’une réaction à un autre commentaire, on trouve souvent le mot « 前方 » (devant) dans la phrase, du fait que les commentaires publiés plus tôt défilent vers la gauche et les commentaires qui viennent plus tard les suivent. Lorsqu’un commentaire réagit (c’est-à-dire, en répondant, ou en contestant) à un ou plusieurs autres commentaires, on annote aussi l’index de ces derniers. La **Figure** 2.1 illustre un exemple de l’annotation pour un commentaire de la classe 2.

a) 片头什么歌?

‘C’est quoi la musique d’ouverture ?’

b) 杀破狼?

‘Shapolang (nom d’une chanson) ?’

id	ts	ct	com	cat	piv	pid	seg
323174481#21	7000	1,50E+12	片头什么歌	2.1			['片头', '什么', '歌']
323174481#30	15000	1,51E+12	杀破狼	2.1	片头什么歌	323174481#21	['杀破', '狼']

Figure 2.1 – Un exemple de l'annotation de la classe 2

Les deux commentaires dans l'exemple sont censés être de la classe 2. Le commentaire (b) constitue une réponse au commentaire (a). Dans ce cas-là, nous nommons le commentaire (a) le *pivot*, car il est le provocateur des interactions. Pour marquer la relation entre un commentaire de pivot et les commentaires qui réagissent là-dessus, il faut mettre le texte du commentaire de pivot dans la colonne « piv » (pour « pivot »), et l'index du pivot (« 323174481#21 » dans l'exemple) dans la colonne « pid » du commentaire (b).

(3) **Réactions**. Ces commentaires manifestent l'intention de leur auteur de partager une certaine réaction au visionnement de la vidéo. Ils ont un caractère « subjectif », dans le sens où leur auteur ne cherche pas à partager un fait indépendant de son propre état mental. Au contraire, il s'agit des commentaires avec l'intervention de la perspective du locuteur sous toute forme, avec au premier plan les cas susceptibles de provoquer un désaccord sans faute. Ce point les distingue de certains commentaires de catégorie 1. Par exemple, l'exemple ii) sous 1.3) a pour but de faire connaître un fait au lecteur. Ce commentaire est donc de catégorie 1 et pas 3. En revanche, l'exemple ii) sous 3.1) n'a vraisemblablement pas pour but de présenter un fait, ou d'informer de ce que les effets spéciaux manquent de budget mais bien de communiquer que l'auteur « juge » (négativement) que les effets spéciaux manquent de budget. Ce dernier commentaire est donc de catégorie 3. On regroupe plusieurs types de réactions dans cette catégorie :

3.1) Des jugements subjectifs, qu'ils soient liés à une certaine scène ou non :

i. 这人太菜了。

'Le gars est trop nul.'

ii. 五毛特效。

'Cet effet spécial ne coûte que 0,50 yuan.'

iii. 这剧还是别看了。

'La série ne mérite pas d'être regardée.'

3.2) Argot Internet : des phrases « toutes faites », fréquemment réemployées sur internet pour exprimer un sentiment :

i. 爷青回。

'Ma jeunesse est de retour (pour exprimer la nostalgie).'

ii. 老铁。

'Pote (en anglais : *Bro/Buddy*).'

3.3) L'expression directe ou indirecte d'un sentiment éprouvé :

i. 好无聊。

'J'en ai marre (de la série ou de la scène actuelle).'

ii. 我要哭了。

'Je vais pleurer (à cause de cette scène).'

3.4) L'interpellation des personnages de la série, souvent par une question. Très souvent dans ce genre de cas, le commentateur émet aussi son jugement ou opinion sur une scène d'une certaine manière. Par exemple, la phrase (i) équivaut à dire : « Le personnage est arrivé bien trop tard. »

i. 人都走光了你们才来?

'Tu es arrivé longtemps après que les gens se sont enfuis ?'

ii. 鬼见愁，你以为你剪了头发我就认不出你了吗?

'Guijianchou (personnage d'une autre série), tu penses que je ne peux pas te reconnaître après que tu t'es fait couper les cheveux ?'

3.5 Le commentateur parle à la place d'un personnage dans la série. En l'occurrence, l'auteur du commentaire exprime son propre jugement sur l'acte d'un personnage, donnant l'impression qu'il s'agit d'un acte décidé par le personnage. L'exemple (i) ci-dessous peut aussi se transformer en « Je pense que le marquis veut dire qu'il va prendre un thé pour se calmer. »

i. 侯爷：我要喝杯茶冷静一下。

'Marquis : je vais prendre un thé pour me calmer.'

ii. 师父，坚持住，我是悟空。

'Maître, tenez, je suis Wukong.'

(4) **Les commentaires non pertinents**. Ces commentaires peuvent être factuels ou subjectifs : ils manifestent l'intention de leur auteur de véhiculer une opinion/un sentiment/une information sans rapport avec ce qui se passe dans la série. Il est à noter que certains commentaires sous cette catégorie peuvent être provoqués par une scène spécifique mais qui n'ont pas beaucoup de liens avec le contenu de la série. On regroupe notamment dans cette catégorie :

4.1) Les rapports des commentateurs sur un fait non lié à la série, qu'il s'agisse d'informations sur le spectateur lui-même, sur le monde réel, ou sur des connaissances communes :

i. 我来了。

'J'arrive.'

ii. 我在看广告。

'Je regarde la coupure publicitaire.'

iii. 土豆现在没有广告了。

'Il n'y a plus de publicités sur Tudou (un autre site de vidéos).'

iv. 听说VIP可以在底部显示弹幕。

'Il paraît que les VIPs peuvent afficher les danmakus en bas.'

v. 2018年1月1日。

'(Je regarde la série) le 01 janvier 2018.'

vi. 我家也有这样的衣柜。

'J'ai aussi chez moi une telle armoire.'

4.2) Les commentaires exprimant le sentiment ou opinion du spectateur provoqué par la série mais non pertinents. Autrement dit, le spectateur peut ressentir en

lui-même un certain sentiment suscité par un scénario particulier, mais ce dont il parle n'a pas beaucoup de lien avec la série :

i. 向所有交警同志致敬。

'Salut à tous les camarades de la police de la circulation.'

ii. 优酷上的弹幕好可爱。

'J'ai trouvé que les danmakus sur Youku (le site de vidéos) étaient tellement mignons.'

4.3) Les commentaires qui servent à tester l'effet d'affichage de danmaku :

i. 最多四十个字，最多四十个字……

'Quarante mots au maximum, quarante mots au maximum...(ce commentaire sert à tester la limite de mots pour un commentaire de danmaku.)'

ii. 我要霸屏……

'Je veux encombrer l'écran (le commentateur peut répéter cette phrase beaucoup de fois en peu de temps pour que sur l'écran il n'y ait plus que ses commentaires, et que ces derniers donnent l'impression d'un mur qui cachent l'image derrière).'

4.4) L'admiration pour une star qui joue dans la série :

i. 靳东好帅。

'Jingdong est vraiment beau.'

ii. 为什么靳东还没有出现，我爱靳东。

'Pourquoi Jin Dong (le nom d'une star) n'est toujours pas apparu. J'adore Jin dong.'

2.2. Annotation manuelle

Nous avons utilisé la taxonomie développée dans les sections précédentes, dans le cadre de deux tâches : 1) classification et annotation manuelle par des humains ; 2) classification automatique par la machine. La première tâche a permis de produire l'ensemble de données

Pertinent	Subjectif	Temporel	Catégorie
1	1	1	3.1 ; 3.2 ; 3.3 ; 3.4 ; 3.5
		0	3.1 ; 3.3
0	1	1	1.1 ; 1.2 ; 1.3
		0	1.3
0	1	1	4.2 ; 4.4
		0	4.2 ; 4.3 ; 4.4
0	1	1	4.1
		0	4.1 ; 4.3

Tableau 2.2 – Distribution de catégories par rapport aux trois axes : la subjectivité, la pertinence et la temporalité

annotées nécessaires à l'entraînement et à l'évaluation d'un modèle computationnel pour résoudre la seconde. Elle nous a permis également d'évaluer la fiabilité des critères du guide d'annotation, ainsi que la clareté et l'absence de chevauchement des définitions en mesurant à quel point différents annotateurs humains sont capables de distinguer les commentaires et de leur assigner une classe de manière cohérente. Dans le cadre de la deuxième tâche, nous avons élargi la portée de notre recherche en cherchant à déterminer si des modèles computationnels peuvent reproduire les annotations des annotateurs humains. Cette section présente la première tâche, la seconde est présentée dans le chapitre 5.

Dans le cadre de cette recherche, nous avons annoté un sous-ensemble des données en suivant le guide d'annotation. Pour mesurer la qualité des annotations faites et tester la solidité de la taxonomie, nous avons aussi fait une co-annotation sur un échantillon des données déjà étiquetées. Etant donné la taille de l'échantillon co-annoté ($\approx 23\%$ de la taille du corpus annoté), nous avons décidé de ne pas procéder à l'adjudication traditionnelle (faire une décision finale entre les choix de tous les annotateurs) car cette pratique exige que tous les annotateurs donnent la même quantité d'annotations. À défaut, nous avons simplement utilisé les

données de l'annotatrice principale comme référence pour les expériences de classification automatique.

489 commentaires ont été utilisés en vue de l'estimation de l'accord inter-annotateurs. Deux locutrices natives de la langue chinoise en maîtrise (dont l'une se spécialisait dans le domaine de la linguistique) se sont engagées dans le travail d'annotation. Elles avaient toutes deux l'habitude de regarder des vidéos en activant les danmakus et connaissaient bien les mots et expressions spéciaux utilisés par la communauté des utilisateurs de danmaku.

La première annotatrice (l'auteure du présent mémoire) a annoté 2 118 commentaires provenant de 5 épisodes différents (voir le **Tableau** 2.3). La deuxième annotatrice en a annoté 489 choisis aléatoirement parmi les commentaires que la première annotatrice avait annotés. Les commentaires qui ont reçu *None* comme étiquette et ceux qui n'ont reçu aucune étiquette de la première annotatrice ont été abandonnés pour l'étape de co-annotation. Par ailleurs, comme l'identification des commentaires de la catégorie 2 repose en grande partie sur les commentaires affichés dans une zone temporelle proche, et qu'il était difficile de conserver ce genre d'informations contextuelles lors de l'échantillonnage et de les communiquer à la seconde annotatrice sans compliquer considérablement sa tâche, nous avons choisi de lui indiquer directement les cas annotés comme la classe 2 par la première annotatrice. Toutefois, l'annotatrice 2 devait toujours décider des autres étiquettes à assigner à un commentaire de catégorie 2. Les scores d'accord inter-annotateurs rapportés permettent donc seulement d'évaluer la robustesse des annotations des catégories 1, 3, 4, et de leurs sous-catégories.

Bien qu'il n'y ait pas de limites sur le maximum d'étiquettes attribuées, les quantités d'étiquettes données par les deux annotatrices sont proches l'une de l'autre (543 pour l'annotatrice 1 et 588 pour l'annotatrice 2). Les deux annotatrices ont donné au maximum trois étiquettes par échantillon, et la plupart du temps, elles ont identifié les commentaires comme appartenant à une seule catégorie. Nous pouvons constater également que pour toutes les deux annotatrices, les étiquettes de la classe 3 dépassent de loin les autres en nombre, et particulièrement la classe 3.1 (voir la **Figure** 2.5). Ainsi, nous avons examiné l'accord inter-annotateurs à partir de trois points de vue distincts. D'abord, nous avons comparé l'accord sur la classe 3 avec celui sur les autres classes ensemble (parce que nous avons

Nom de série	# Annotation
A Step into the Past Épisode 30	97
Chinese paladin Épisode 7	297
Nirvana in Fire Épisode 30	322
SWAT Épisode 1	965
SWAT Épisode 2	437
Total	2 118

Tableau 2.3 – Constitution de commentaires annotés dans les séries

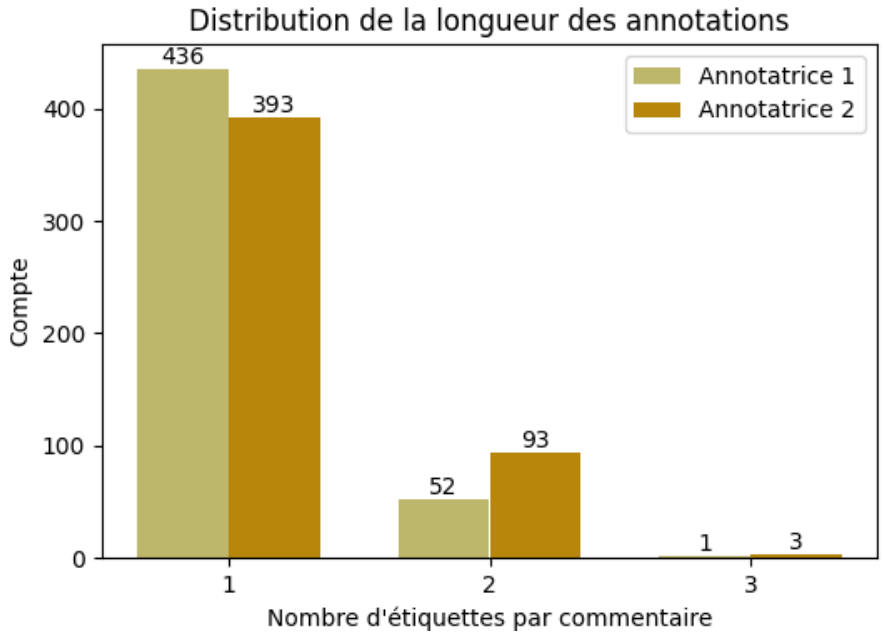

Figure 2.2 – Distribution des étiquettes assignées

voulu commencer par examiner, à l'étape de la classification automatique, si les modèles pouvaient fonctionner d'une manière plus efficace que le simple hasard.). Pour ce faire, nous avons converti les étiquettes en 1 (pour la classe 3) et 0 (pour les classes autres que 3). Les étiquettes dupliquées ont été supprimées après cette transformation. Par exemple, si un commentaire a été annoté avec les catégories « 3.1, 3.2, 4.1 », la transformation produit

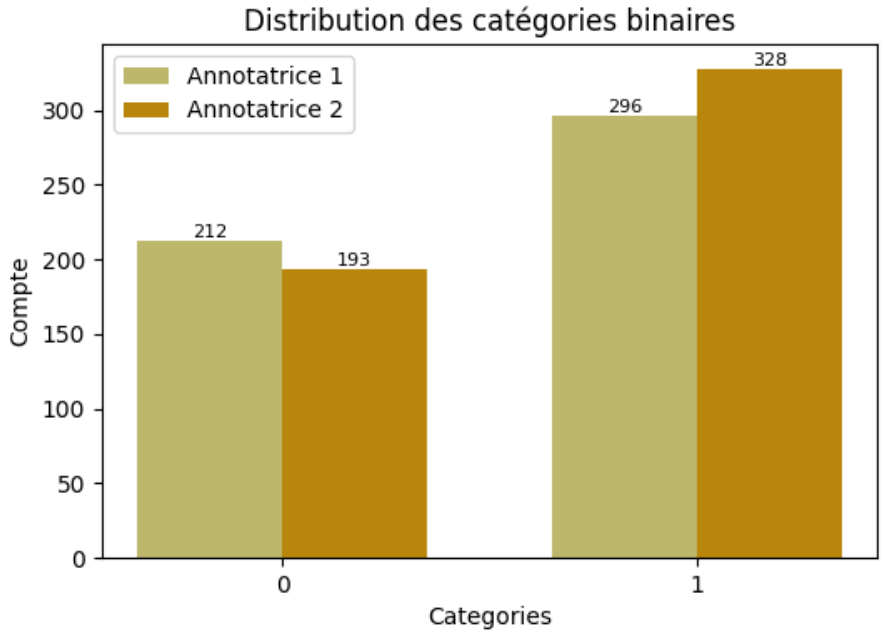

Figure 2.3 – Distribution des catégories binaires : « 1 » représente la classe 3 et « 0 » représente les classes autre que 3.

la liste « 1, 1, 0 », puis « 1, 0 » après la supression des doublons. La distribution des étiquettes après la conversion est comme le montre la **Figure** 2.3. Nous avons au total eu 1 029 étiquettes binaires. Ensuite, nous avons regardé comment les annotatrices étaient d'accord sur les classes principales (catégories à granularité grosse). Cette fois, nous n'avons conservé qu'une seule étiquette de la même classe principale lorsque les annotatrices ont donné plus d'une étiquette appartenant à la même catégorie principale. Par exemple, si un commentaire reçoit comme étiquettes « 3.1, 3.2, 4.1 », celles-ci vont être transformées en « 3, 4 ». Le nombre total d'étiquettes à granularité grosse est de 1 070. **La Figure** 2.4 présente la distribution des étiquettes de classes principales. Enfin, nous avons scruté également l'accord inter-annotateurs sur les classes fines, dont la distribution est visible sur la **Figure** 2.5.

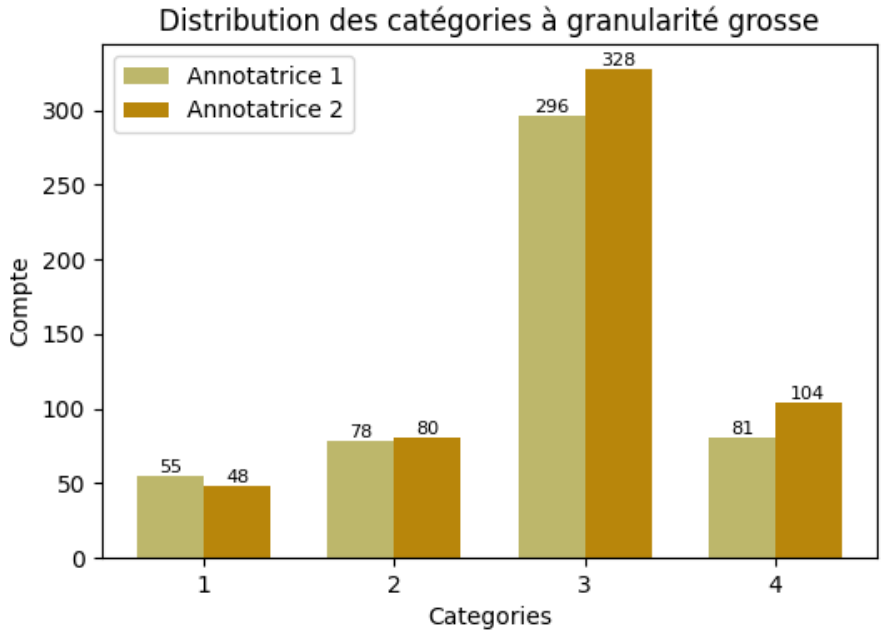

Figure 2.4 – Distribution des catégories à granularité grosse

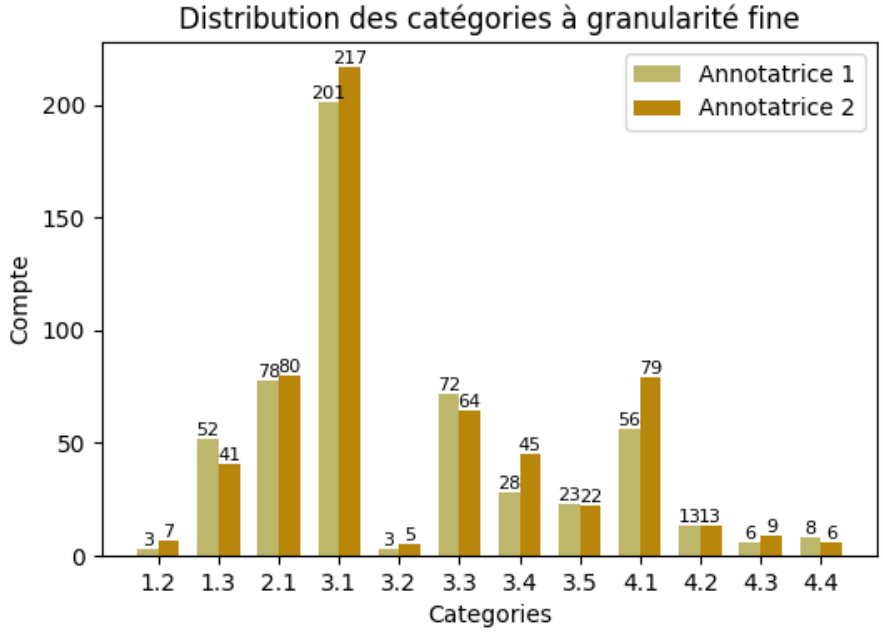

Figure 2.5 – Distribution des catégories à granularité fine

2.3. Accord inter-annotateurs pour une classification multi-étiquettes

Vu que notre tâche de classification admet plusieurs étiquettes pour un seul commentaire, l'évaluation de l'accord entre les annotateurs n'est pas triviale. Il existe une multitude de mesures pour calculer le degré d'accord inter-annotateurs. Elles ont cependant une formulation commune, qui se représente par le coefficient d'accord, c'est-à-dire, la différence normalisée entre un accord estimé et un accord observé :

$$A(c) = \frac{A(o) - A(e)}{1 - A(e)} \tag{2.3.1}$$

Ici, A(e) dénote l'accord au hasard, et A(o) l'accord observé (l'accord réel mesuré sur l'annotation). De cette façon, quand l'accord observé atteint 1 (quand les annotateurs sont d'accord sur tous les cas), le coefficient d'accord atteint aussi 1. Par contre, lorsque l'accord observé égale l'accord estimé (l'accord entre les annotateurs est aussi faible que l'accord au hasard), le coefficient d'accord égale 0. Si l'accord observé est plus petit que l'accord estimé, on obtient un accord pire qu'au hasard.

Le Kappa de Cohen (Cohen, 1960) est une méthode courante destinée au calcul d'accord entre deux annotateurs sur la classification à étiquette unique. Elle ne peut pas être appliquée directement sur une tâche de classification multi-étiquettes. Marchal et al. (2022) ont discuté de quelques méthodes existantes pour l'évaluation de l'accord observé dans la classification multi-étiquettes et ils ont proposé une estimation de l'accord au hasard par le ré-échantillonnage avec remise (en anglais : *bootstrapping*).

(1) La méthode la plus simple serait peut-être de considérer une annotation multi-étiquettes comme une annotation à étiquette unique dans laquelle l'étiquette annotée est en fait un ensemble. Un accord n'est alors observé que lorsque les deux ensembles d'étiquettes assignées par les deux annotateurs sont égaux. Par exemple, il n'existe pas d'accord entre les étiquettes « 1 » et « 0, 1 », cette dernière annotation n'est en accord qu'avec l'étiquette « 0, 1 ». Si les annotations sont prises en compte de cette manière, il nous est possible de calculer le coefficient d'accord à travers le Kappa de

Cohen. L'accord observé est tout simplement le nombre de fois où les deux annotateurs donnent la même étiquette parmi toutes les instances d'annotations. L'accord estimé se calcule de la façon suivante :

$$\kappa = \frac{\sum_{c \in C} n_{c1} n_{c2}}{N^2} \tag{2.3.2}$$

où C représente le nombre total de catégories ; N le nombre total de commentaires ; n_{c1}, le nombre de cas où l'annotatrice 1 assigne l'étiquette c, et n_{c2}, le nombre des cas où l'annotatrice 2 assigne l'étiquette c. Cela revient au même que de calculer d'abord la probabilité jointe que chaque annotateur attribue (simultanément) l'étiquette c pour ensuite estimer la probabilité globale que les annotateurs soient d'accord. Enfin, le coefficient d'accord s'obtient par la formule 2.3.1. Cette manière d'évaluation est très stricte dans le sens où un accord n'est compté que s'il y a un chevauchement complet d'annotations pour une instance.

(2) Une alternative plus relâchée s'appelle *Accord souple* (en anglais : *soft-match*) (Crible & Degand, 2019). L'idée est que l'on considère, pour une instance, qu'un accord existe à condition que l'une des étiquettes soit commune aux deux annotations. Par exemple, pour un commentaire A, si l'annotateur 1 le classe comme « 0 », alors que l'annotateur 2 lui assigne « 0, 1 », nous considérons qu'il y a un accord en ignorant le divergence sur la classe 1. S'il y a plus d'une étiquette commune, nous traitons le cas comme s'il n'y avait qu'un seul accord. Il est particulièrement important d'estimer correctement l'accord par hasard, sinon le score obtenu risque d'être surestimé. Par exemple, si l'un des annotateurs assigne aveuglement toutes les étiquettes possibles à tous les points de données, l'accord observé égale 1. Mais dans un tel cas, un accord par hasard prenant en compte la tendance de cet annotateur égalerait aussi 1 [1]. On peut obtenir une bonne estimation de l'accord par hasard via le

1. Dans ce cas limite ($A_e = 1$), le coefficient d'accord n'est pas défini, mais dans des cas plus réalistes ou $A_e < 1$ et $A_e \approx A_o$, le coefficient d'accord serait proche de 0.

ré-échantillonnage avec remise proposé par Marchal et al. (2022), ce qui nous permet de mitiger l'influence des cas extrêmes comme ceux que nous avons mentionnés plus haut. Nous allons parler plus tard des détails du ré-échantillonnage.

(3) *Kappa augmenté*. Cette métrique a été proposée par Rosenberg et Binkowski (2004). Elle pénalise l'incertitude des annotateurs, de sorte que le score d'accord mesuré pour un commentaire est inversement proportionnel au nombre d'étiquettes fournies par un annotateur donné : plus un annotateur attribue d'étiquettes, plus il est incertain de la classe du commentaire. L'accord observé pour un commentaire i annoté par n annotateurs est défini comme ci-dessous :

$$A_o^i = \left| \bigcap_j C_j \right| \cdot \prod_j W_j^i \tag{2.3.3}$$

où C_j est l'ensemble des étiquettes assignées par l'annotateur j au commentaire i, et W_j^i est le poids d'étiquettes, défini comme l'inverse du nombre total des étiquettes attribuées par i au commentaire j ($W_j^i = \frac{1}{|C_j|}$). De cette manière, pour un annotateur déterminé, l'accord observé augmente quand il y a plus de classes convenues, mais diminue quand un annotateur assigne plus d'étiquettes pour un seul commentaire. Le problème avec cette méthode est qu'elle suppose que les cas de multi-annotations sont dûs à l'incertitude des annotateurs et néglige la vraie ambiguïté résidant dans les points de données. Étant donné que les danmakus sont en soi ambigus, il est parfois légitime de leur assigner plusieurs classes, et cette mesure d'évaluation ne s'adapte pas à notre situation.

(4) *Score F*. La métrique est couramment utilisée pour évaluer la performance des modèles. Elle mesure la qualité de la prédiction par rapport à la réalité (les vraies étiquettes). La **Figure** 2.6 montre une matrice de confusion classique qui met en parallèle les vraies étiquettes et les étiquettes prédites. Le VP (vrai positif) et VN (vrai négatif) désignent respectivement les cas en réalité positifs ou négatifs qui sont correctement prédits comme positifs ou négatifs. FP (faux positif) correspond, par contre, aux cas négatifs mal identifiés par un modèle comme positifs, et FN (faux

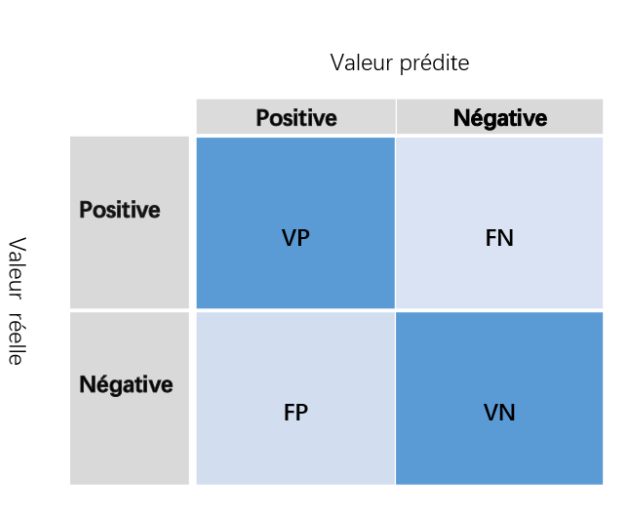

Figure 2.6 – Matrice de confusion

négatif) se réfère à la situation contraire. Le score F est exprimé de la façon suivante :

$$F = \frac{2 \cdot VP}{2 \cdot VP + FP + FN} \tag{2.3.4}$$

Cependant, l'utilisation de la métrique F ne se limite pas à l'estimation du modèle : elle s'applique aussi au calcul de l'accord entre deux annotateurs si les annotations de l'un des annotateurs sont considérées comme étant la réalité. Vu que notre objet à cette étape est d'évaluer la robustesse des annotations manuelles de l'annotatrice 1, nous avons considéré l'annotation faite par cette dernière comme la vérité de terrain pour calculer le score F. Ainsi, la case VP correspond aux accords des annotatrices sur la classe intéressante ; VF aux accords sur la classe non intéressante ; FP fait référence aux cas où l'annotatrice 2 opte pour la classe intéressante contrairement à l'annotatrice 1. FN désigne les cas opposés aux cas FP. Nous avons adopté la moyenne micro des scores de F de chaque catégorie.

$$F_{micro} = \frac{2 \cdot \sum_{c \in C} VP_c}{2 \cdot \sum_{c \in C} VP_c + \sum_{c \in C} FP_c + \sum_{c \in C} FN_c} \tag{2.3.5}$$

c représente l'étiquette en question et C toutes les catégories. Les étiquettes assignées par l'annotatrice 2 qui sont différentes de celles qui sont données par l'annotatrice 1, y compris les classes ignorées ou supplémentaires du point de vue de l'annotatrice 1,

seront pénalisées. Pourtant, cette technique est aussi problématique pour la même raison que l'accord souple traditionnel : elle ignore l'accord au hasard. Ainsi, nous avons également appliqué le ré-échantillonnage avec remise pour estimer l'accord par hasard et calculer ensuite le coefficient d'accord.

En raison du manque de considération de l'accord au hasard dans les méthodes susmentionnées telles que l'accord souple et le score F, nous nous sommes servis du ré-échantillonnage avec remise testé par Marchal et al. (2022) afin d'évaluer l'accord par hasard. Le ré-échantillonnage avec remise consiste à réassigner aléatoirement une (multi-) annotation à chaque commentaire, pour chaque annotateur, en conservant certaines caractéristiques de la distribution observée pour cet annotateur sur l'échantillon original. Dans notre cas, l'échantillon original consistait en 489 commentaires co-annotés par les deux annotatrices et nous avons exécuté 10 étapes de ré-échantillonnage, en gardant inchangées la proportion des différentes étiquettes assignées par chaque annotatrice ainsi que la proportion de commentaires ayant reçu n étiquettes (même proportion d'étiquettes/simples doubles/triples). Par exemple, dans le cas de la classification binaire, les annotations d'étiquettes simples et doubles (dans la classification binaire, il n'existe que des annotations d'étiquettes simples ou doubles (voir la section 2.2)) représentent respectivement 96% et 4% du total d'annotations (489). À chaque ré-échantillonnage, le nombre d'annotations d'étiquettes simples est d'environ 469, alors que celui d'étiquettes doubles est près de 20. Quand nous avons généré les données pour représenter les annotations faites par l'annotatrice 1, nous avons maintenu à chaque fois la proportion des étiquettes 0 et des étiquettes 1 à respectivement 42% et 58% (l'annotatrice 1 a assigné 212 fois l'étiquette 0 parmi 508 étiquettes et 296 fois l'étiquette 1 dans le jeu de données original). Le même principe s'applique aux ré-échantillonnage simulant les annotations effectués par l'annotatrice 2. Pour chaque métrique (le score F et l'accord souple), le résultat obtenu sur l'échantillon original constitue la mesure de l'accord observé, et la moyenne des résultats obetnus sur les 10 ré-échantillonages constitue la mesure de l'accord par hasard. Avec ces deux mesures, nous avons pu calculer le coefficient d'accord en ayant recours à la formule 2.3.1. Pour l'accord souple, A(o) représente la proportion des cas d'accord partiel parmi l'ensemble des 489 points de données et, A(e) la moyenne de cette

proportion sur les 10 ré-échantillonnages avec remise. Dans le cas du score F, A(o) est le score F micro de l'échantillon original, alors que A(e) constitue la moyenne des scores de F micro des 10 ré-échantillonnages. Le **Tableau** 2.4 montre les résultats obtenus en appliquant toutes les techniques susmentionnées :

	κ **de Cohen**	**Accord souple**	**Score F**
Binaire	0,62	0,76	0,69
Multi-grosse	0,51	0,75	0,65
Multi-fine	0,39	0,64	0,53

Tableau 2.4 – Accord inter-annotateurs pour la classification binaire (entre la classe 3 et les autres classes), la classification multiclasses à granularité grosse et la classification multiclasses à granularité fine.

2.4. Analyse des cas difficiles dans l'annotation et la co-annotation

Les scores d'accord, mis à part ceux qui sont obtenus avec le kappa de Cohen, sont globalement bons. Si on regarde horizontalement, les scores obtenus par accord souple sont les meilleurs, ce qui n'est pas surprenant, vu que c'est la mesure la plus tolérante parmi les trois. En revanche, les accords sont les plus faibles avec le coefficient de Kappa de Cohen, la mesure la plus rigide. Verticalement, les coefficients de Kappa de Cohen diminuent d'une manière significative avec l'augmentation de la finesse des catégories, tandis que les scores de l'accord souple et de F ne varient presque pas entre la classification binaire et la classification à granularité grosse, mais chutent considérablement pour la classification à granularité fine.

À l'examen des résultats obtenus avec le kappa de Cohen, nous avons remarqué le kappa de Cohen peut être problématique pour notre corpus co-annoté, parce que les accords sont inversement proportionnels au nombre d'étiquettes : les accords sur les annotations d'étiquette unique concernent une écrasante majorité des cas (la plupart des accords sont sur la

classe 3, plus précisément la classe 3.1) ; les accords sur les annotations d'étiquette double sont beaucoup plus rares, alors qu'il n'y pas d'accord sur les annotations d'étiquettes triples. Autrement dit, le kappa de Cohen est susceptible d'ignorer les accords partiels dans les annotations d'étiquettes multiples qui représentent tout de même un nombre non négligeable de cas.

L'accord souple s'adapte mieux à notre situation dans le sens où il est capable, contrairement au kappa de Cohen, de profiter plus ou moins des accords partiels, en particulier, ceux qui sont sur les classes minoritaires. Par conséquent, les consensus observés augmentent d'une façon notable avec cette mesure.

Le coefficient d'accord basé sur le score F nous donne plus de détails sur les options des deux annotatrices. Cette méthode montre que ces dernières étaient généralement d'accord dans la classification binaire, et que l'annotatrice 2 avait plus tendance à assigner la classe 3 par rapport à l'annotatrice 1. En particulier, elle n'a pas reconnu des cas identifiés par l'annotatrice 1 comme appartenant à la classe 1, peut-être en raison d'un biais envers la classe majoritaire pour l'annotation des exemples délicats. Nous trouvons que les exemples classifiés comme étant 1.3, 3.3, 3.1, 3.4 et 4.1 par l'annotatrice 1 sont les plus souvent différemment classifiés par l'annotatrice 2.

En examinant les exemples conflictuels, nous avons émis des hypothèses sur les raisons à l'origine des divergences d'annotations :

(1) Manque de contexte. L'échantillonnage aléatoire a privé certains commentaires de leur contexte, c'est-à-dire, de la position temporelle dans la vidéo ou des commentaires concomitants. Ce qui complique la catégorisation des commentaires dont l' interprétation dépend fortement du contexte. Par exemple, distinguer les catégories 1 vs 4 et 3 vs 4, implique la pertinence. Cela peut être illustré par (1a) et (1b) :

a. mx1014是TM半自动。
'Le MX1014 (modèle d'armes à feu) est semi-automatique.'

b. 我饿了。
'J'ai faim.'

En (1a), l'annotatrice 1 croyait que le commentaire concernait une information générale provoquée par la série, alors que l'annotatrice 2 pensait que ce commentaire parlait de ce qui se passait sur l'écran, par conséquent, elles l'ont annoté respectivement comme « 4.1 » et « 1.3 ». En (1b), nous ne savons pas si le spectateur se réfère à sa sensation réelle, ou bien s'il voulait dire qu'une image sur l'écran (de la nourriture peut-être) lui donnait faim. Dans le premier cas, le commentaire appartient à la classe 4.1, tandis que dans le dernier cas, le commentaire appartient à la classe 3.3.

(2) Différentes conceptions de l'interlocuteur. Les commentateurs peuvent envisager plusieurs interlocuteurs potentiels : soit ils parlent à un autre commentateur (comme les commentaires de la classe 2.1), soit ils s'adressent à un personnage dans une série (comme les commentaires de la classe 3.4) ou à une star (comme les commentaires de classe 4.4) ; ou bien il s'agit d'un monologue (comme les commentaires des classes 3.1 ou 3.3), les spectateurs ne veulent qu'exprimer un sentiment. Les phrases suivantes ont été étiquetées différemment par les deux annotatrices en fonction de leur perception de l'interlocuteur de l'auteur de ces commentaires.

a. 我的灵儿啊。

'Ma Ling'er (un personnage dans une série).'

b. 现场有很多群众，就那么两个人？？

'Une grande foule sur place, alors il n'y que deux personnes ? ?'

L'annotatrice 1 a catégorisé (2a) comme étant l'expression d'une sympathie sincère, soit la classe 3.3 ; alors que l'annotatrice 2 l'a mis dans la classe 3.4, signifiant une parole adressée à un personnage. En (2b), comme la phrase n'est pas formulée d'une manière explicite, nous ignorons si l'auteur du commentaire a pour seul objectif d'exprimer son jugement sur une ligne de dialogue (la classe 3.1) sans avoir une audience spécifique ou s'il fait semblant de parler avec un personnage (la classe 3.4).

(3) Nuances entre le jugement et l'émotion. Il arrive souvent que l'une des annotatrices perçoive la communication d'une sorte d'émotion dans un commentaires tels que (3a), alors que l'autre y percevait davantage du jugement. C'est probablement parce

qu'émettre un jugement est souvent suivi d'une réaction émotionnelle, et vice-versa. Par exemple, (3b) peut transmettre de la mélancolie éprouvée au visionnement d'une scène émouvante.

a. 66666。

'Bravo (argot internet).'

b. 我要哭了。

'Je vais pleurer.'

(4) Nuances entre l'information objective sur la série ou le jugement subjectif. Cela dépend des connaissances des annotatrices sur la série. À titre d'exemple, le commentaire sera étiqueté « 1.3 » si on le comprend comme la réalité, mais sera annoté « 3.1 » si nous le considérons comme la vision subjective du commentateur. Le commentaire (4a) a été identifié respectivement comme une révélation d'intrigue par l'annotatrice et une déduction par l'auteur du commentaire.

a. 真的没有回来。

'Il (elle) n'est vraiment pas revenu(e).'

(5) Il existe également des commentaires qui ont plus d'une fonction. Ainsi ces commentaires-là peuvent recevoir une ou plusieurs étiquettes, ce qui mène aussi à ce que les annotatrices assignent des étiquettes distinctes.

Pour conclure, les annotatrices ont parvenu, en termes de l'accord souple, à une concordance substantielle dans la classification binaire (0,76) et la classification à granularité grosse (0,75). Cela suggère d'une certaine manière que la définition des catégories principales étaient solide et que les données annotées par l'annotatrice principale sont admissibles aux classifications automatiques binaire et multiclasses à granularité grosse. Pourtant, les accords à granularité fine sont trop faibles qu'il serait peu probable pour la machine de distinguer les sous-catégories les unes des autres dans une classification automatique à granularité fine.

Chapitre 3

Classification automatique de danmakus

Dans cette section, nous aborderons les expériences de classification automatique menées sur les données annotées par l'annotatrice principale. Pour toutes les expériences d'apprentissage automatique, nous avons utilisé une répartition des données de 80% pour l'ensemble d'entraînement et 20% pour l'ensemble de test. Les jeux de test n'ont pas été équilibrés afin que les résultats puissent représenter la performance des modèles sur les données réelles.

3.1. Distribution des données annotées

Le corpus annoté présente une distribution déséquilibrée entre les catégories. La catégorie 3 (commentaires subjectifs pertinents) représente une proportion prédominante. À cet

	Catégorie principale				
Nom de série	**1**	**2**	**3**	**4**	**Total**
A Step into the Past Épisode 30	9	20	49	30	**108**
Chinese paladin Épisode 7	19	91	90	113	**313**
Nirvana in Fire Épisode 30	22	59	145	106	**332**
SWAT Épisode 1	125	122	715	46	**1 008**
SWAT Épisode 2	52	56	344	16	**468**
Total	**227**	**348**	**1 343**	**311**	**2 229**

Tableau 3.1 – Distribution de catégories principales des données annotées

égard, la classification automatique se décline en plusieurs sous-tâches afin de faire une classification progressive. Nous avons commencé par une classification binaire entre la catégorie 3 et les autres catégories, puis une classification à granularité grosse pour distinguer les catégories principales, enfin et sur la base des expériences précédentes, nous avons essayé de classer les commentaires en classes plus fines.

3.1.1. Classification binaire

De la même manière que nous avons fait pour évaluer l'accord inter-annotateurs dans la classification binaire, nous avons converti les étiquettes à granularité fine en étiquettes binaires : 1 pour la classe 3 et 0 pour les classes autres que la classe 3. Puis, les répétitions d'étiquettes dans une instance d'annotation ont été enlevées (voir le chapitre 2 section 2.2). Ensuite, nous avons créé des copies de commentaires comprenant deux étiquettes, de sorte que chaque commentaire et sa copie correspondent à une étiquette. Par exemple, le commentaire « 灵儿 » (Ling'er, nom d'un personnage) porte deux étiquettes « 1, 0 » (voir le **Tableau** 3.2). Après la duplication, le commentaire et sa copie reçoivent respectivement les étiquettes « 1 » et « 0 » (Voir le **Tableau** 3.3).

Commentaire	**Étiquette**
灵儿	1; 0

Tableau 3.2 – Un commentaire portant deux étiquettes en cas de la classification binaire.

Commentaire	**Étiquette**
灵儿	1
灵儿	0

Tableau 3.3 – Le commentaire et sa copie portent chacun une étiquette différente après la transformation.

Après toutes les transformations, il reste 2 211 commentaires. Deux expériences ont été faites à cette étape : l'une avec les données de taille originale, l'autre avec les données sous-échantillonnées (voir **Tableau** 3.4). Pour les expériences de sous-échantillonage, une partie des commentaires de la classe 3 a été exclue aléatoirement du jeu d'entraînement original de sorte que les commentaires de classe 3 soient aussi nombreux que ceux des autres catégories.

	Entraînement		**Test**		**Total**
	1	0	1	0	
Données originales	1 074	694	269	174	**2 211**
Données sous-échantillonnées	694	694	269	174	**1 831**

Tableau 3.4 – Distribution des données dans les jeux d'entraînement et de test pour la classification binaire

3.1.2. Classification multiclasses à granularité grosse

Pour la classification multiclasses à granularité grosse, nous avons d'abord converti les étiquettes à granularité fine en étiquettes à granularité grosse. Après cette conversion, les étiquettes dupliquées assignées au même commentaire ont été supprimées. Ensuite, la réplication des commentaires comportant plusieurs étiquettes a été effectuée de la même manière que dans le cas de la classification binaire. Le **Tableau** 3.5 représente la distribution des données pour la classification à granularité grosse.

3.1.3. Classification à granularité fine

En raison de la taille limitée du corpus annoté, du déséquilibre des données annotées, et des scores d'accord inter-annotateur obtenus, il nous a paru probable qu'une classification à granularité fine entre toutes les sous-catétgories ne génère pas de bons résultats. Nous avons toutefois essayé des classifications à granularité fine sur une partie des sous-catégories pour voir si la machine pouvait mieux faire ou au contraire, elle se comportait pire que les humains dans la distinction de ces classes difficiles.

	Entraînement	Test	Total
Catégorie principale			
1	182	45	
2	278	70	
3	1074	269	
4	249	62	
Total	1783	446	2229

Tableau 3.5 – Distribution des données dans les jeux d'entraînement et de test pour la classification à granularité grosse

	Entraînement	Test	Total
Sous-catégorie			
3.1	751	188	
3.3	268	67	
3.4	91	23	
1.3	173	43	
Total	1283	321	1604

Tableau 3.6 – Distribution des données dans les jeux d'entraînement et de test pour la classification à granularité fine

3.2. Modèles

3.2.1. Représentation vectorielle

Habituellement, les classificateurs d'apprentissage automatique prennent en entrée des vecteurs de longueur fixe. De ce fait, il nous était nécessaire de transformer les données textuelles en vecteurs numériques. L'un des moyens les plus courants est d'utiliser des sac

de mots (SDM). Dans le cadre de notre recherche, nous avons appliqué trois méthodes de représentation vectorielle de mots, reposant sur les SDM avec différentes caractéristiques :

(1) Sac-de-mots de Fréquence (SDM Fréquence)

(2) Sac-de-mots de TF-IDF (SDM TF-IDF)

(3) Sac-de-mots Word2Vec (SDM Word2Vec)

Le sac de mots (SDM) (Harris, 1954) est une méthode de représentation de données textuelles sous forme de vecteurs de longueur fixe (le plus souvent, la taille du vocabulaire du corpus). Dans la méthode SDM Fréquence, on associe d'abord chaque mot du vocabulaire à un encodage un parmi n (Onehot). Ainsi, chaque mot est représenté par un vecteur binaire comprenant autant de composante que la taille du vocabulaire, dans lequel on assigne à chaque composante une valeur binaire : la valeur 1 à la composante correspondante au mot à encoder, la valeur 0 dans les autres composantes. Ensuite, le vecteur d'un commentaire entier est formé en sommant les vecteurs des mots du commentaire. De cette manière, les valeurs dans le vecteur obtenu décrivent les nombres d'occurrences des mots dans le commentaire. Comme les mots vides, dont la distribution est uniforme à travers le corpus, ne contribuent pas grandement à la discrimination des commentaires, nous les avons enlevés avant la vectorisation. La liste des mots vides que nous avons utilisée était l'union des listes des mots vides de HIT et de Baidu (2016). Au final, nous avons utilisé un vocabulaire de 2 555 mots. Comme les commentaires sont globalement courts, les vecteurs de commentaires produits sont creux.

SDM TF-IDF est une approche similaire à SDM Fréquence qui consiste à pondérer les vecteurs One-hot par le score fréquence de terme-fréquence inverse de document (TF-IDF) de chaque mot dans un document (un commentaire). Cela permet, d'une part, de donner un poids élevé aux mots spécifiques à un document, et d'autre part, d'atténuer l'influence des mots vides de sens non filtrés ou des mots qui ne sont pas des mots vides mais qui sont trop communs dans le corpus. TF, soit la fréquence de terme, mesure le nombre d'occurrences d'un mot dans un document par rapport au nombre total des mots dans ce document. Intuitivement, plus un terme est fréquent, plus il est important dans le document. Cependant,

l'utilisation seule de la fréquence de terme risque de biaiser les résultats, car les mots qui se répètent souvent dans un document sont possiblement des mots également omniprésents dans l'ensemble du corpus (comme les mots vides), ce qui les rend moins discriminants. IDF, la fréquence de terme inverse de document, permet de régler ce problème en prenant en compte l'inverse de la proportion de documents contenant le terme intéressé par rapport au nombre total de documents dans le corpus. Ainsi, la valeur de TF-IDF d'un mot i dans un document j est calculé par la formule :

$$TF\text{-}IDF_{i,j} = \frac{n_{i,j}}{\sum_k^m n_{k,j}} \cdot \log \frac{|D|}{|\{j' \in D : i \in j'\}|} \tag{3.2.1}$$

où $n_{i,j}$ est le nombre d'occurrences du terme i dans le document j. m est le nombre total de termes dans le document j. $n_{k,j}$ est le nombre d'occurrences du terme k dans le document j. D est l'ensemble des documents dans le corpus.

Comme ces deux premières méthodes génèrent des vecteurs de grande dimension, à savoir 2 555, nous nous sommes aussi servis de l'analyse en composantes principales (ACP), une technique de réduction de dimension qui consiste à projeter les données originales sur n axes tout en conservant le plus de variance possible. Nous avons ainsi réduit la dimension à 50.

Les désavantages de la méthode de SDM formés par l'agrégation des vecteurs binaires résident dans la production des vecteurs de dimension très élevée quand le corpus incorpore un vocabulaire très large et dans l'ignorance de l'ordre des mots. De plus, Onehot néglige les liens sémantiques (Le & Mikolov, 2014). Supposons que nous avons un corpus comprenant un seul commentaire « 你叫谁？» (Qui est-ce que tu appelles ?). Les trois tokens qui composent cette phrase seront réprésentés respectivement par les vecteurs binaires $\begin{bmatrix}1 & 0 & 0\end{bmatrix}$, $\begin{bmatrix}0 & 1 & 0\end{bmatrix}$, $\begin{bmatrix}0 & 0 & 1\end{bmatrix}$. Dans ce cas-là, le vecteur du commentaire sera alors $\begin{bmatrix}1 & 1 & 1\end{bmatrix}$. Mais si nous inversons l'ordre des tokens, maintenant la phrase est « 谁叫你？» (Qui t'appelle ?). Ainsi, le sens de la phrase change complètement, alors que le vecteur demeure intact.

En revanche, SDM Word2Vec diffère des deux premières méthodes par sa capacité à créer des vecteurs denses et à prendre en considération la propriété distributionnelle des mots. Word2vec (W2V) (Mikolov et al., 2013) appartient à l'apprentissage auto-supervisé. Il

est constitué de deux structures nommées respectivement sac de mots continu (en anglais: *continuous bag of words*) (CBOW) et skip-gram. CBOW prédit un mot en fonction de son contexte tandis que skip-gram se sert d'un certain mot pour prédire les mots qui l'entourent (la portée est déterminée par une longueur de fenêtre qui est un hyper-paramètre).

Étant donné la particularité des mots et expressions utilisés dans les danmakus, nous avons pré-entraîné notre modèle sur l'ensemble du corpus non annoté, à savoir tous les commentaires segmentés dans 8 156 épisodes (fichiers) à l'aide du module `models.word2vec` de la bibliothèque Python `gensim`. De ce fait, il n'y avait pas de mots hors-vocabulaire. Nous avons appliqué le modèle CBOW qui a été entraîné avec une taille de vecteur configurée à **300** et une valeur de fenêtre à **5**. Le compte minimal a été fixé à 1 pour inclure tous les tokens du corpus.

Les plongements lexicaux que nous avons obtenus sont en mesure de saisir d'une certaine manière les liens sémantiques entre les tokens. Par exemple, les cinq plongements lexicaux les plus proches du mot « 枪 » (fusil) étaient ceux de « 这枪 » (ce fusil), « 机枪 » (mitrailleuse), « ak/AK » (un type d'armes à feu), « 拿枪 » (tenir le fusil) et « 刀 » (couteau). Apparemment, il s'agissait des mots liés à « fusil » en quelque sorte.

Les plongements lexicaux obtenus ont ensuite été concaténés ou agrégés afin de former les vecteurs de commentaire. Les détails de cette construction sont expliqués dans la suite :

Agrégation : un commentaire de n tokens (w_1,w_2...,w_n) est représenté par la combinaison linéaire de la fréquence des mots dans la phrase $(f_1, f_2, \ldots, f_n)$ et les plongements lexicaux des mots correspondant. Le vecteur résultant est d'une dimension de *300*.

$$\sum_{i=1}^{n} f_i \cdot \mathbf{w}_i, \mathbf{w_i} \in \mathbb{R}^{300} \tag{3.2.2}$$

Concaténation : La concaténation consiste à mettre bout à bout les plongements lexicaux de tous les mots dans un commentaire. Cependant, les commentaires sont de longueurs variées, nous avons donc traité les commentaires de deux manières pour avoir des vecteurs de même taille : (1) La dimension de tous les vecteurs a été fixée à 50, **la longueur maximale de commentaire**. En l'occurrence, nous avons ajouté des remplissages (en anglais : *paddings*) de zéro pour toutes les séquences ayant moins de 50 tokens. (2) Nous avons pris **la**

longueur moyenne de commentaire, qui était de **10**, comme la longueur de vecteur résultant. Si un commentaire de longueur m avait plus de 10 ($m > 10$) mots, les derniers $m - 10$ mots étaient enlevés, et les dix mots qui restaient étaient représentés par leurs plongements lexicaux concaténés. À l'opposé, nous avons rempli les places blanches par zéro quand la longueur d'un commentaire était plus petite que 10. Pour illustrer, si le commentaire contenait n ($n \leq 10$) tokens, les premiers n tokens seront représentés par leurs plongements lexicaux de dimension de 300, alors que les $10-n$ places qui restaient ont été remplies par une matrice de zéro de dimension $(n-1) \times 300$. Après l'aplatissement, les vecteurs de commentaires étaient de dimension 3000. Enfin, l' ACP a été appliquée afin de réduire la dimension des vecteurs à 50 (nombre de composant pris d'une manière arbitraire).

3.2.2. Classificateurs

Cette partie se consacre à la présentation des classificateurs utilisés dans cette recherche. Il s'agit d'algorithmes d'apprentissage automatique, incluant des méthodes d'apprentissage supervisé et semi-supervisé. Les principaux algorithmes appliqués sont les suivants :

Séparateur à vaste marge (SVM) un modèle linéaire qui cherche à trouver une frontière de séparation qui maximise la distance entre les échantillons les plus proches et la frontière de séparation (*marge*). Les échantillons les plus proches sont appelés *vecteurs de support* (**Figure** 3.1). La fonction de décision se formule comme suit :

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} w_i \times x_i + b\right) \tag{3.2.3}$$

où w_i est le poids ou le coefficient pour le trait x_i, et b représente le biais. Dans un SVM, la magnitude du poids est reliée à l'importance du trait correspondant : plus un poids est grand, plus le trait correspondant contribue à la détermination de la frontière. Le vecteur de coefficients est orthogonal à la fonction de décision. SVM est réputé pour son niveau performances dans des contextes où la quantité de données est limitée, ce qui motive notre choix. De plus, SVM est aussi capable de traiter les données non linéairement séparables, en cherchant les séparer dans un espace de plus grande dimension avec l'astuce de noyau.

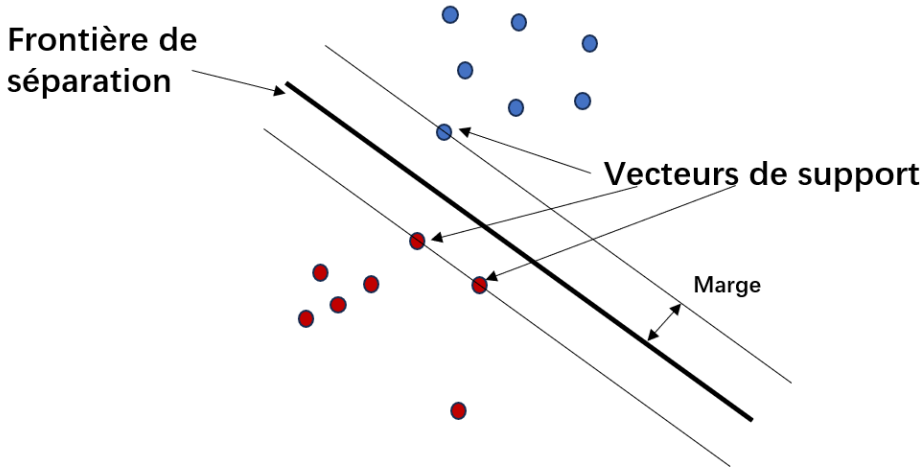

Figure 3.1 – SVM linéaire

SVM avec la fonction de base radiale (FBR) une méthode permettant de traiter le cas où les données sont non linéairement séparables. La FBR transforme les données dans un espace de plus haute dimension dans le but de trouver une frontière de séparation linéaire (**Figure** 3.2).

Auto-formation une méthode d'apprentissage semi-supervisé profitant des données partiellement annotées pour améliorer la performance du classificateur. L'hypothèse centrale derrière l'apprentissage semi-supervisé est que deux points proches dans une région de densité élevée doivent avoir les mêmes étiquettes (Chapelle et al., 2010). Au cours de chaque itération, le modèle Auto-formation attribue à des exemples non étiquetés les étiquettes prédites lorsque la confiance de la prédiction dépasse un seuil prédéfini (ces étiquettes sont nommées *pseudo-étiquettes*). Ensuite, les échantillons pseudo-étiquetés seront combinés avec les données annotées pour entraîner de nouveau le classificateur. Les itérations se terminent lorsque les données sans étiquettes sont toutes pseudo-étiquetées ou qu'il n'y a plus de pseudo-étiquettes à ajouter.

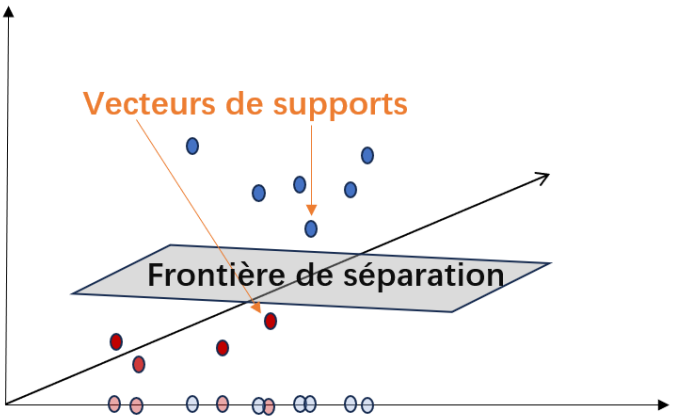

Figure 3.2 – SVM avec la fonction de noyau FBR : les points de couleur claire sur un espace d'une dimension sont les points originaux non linéairement séparables ; les points sont susceptibles de se discriminer par un hyperplan optimal après la transformation.

3.2.3. Méthodes d'apprentissage supervisé

Le SVM est adopté comme modèle de référence dans cette recherche, vu la grande quantité de traits et une quantité limitée d'exemples (2 118 commentaires annotés). Le modèle est utilisé avec un noyau linéaire et un nombre maximal d'itérations de 1 000. On équilibre également les classes : la classe minoritaire est affectée d'un poids plus fort pour limiter l'influence du déséquilibre des données d'entraînement. Voici les modèles de SVM utilisés dans nos expériences :

-SVM$_{onehot\_occ}$: modèle SVM linéaire avec les traits de Onehot agrégés.

-SVM$_{onehot\_tfidf}$: modèle SVM$_{onehot_occ}$ avec les traits pondérés par TF-IDF.

-SVM$_{occ\_SE}$: modèle SVM$_{onehot_occ}$ avec les données d'entraînement sous-échantillonnées.

-SVM$_{concat\_base}$: modèle SVM linéaire avec les traits constitués des plongements lexicaux de mot concaténés.

-SVM$_{aggr\_base}$: modèle SVM linéaire avec les traits constitués des plongements de mot agrégés.

-SVM$_{concat\_SE}$:modèle SVM$_{concat_base}$ appliqué sur les données d'entraînement sous-échantillonnées.

-SVM$_{aggr\_SE}$: le modèle SVM$_{aggr_base}$ appliqué sur les données d'entraînement sous-échantillonnées.

3.2.4. Méthodes d'apprentissage semi-supervisé

Dû à la taille limitée du corpus annoté, les modèles semi-supervisés ont été envisagés pour profiter également des données non-annotées.

Nous avons eu recours au modèle d'auto-formation dans le cadre de cette recherche. 6 000 commentaires sans étiquettes ont été ajoutées dans le jeu d'entraînement. Le seuil de fiabilité a été configuré à 0.7 : seules les prédictions avec une probabilité au dessus de 0.7 étaient valides et les pseudo-étiquettes attribuées pourraient ainsi être acceptées et utilisées au cours de la classification. Le classificateur SVM avec la FBR dans ce cas-là discrimine les données non linéairement séparables. L'itération maximale a été fixée à quatre. Nous avons expérimenté avec les modèles d'auto-formation suivants :

-AF$_{aggr\_base}$: le modèle auto-formation avec les traits de plongements lexicaux de mot agrégés.

-AF$_{aggr\_SE}$: le modèle AF$_{aggr_base}$ appliqué sur les données d'entraînement sous-échantillonnées.

Chapitre 4

Résultats et discussion

4.1. Classes à granularité grosse

4.1.1. Classification binaire

En vue de mesurer les résultats et, compte tenu de notre jeu de données déséquilibré comportant une fréquence plus importante de commentaires de la classe 3, nous nous sommes servis d'une collection de métriques en plus de l'exactitude. Les **Tableaux** de 4.1 à 4.3 montrent les résultats de toutes nos expériences avec les meilleurs résultats en gras. Les **Tableaux** 4.1 et 4.2 rapportent les performances des modèles d'apprentissage supervisé ; le **Tableau** 4.3 concerne les modèles d'apprentissage semi-supervisé. Pour la classification binaire, nous avons désigné la classe majoritaire (classe 3 dans notre taxonomie) avec l'étiquette 1 et l'union des autres classes avec l'étiquette 0. Ainsi P0, R0 et F0 constituent respectivement la précision, le rappel et le score F pour les classes minoritaires réunies ; P1, R1 et F1 représentent la précision, le rappel et le score F pour la classe majoritaire ; E se réfère à l'exactitude ; MaP, MaR et MaF désignent la précision macro, le rappel macro et le score F macro respectivement. Comme la moyenne micro des métriques sont identiques à l'exactitude dans le cas de la classification binaire, nous ne les présentons pas dans les résultats.

TF-IDF n'a pas bien fonctionné dans l'expérience (voir le **Tableau** 4.1). En vérifiant les mots les plus importants dans les matrices formées respectivement de l'occurrence des mots (SVM_{onehot_occ}) et des scores TF-IDF ($\text{SVM}_{onehot\_tfidf}$), nous avons remarqué que des mots considérés comme importants pour le modèle SVM_{onehot_occ} sont dévalorisés dans le modèle

Modèle	P0	R0	F0	P1	R1	F1	E	MaP	MaR	MaF
SVM_{onehot_occ}	**0,49**	0,40	0,44	0,65	**0,73**	**0,69**	**0,60**	0,57	**0,57**	**0,57**
SVM_{onehot_tfidf}	0,42	0,48	0,45	0,63	0,57	0,60	0,54	0,53	0,53	0,53
SVM_{occ_SE}	0,45	**0,77**	**0,56**	**0,72**	0,38	0,50	0,53	**0,58**	**0,57**	0,53

Tableau 4.1 – SVM avec représentation de mots par Onehot

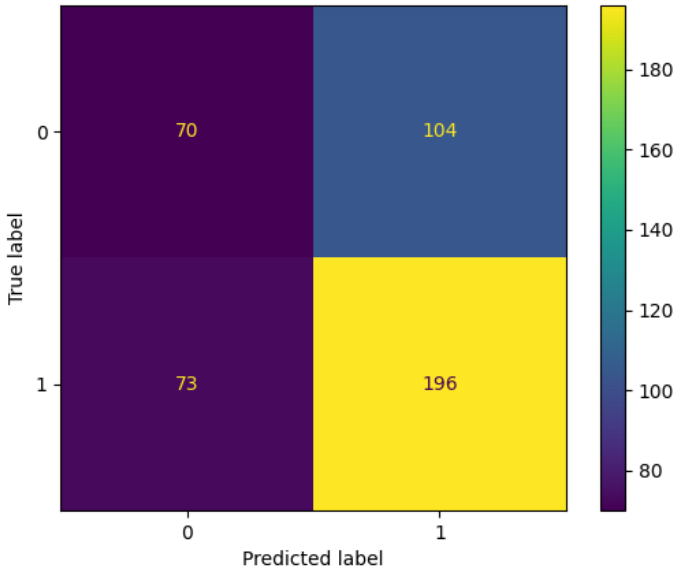

Figure 4.1 – Matrice de confusion pour le modèle SVM_{onehot_occ}

SVM_{onehot_tfidf}. Par exemple, le mot « 霸屏 » (occuper l'écran) est parmi les mots les plus significatifs en termes d'occurrences, et c'est un mot discriminant pour le commentaire qui le contient car il ne se présente que dans ce commentaire. Pourtant, ce même commentaire se répète à travers le corpus, ce qui induit une valeur de TF-IDF très basse pour ce mot. La diminution de la valeur d'un mot critique peut rendre plus difficile la détection des classes.

Dans le **Tableau** 4.2, nous constatons que le classificateur SVM appliqué aux vecteurs formés par l'agrégation des plongements lexicaux de mot fonctionne mieux que sur ceux formés par concaténation. Parmi ces modèles, le modèle SVM_{aggr_base} a obtenu les meilleurs

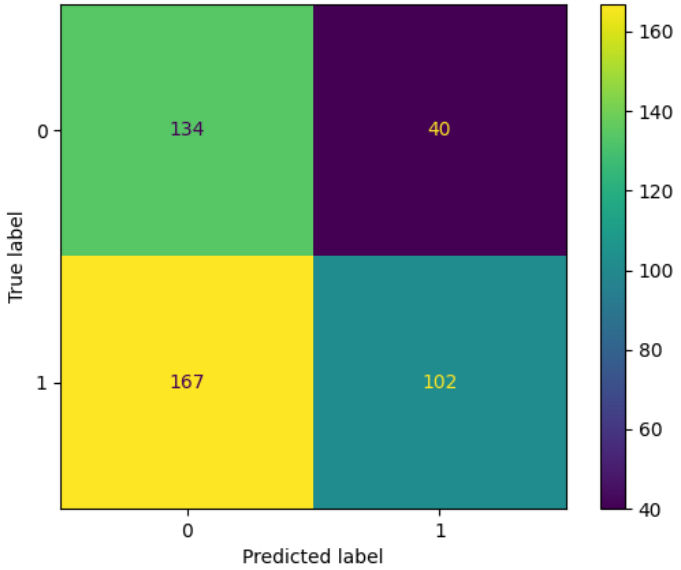

Figure 4.2 – Matrice de confusion pour le modèle SVM_{occ_SE}

Modèle	P0	R0	F0	P1	R1	F1	E	MaP	MaR	MaF
SVM_{concat_base}	0,46	0,48	0,47	0,66	0,64	0,65	0,58	0,56	0,56	0,56
SVM_{concat_SE}	0,44	0,57	0,50	0,66	0,54	0,59	0,55	0,55	0,55	0,55
SVM_{aggr_base}	**0,62**	0,71	**0,66**	0,79	**0,72**	**0,76**	**0,72**	**0,71**	**0,72**	**0,71**
SVM_{aggr_SE}	0,60	**0,74**	**0,66**	**0,80**	0,68	0,74	0,70	0,70	0,71	0,70

Tableau 4.2 – SVM avec représentation de mots par W2V

résultats en comparaison avec les autres. De même, ce modèle a une plus grande tendance à assigner la classe majoritaire, mais le score F macro indique que le modèle a atteint un équilibre de performance entre la prédiction des deux classes. À l'aide de la matrice de confusion de ce modèle (**Figure** 4.3), nous pouvons voir que la plupart des exemples ont été correctement classés. Ainsi, nous avons utilisé ce modèle comme base pour les expériences

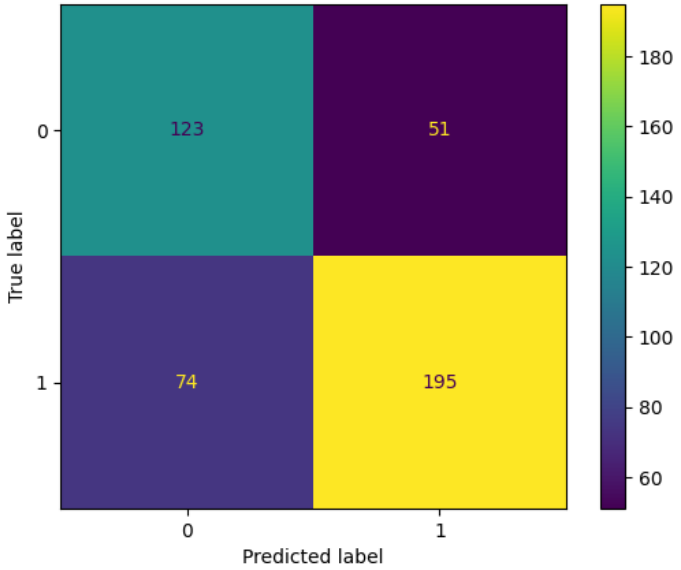

Figure 4.3 – Matrice de confusion pour le modèle SVM_{aggr_base}

ultérieures. Par contre, la technique de sous-échantillonage n'a pas réussi à améliorer davantage les résultats : le classificateur a distribué plus points de données dans la classe 0 en sacrifiant les bonnes prédictions de la classe 1 (voir la **Figure** 4.4).

Modèle	**P0**	**R0**	**F0**	**P1**	**R1**	**F1**	**E**	**MaP**	**MaR**	**MaF**
AF_{aggr_base}	**0,70**	0,52	0,60	0,73	**0,86**	**0,79**	**0,72**	**0,72**	0,69	0,69
AF_{aggr_SE}	0,64	**0,67**	**0,65**	**0,78**	0,75	0,77	**0,72**	0,71	**0,71**	**0,71**

Tableau 4.3 – Résultats pour les modèles d'entraînement automatique

Concernant les modèles semi-supervisés (en voir les résultat dans le **Tableau** 4.3), le modèle AF_{aggr_SE} a eu une performance légèrement meilleure que celui sans sous-échantillonage en termes de score F macro. Le modèle AF_{aggr_base} a obtenu un R1 notable. Par rapport à notre meilleur résultat obtenu par SVM_{aggr_base}, le modèle AF_{aggr_base} est encore plus sensible à la classe 1 et a vu également plus de bonnes prédictions de la classe 1, et cette méthode d'apprentissage semi-supervisé a détecté moins de cas comme étant la classe 0, ce

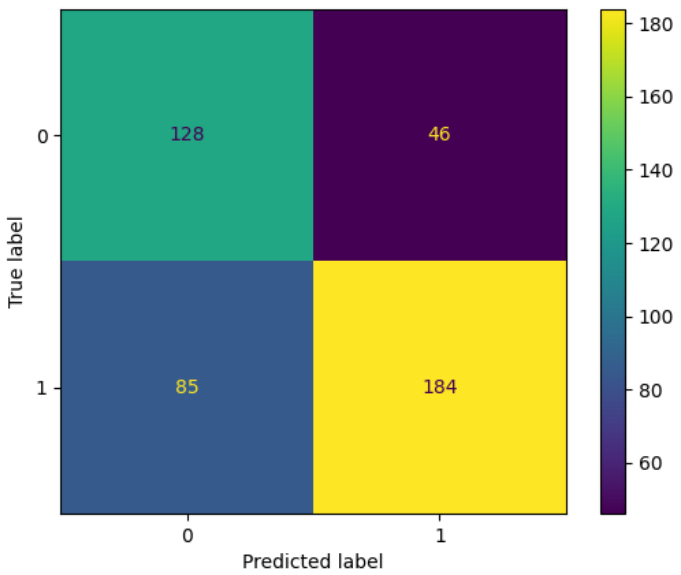

Figure 4.4 – Matrice de confusion pour le modèle SVM_{aggr_SE}

qui se traduit par une augmentation de la précision et une diminution du rappel manifeste pour la classe 0. Mais comme susmentionné, le sous-échantillonage peut aider à identifier plus de cas de la classe 0 par rapport au modèle de base : le modèle AF_{aggr_SE} est arrivé à augmenter le R0 et P1 par rapport au modèle AF_{aggr_base}. Bien que le modèle AF_{aggr_SE} se présente comme le meilleur parmi les modèles semi-supervisés pour le score F macro, il demeure inférieur au meilleur modèle avec vecteurs agrégés de W2V, soit SVM_{aggr_base}. L'apprentissage semi-supervisé n'a donc pas apporté une grande optimisation. Le modèle SVM_{aggr_base} a tendance à prédire la classe 0 tandis que AF_{aggr_SE} prédit plus souvent la classe 1.

Après avoir obtenu le meilleure modèle (SVM_{aggr_base}), nous avons entrepris une recherche par grille pour voir s'il était possible d'apporter une amélioration supplémentaire à la performance des classificateurs en changeant la valeur des hyper-paramètres. Avec la recherche par grille, nous avons parcouru toutes les combinaisons possibles de différentes valeurs d'hyper-paramètres, et retenu le meilleur résultat.

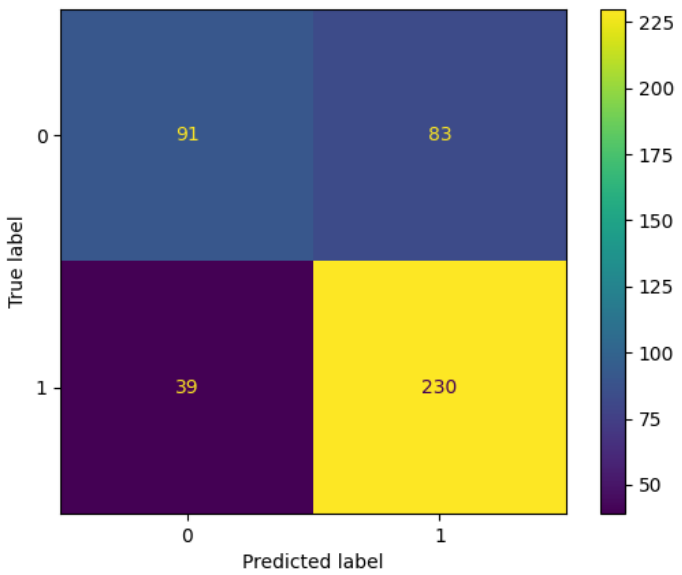

Figure 4.5 – Matrice de confusion pour le modèle AF_{aggr_base}

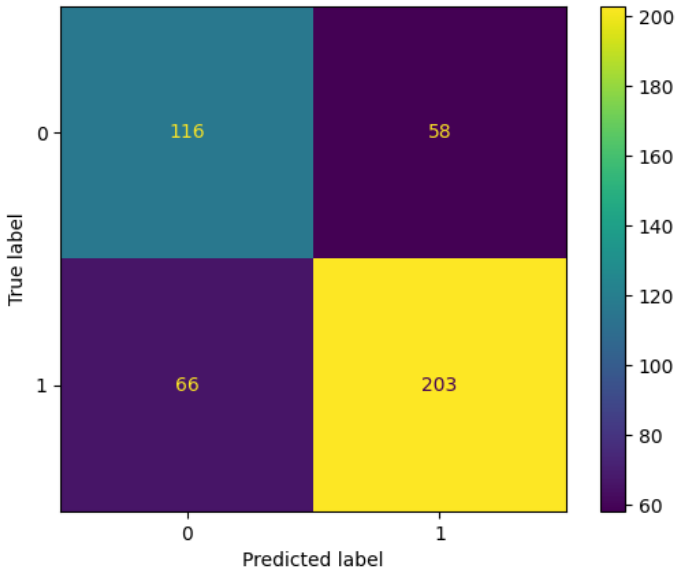

Figure 4.6 – Matrice de confusion pour le modèle AF_{aggr_SE}

Pour SVM, les hyper-paramètres les plus importants sont C, qui est inversement proportionnel à la force de régularisation (quand C est petit, la pénalité pour les mauvaises

classifications est aussi petite, et une marge plus grande va être sélectionnée, et vice-versa) ; le type de noyau (FBR, linéaire ou polynomial) ; et γ, qui contrôle le rayon d'influence des vecteurs de support. Plus γ est grand, plus le rayon d'influence des vecteurs de support est petit, et plus il y a le risque de sur-apprentissage. À l'opposé, si γ est trop petit, la détection de motifs statistiques dans les données est difficile. γ est utilisé quand on se sert du noyau FBR.

Le fait que la recherche par grille essaie toutes les combinaisons d'hyper-paramètres rend cette technique coûteuse en termes de calcul. Alternativement, la recherche aléatoire accélère largement le processus, mais risque de rater les meilleurs résultats. Ainsi, nous avons combiné les deux techniques : une recherche aléatoire en premier afin de trouver la plage optimale, suivie par une recherche sur grille qui a itéré sur des combinaisons d'hyper-paramètres plus précises.

Le **Tableau** 4.4 montre les valeurs des hyper-paramètres de SVM testés :

Hyper-paramètre	**Valeur**				
C	0,1	1,0	10,0	100,0	
Gamma	0,1	1,0	10,0	scale	auto
Noyau	rbf	poly			

Tableau 4.4 – Recherche sur grille pour SVM

Quand la valeur de gamma égale « scale » et « auto », gamma se calcule respectivement par $\frac{1}{N_{caracteres} \times Variance_{caracteres}}$, et $\frac{1}{N_{caracteres}}$, ce qui permet d'ajuster la valeur de gamma en fonction des données d'entraînement , et d'éviter ainsi le sur/sous-apprentissage.

Nos données d'entraînement ont été divisées en trois plis d'une manière stratifiée (en gardant la même distribution des classes dans chaque pli). Les meilleurs hyper-paramètres trouvés par la recherche aléatoire sont présentés dans le **Tableau** 4.5 :

Et le meilleure score F macro obtenu pour les données de test était 0,75, représentant un gain de 0,04 par rapport au modèle original (avec le noyau linéaire et $C = 1{,}0$).

Hyper-paramètre	Valeur
Noyau	rbf
Gamma	auto
C	10,0

Tableau 4.5 – Meilleurs hyper-paramètres avec l'application de la recherche aléatoire sur SVM_{aggr_base}

Les meilleurs résultats restaient inchangés quand nous avons appliqué la recherche par grille en rétrécissant graduellement la plage de valeurs de C.

4.1.2. Classification multiclasses

Les résultats obtenus au cours de la classification binaire montrent que généralement les modèles *base* retournent les meilleurs résultats en termes de score F macro (sauf dans le cas des modèles d'entraînement semi-supervisé où le score F macro du modèle *sous-échantillonnage* est légèrement plus élevé que celui du modèle *base*). Nous avons donc choisi d'appliquer les modèles *base* à la classification multiclasses pour voir s'ils sont susceptibles de faire des distinctions plus approfondies. Les **Tableaux** 4.6 à 4.8 montrent les résultats des modèles *base* et la distribution des étiquettes pour la classification à gros grains (P pour la précision, R pour le rappel, F pour le score F macro, et Support pour le nombre de commentaires concernés).

Le **Tableau** 4.6 montre que le modèle SVM_{onehot_occ} a une probabilité faible d'assigner la classe 4 à tort (avec une précision aussi élevée que 0,80 pour la classe 4), mais également, qu'il a très peu réussi à reconnaître les commentaires qui appartiennent effectivement à la classe 4 (avec un rappel modeste de 0,53 pour la classe 4). Autrement dit, le modèle n'est pas très sensible à la classe 4. En revanche, les commentaires de la classe 3 sont beaucoup plus identifiables pour le classificateur. D'ailleurs, il est remarquable que les scores pour la classe 1 restent toujours très faibles, ce qui découle vraisemblablement de la sous-représentation de cette classe.

Le modèle SVM_{aggr_base} est parvenu à améliorer les scores F macro de la classe 1 et de la classe 2 en élevant considérablement les rappels de ces deux classes (une augmentation de 0,16 pour la classe 1, et une augmentation de 0,27 pour la classe 2), tout en maintenant stables les scores F macro de la classe 3 et de la classe 4 (voir le **Tableau** 4.7).

Par contre, avec le modèle AF_{aggr_base}, les scores de la classe 1 ont diminué davantage : beaucoup d'exemples appartenant à classe 1 ont été transférés dans la classe 3 car le classificateur a identifié presque tous les commentaires appartenant à la classe 3. Cela pourrait expliquer également la croissance substantielle de la précision de la classe 4 (voir le **Tableau** 4.8).

Classe	**P**	**R**	**F**	**Support**
C_1	0.29	0.33	0.31	45
C_2	0,45	0,40	0,42	70
C_3	0,73	**0,78**	**0,75**	269
C_4	**0,80**	0,53	0,64	62
Moyenne macro	0,57	0,51	0,53	446

Tableau 4.6 – Classification multiclasses à granularité grosse avec le modèle SVM_{onehot_occ}

Classe	**P**	**R**	**F**	**Support**
C_1	0,36	0,49	0,42	45
C_2	0,49	0,67	0,57	70
C_3	**0,83**	0,67	**0,74**	269
C_4	0,58	**0,69**	0,63	62
Moyenne macro	0,57	0,63	0,59	446

Tableau 4.7 – Classification multiclasses à granularité grosse avec le modèle SVM_{aggr_base}

Parmi les trois modèles *base*, le modèle AF_{aggr_base} présente les performances les plus médiocres, alors que le modèle SVM_{aggr_base} l'emporte sur les autres. Étant donné que

Classe	P	R	F	Support
C_1	0,18	0,04	0,07	45
C_2	0,68	0,37	0,48	70
C_3	0,69	**0,94**	**0,79**	269
C_4	**0,84**	0,44	0,57	62
Moyenne macro	0,60	0,45	0,48	446

Tableau 4.8 – Classification multiclasses à granularité grosse avec le modèle AF_{aggr_base}

nous avons appliqué, dans la classification binaire, la recherche par grille pour une optimisation plus poussée sur le meilleur modèle, nous avons fait de même pour la classification multiclasses en utilisant les combinaisons dans le **Tableau** 4.4. Les meilleures hyper-paramètres trouvés sont identiques aux hyper-paramètres optimaux dans la classification binaire (**Tableau** 4.5). Le modèle optimisé sera nommé $\text{SVM}^{RBF}_{aggr\_base}$, et ses résultats sont rapportés dans le **Tableau** 4.9.

Classe	P	R	F	Support
C_1	0,45	0,22	0,30	45
C_2	0,59	0,57	0,58	70
C_3	0,77	**0,88**	**0,82**	269
C_4	**0,85**	0,65	0,73	62
Moyenne macro	0,67	0,58	0,61	446

Tableau 4.9 – Classification multiclasses à granularité grosse avec le modèle $\text{SVM}^{RBF}_{aggr\_base}$

En matière du score F, ce modèle a eu les résultats les plus satisfaisants dans la catégorisation de la classe 3, et la majorité des commentaires appartenant à la classe 3 ont été bien détectés. En ce qui concerne les classes qui restent, le modèle performe le mieux dans la détection de la classe 4 avec un score F de 0,73, suivi par celle de la classe 2 dont le score de F est de 0,58. L'identification de la classe 1 est relativement difficile : beaucoup de

commentaires de la classe 1 n'ont pas été repérés (le rappel de la classe 1 est seulement de 0,22). La classe 4 a le moins de cas faux positifs par rapport aux autres classes. (la précision de la classe 4 est la plus élevée, atteignant une valeur de 0,85). Par rapport à SVM_{aggr_base}, la moyenne des scores de F macro s'est légèrement améliorée : à part la classe 1, toutes les autres ont présenté une amélioration de degré variable. Le fait que tous les rappels sauf celui de la classe 3 descende montre que le modèle a redistribué les commentaires précédemment assignés aux autres classes dans la classe 3.

À l'aide de la matrice de confusion (**Figure** 4.7), nous pouvons constater que le modèle a rarement confondu les classes 1, 2 et 4. Particulièrement, les classes 1 et 4 n'ont presque jamais été mélangées (les commentaires de la classe 4 n'ont jamais été prises pour ceux de la classe 1). Par contre, ces commentaires ont souvent été identifiés par erreur comme étant de la classe 3. Cette confusion est prépondérante pour les commentaires de classe 1 et plus modérée pour ceux de classe 4.

En conclusion, le modèle semble capable de plutôt bien différencier les commentaires de la classe 4 des autres. Ce point est particulièrement intéressant puisque les commentaires de classe 4 incluent les commentaires non pertinents alors que les classes 1 et 3 ne comprennent que des commentaires pertinents.

Comme les modèles démontrent globalement une tendance de prédiction similaire —ils présentent tous la meilleure performance dans la catégorisation des classes 3 et 4, suivie par la classe 2, et ont tous une performance relativement modeste quand à la classification des commentaires de la classe 1, nous pourrions peut-être tirer parti des coefficients du modèle SVM_{onehot_occ}, qui ont l'avantage d'être interprétables, pour examiner les mots les plus critiques qui influent sur la décision des modèles.

Les coefficients de la fonction de décision du SVM linéaire révèlent l'impact des caractéristiques (tokens) dans la détermination du séparateur : plus la magnitude des coefficients est grande, plus les caractéristiques correspondantes contribuent à la prédiction d'une classe spécifique.

Pour illustrer, les tokens comme « 鬼见愁 » (Guijianchou, nom d'un personnage dans une série), « 奉先 » (Fengxian, nom d'un personnage dans une autre série) et « 旅座 » (titre

militaire d'un personnage dans une autre série, signifiant « commandant de la brigade ») sont parmi les tokens les plus importants pour identifier la classe 1, car ces mots-là parlent de l'expérience scénique d'un acteur dans d'autres productions. Les mots tels que « 有人 » (Il y a quelqu'un) , « 没人 » (Il n'y a personne) sont parmi les mots critiques pour signaler la classe 2, qui implique les commentaires visant à parler ou à répondre aux autres spectateurs ; les mots exprimant l'émotion incluant « 哭 » (Pleurer), « 卧槽 » (Oh mon dieu), « 呵呵 » (Héhé) sont les plus représentatifs pour indiquer la classe 3 ; et les noms de personne « 陈翔 » (Chen Xiang, nom d'une star, « 庄妍 » (Zhuangyan, nom d'une personne dans la vie réelle) ou bien « 会员 » (Membre du site de vidéo) contribuent à la détection de la classe 4, qui incorpore les commentaires dont le contenu n'a pas de lien direct avec la série.

Cependant, le classificateur accorde aussi de l'attention à des mots qui ne semblent a priori être de bons indicateurs d'une classe spécifique. Par exemple, « 本身 » (lui-même) dans « 他本身就是狼牙特战旅的 » (Il faisait lui-même parti de la brigade des forces spéciales) n'introduit pas nécessairement d'informations préalables ou de contexte par rapport à la série. De même pour certains mots communs tels que « 冲突 » (conflit) dans « 有冲突 » (Il y a un conflit). Le sens de ces mots semble compatible aussi bien avec un usage factuel que subjectif. La raison pour laquelle l'algorithme a attribué un poids fort à ces mots est peut-être le manque d'exemples de classe 1 : le modèle a relevé certains biais accidentels dans les exemples de classe 1 et relie certains mots spécifiques à cette classe, sans que cette association se généralise aux données de test. À titre de comparaison, nous avons analysé les commentaires classés dans la classe 3 par le modèle, dont la proportion était beaucoup plus grande : les mots exprimant l'émotion ou le jugement (la subjectivité) étaient les plus fréquents parmi les vingt mots les plus influents sur la décision du classificateur, ce qui indique que ce dernier était susceptible de s'appuyer sur la présence de ces mots-là pour identifier la classe 3. En revanche, les mots qui ne sont pas nécessairement liés sémantiquement à la définition de la classe 3 étaient plus rares et souvent, ils se rapportaient au scénario.

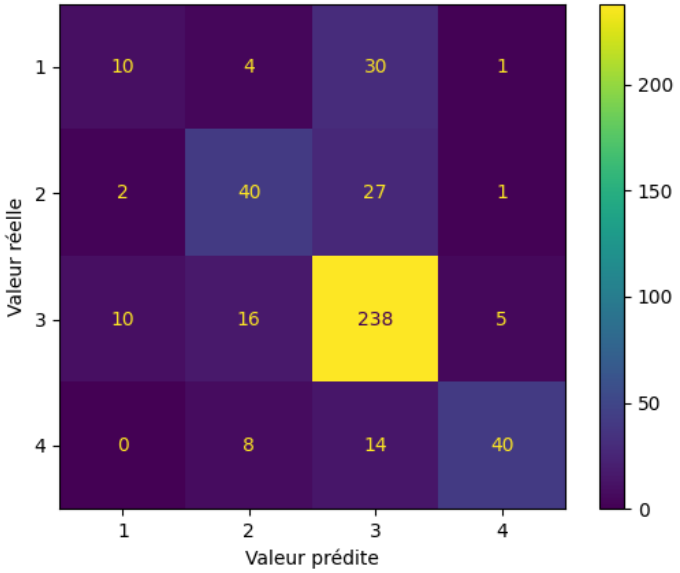

Figure 4.7 – Matrice de confusion pour le modèle $\text{SVM}^{RBF}_{aggr\_base}$ appliqué dans la classification multiclasses

4.2. Classes à granularité fine

Dans les classifications à granularité grosse, la classe 3 était toujours la plus reconnaissable pour la machine. Cependant, nous avons observé beaucoup de désaccords entre les annotatrices concernant les sous-catégories de la classe 3, notamment concernant les paires suivantes :

- 3.1 et 3.3 ;
- 3.1 et 3.4 ;
- 3.1 et 1.3 ;
- 3.3 et 3.4.

Dans notre corpus annoté, la classe 3 représente la classe majoritaire, à l'intérieur de laquelle la sous-catégorie 3.1 occupe une proportion prédominante ($\approx$ 39% de l'ensemble des données et, $\approx$ 63% des données de la classe 3). En phase de co-annotation, des divergences sont survenues entre la classe 3.1 et plusieurs sous-catégories. On se demandait ainsi si ces

divergences étaient également saisissables pour la machine. Nous avons exercé des classifications binaires dont trois entre la sous-catégorie 3.1 et les sous-catégories concurrentes (voir la liste ci-dessus) et une entre les sous-catégories 3.3 et 3.4. Le modèle dont on s'est servi était le meilleur modèle obtenu dans les expériences des sections précédentes—$\text{SVM}^{RBF}_{aggr\_base}$. Les résultats sont résumés dans les **Tableaux** 4.10 à 4.13, où 1 représente la classe majoritaire, et 0 la classe minoritaire. Par exemple, dans le **Tableau** 4.10, P0 constitue la précision de la sous-catégorie 3.3 dont le nombre est inférieur.

P0	**R0**	**F0**	**P1**	**R1**	**F1**	**E**	**MaP**	**MaR**	**MaF**
0,60	0,31	0,41	0,79	0,93	0,85	0,76	0,70	0,62	0,63

Tableau 4.10 – Résultats pour la classification entre les sous catégories 3.1 et 3.3 avec le modèle $\text{SVM}^{RBF}_{aggr\_base}$

En premier, nous avons constaté que la sous-catégorie 3.3 est difficilement perceptible par rapport à la classe 3.1. Comme le démontre le rappel des deux sous-catégories (voir le **Tableau** 4.10), le classificateur ne ratait presque jamais les exemples appartenant à 3.1, mais ignorait fréquemment ceux dans 3.3. La distinction entre ces deux sous-catégories est la plus délicate.

P0	**R0**	**F0**	**P1**	**R1**	**F1**	**E**	**MaP**	**MaR**	**MaF**
0,79	0,48	0,59	0,94	0,98	0,96	0,93	0,86	0,73	0,78

Tableau 4.11 – Résultats pour la classification entre les sous catégories 3.1 et 3.4 avec le modèle $\text{SVM}^{RBF}_{aggr\_base}$

Dans la classification des classes 3.1 et 3.4 (en voir les résultats dans le **Tableau** 4.11), le classificateur a identifié presque tous les commentaires de la classe 3.1, mais a performé d'une façon médiocre dans la détection de la classe 3.4, ce qui est probablement influencé par le nombre limité des commentaires de la classe 3.4. Dans ce cas-là, même s'il y avaient seulement quelques exemples non identifiés, le rappel de classe 3.4 était petit.

P0	R0	F0	P1	R1	F1	E	MaP	MaR	MaF
0,70	0,44	0,54	0,88	0,96	0,92	0,86	0,79	0,70	0,73

Tableau 4.12 – Résultats pour la classification entre les sous-catégories 3.1 et 1.3 avec le modèle $\text{SVM}^{RBF}_{aggr\_base}$

La distinction entre les sous-catégories 3.1 et 1.3 a produit des résultats similaires, mais d'une façon plus modérée (voir le **Tableau** 4.12) : le rappel de la classe minoritaire n'est pas aussi bas que dans le cas précédent. Les exemples dans la classe 1.3 sont légèrement plus nombreux que dans la classe 3.4.

P0	R0	F0	P1	R1	F1	E	MaP	MaR	MaF
0,84	0,70	0,76	0,90	0,96	0,93	0,89	0,87	0,83	0,84

Tableau 4.13 – Résultats pour la classification entre les sous-catégories 3.3 et 3.4 avec le modèle $\text{SVM}^{RBF}_{aggr\_base}$

Dans le **Tableau** 4.13, nous pouvons constater que bien que la différence de quantité entre les exemples dans les classes 3.3 et 3.4 soient proche de celle entre les commentaires dans les classes 3.1 et 3.3 (dans ces deux cas, la classe la moins fréquente représente $\approx 35\%$ de l'union des deux classes), les scores respectifs des classes 3.3 et 3.4 sont moins disparates que dans les cas précédents. On peut voir que le classificateur excellait dans la détection de la classe 3.3 mais qu'il ne s'est pas comporté si mal à trouver les échantillons appartenant à la classe 3.4. Le score F signale aussi que le modèle a plutôt bien discriminé les deux sous-catégories.

4.3. Analyse des erreurs

Dans cette section, nous procédons à l'analyse des cas où les prédictions de la machine (le modèle $\text{SVM}^{RBF}_{aggr\_base}$) diffèrent des annotations manuelles dans les classification à granularité

fine. L'origine de ces différences est diverse, et nous résumons les raisons principales comme suit :

(1) L'acte illocutoire indirect ;

a. 这什么雷威力这么大?
'De quel type est cette grenade d'une puissance aussi forte ?'
(annotation manuelle : 3.3, prédiction : 3.1)

b. 连枪的名字都叫不对。
'Même le nom de fusil a été pris à tort.'
(annotation manuelle : 3.3, prédiction : 3.1)

L'annotatrice a pensé que les commentaires (1a) et (1b) impliquent l'émotion, alors que le modèle les a classés comme jugement. À première vue, les deux exemples n'incluent pas de mots signalant manifestement une émotion. Il faut que l'existence de l'émotion soit expliquée par l'acte illocutoire indirect des phrases. Dans la phrase (1a), il ne s'agit pas d'une interrogation, mais plutôt d'une expression de surprise face à une scène exagérée (le spectateur ne croit pas qu'une grenade peut causer de tels dégâts). Pareillement, la phrase (1b) ne consiste pas à *raconter* le fait que le rôle dans la série s'est trompé du nom du fusil, mais à *blâmer* le personnage, en tant que policier, de mélanger le nom du fusil.

(2) Les actes illocutoires multiples ;

a. 等你们破了案人早跑了重案组。
'Après que vous, l'unité d'enquête criminelle, résolvez l'affaire, la personne est déjà partie depuis longtemps.'
(annotation manuelle : 3.4, prédiction : 3.1)

b. 捡枪啊。
'Ramassez le pistolet.'
(annotation manuelle : 3.4, prédiction : 3.3)

c. 她比火凤凰里的角色好多了。

‘Son rôle (dans cette série) est bien meilleur que dans *Phénix en feu*.’

(annotation manuelle : 3.1, prédiction : 1.3)

d. 又是我是特种兵那些人吗?

‘Ce sont encore les gens qui ont joué dans *Je suis un soldat d’opérations spéciales* ?’

(annotation manuelle : 1.3, prédiction : 3.1)

e. 飞鸿。

‘Feihong (nom d’un personnage).’

(annotation manuelle : 3.3, prédiction : 3.4)

f. 这位姐姐。

‘Cette grande soeur.’

(annotation manuelle : 3.3, prédiction : 3.4)

Certains commentaires peuvent exercer plus d’une fonction (acte illocutoire). En (2a) et (2b), il est évident qu’il existe un interlocuteur potentiel. Mais il est vrai que ces commentaires expriment aussi, d’une certaine manière, un jugement : en (2a), le locuteur pourrait se moquer du retard dans la réaction de l’unité d’enquête criminelle. Dans la phrase (2b), le locuteur est susceptible de critiquer le fait que le personnage ne ramasse pas le pistolet au moment opportun. La phrase (2c) peut être interprétée comme la divulgation d’informations sur l’expérience de performance de l’actrice (qu’elle a joué dans une autre série) et sur le rôle qu’elle a joué dans cette série précédente (qui n’était pas à la hauteur). En même temps, ce commentaire constitue aussi un jugement positif sur le rôle dans la série actuelle. La même chose pour le commentaire (2d). D’une part, le commentaire révèle que les acteurs ont joué aussi dans une autre série. D’autre part, le spectateur exprime une sorte d’agacement sur le groupe d’acteurs immuables dans de différentes séries par le même réalisateur. Nous trouvons aussi dans notre corpus une sorte de commentaires ayant un sens sémantique complet, mais leur longueur est trop court pour déterminer les actes

illocutoires. Par exemple, (2e) et (2f) peuvent servir à interpeller à un personnage, ou exprimer un sentiment/émotion envers ce dernier.

(3) Manque de contexte.

a. 小虫子在燕子心里。

'Le Petit Insecte (surnom d'un personnage dans la série) est dans le coeur d'Hirondelle (surnom d'un autre personnage dans la série).'

(annotation manuelle : 3.1, prédiction : 1.3)

À cause du manque d'informations contextuelles dans (3a), il est difficile de déterminer si ce qui est concerné dans le commentaire s'est déjà passé ou non. Dans le premier cas, il s'agit d'une divulgation de l'intrigue, alors que dans le dernier cas, il est possible qu'il s'agisse d'un jugement subjectif du spectateur : il *juge* que le personnage Hirondelle est amoureux de Petit Insecte.

L'analyse d'erreurs qui précède a établi une liste des cas typiques où les décisions des êtres humains et la machine pourraient diverger. La complexité des actes illocutoires peut entraîner la difficulté dans la distinction des sous-catégories de la classe 3. Et le manque du contexte influence la détermination des classe 1.3 et 3.1.

4.4. Conclusion

Dans ce chapitre, nous avons fait l'état des modèles en vue d'une classification automatique. D'un point de vue général, le modèle SVM_{aggr_base} présente le meilleur résultat tant en classification binaire (**Tableau** 4.2) qu'en classification multiclasses (**Tableau** 4.7). Il était en mesure de faire un équilibre entre la prédiction de la classe majoritaire (la classe 3) et celle des autres classes. Dans la classification binaire, la méthode de sous-échantillonage permet de mitiger dans une certaine mesure l'impact de la classe dominante : les classificateurs ont réussi à détecter plus de cas minoritaires, mais ce, aux dépens de la performance dans la catégorisation de la classe prépondérante (**Tableaux** 4.1 à 4.3). La technique semi-supervisée n'a pas apporté une amélioration significative, car elle a simplement inclus plus de cas dans la classe dominante (**Tableau** 4.3). Particulièrement, au cours de la classification

multiclasses, le modèle *base* d'auto-formation a attribué trop de données dans la classe 3, si bien que les scores des autres classes, notamment ceux de la classe 1, se sont fortement abaissés (**Tableau** 4.8). Les recherches par grille ont légèrement augmenté le score F macro sur la base du modèle SVM_{aggr_base} en utilisant la même combinaison d'hyper-paramètres dans les classifications binaire et multiclasses (**Tableau** 4.9).

Il résulte de nos expériences que la classification binaire automatique semble possible : le meilleur modèle obtenu avec la recherche sur grille a vu un bon score F macro de 0,75 (**Tableau** 4.2). À partir de là, la classification multiclasses était capable d'aller plus loin. Bien que la distinction des classes à granularité grosse reste difficile (surtout, le modèle a éprouvé des difficultés dans la distinction de la classe 1 dont les exemples sont limités en nombre), il semble au moins possible de discriminer la classe 4 des classes 1 et 3, soit les commentaires non pertinents des commentaires pertinents : contrairement aux autres classes principales, les commentaires de la classe 4 n'étaient pas fréquemment mal classifiés en tant qu'appartenant à la classe 3. En examinant les coefficients du modèle SVM_{onehot_occ}, nous avons constaté que les mots ayant joué un rôle critique dans la décision du classificateur pouvaient être discriminants pour chaque catégorie principale telle que les noms de personnage pour la classe 1, ainsi que les mots transmettant la subjectivité pour la classe 3.

Les résultats des classifications à granularié fine (**Tableaux 4.10 à 4.13**) démontrent qu'à l'intérieur de la classe 3, la sous-catégorie 3.1 était la plus facile à détecter pour la machine en raison de sa quantité prédominante. Parmi toutes les expériences, la distinction entre 3.1 et 3.3 était la plus difficile bien que la classe 3.3 ne soit pas la classe la moins représentée. Cela est probablement dû à la nuance entre le jugement et l'émotion. Dans la plupart des cas, l'émission d'un jugement s'accompagne d'émotion. Étant donnée la petite quantité de la classe 3.4, la machine a eu du mal à repérer les exemples de cette classe. Et le modèle distingue bien les sous-catégories 3.3 et 3.4. La performance de cette expérience était la meilleure parmi les expériences sur les sous-catégories.

À travers l'analyse des cas d'erreurs, nous avons trouvé des raisons potentielles conduisant aux difficultés de la classification automatique, à savoir la complexité dans l'interprétation de l'acte illocutoire et le manque du contexte.

Chapitre 5

Conclusion et recherche future

5.1. Conclusion

Ce mémoire se consacre à la classification des danmakus, un type de commentaire synchronisé, dans le but de filtrer les commentaires qui ne sont pas intéressants pour les spectateurs et de porter ainsi une amélioration sur l'expérience de visionnement. Nous avons parvenus à construire un filtrage qui donne la possibilité de masquer à certain degré les commentaires non désirés par l'audience. Nos contributions principales sont :

(1) la création d'un corpus annotées de multi-étiquettes ;

(2) l'élaboration d'une taxonomie basée sur la théorie des actes de parole de manière à catégoriser les commentaires en référence à leurs fonctions et motivations ;

(3) les expériences de référence avec des modèles d'apprentissage supervisé et semi-supervisé aux fins de classification automatique à granularités différentes ;

Dans le cadre de cette recherche, nous avons d'abord établi une taxonomie hiérarchique avec quatre catégories principales ainsi que leurs sous-catégories respectives. Sur la base, une partie du corpus a été annotée et co-annotée. Vu les bons scores inter-annotateurs dans les annotations à granularité grosse (binaire et multiclasses), nous avons réalisé plusieurs expériences d'apprentissage automatique à granularité grosse sur l'ensemble des données annotées à l'aide des méthodes supervisées et semi-supervisées.

Le meilleur modèle ($\text{SVM}^{RBF}_{aggr\_base}$) était, dans la classification binaire, en mesure de distinguer d'une certaine manière les commentaires pertinents (dont le contenu est en lien direct

avec la vidéo) et subjectifs (intégrant la perspective du commentateur dans le commentaire), à savoir ceux qui appartiennent à la classe 3, des autres commentaires.

Dans le cas de la classification multiclasses à granularité grosse, la classe 4 (les commentaires non pertinents) a été bien séparée des classes 1 et 3 (les commentaires pertinents). Particulièrement, le meilleur modèle prend rarement à tort les commentaires pertinents comme non pertinents. Cela est un résultat encourageant pour notre recherche, car nous ne voulons pas filtrer par erreur les commentaires pertinents. En revanche, nous acceptons que certains commentaires non pertinents soient conservés.

Les expériences de classification binaires entre les sous-catégories controversées dans le cas de co-annotation ont démontré que les données étaient aussi biaisées à l'intérieur de la classe 3. La classe 3.1 représentait une majorité absolue, et les autres sous-catégories n'étaient pas bien reconnaissables quand elles étaient confrontées à la classe 3.1. Mais la différenciation entre les classes 3.3 et 3.4 était relativement facile pour le modèle. Cela suggère que les définitions des sous-catégories seraient problématiques dans le sens où elles sont trop minutieuses et qu'il n'y a pas suffisamment d'exemples dans chaque sous-catégorie. En outre, la détermination des sous-catégories est basée sur l'acte de parole des commentaires, qui est souvent compliqué à décider.

5.2. Discussion et recherche future

Dû à la limite du temps et des ressources, il y a des aspects à améliorer dans notre recherche.

En premier lieu, notre guide d'annotation pourrait être amélioré davantage en fusionnant les sous-catégories dont les différences de définition ne sont pas manifestes, notamment dans la classe 3, ce qui se traduit par des résultats médiocres au cours des classifications manuelles et automatiques.

Ensuite, nous n'avons annoté qu'une petite partie du corpus original. De cette manière, nous n'avons pas pu obtenir les étiquettes de *standard de référence* basées sur les résultats de l'adjudication qui sont plus fiables et solides.

Au cours de la classification automatique, pour les instances portant multi-étiquettes, nous avons fait des copies d'instances de sorte que chacune de ces dernières puisse correspondre à une étiquette. Pourtant, cette opération avait le risque de créer des doublons et d'influencer les résultats des expériences. Nous devrons faire appel dans le futur à une alternative pour traiter les données.

Afin d'optimiser les résultats des classificateurs, plusieurs approches méritent d'être explorées : d'abord, il serait intéressant de se servir des lexiques spécifiques pour certaines catégories : comme la classe 3 est marquée par la subjectivité, il serait utile de faire la détection de l'existence de mots d'émotion pour voir si le commentaire implique de la subjectivité. De même, l'inclusion des informations sur les séries, telles que le nom de la série, le groupe d'acteurs, les musiques, etc., pourrait contribuer à distinguer les commentaires pertinents (classes 1 et 3) et non pertinents (classe 4). En plus, au cours des classifications automatiques et manuelles, nous avons trouvé que le manque de contexte peut mener à des difficultés de classification des classes telles que 1.3 et 3.1. Afin de tenir compte du contexte des commentaires, à savoir l'intrigue environnante, il est possible d'appliquer la méthode proposée par Bai et al. (2021) , qui vise à établir une correspondance entre les danmakus et la trame narrative de la vidéo. Cela permettra de déterminer si le commentaire parle de l'intrigue actuelle. En fin de compte, il est intéressant d'essayer d'autres modèles comme les architectures de réseau neurone —Représentations bidirectionnelles d'encodeurs à partir de transformateurs (en anglais: *bidirectional Encoder Representations from Transformers)* (BERT) (Devlin & Chang, 2018) ou Réseaux de Mémoire à Long Terme et de Court Terme (en anglais : Long Short-Term Memory (LSTM) networks) (LSTM) (Sepp & Schmidhuber, 1997) qui sont capables de saisir les nuances contextuelles et de mieux comprendre la sémantique du texte.

Comme dans le guide d'annotation, nous avons aussi discriminé les catégories selon les axes de subjectivité, la pertinence et la temporalité, ce serait intéressant de faire des classifications des données distinguées en fonction de ces trois axes.

Enfin, dans notre corpus annoté, la classe 2 constitue une classe particulière dans le sens où elle ne se définit pas par les trois axes susmentionnés. Par contre, cette classe se caractérise par l'interactivité explicite. Nous avons annoté les informations qui seraient utiles pour la

catégorisation de cette classe, à savoir, les commentaires qui interagissent et leur indexe. Nous espérons que ces données pourraient être exploitées dans les recherches ultérieures.

Bibliographie

Austin, J. (1962). *How to Do Things with Words*. Oxford University Press.

Bai, Q., Hu, Q. V., Ge, L., & He, L. (2019). Stories That Big Danmaku Data Can Tell as a New Media. *IEEE Access*, *7*, 53509-53519. https://doi.org/10.1109/ACCESS.2019.2909054

Bai, Q., Wu, Y., Zhou, J., & He, L. (2021). Aligned variational autoencoder for matching danmaku and video storylines. *Neurocomputing*, *454*, 228-237. https://doi.org/https://doi.org/10.1016/j.neucom.2021.04.118

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Chapelle, O., Scholkopf, B., & Zien, A. (2010). *Semi-Supervised Learning*. The MIT Press.

Chierchia, G. S. M.-G. (1990). *Meaning and Grammar: An Introduction to Semantics*. MA: MIT Press.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Crible, L., & Degand, L. (2019). Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory*, *15*(1), 71-99. https://doi.org/10.1515/cllt-2016-0046

(2016, juillet 5). http://www.datatang.com/data/19300/.

Devlin, J., & Chang, M.-W. (2018). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. *Google AI Blog*.

Fraser, B. (1999). What are discourse markers? [Pragmatics: The Loaded Discipline?]. *Journal of Pragmatics*, *31*(7), 931-952. https://doi.org/https://doi.org/10.1016/S0378-2166(98)00101-5

Grice, H. P. (1975). Logic and Conversation. In *Syntax and Semantics, Vol. 3, Speech Acts*. Academic Press.

Han, C.-h. (2002). Interpreting interrogatives as rhetorical questions. *Lingua*, *112*(3), 201-229. https://doi.org/https://doi.org/10.1016/S0024-3841(01)00044-4

Harris, Z. (1954). Distributional structure. *Word*.

He, Q. (2014). A Study of the Subjectification of the Chinese Word Suoyi. *Open Journal of Modern Linguistics*, (3), 399-406. https://doi.org/10.4236/ojml.2014.43033

HIT Stopwords. (2019). https://github.com/goto456/stopwords

japandict.com. (s. d.). https://www.japandict.com/?s=danmaku+&lang=eng

Katz, E., & Blumler, J. G. (1974). *The uses of mass communications : current perspectives on gratifications research*. Beverley Hills : Sage.

Katz, E., Haas, H., & Gurevitch, M. (1973). On the Use of the Mass Media for Important Things. *American Sociological Review*, *38*(2), 164-181. http://www.jstor.org/stable/2094393

Khan, M. L. (2017). Social Media Engagement:What motivates user participation and consumption on YouTube? *Comput. Hum. Behav.*, *66*(100), 236-247. https://doi.org/10.1016/j.chb.2016.09.024

Kölbel, M. (2003). Faultless Disagreement. *Proceedings of the Aristotelian Society*, *104*(1), 53-73. https://doi.org/10.1111/j.0066-7373.2004.00081.x

Langacker, R. W. (1985). Observations and Speculations on Subjectivity. In J. Haiman (Éd.), *Iconicity in Syntax* (p. 109-150). Benjamins. https://doi.org/http://dx.doi.org/10.1075/tsl.6.07lan

Langacker, R. W. (1990). Subjectification. *Cognitive Linguistics*, *1*(1), 5-38. https://doi.org/doi:10.1515/cogl.1990.1.1.5

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents.

Livingstone, S. (2003). The Changing Nature of Audiences: From the Mass Audience to the Interactive Media User. In *A Companion to Media Studies* (p. 337-359). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9780470999066.ch17

Marchal, M., Scholman, M., Yung, F., & Demberg, V. (2022). Establishing Annotation Quality in Multi-label Annotations. *Proceedings of the 29th International Conference on Computational Linguistics*, 3659-3668. https://aclanthology.org/2022.coling-1.322

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781

Rosenberg, A., & Binkowski, E. (2004). Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. *Proceedings of HLT-NAACL 2004: Short Papers*, 77-80. https://aclanthology.org/N04-4020

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Searle, J. R. (1979). Indirect speech acts. In *Expression and Meaning: Studies in the Theory of Speech Acts* (p. 30-57). Cambridge University Press. https://doi.org/10.1017/CBO9780511609213.004

Sepp, H., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation*. https://doi.org/10.1162/neco.1997.9.8.1735

Smet, H. D., & Verstraete, J.-C. (2006). Coming to terms with subjectivity. *Cognitive Linguistics*, *17*(3), 365-392. https://doi.org/doi:10.1515/COG.2006.011

Sun, J. (2019). jieba [Python package version 0.42.1]. https://github.com/fxsjy/jieba

Traugott, E. C., & Dasher, R. B. (2002). *Regularity in Semantic Change*. Cambridge University Press.

Traugott, E. C. (1989). On the Rise of Epistemic Meanings in English: An Example of Subjectification in Semantic Change. *Language*, *65*(1), 31-55. http://www.jstor.org/stable/414841

Yule, G. (1996). *Pragmatics*. Oxford university press.

Zhao, S. (2019). 中国网络评论发展报告（2019）Annual report on China's online commentary development （2019）. https://www.pishu.com.cn/skwx_ps/bookDetail?SiteID=14&ID=11414566