Traitement automatique des langues

Abusive Language: Linguistic Resources, Methods and Applications

sous la direction de
Delphine Battistelli
Farah Benamara
Viviana Patti

Vol. 65 - n°3 / 2024

ATALA

Abusive Language: Linguistic Resources, Methods and Applications

ATALA
Revue de l'Association pour le Traitement Automatique des Langues

TAL
Vol.
65

n°3
2024

Abusive Language: Linguistic Resources, Methods and Applications

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS.

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Traitement automatique des langues

Volume 65 – n°3 / 2024

ABUSIVE LANGUAGE:LINGUISTIC RESOURCES, METHODS AND APPLICATIONS

Table des matières

Introduction to the Special Issue of the TAL Journal on Abusive Language: Linguistic Resources, Methods and Applications

Delphine Battistelli* — **Farah Benamara**** — **Viviana Patti*****

* *MoDyCo-CNRS, Université Paris Nanterre*
** *Université de Toulouse, IRIT-CNRS, Toulouse INP and IPAL-CNRS Singapore*
*** *Dipartimento di Informatica, Università degli Studi di Torino, Italy*

ABSTRACT. *Abusive language and the propagation of harmful stereotypes have unfortunately become commonplace occurrences on various social media platforms, partly due to users' freedom, anonymity and the lack of regulation provided by these platforms. The sheer volume and often implicit nature of such unwanted content make manual moderation of these user spaces a formidable task. Various scientific communities (Computational Social Science, Natural Language Processing and Computational Linguistics) interested in its at least partial automation have taken up the problem over the past ten years. This special issue aims to encourage interdisciplinary submissions in the field of abusive language detection discussing the limitations of the current approaches and directions for future work.*

KEYWORDS: *Abusive language, Linguistic resources, Automatic detection*

TITRE. *Introduction au numéro spécial de la revue TAL sur le discours de haine : ressources linguistiques, méthodes et applications*

RÉSUMÉ. *Les discours de haine ainsi que la propagation de stéréotypes qui les accompagnent bien souvent sont légion sur les médias sociaux en raison de l'anonymat de leurs utilisateurs, mais aussi du fait du manque de réglementation fournie par les plateformes. Le volume considérable et la nature souvent implicite de ces contenus indésirables font de la modération manuelle une tâche extrêmement complexe. Les sciences sociales computationnelles, le traitement automatique des langues et la linguistique computationnelle se sont emparées de la problématique depuis une dizaine d'années. Ce numéro spécial a pour objectif d'encourager les soumissions interdisciplinaires autour de la tâche de détection de discours de haine tout en abordant les limites des approches actuelles ainsi que les orientations futures.*

MOTS-CLÉS : *Discours de haine, Ressources linguistiques, Détection automatique*

1. Introduction

1.1. *Abusive Language Detection: a Well Established Interdisciplinary Research Field*

Abusive language, hate speech, and the propagation of harmful stereotypes have unfortunately become commonplace occurrences on various social media platforms, due to users' freedom and anonymity and the absence of regulation provided by these platforms. The sheer volume and often implicit nature of such unwanted content make manual moderation of these user spaces a formidable task. Consequently, the Computational Social Science, Natural Language Processing (NLP) and Computational Linguistics communities have proposed numerous works to create resources, datasets, and models aimed at automating the task of abusive language detection (henceforth ALD) (Talat and Hovy, 2016; Fortuna and Nunes, 2018; Vidgen *et al.*, 2019; Fortuna *et al.*, 2020), making it a significant and well-established research area in NLP, with a substantial body of literature.

At the international level, many dedicated workshops have been organized, such as the workshop on Online Abuse and Harms (WOAH) @ACL 2022, ACL 2023, NAACL 2024 (47, 55, 56 submissions respectively) and the workshop on Trolling, Aggression and Cyberbullying @LREC 2020 (70 submissions). We also cite well-attended shared tasks such as HateEval (Basile *et al.*, 2019), OffensEval (Zampieri *et al.*, 2019; Zampieri *et al.*, 2020) and ToxicSpan@ SemEval 2019, 2020 and 2021. For example, 74 (resp. 70) teams submitted papers at HateEval (resp. OffensEval), HateEval being co-organized by one of the coordinator of this special issue. Finally, two special issues of the Journal of Online Social Networks and Media, volume 27, 2022 (Detecting, Understanding and Countering Online Harms) and the Journal of Personal and Ubiquitous Computing, volume 27 (2023) (Intelligent Systems for Tackling Online Harms).

At the national level (i.e., France), most special issues/workshops are multidisciplinary, with a particular focus on approaches from social science and linguistics. For example, the Draine multidisciplinary workshops organized by a French consortium on combating extreme and hate discourse. [1] We also cite "Analyse et exploration des données sociales" (analysis and exploration of social data) (ALIAS) workshop series @TALN 2018 and INFORSID 2019 proposed by the ALIAS GDR-MADICS, a CNRS action on cyberviolence and extreme ideology in social media, [2] founded by the two French coordinators of this special issue. We finally cite a special issue of the Journal MOTS in 2021. [3]

1. `https://groupedraine.github.io/`.
2. `https://www.madics.fr/event/1520426929-3916/`.
3. `https://shs.cairn.info/journal-mots?lang=en`.

1.2. *Abusive Language: a Complex Phenomenon*

Following Poletto *et al.,* (Poletto *et al.*, 2021a), we use here "Abusive Language" (AL) as an umbrella term to refer to the various forms of harmful language, such as toxic, offensive language, hate speech, and stereotypes. The reader can refer to Vidgen *et al.,* and Madukwe *et al.,* for a discussion on the lack of universal definitions and its impact on automatic detection (Vidgen *et al.*, 2019; Madukwe *et al.*, 2020). For comprehensive overviews of this field, we recommend surveys such as Schmidt *et al.,* Fortuna and Nunes, Vidgen and Derczynski, Poletto *et al.,* Yin and Zubiaga, and Pamungkas *et al.,* (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Vidgen and Derczynski, 2020; Poletto *et al.*, 2021a; Yin and Zubiaga, 2021; Pamungkas *et al.*, 2023).

AL is topically focused (misogyny, sexism, racism, xenophobia, homophobia, etc.), and each specific manifestation of hate speech targets different vulnerable groups based on characteristics such as gender (misogyny, sexism), ethnicity, race, religion (xenophobia, racism, islamophobia), sexual orientation (homophobia), and so on. Most automatic abusive language detection approaches cast the problem into a binary classification task by neglecting three crutial aspects: (1) the topical focus or the target-oriented nature of hate speech ; (2) the degree of engagement of users in toxic content (denunciation, approbation, reporting and neutral attitude, etc.) ; (3) the question of stereotypes. Furthermore, most of the work (resources, classifiers) is developed for English.

Thus, the scientific challenges are numerous. For Computational Linguistics, the challenge is also linked to the development of methods capable of processing heterogeneous (different topics, various structures and volume) and noisy content (possible presence of abbreviations, smileys or even-sentences in several languages). For Machine Learning methods, a major challenge remains adaptation to a field in constant evolution both in its content (e.g., emerging topics in propaganda rhetoric) and its form. A transverse lock is the constitution of a coherent knowledge base supported by a formal model highlighting both the indices and risk factors provided by sociological models as well as their linguistic anchoring in the content retrieved from the Internet. These challenges show that addressing the threat posed by message and idea propagation to societal security requires a deeper understanding of the linguistic and extra-linguistic content (see for instance Ricardo *et al.,* Chiril *et al.,* Dragos *et al.,* (Ricardo *et al.*, 2018; Chiril *et al.*, 2022; Dragos *et al.*, 2022) on the role of emotion categories in toxic languages; or Poletto *et al.,* and Batistelli et *al.,* about the question of degrees of or engagement in hatefulness (Poletto *et al.*, 2019; Battistelli *et al.*, 2024)).

2. Abusive Language Detection: Current Research and Future Directions

As we said before, ALD has received a growing attention within the field of NLP (Poletto *et al.*, 2021b; Plaza-del Arco *et al.*, 2023; Röttger *et al.*, 2021; Malik

et al., 2024; Nozza *et al.*, 2022), emerging as a fundamental tool for many purposes. Such purposes are ranging from the development of platforms for hate speech monitoring in social media to map vulnerable groups and support policy actions (Capozzi *et al.*, 2020; Laurent, 2020), to the recognition of new targets and vulnerable identities that may become targets of hate speech at a certain historical moment or social climate (Guillén-Pacho *et al.*, 2024); from supporting anti-discrimination educational programs in schools (D'Errico *et al.*, 2024; Cignarella *et al.*, 2023; Cignarella *et al.*, 2024) to moderating online content to prevent the proliferation of hate speech before it causes harm, a purpose as relevant as ever, also considering the recent integration of the Revised EU Code of Conduct on Countering Illegal Hate Speech Online (*Code of Conduct+*) into the Digital Services Act (DSA) regulatory framework, which imposes stricter obligations on online platforms regarding the detection and removal of illegal hate speech. [4]

Recently, the adaptability and flexibility of transformer-based models and Large Language Models (LLMs) led various scholars to focus more and more on exploring and detecting the different nuances that AL could assume depending on diverse contexts, topical focuses and targets. This has encouraged the development of increasingly precise models capable of capturing the specific shapes that AL assumes depending on the affected target, such as misogyny (Rehman *et al.*, 2025; Jiang *et al.*, 2024; Hashmi *et al.*, 2025; Muti *et al.*, 2024; Mohasseb *et al.*, 2025; Pamungkas *et al.*, 2020b), sexism (Plaza *et al.*, 2023; Kirk *et al.*, 2023), homophobic and transphobic discourses (Nozza *et al.*, 2023; Gómez-Adorno *et al.*, 2024). However, even though this research field is now widespread and state-of-the-art models achieve good results, detecting and moderating online abuse remains a complex task, with an increasing awareness of the intertwining of technical, social, legal, and ethical challenges (Cao *et al.*, 2024; Dong *et al.*, 2024; Elesedy *et al.*, 2024).

It remains challenging to provide a univocal definition of what constitutes hate speech (Korre *et al.*, 2025) and to determine the extent to which certain terms should be considered harmful. Different scholars highlighted that AL is commonly a context-dependent phenomenon (Anderson and Barnes, 2022; Brown, 2017; Yoder *et al.*, 2022), and it is often simplistic to classify hate speech using clear-cut boundaries (Parker and Ruths, 2023; Draetta *et al.*, 2024), noting that some terms can assume different meanings depending on the background and the communicative intent of the speaker (Pamungkas *et al.*, 2020a; Pamungkas *et al.*, 2023; Zsisku *et al.*, 2024). For instance, contrastive non-hate variations, such as counter-speech (Yu *et al.*, 2022; Cepollaro *et al.*, 2023; Bonaldi *et al.*, 2024), often blur the line between harmful and not-harmful language.

To properly support content moderation, AL detection systems must be sophisticated enough to identify also hard cases. Recent studies (Dias Oliva *et al.*, 2021; Zsisku *et al.*, 2024; Sap *et al.*, 2019) highlighted that state-of-the-art ALD models

4. `https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online`.

are at risk of both over-moderation (i.e., classifying non-hateful content as hateful) and under-moderation (i.e., failing to detect and classify hateful content), potentially leading to the removal of not abusive speech and, paradoxically, contributing to the marginalization of vulnerable groups. This can be also related to the fact that models still struggle in distinguishing between abusive and not-abusive swearing contexts (Pamungkas *et al.*, 2023), disregarding the multifaceted nature of slurs, which are often used with positive social functions (Jay, 2009). In line with this, a still open challenge is the detection of reclamatory uses of slurs (Cepollaro and de Sa, 2023), where members of target groups re-purpose the terms historically used to derogate their group, to express belonging and identity, manifesting solidarity and subverting structures of discrimination. This phenomenon is mostly overlooked in NLP (Zsisku *et al.*, 2024; Draetta *et al.*, 2024; Röttger *et al.*, 2021), and this feeds into the risk of removing legal speech in content moderation, with the paradoxical outcome of hurting the categories of users that one would like to protect. This can be taken as a concrete example for the need to develop socially relevant AL detection models, able to recognize authentic uses in different contexts, embracing new practice of inclusive design in the development of ALD corpora.

Other open challenges, also relevant for mitigating the under-moderation risks in the current ALD systems, are related to the need for a deeper exploration of the nuanced ways online harms manifest (for instance analyzing the relationship between the linguistic expressions of gender-based violence (GBV) in news and responsibility perception (Minnema *et al.*, 2022; Ferrando *et al.*, 2024), and the capability of the ALD systems to recognize also implicit manifestations of abusive language, as the ones featured by the presence of figurative language and sarcastic devices (Frenda *et al.*, 2022; Frenda *et al.*, 2023)).

Looking at the challenge of monitoring users' opinions and hate in online social platforms across time, the availability of large annotated corpora from social media and the development of powerful classification approaches have contributed in an unprecedented way to tackle the issue. Such linguistic data are strongly affected by events and topic discourse, and this aspect is crucial when detecting phenomena such as hate speech, especially from a diachronic perspective. In this context, temporal robustness of hate speech detection and monitoring systems is still a challenge. First findings on data from the real case study of the "Contro l'Odio" platform for monitoring hate speech against immigrants in the Italian Twittersphere (Florio *et al.*, 2020) highlighted the limits that supervised classification models encounter on data that are heavily influenced by events. Future approaches to be investigated could rely, on the one hand, on computational approaches to lexical semantic change detection (Tahmasebi *et al.*, 2018), on the other hand, on techniques of Longitudinal Evaluation of Model Performance that have been recently applied in the context of sentiment analysis in the LongEval CLEF 2023 challenge (Alkhalifa *et al.*, 2023).

Finally, interdisciplinary research and involvement of social scientists, cultural scholars, and practitioners seem to be more and more the key to address the NLP challenges related to the positive final aim of promoting inclusive and fair language,

as several ethical questions arise, particularly concerning how certain linguistic uses are perceived by the target communities. Understanding such perceptions and integrating participatory design methods (Caselli *et al.*, 2021) is crucial for achieving, on the one hand, a more accurate representation of language in ALD datasets (revising common practices in data collection and annotation (Frenda *et al.*, 2024; Madeddu *et al.*, 2023)) and, by extension, cultural diversity in NLP models.

3. Submission Topics

Motivated by the interest of the community in the problem of ALD, we invited papers from Natural Language Processing, Machine Learning, Computational Social Sciences, and Linguistics. We explicitly encouraged interdisciplinary submissions including linguistics resources, methods, end-user applications but also position papers on the actual state of the art in the field discussing the limitations of the current approaches and directions for future work. The topics covered by the special issue include, but are not limited to:

– linguistic resources and evaluation: annotation scheme, corpus linguistics studies, new datasets, with a particular interest on the French language and/or multilingual resources;

– formal/conceptual approaches for AL as inspired by sociological and psychological models;

– models and methods: supervised and non supervised approaches, including LLMs;

– role of contextual phenomena, including discourse, extra-linguistic (e.g., cultural aspects) context;

– models for cross-lingual and multimodal detection;

– new approaches beyond binary classification: target-oriented ALD, degree of user engagement;

– dynamics of online AL in social media, propaganda propagation;

– bias detection and removal in resource creation, datasets and methods;

– application of ALD tools in education, social media content moderation, etc.;

– social, legal, and ethical implications of detecting, monitoring and moderating AL.

The call for papers for this special issue has been launched in February 2024, with a deadline fixed to mid-June 2024.

4. Reviewing and Selection of Papers

Five papers (two in French and three in English) have been submitted, covering a large spectrum in the field ranging from linguistic resource creation, corpus-based

analysis, and automatic detection. We received submissions from Senegal, India, Germany, Italy and France. Each article has been reviewed by three experts: two members of the special issue scientific committee and one member of the journal editorial board. The first round of reviews has been discussed with the editorial board and the guest editors, resulting in the selection of three papers for a second round of reviews, among which two are in English. The final decisions were made in February 2025 where three papers have been accepted, resulting in a selection rate of 60%.

5. Accepted Papers

The aim of this special issue was to report on some recent and innovative methodological angle of attack of what is referred to as Abusive Language circulating on the internet. The accepted papers contribute to this end. Each of them proposes a new dataset related to abusive language (one for French, one for German and one for English) with rigorous indications about the ways the resource has been built. They also offer a set of classification experiments aiming at characterizing and distinguishing abusive language from other types of language. It appears clearly that the classification tasks are necessarily closely linked to how the datasets have been constructed; thus, the ways of investigating correlations between linguistic characteristics and abusive language are necessarily different but offer both interesting results. In a little more detail, the content of the articles is as follows:

– *CyberAgressionAdo-Large: French Multiparty Chat Dataset to Address Online Hate* (by Anaïs Ollagnier, Elena Cabrio, Serena Villata and Valerio Basile) describes a dataset of conversations written by French teenagers involving cyberbullying situations. The adopted methodology for building this dataset consisted in organizing, in close collaboration with sociologists and experts in education, role-playing games addressing different topics (homophobia, religion, etc.) and with annotations belonging to several dimensions (hate, target, verbal abuse, etc.). The paper details the annotation procedure and presents statistical insights to measure divergences across groups of annotations and a study of the most frequent annotated patterns;

– *Comparaison de méthodes pour la détection du discours des incels sur Reddit (Comparing methods for detecting incels' speech on Reddit)* (by Camille Demers and Dominic Forest) addresses the problem of analyzing and then detecting incels' comments in English-speaking forum Reddit. The hypothesis is that incel communities' speech can be violent and therefore be considered as abusive language, particularly against women. The dataset is created by labelling comments according to their community label, not according to their content. Then the learning experiments are based on the Bag-of-Communities method in which a comment is labeled according to the subreddit it originated from. The authors also proposes a set of lexical-based analysis to identify specific lexical units that are more likely to be predictive of incel-type content;

– *Automated Speech Act Classification in Offensive German Language Tweets* (by Melina Plakidis, Elena Leitner and Georg Rehm) presents a manually annotated

dataset of tweets in German according to speech act theory. The hypothesis is that the annotation of speech acts could improve the detection of abusive language. The data used come from the 2018 and 2019 editions of GermEval, a community shared task that focuses on abusive language phenomena. The authors present a correlation study between speech acts and hate-speech annotations with the annotations from the shared task, observing a difference in distribution between the categories. The dataset is then used to train a classifier.

Acknowledgements

We gratefully thank the TAL editors-in-chief, especially Maxime Amblard and Cécile Fabre, for inviting us to coordinate this special issue, and for their supports and guidelines along the whole editing process. We also thank the journal editorial board and the special issue reviewing committee for their work and reactivity: Elena Cabrio (University of Côte d'Azur), Marie Candito (University of Paris Cité), Tommaso Caselli (Faculty of Arts, Rijksuniveristeit Groningen), Vincent Claveau (CNRS IRISA), Valentina Dragos (ONERA), Benoît Favre (University of Aix-Marseille), Claire Hugonnier (University of Grenoble Alpes), Irina Illina (University of Lorraine), Véronique Moriceau (Toulouse University), Frédérique Segond (Inria and INALCO), Didier Schwab (University of Grenoble Alpes), Mathieu Valette (INALCO), Samuel Vernet (University of Aix-Marseille), and François Yvon (CNRS and Sorbonne University).

The work of Farah Benamara is partially supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. The work of Viviana Patti was also partially supported by "HARMONIA" project – M4-C2, I1.3 Partenariati Estesi – Cascade Call – FAIR – CUP C63C22000770006 – PE PE0000013 under the NextGenerationEU programme. Farah Benamara and Viviana Patti have also been supported by the International project STERHEOTYPES (Studying European Racial Hoaxes and sterEOTYPES) funded by the Compagnia di San Paolo and VolksWagen Stiftung under the Challenges for Europe call for Project (CUP: B99C20000640007). The work of Delphine Battistelli was partially supported by the project FLYER (Artificial intelligence for extremist content analysis) – ANR-19-ASTR-0012.

6. References

Alkhalifa R., Bilal I., Borkakoty H., Camacho-Collados J., Deveaud R., El-Ebshihy A., Espinosa-Anke L., Gonzalez-Saez G., Galuščáková P., Goeuriot L. *et al.*, "Extended Overview of the CLEF-2023 LongEval Lab on Longitudinal Evaluation of Model Performance", *CEUR Workshop Proceedings*, vol. 3497, CEUR-WS, p. 2181-2203, 2023.

Anderson L., Barnes M. R., "Hate Speech", *in* E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, The Metaphysics Research Lab, Philosophy Department, Stanford University, 2022.

Basile V., Bosco C., Fersini E., Nozza D., Patti V., Rangel Pardo F. M., Rosso P., Sanguinetti M., "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter", *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, p. 54-63, June, 2019.

Battistelli D., Dragos V., Mekki J., "Annotating social data with speaker/user engagement. Illustration on online hate characterization in French", *Fortino, G., Kumar, A., Swaroop, A., Shukla, P. (eds) Proceedings of Third International Conference on Computing and Communication Networks: ICCCN 2023*, Lecture Notes in Networks and Systems, vol 917. Springer, Singapore, p. 317-330, 2024.

Bonaldi H., Chung Y.-L., Abercrombie G., Guerini M., "NLP for Counterspeech against Hate: A Survey and How-To Guide", *in* K. Duh, H. Gomez, S. Bethard (eds), *Findings of the Association for Computational Linguistics: NAACL 2024*, Association for Computational Linguistics, Mexico City, Mexico, p. 3480-3499, June, 2024.

Brown A., "What is hate speech? Part 2: Family resemblances", *Law and Philosophy*, vol. 36, p. 561-613, 2017.

Cao Y. T., Domingo L.-F., Gilbert S., Mazurek M. L., Shilton K., Daumé Iii H., "Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators through a User-Centric Method", *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, p. 3567-3587, November, 2024.

Capozzi A. T., Lai M., Basile V., Poletto F., Sanguinetti M., Bosco C., Patti V., Ruffo G., Musto C., Polignano M., Semeraro G., Stranisci M., "Contro L'Odio: A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media", *IJCoL (Torino)*, vol. 6, n^{o} 1, p. 77-97, 2020.

Caselli T., Cibin R., Conforti C., Encinas E., Teli M., "Guiding Principles for Participatory Design-inspired Natural Language Processing", *in* A. Field, S. Prabhumoye, M. Sap, Z. Jin, J. Zhao, C. Brockett (eds), *Proceedings of the 1st Workshop on NLP for Positive Impact*, Association for Computational Linguistics, Online, p. 27-35, August, 2021.

Cepollaro B., de Sa D. L., "The successes of reclamation", *Synthese*, vol. 202, n^{o} 6, p. 205, 2023.

Cepollaro B., Lepoutre M., Simpson R. M., "Counterspeech", *Philosophy Compass*, vol. 18, n^{o} 1, p. e12890, 2023.

Chiril P., Pamungkas E., Benamara F., Moriceau V., Patti V., "Emotionally Informed Hate Speech Detection: A Multi-target Perspective", *Cognitive Computation*, vol. 14, p. 322–352, 2022.

Cignarella A. T., Chierchiello E., Ferrando C., Frenda S., Lo S. M., Marra A., "From Hate Speech to Societal Empowerment: A Pedagogical Journey Through Computational Thinking and NLP for High School Students", *in* S. Al-azzawi, L. Biester, G. Kovács, A. Marasović, L. Mathur, M. Mieskes, L. Weissweiler (eds), *Proceedings of the Sixth Workshop on Teaching NLP*, Association for Computational Linguistics, Bangkok, Thailand, p. 54-65, August, 2024.

Cignarella A. T., Frenda S., Lai M., Patti V., Bosco C., "DeactivHate: An Educational Experience for Recognizing and Counteracting Online Hate Speech", *IJCoL (Torino)*, vol. 9, n^{o} 2, p. 1007-1023, 2023.

D'Errico F., Bosco C., Paciello M., Benamara F., Cicirelli P. G., Patti V., Moriceau V., Taulé M., "SteRHeotypes Project. Detecting and Countering Ethnic Stereotypes emerging from Italian, Spanish and French Racial hoaxes", *in* A. Bonet-Jover, R. Sepúlveda-Torres, R. M. Guillena, E. Martínez-Cámara, E. L. Pastor, Á. Rodrigo-Yuste, A. Atutxa (eds), *Proceedings of the Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations (SEPLN-CEDI-PD 2024) co-located with the 7th Spanish Conference on Informatics (CEDI 2024), A Coruña, Spain, June 19-20, 2024*, vol. 3729 of *CEUR Workshop Proceedings*, CEUR-WS.org, p. 77-81, 2024.

Dias Oliva T., Antonialli D. M., Gomes A., "Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online", *Sexuality & Culture*, vol. 25, p. 700-732, 2021.

Dong Z., Zhou Z., Yang C., Shao J., Qiao Y., "Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey", *in* K. Duh, H. Gomez, S. Bethard (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, Mexico, p. 6734-6747, June, 2024.

Draetta L., Ferrando C., Cuccarini M., James L., Patti V., "ReCLAIM Project: Exploring Italian Slurs Reappropriation with Large Language Models", *in* F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (eds), *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, CEUR Workshop Proceedings, Pisa, Italy, p. 335-342, December, 2024.

Dragos V., Battistelli D., Étienne A., Constable Y., "Angry or Sad? Emotion Annotation for Extremist Content Characterisation", *LREC*, European Language Resources Association, p. 193-201, 2022.

Elesedy H., Esperanca P. M., Oprea S. V., Ozay M., "LoRA-Guard: Parameter-Efficient Guardrail Adaptation for Content Moderation of Large Language Models", *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, p. 11746-11765, November, 2024.

Ferrando C., Madeddu M., Patti V., Lai M., Pasini S., Telari G., Antola B., "Exploring YouTube Comments Reacting to Femicide News in Italian", *in* F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (eds), *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, CEUR Workshop Proceedings, Pisa, Italy, p. 356-365, December, 2024.

Florio K., Basile V., Polignano M., Basile P., Patti V., "Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media", *Applied Sciences*, 2020.

Fortuna P., Nunes S., "A Survey on Automatic Detection of Hate Speech in Text", *ACM Computing Surveys*, vol. 51, n° 4, p. 85:1-85:30, July, 2018.

Fortuna P., Soler J., Wanner L., "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets", *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 6786-6794, May, 2020.

Frenda S., Abercrombie G., Basile V., Pedrani A., Panizzon R., Cignarella A. T., Marco C., Bernardi D., "Perspectivist approaches to natural language processing: a survey", *Language Resources and Evaluation*, vol. 59, p. 1-28, 2024.

Frenda S., Cignarella A. T., Basile V., Bosco C., Patti V., Rosso P., "The unbearable hurtfulness of sarcasm", *Expert Systems with Applications*, vol. 193, p. 116398, 2022.

Frenda S., Patti V., Rosso P., "When Sarcasm Hurts: Irony-Aware Models for Abusive Language Detection", *in* A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (eds), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, vol. 14163 of *Lecture Notes in Computer Science*, Springer, p. 34-47, 2023.

Gómez-Adorno H., Bel-Enguix G., Calvo H., Ojeda-Trueba S., Andersen S. T., Vásquez J., Alcántara T., Soto M., Macias C., "Overview of homo-mex at iberlef 2024: Hate speech detection towards the Mexican Spanish speaking LGBT+ population", *Procesamiento del Lenguaje Natural*, vol. 73, p. 393-405, 2024.

Guillén-Pacho I., Longo A., Stranisci M. A., Patti V., Badenes-Olmedo C., "The Vulnerable Identities Recognition Corpus (VIRC) for Hate Speech Analysis", *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, CEUR, 2024.

Hashmi E., Yayilgan S. Y., Yamin M. M., Ullah M., "Enhancing misogyny detection in bilingual texts using explainable AI and multilingual fine-tuned transformers", *Complex & Intelligent Systems*, vol. 11, nº 1, p. 39, 2025.

Jay T., "Do offensive words harm people?", *Psychology, public policy, and law*, vol. 15, nº 2, p. 81, 2009.

Jiang A., Vitsakis N., Dinkar T., Abercrombie G., Konstas I., "Re-examining Sexism and Misogyny Classification with Annotator Attitudes", *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, Miami, Florida, USA, p. 15103-15125, November, 2024.

Kirk H., Yin W., Vidgen B., Röttger P., "SemEval-2023 Task 10: Explainable Detection of Online Sexism", *in* A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (eds), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Association for Computational Linguistics, Toronto, Canada, p. 2193-2210, July, 2023.

Korre K., Muti A., Ruggeri F., Barrón-Cedeño A., "Untangling Hate Speech Definitions: A Semantic Componential Analysis Across Cultures and Domains", *Findings at NAACL 2025*, 2025.

Laurent M., "Project Hatemeter: helping NGOs and Social Science researchers to analyze and prevent anti-Muslim hate speech on social media", *Procedia Computer Science*, vol. 176, p. 2143-2153, 2020. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020.

Madeddu M., Frenda S., Lai M., Patti V., Basile V., "DisaggregHate It Corpus: A Disaggregated Italian Dataset of Hate Speech", *in* F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (eds), *Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023*, vol. 3596 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

Madukwe K., Gao X., Xue B., "In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets", *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Association for Computational Linguistics, Online, p. 150-161, November, 2020.

Malik J. S., Qiao H., Pang G., van den Hengel A., "Deep learning for hate speech detection: a comparative study", *International Journal of Data Science and Analytics*, vol. 12, p. 1-16, 2024.

Minnema G., Gemelli S., Zanchi C., Caselli T., Nissim M., "Dead or Murdered? Predicting Responsibility Perception in Femicide News Reports", *in* Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (eds), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online only, p. 1078-1090, November, 2022.

Mohasseb A., Amer E., Chiroma F., Tranchese A., "Leveraging Advanced NLP Techniques and Data Augmentation to Enhance Online Misogyny Detection", *Applied Sciences*, vol. 15, nº 2, p. 856, 2025.

Muti A., Ruggeri F., Khatib K. A., Barrón-Cedeño A., Caselli T., "Language is Scary when Over-Analyzed: Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts", *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, p. 21091-21107, November, 2024.

Nozza D., Bianchi F., Attanasio G., "HATE-ITA: Hate Speech Detection in Italian Social Media Text", *in* K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, Z. Talat (eds), *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Association for Computational Linguistics, Seattle, Washington (Hybrid), p. 252-260, July, 2022.

Nozza D., Cignarella A. T., Damo G., Caselli T., Patti V., "HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian", *in* M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (eds), *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023*, vol. 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

Pamungkas E. W., Basile V., Patti V., "Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media", *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 6237-6246, May, 2020a.

Pamungkas E. W., Basile V., Patti V., "Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study", *Inf. Process. Manag.*, vol. 57, nº 6, p. 102360, 2020b.

Pamungkas E. W., Basile V., Patti V., "Towards multidomain and multilingual abusive language detection: a survey", *Pers. Ubiquitous Comput.*, vol. 27, nº 1, p. 17-43, 2023.

Parker S., Ruths D., "Is hate speech detection the solution the world wants?", *Proceedings of the National Academy of Sciences*, vol. 120, nº 10, p. e2209384120, 2023.

Plaza-del Arco F. M., Nozza D., Hovy D. *et al.*, "Respectful or toxic? using zero-shot learning with language models to detect hate speech", *The 7th Workshop on Online Abuse and Harms (WOAH)*, Association for Computational Linguistics, 2023.

Plaza L., Carrillo-de Albornoz J., Morante R., Amigó E., Gonzalo J., Spina D., Rosso P., "Overview of exist 2023–learning with disagreement for sexism identification and characterization", *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, p. 316-342, 2023.

Poletto F., Basile V., Sanguinetti M., Bosco C., Patti V., "Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review", *Language Resources and Evaluation*, vol. 55, nº 2, p. 477-523, June, 2021a.

Poletto F., Basile V., Sanguinetti M., Bosco C., Patti V., "Resources and benchmark corpora for hate speech detection: a systematic review", *Language Resources and Evaluation*, vol. 55, p. 477-523, 2021b.

Poletto F., Valerio B., Bosco C., Patti V., Stranisci M., "Annotating hate speech: Three schemes at comparison", *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, vol. 2481 of CEUR Workshop Proceedings, 2019.

Rehman M. Z. U., Zahoor S., Manzoor A., Maqbool M., Kumar N., "A context-aware attention and graph neural network-based multimodal framework for misogyny detection", *Information Processing & Management*, vol. 62, nº 1, p. 103895, 2025.

Ricardo M., Marco G., João A. J., Paulo N., Pedro H., "Hate Speech Classification in Social Media Using Emotional Analysis", *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, p. 61-66, 2018.

Röttger P., Vidgen B., Nguyen D., Waseem Z., Margetts H., Pierrehumbert J., "HateCheck: Functional Tests for Hate Speech Detection Models", *in* C. Zong, F. Xia, W. Li, R. Navigli (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online, p. 41-58, August, 2021.

Sap M., Card D., Gabriel S., Choi Y., Smith N. A., "The risk of racial bias in hate speech detection", *Proceedings of the 57th annual meeting of the association for computational linguistics*, p. 1668-1678, 2019.

Schmidt A., Wiegand M., "A Survey on Hate Speech Detection Using Natural Language Processing", *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Association for Computational Linguistics, Valencia, Spain, p. 1-10, April, 2017.

Tahmasebi N., Borin L., Jatowt A., "Survey of Computational Approaches to Diachronic Conceptual Change", *CoRR*, 2018.

Talat Z., Hovy D., "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter", *Proceedings of the NAACL Student Research Workshop*, Association for Computational Linguistics, San Diego, California, p. 88-93, June, 2016.

Vidgen B., Derczynski L., "Directions in Abusive Language Training Data, a Systematic Review: Garbage in, Garbage Out", *PLOS ONE*, vol. 15, nº 12, p. e0243300, December, 2020.

Vidgen B., Harris A., Nguyen D., Tromble R., Hale S., Margetts H., "Challenges and Frontiers in Abusive Content Detection", *Proceedings of the Third Workshop on Abusive Language Online*, Association for Computational Linguistics, Florence, Italy, p. 80-93, August, 2019.

Yin W., Zubiaga A., "Towards Generalisable Hate Speech Detection: A Review on Obstacles and Solutions", *PeerJ Computer Science*, vol. 7, p. e598, June, 2021.

Yoder M., Ng L., Brown D. W., Carley K., "How Hate Speech Varies by Target Identity: A Computational Analysis", *in* A. Fokkens, V. Srikumar (eds), *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), p. 27-39, December, 2022.

Yu X., Blanco E., Hong L., "Hate Speech and Counter Speech Detection: Conversational Context Does Matter", *in* M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (eds), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, United States, p. 5918-5930, July, 2022.

Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R., "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)", *in* J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (eds), *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, p. 75-86, June, 2019.

Zampieri M., Nakov P., Rosenthal S., Atanasova P., Karadzhov G., Mubarak H., Derczynski L., Pitenis Z., Çöltekin Ç., "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)", *in* A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (eds), *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), p. 1425-1447, December, 2020.

Zsisku E., Zubiaga A., Dubossarsky H., "Hate Speech Detection and Reclaimed Language: Mitigating False Positives and Compounded Discrimination", *Proceedings of the 16th ACM Web Science Conference*, p. 241-249, 2024.

CyberAgressionAdo-Large: French Multiparty Chat Dataset to Address Online Hate

Anaïs Ollagnier* — **Elena Cabrio*** — **Serena Villata*** — **Valerio Basile****

* *Université Côte d'Azur, Inria, CNRS, I3S, 930 route des Colles, BP 145, 06903 Sophia Antipolis Cedex, France*
** *Department of Computer Science, University of Turin, Corso Svizzera, 185, 10149 Torino, Piemonte, Italy*

ABSTRACT. *This paper presents an extended version of CyberAgressionAdo, a French open-access dataset for online hate detection in multiparty conversations. The annotation process was improved with refined guidelines and a two-phase inter-annotator agreement study. A new adaptation of the Weirdness Index is introduced to analyze annotator disagreements. Now structured as a perspectivist corpus, with annotations provided by multiple annotators, CyberAgressionAdo-Large constitutes an enriched resource for the computational analysis of online hate situations in French.*

KEYWORDS: *Cyber-aggression, Annotation scheme, Multiparty chats.*

TITRE. *CyberAgressionAdo-Large: Jeux de données français de conversations multipartites pour l'étude de la haine en ligne*

RÉSUMÉ. *Cet article présente une version étendue de CyberAgressionAdo, un jeu de données français en accès libre destiné à la détection de la haine en ligne dans des conversations multipartites. Le processus d'annotation a été amélioré grâce à des directives affinées et à une étude en deux phases de l'accord inter-annotateurs. Une nouvelle adaptation de l'indice de « Weirdness » est présentée afin d'analyser les désaccords entre annotateurs. Désormais structuré comme un corpus perspectiviste, avec des annotations réalisées par plusieurs annotateurs, CyberAgressionAdo-Large constitue une ressource enrichie pour l'analyse computationnelle des situations de haine en ligne en français.*

MOTS-CLÉS : *Cyber-agression, Schéma d'annotation, Conversation multipartites.*

1. Introduction

Social media platforms have become essential components of contemporary communication, fostering freedom of expression and enabling the exchange of diverse ideas. This rapid expansion has also led to a rise in harmful, abusive, and degrading content, exposing individuals across various demographics to unsafe and detrimental interactions, thereby posing significant risks to their mental health and overall well-being. In response, automatic detection of online hate has emerged as a prominent research area in Natural Language Processing (NLP), with extensive studies presented at leading conferences, specialized workshops, and shared tasks (Alkomah and Ma, 2022). The majority of research efforts focus on popular social media platforms such as Twitter and Facebook, with an increasing number of techniques specifically developed for identifying harmful content in a monolingual setting, primarily English. However, recent studies indicate that private messaging platforms and chat rooms are significant environments for cyberbullying, particularly among adolescents (Alhashmi *et al.*, 2023). Due to the private nature of exchanges on these platforms and privacy policies that restrict data collection, few datasets capture aggressive interactions suitable for computational analysis (Cecillon *et al.*, 2020). Recently developed resources that simulate online aggression through role-playing games—where participants take on fictional roles to replicate cyber-aggression situations occurring in multiparty conversational settings—have contributed to addressing this gap (Gamal *et al.*, 2023). Notably, two recent French-language datasets (Ollagnier *et al.*, 2022; Ollagnier, 2024) provide valuable resources for addressing multiple online hate-related detection subtasks in conversational contexts, while also contributing to the exploration of linguistic diversity and cultural factors in non-English languages. Despite their significance, these datasets are relatively small in size (19 conversations, 2,921 messages) and exhibit limited topic diversity, resulting in an uneven representation of sensitive themes. Additionally, since participants in these role-playing scenarios have the freedom to influence the direction of their group's story, expanding the scope of scenarios and increasing data collection are essential to better capture the breadth of bullying practices observed in real-world contexts.

In this paper, we present the *CyberAgressionAdo-Large* dataset [1], an extended version of the dataset introduced in Ollagnier (2024). To our knowledge, *CyberAgressionAdo-Large* is currently the largest publicly available French-language dataset of aggressive conversations. It is distinguished by its size, diversity (covering four sensitive topics), and its in-depth linguistic analysis, featuring six layers of annotation designed to computationally address multiple online hate-related detection sub-tasks. Building on both existing materials (Ollagnier *et al.*, 2022; Ollagnier, 2024) and newly developed resources, this paper consolidates all information regarding the experimental process designed to collect conversations mimicking cyber-aggression in schools, including the experimental setup and role-play scenarios. It offers a compre-

1. The dataset is publicly available at: `https://anonymous.4open.science/r/CyberAgression-Large-C71C/`.

hensive presentation of the annotation tagset, supported by inter-annotator agreement experiments, highlighting its scalability and applicability on a larger scale. Additionally, the paper provides a comprehensive overview of the annotation tagset, supported by inter-annotator agreement experiments, to demonstrate its scalability and applicability on a larger scale. Additionally, we adapt the Weirdness Index (Basile, 2020) to facilitate an in-depth analysis of annotator disagreements. Specifically, polarized Weirdness scores are integrated into a visualization method to investigate potential semantic shifts that could explain divergences in human-provided labels. Furthermore, the released dataset includes annotations from three distinct annotators, supporting computational approaches aligned with the perspectivism trend in machine learning (Cabitza *et al.*, 2023). In summary, our contributions are as follows:

– with 5,789 annotated messages and six layers of annotation, *CyberAgressionAdo-Large* serves as a valuable resource for the French online hate-detection community, supporting diverse research and applications beyond the traditional conceptualization of online hate-related sub-tasks;

– we provide a detailed methodology for collecting naturalistic interactions and a comprehensive description of the annotation tagset to ensure reproducibility;

– we conduct (dis)agreement analysis experiments to ensure reliability and scalability, while offering a methodology to explore divergences in annotators' opinions;

– we apply pattern mining to deepen the understanding of complex cyberbullying communication practices commonly observed in multiparty settings;

– we release data annotated by multiple individuals to support perspectivist computational approaches.

2. Related work

From 2016 onward, a high number of resources and benchmark corpora have been developed to address various tasks related to online hate detection. Most of the research has focused on detecting harmful content such as offensive, toxic, abusive, and hateful speech, mainly from social media platforms like Twitter, Facebook, Gab, and Reddit (see Fortuna *et al.* (2020), for definitions). Recently, the diversity of data sources has expanded, with corpora including comments from news media, Wikipedia, chat rooms, forums, as well as messaging services like WhatsApp. Several studies have provided an organized overview of the domain by cataloging existing datasets (Madukwe *et al.*, 2020; Alkomah and Ma, 2022). Despite recent advances, very few datasets collecting aggressive conversations are available for computational analysis of cyberbullying situations. Table 1 lists all datasets related to conversations in which this phenomenon is observed. By "conversation", we refer to interactions involving at least two human speakers who alternate turns, resulting in non-linear and intertwined discourse. Formally, we considered datasets providing at least one sentence with its preceding or following context, as well as a turn of speech (a complete interaction between two speakers). The collected resources are analyzed and com-

pared across six dimensions: language, number of entries, source of collected data, provided annotation layers, number of speakers per conversation, and resource format.

Dataset	Language	Entries	Source	Annotation	Structure	Type
ConvAbuse	English	4,185	Chatbots	Abusive Type Severity Target Directness	Two-party	Utterance + context (the agent's turn plus the previous turn of both user and agent)
TOXICCHAT	English	10,166	Chatbots	Toxicity Jailbreaking	Two-party	User prompt + agent's turn
Wikipedia Abuse Corpus	English	382,665	Wikipedia comments	Personal attack Aggression Toxicity	Multiparty	Reconstructed conversations
CyberAgression Ado-v1 & -v2	French	3,552	Scripted conversations	v1/v2 Hate v1 Role v1 Target v1 Verbal abuse v1 Humour v2 Intention v2 Context	Multiparty	Full conversations
WhatsApp Dataset	Italian	2,066	Scripted conversations	Cyberbullying Role Type Sarcasm Offensive	Multiparty	Full conversations

Table 1. *Conversational datasets available for tasks related to online hate detection.*

ConvAbuse consists of data from two-party conversations (1 human interlocutor and 1 conversational agent) in English between users and three different conversational AI systems (ELIZA, CarbonBot, and Alana v2) (Curry *et al.*, 2021). Each entry is provided in context, including the target input associated with the system's output and the preceding turns (when available) from both the user and the system. This dataset publicly available provides around 4,000 annotated entries using a hierarchical annotation schema identifying whether the target content is abusive or not, the severity level, type, target, and the presence of implicit or explicit content. *TOXICCHAT* (Lin *et al.*, 2023) is a dataset consisting of 10,166 user prompts, with 5,634 manually annotated for toxicity and jailbreaking. The dataset is based on pre-processed user interactions collected from a demo of the popular open-source chatbot Vicuna, comprising both user prompts and the corresponding agent responses. *Wikipedia Abuse Corpus* (WAC) consists of conversations in English reconstructed from comments posted on Wikipedia talk pages, which are web pages associated with Wikipedia articles where editors can interact. Editors typically write explanatory messages (on average 1,000 characters) about changes made to the articles. The reconstructed conversations are based on discussions in response to these comments. These discussions are relatively short, with the majority consisting of fewer than 20 messages. A dataset of approximately 193,000 conversations, including 383,000 messages annotated as abusive or not, is publicly available (Cecillon *et al.*, 2020). *CyberAgressionAdo V1 & V2* are

two datasets presented respectively in Ollagnier *et al.* (2022) and Ollagnier (2024), consisting of multiparty conversations in French mimicking cyberbullying situations that may occur among adolescents. This data was collected during role-playing games in several high-schools and middle schools and annotated considering various layers, such as participant roles, the presence of hate speech, the type of verbal abuse, the authors' intentions (what they aim to accomplish or convey through their messages), the context in which these messages are situated as responses, etc. The released corpus includes about 3,000 entries, with 19 conversations of approximately 187 messages each exchanged between 5 to 7 adolescents. *WhatsApp Dataset* is a corpus of interactions in Italian, featuring cyberbullying situations, collected during an experiment on WhatsApp with high-school students. It has been annotated in terms of cyberbullying roles, types of cyberbullying, presence of sarcasm, and whether it is an offensive message or simply a joke (Sprugnoli *et al.*, 2018). This corpus includes around 2,000 messages exchanged among about 10 adolescents in 10 different conversations, each comprising approximately 207 messages. Two other corpora similar to conversations, as defined above, have been released; however, for the *Space-Origin* corpus (Papegnies *et al.*, 2017), it is based on proprietary data, and for the *Hateful Messages* corpus (Fillies *et al.*, 2023), no access link is provided, explaining their absence in Table 1. *Ruddit* (Hada *et al.*, 2021), the *Reddit Contextual Abuse Dataset* (Vidgen *et al.*, 2021), and the dataset introduced in Tufa *et al.* (2024) provide conversation threads extracted from Reddit, annotated specifically for online hate-related detection tasks. Similarly, *DeTox* consists of 10,278 annotated German social media tweets, half of which are part of coherent conversation segments (reply trees) annotated for toxicity, criminal relevance, and discrimination types. However, these datasets cannot be considered as conversations due to the dynamics of thread-based interactions on these platforms, which differ substantially from turn-taking, non-linear, and interwoven discourse. On Reddit and Twitter, conversation threads are structured hierarchically around an initial post, with comments grouped beneath it. Each comment can reply to a previous one, forming a tree-like structure. These branches often remain unrelated and involve different users, making it impractical to treat such thread structures as authentic conversations. Literature also reports the use of conversational datasets in other contexts (Ganesh *et al.*, 2023); however, these resources do not include annotations allowing for the development of methods dedicated to online hate detection.

From this analysis, it appears that access to conversational data from real-world applications, curated and annotated for the development of online hate detection tools, is more than limited. Each dataset offers solutions to overcome social media privacy policies that restrict the collection of such data. For example, the *WhatsApp Dataset* and *CyberAgressionAdo V1 & V2* are based on a data-collection methodology that closely resembles natural interactions. Indeed, human-machine interactions, such as those provided in the *ConvAbuse* and *TOXICCHAT* corpora, cannot fully replicate the complexity and dynamics of human-to-human interactions. Similarly, reconstructing conversations based on comments like those in the WAC corpus does not replicate the intrinsic nature and dynamics of such data. Furthermore, conclusions from pre-

vious studies support that role-playing games are a more valid measure of authentic language use than more traditional data collection methods such as interviews or self-assessment questionnaires (Kasper, 1999; Tran, 2006). Beyond collection methodologies, the provided annotations mainly focus on identifying and characterizing abuse, thus limiting the computational analysis of cyberbullying situations in this context. Indeed, online hate remains a complex and multifaceted phenomenon shaped by a multitude of linguistic, contextual, and social factors in general (Baider, 2020). This observation is consistent with the research presented in Kumar *et al.* (2022) and Ollagnier (2024), where the utilization of pragmatic-level information provides descriptors to enhance the understanding of this phenomenon. Moreover, another characteristic of conversational data, especially multiparty ones, is the presence of multiple participants or interlocutors, whose identification is a separate task in addition to identifying abuse. This aspect is even more crucial in cyberbullying situations where participants are involved differently. It may involve the victim, the harasser, and bystanders, which is important to distinguish to nuance interpretations of abuse, as demonstrated in these two studies (Ollagnier *et al.*, 2023a; Ollagnier *et al.*, 2023b). In conclusion, the *CyberAgressionAdo V1 & V2* datasets appear to be the most suitable for the computational analysis of this type of phenomenon occurring within conversations. Additionally, the *V2* version provides annotations allowing for the study of the interplay of different aspects related to the practices underlying the operationalization of cyberbullying situations. While the aforementioned datasets are valuable, their relatively small size and limited topic diversity constrain their capacity to comprehensively capture the breadth of bullying practices, sensitive themes, and participant roles observed in real-world online aggression scenarios. To address this limitation, we build on both existing materials and newly developed resources to introduce *CyberAgressionAdo-Large*, a corpus comprising 36 conversations and a total of 5,789 entries. To our knowledge, this corpus stands out as the largest publicly available dataset of its kind, distinguished by its diversity (covering four sensitive topics) and its in-depth analysis, featuring six layers of annotation addressing multiple analytical dimensions. Additionally, we provide annotations from each annotator to explore the possibilities of developing perspectivist approaches, aiming to preserve the divergence of opinions and integrate them into the process of developing machine-learning methods (Cabitza *et al.*, 2023).

3. CyberAgressionAdo-Large: construction

The *CyberAgressionAdo-Large* dataset was developed following the collection process and annotation scheme introduced in the initial works presented in Ollagnier *et al.* (2022) and Ollagnier (2024). The experimental setting conducted in schools, along with the scenarios used for the role-playing games and the guidelines for applying the multi-label, fine-grained tagset, adhered to the same protocol, as detailed below.

3.1. *Data collection*

CyberAgressionAdo-Large was created from multiple data collection efforts conducted at four French high-schools and one middle school, involving approximately 243 participants. Our intervention in schools was part of a broader effort to raise awareness about cyberbullying and hate speech, aiming to provide students with additional means to understand and better address this phenomenon. The initial contact with students involved introducing them to artificial intelligence and its potential role in detecting harmful online messages (1.5 hours). Then, students were asked to complete an anonymous questionnaire elaborated by the sociologist involved in the study, aiming to collect data on their online behavior (e.g., time spent on the web, on social media) and their perception of cyberbullying phenomena (10 to 15 minutes). Researchers then introduced the practical phase, during which students participated in a role-playing game that mimicked cyberbullying situations occuring on instant messaging platforms. Each student had a computer to work with and had to log into an Internet Relay Chat (IRC) with a pseudonym provided at their discretion during each game, ensuring fully anonymous data collection. Each role-playing game lasted on average 45 minutes. Teachers were present in the room but were in no way involved in the role-plays. A few weeks after the experimentation, based on feedback from the sociologist's survey, a two-hour meeting with the students was organized to discuss cyberbullying issues and online hate speech with them and their teachers. During this meeting, students could exchange ideas and share their personal experiences and feelings about conducting this experiment. Regarding the latter point, students were asked to fill out a second questionnaire to share their perceptions of the advantages and disadvantages of this experimental method. Since young people are the actors and experts of their own lives, we deemed it relevant to consult them to avoid misinterpretations or to confine them to our own representations, which, in the context of developing tools for detecting and preventing this phenomenon, could lead to biases (Alderson and Morrow, 2011).

3.2. *Scenarios*

Created in collaboration with a sociologist and an expert in education sciences, the scenarios address topics commonly reported during cyberbullying incidents, including cyberhate related to ethnic origin, religion, obesity, and homophobia. Table 2 presents some examples of scenarios proposed to students. These scenarios were developed based on interviews and case studies conducted in French secondary schools reported in Blaya and Audrin (2019), thus relying on authentic negative experiences encountered by young people. We included different types of situations: obesity, religion, ethnic origin, and homophobia. These situations were selected based on research showing that overweight students (Puhl *et al.*, 2017) and LGBT+ individuals are more likely to be discriminated against and harassed (online) (Bucchianeri *et al.*, 2014), and that cyberhate based on origin and religion is one of the types of victimization that has increased the most in recent decades (Blaya and Audrin, 2019; Llorent

et al., 2016; Räsänen *et al.*, 2016), and that the processes of exclusion and discrimination related to weight are similar to racism, sexism, and gender-based harassment (Van Amsterdam *et al.*, 2012). Obese and overweight students are more likely to be victims of bullying (Kahle and Peguero, 2017).

In these role-playing scenarios, participants were assigned specific active roles reflecting varying levels of involvement in cyberbullying situations. These roles included: the bully, who initiates the harassment; the victim, who is the target of the harassment; the victim supporter, who defends the victim; the bully supporter, who assists or encourages the bully's actions; and the conciliator, a mutual friend of both the bully and the victim who intervenes to mediate and resolve the conflict. Additionally, a moderator role was introduced to ensure that the interactions adhered to the rules of the role-playing game. This role, which remained passive and observational, was fulfilled by one of the researchers present during the data collection process. Since the role-playing game represents a protected space to experiment with cyber violence, we avoided having students play the victims, with the victims always being represented by researchers from our team who were not physically present in the experimentation room. In order to involve all students in the role-playing game, some roles were duplicated and embodied in the same scenario. The number of bullies could vary between 1 and 2, the victim supporter role between 1 and 3, and the bully supporter role between 2 and 3. All other actively involved participants, i.e., the victim and the conciliator, were played by one person per scenario. In general, each scenario was played by 5 to 7 people. Students were randomly assigned to a scenario and a role (regardless of their gender). In a few cases, teachers advised us to avoid assigning a certain role to a student considering previous class dynamics and the student's behavior or personal characteristics.

4. CyberAgressionAdo-Large: annotation

The annotation scheme utilized in this study builds on the schema introduced in Ollagnier (2024). This multi-label, fine-grained tagset encompasses six distinct annotation layers, including participant roles, the presence of hate speech and the type of verbal abuse. Furthermore, it incorporates a detailed hierarchical structure aimed at capturing the communicative intentions behind each message and the contextual factors influencing its production. Table 3 presents the statistical properties of the *CyberAgressionAdo-Large* dataset, while Table 4 provides a detailed description of the annotation schema. The complete annotation guidelines are publicly accessible on the *CyberAgressionAdo-Large* project webpage [2].

2. `Https://github.com/aollagnier/CyberAgression-Large/`.

Scenario	Topic
Julie and Léa use to hang out together and are walking in the schoolyard holding hands. Emilie, jealous of Julie, posts their photo on Snapchat and makes mean comments about their relationship, insinuating that they are lesbians. Marie tries to intervene to defend Julie and Léa, but Emilie brings her best friends, Elodie and Anna, with her, and they try to exclude them from their friend group in class and on social media. Arthur, who is friend with both Julie, Léa, and Emilie, tries to intervene by explaining to them that it's pointless and that they should stop arguing.	homophobia
In the cafeteria, Paul, who is a bit overweight, has his dessert stolen by his table neighbor, Brice, who is also in his class. Brice tells him he's already fat enough and doesn't need to eat, while he eats the dessert. Meanwhile, Julien films the scene and shares it on social media, commenting on Paul's appearance, his gluttony, and his lack of control, which makes everyone laugh. Justine and Thibaut try to defend him, and Pierre, a friend of Paul's but also of Brice and Julien, tries to stop the teasing.	obesity
Justine is Jewish. On her profile, she posts a picture of her little brother's Bar Mitzvah. Léo and Guillaume, Justine's classmates, share the photo with harmful comments against Jews, including caricatures. Aurélie and Isabelle, while looking at the photo, also laugh. Léa and Anna, friends of Justine, try to defend her in the chat with the help of Amine to put an end to the harassment against Justine and her religion.	religion
Sophie and Lucas have been together for a few months and attend the same school. During a school trip, taking advantage of Sophie's absence, who stayed home with the flu, Lucas secretly kisses Silvia, a classmate of Sophie. Sophie discovers Lucas's betrayal through her friend Adrien, who witnessed the scene. Thinking that Silvia had flirted with Lucas, Adrien starts insulting Silvia on the WhatsApp chat, aided by Théo, Diana, and Camille: "She's here because they didn't want her at home! She has no business being here. She came to steal other people's boyfriends. Besides, they're all thieves." Soan, a classmate, decides to defend Silvia by blaming Lucas. Herbert, a friend of both Adrien and Silvia, intervenes to put an end to the harassment between Adrien and Silvia.	ethnic origin

Table 2. *Examples of role-playing scenarios on each sensitive topics proposed to students.*

Metric	Value
Number of conversations	36
Number of lines	5,789
Number of tokens	36,299
Average messages per conversations	156.45
Average length of messages (tokens)	6.47

Table 3. *Statistics of the CyberAgressionAdo-Large.*

Aggression		
Code	**Aggression Level**	**TAG**
1.1	Overtly Aggressive	OAG
1.2	Covertly Aggressive	CAG
1.3	Non-Aggressive	NAG
Role/Target		
Code	**Attribute**	**TAG**
1.A 1.1	victim	victim
1.A 1.2	victim support	victim_support
1.A 1.3	bully	bully
1.A 1.4	bully support	bully_support
1.A 1.5	conciliator	conciliator
Verbal Abuse		
Code	**Attribute**	**TAG**
1.B 1.1	Blaming	BLM
1.B 1.2	Name-calling	NCG
1.B 1.3	Threat / Coercion	THR
1.B 1.4	Denigration	DNG
1.B 1.5	Aggression-other	OTH
Discursive Level		
Code	**Intention/Context**	**TAG**
2.1	Attack	ATK
2.2	Defend	DFN
2.3	Counterspeech	CNS
2.4	Abet and Instigate	AIN
2.5	Gaslighting	GSL
2.6	Conflict-resolution	CR
2.7	Empathy	EMP
2.8	Other	OTH

Table 4. *The CyberAgressionAdo-Large tagset.*

4.1. *Aggression level*

This label is based on a multiclass schema comprising the categories OAG, CAG, and NAG. Label assignment is performed by interpreting aggression within its context, requiring annotators to consider extralinguistic knowledge and the perspectives of both the author and the recipient, including their roles and discursive postures. Detailed definitions and corresponding examples for each aggression label are provided below.

1.1 **Overt Aggression (OAG)**: This refers to communication, whether in speech or text, where aggressive behavior is explicitly expressed. It often involves offensive

or hostile language, explicit threats, hate speech, derogatory terms, or direct insults. Overt aggression may also arise from specific lexical items, features, or syntactic structures whose aggressive nature becomes apparent when contextualized with extralinguistic knowledge and the perspectives of both the author and recipient.

Example: The French sentence “woaa!! mate le cachalot” (EN: “woah !! look at the whale”) demonstrates overt aggression in the context of cyberbullying related to obesity. Here, “le cachalot” (the whale) is derogatory and offensive, mocking someone based on their weight. The phrase “mate” (look at) adds a mocking tone, inviting others to ridicule the individual. The exclamation marks and overall tone further emphasize the aggressive nature of the statement.

1.2 **Covert Aggression (CAG)**: This form of communication employs linguistic strategies to mask aggression beneath subtle or indirect expressions, avoiding explicit threats or derogatory language. While covert aggression is often subtle, it can also include non-subtle expressions that still convey aggressive intent despite an attempt to conceal it. Common strategies include figurative language (e.g., sarcasm, irony, black humor, exaggeration, metaphor), rhetorical questions, euphemisms, fallacies, or circumlocution.

Example: The sentence “T’as vraiment des fringues de ouf, mec, personne peut rivaliser avec ton style” (EN: “You’ve got some crazy clothes, dude, nobody can compete with your style”) appears to be a compliment. However, the phrase “des fringues de ouf” (crazy clothes) and “personne peut rivaliser” (nobody can compete) carry a sarcastic and mocking tone, revealing covert aggression.

1.3 **Non-Aggression (NAG)**: This category includes any text or speech devoid of hostile or harmful intent. It excludes explicit derogatory language, threats, or expressions of harm towards individuals or groups, as well as linguistic strategies that might subtly imply aggression or intimidation.

4.2. *Role/target*

Five specific active roles are used to represent varying levels of involvement in cyberbullying situations, depicting both the fictional roles embodied by participants during the scenarios and the target(s) of online hate. Target annotations are applied exclusively to messages identified as OAG or CAG. As described in Section 3.2, these roles include: (1.A 1.1) the victim, who is the individual being harassed; (1.A 1.2) the supporter of the victim, who defends him; (1.A 1.3) the bully, who initiates the harassment; (1.A 1.4) the supporter of the bully, who collaborates in or supports the bully’s actions; and (1.A 1.5) the conciliator, a mutual friend of the bully and the victim who intervenes to mediate and resolve the conflict.

4.3. *Verbal abuse*

Cyberbullying can take many forms, with verbal abuse being prevalent among them. It may include harassment, which involves sending repetitive and offensive messages to a target, cyberstalking (sending repetitive threatening communications), flaming, which entails sending messages containing abusive and vulgar terms such as insults, gossip, or mockery, and denigration (Bauman, 2014; Tokunaga, 2010; Watts *et al.*, 2017). Five types commonly encountered in written language are annotated here, and these are exclusively assigned to messages identified as OAG or CAG:

1.B 1.1 **Blaming (BLM)**: This involves making the individual believe they are responsible for the abuse they are experiencing, attributing it to their actions, words, or behavior. **Example:** "*on la traiterait pas de truie si elle avait pas autant de graisse*" ("she wouldn't be called a pig if she didn't have so much fat").

1.B 1.2 **Name-calling (NCG)**: Refers to abusive, insulting, or derogatory language aimed at undermining the self-esteem, personal worth, and self-perception of the targeted individual. **Example:** "*té qu1 putain de mongol*" ("you're such a fucking retard").

1.B 1.3 **Threat (THR)**: These statements are intended to intimidate, control, or manipulate the victim, coercing them into submission. **Example:** "*je vais venir en bas de chez toi, tu vas voir qui va plus parler*" ("I'm going to come to your house, and you'll see who won't be talking anymore").

1.B 1.4 **Denigration (DNG)**: Disparaging remarks aimed at attacking the reputation of the targeted person, belittling, discrediting, and tarnishing their image. These remarks are deliberately hurtful, non-constructive, and malicious. **Example:** "*les filles comme toi, ça me dégoûte*" ("girls like you disgust me").

1.B 1.5 **Other aggression (OTH)**: Covers content that includes deliberately harmful, abusive, insulting, or derogatory language that does not align with the other defined categories. **Example:** "*va crevé en enfer*" ("go die in hell").

4.3.1. *Discursive level*

The intention and context categories form two distinct layers, encompassing classifications such as attack (ATK), defense (DFN), counter-speech (CNS), instigation (AIN), gaslighting (GSL), conflict resolution (CR), and empathy (EMP). The purpose of label assignment is to decipher the discursive function of exchanged messages based on their underlying intentions, covering both aggressive and non-aggressive utterances. This annotation serves a dual purpose: first, to uncover the authors' intentions (what they aim to achieve or convey through their messages), and second, to establish the contextual framework in which these messages function as responses. Below, we provide the definitions and examples for each label.

2.1 **Attack (ATK)**: Any form of communication that intentionally exhibits overt or covert aggression towards victims, their supporters, or even conciliators. Such communication may involve insults, threats, mockery, exclusion, taunting, and dis-

crediting. This behavior is exclusive to bullies and their supporters and can manifest either as a deliberate act aimed at inflicting harm or as a means to escalate the level of violence.

```
User1: [ATK] ALLEZ MANIFESTE TOI GROS PORCS. / (EN) GO
    ON, SHOW YOURSELF, YOU FAT PIGS.
User2: [ATK] User3 le cachalot. / (EN) User3 the sperm
    whale.
```

2.2 **Defend (DFN)**: Any text/speech aiming to protect oneself or others from perceived attacks. It is characterized as an impulsive and non-deliberate response, which can be either aggressive or non-aggressive, and may be in retaliation for real or perceived attacks. This behavior is exclusive to victims, their supporters, or conciliators and may involve strategies such as challenging and refuting the abuser's messages.

```
User1: [ATK] jalouse de quoi mon pote tu me dégoute.
/ (EN) jealous of what my friend you disgust me.
User2: [DFN] t'es blanche comme un c*l tu crois t mieux
    User1? / (EN) you're as pale as an *ss do you think
    you're better User1?
```

2.3 **Counterspeech (CNS)**: Any non-aggressive response to harmful speech, aiming to undermine it. It employs strategies like presenting facts, highlighting contradictions, warning of consequences, and denouncing hate. It is initiated by victims, supporters, or conciliators.

```
User1: [DFN] tu sais dire d'autres choses à part ça ? /
    (EN) Do you know how to say anything else apart from
     that?
User2: [CNS] ça se fait pas en plus de prendre en photos
     / (EN) It's not right in addition to taking
    pictures.
```

2.4 **Abet/instigate (AIN)**: Messages supporting, encouraging, or validating previous negative messages, inciting aggression either beforehand (instigation) or during/after the act (abetment). These messages typically escalate conflicts or foster a hostile atmosphere, often initiated by bullies and their supporters.

```
User1: [ATK] qui les supp du groupe la / (EN) Who
    removes them from the group there?
User2: [AIN] je vais les supprimer / (EN) I am going to
    delete them.
```

2.5 **Gaslighting (GSL)**: Any text/speech minimizing or distorting another person's trauma or memory, aiming to manipulate their perception of reality and exert control. This includes tactics like denying or downplaying harm, blaming the victim, questioning their memory, invalidating their feelings, and using group consensus to make them doubt themselves.

```
User1: [ATK] wsh tu parle pas comme ca je vais te
    dechire / (EN) Hey don't talk like that I'm going to
     tear you apart.
User2: [GSL] User1 t es changer wsh / (EN) User1 you've
    changed seriously.
```

2.6 **Conflict-Resolution (CR)**: Any communication aiming to resolve conflicts and de-escalate situations without resorting to aggression. This includes mediation to resolve conflicts, mitigation to lessen the impact of cyberbullying, and education to promote appropriate online behavior. CR messages are consistently non-aggressive and are typically initiated by victim supporters and conciliators.

```
User1: [GSL] c toi ta un problème grosse p*te / (EN) You
    're the one with a problem you big sl*t.
User2: [CR] mais calmez-vous chaqu'un s'est préférence /
     (EN) calm down, everyone has their preferences.
```

2.7 **Empathy (EMP)**: Messages that demonstrate understanding, compassion, and support for those affected by cyberbullying. These messages may express sympathy, offer assistance or resources, validate emotions, or include self-empathy when victims acknowledge their own distress. This behavior is exclusive to victims, their supporters, or conciliators.

```
User1: [DFN] Elles sont juste immature de faire ca,
    preuve que c'est des gamines / (EN) They are just
    immature to do this, proof that they are kids.
User1: [EMP] User3 tu vaux mieux que sa / (EN) User3 you
    're worth more than this.
```

2.8 **Other (OTH)**: This category applies to cases where the appropriate tag for a message is unclear. It includes neutral utterances (messages without explicit or implicit harm), non-standard utterances such as incomplete sentences, one-word responses, sentence fragments, or emoticons and emojis used to convey emotions, attitudes, or reactions.

```
User1: [CR] Ca sert a rien de se prendre la tete
    franchement / (EN) There's no point in getting
    worked up, honestly.
User2: [OTH] quelle sexplique / (EN) What does it mean?
```

5. Disagreement vs. perspectives

A thorough analysis of the causes of disagreement among annotators, established in Ollagnier (2024), revealed that the various sources of disagreement stemmed from

(a) the clarity of annotation labels (i.e., their applicative scope), (b) text ambiguity, and (c) differences among annotators (i.e., their individual viewpoints), with the latter two being the most frequent causes. Following these findings, the annotation guidelines were improved by providing a precise description of the application cases for each annotation layer and corresponding labels. Based on these new guidelines, *CyberAgressionAdo-Large* was manually annotated by three experts from a text annotation specialized company. Table 5 presents the results of inter-annotator agreement obtained through the measurement of Krippendorff's Alpha across all conversations.

Label	**Score**
Hate	0.83
Target	0.88
Verbal abuse	0.82
Intention	0.87
Context	0.81

Table 5. *Measurement of Inter-Annotator Agreement on CyberAgressionAdo-Large.*

The obtained scores demonstrate a significant increase in inter-annotator agreement across all labels compared to those presented in Ollagnier (2024), and this on a dataset twice as large. This underscores the value of clear guidelines and discussions around challenging situations. Feedback from the annotators highlights that the main source of the remaining disagreement primarily revolves around varied interpretations arising from individual perceptions. Due to this finding we do not conduct here an analysis of annotator disagreements consisting in categorizing potential reasons behind conflicting annotations (Sandri *et al.*, 2023), such as sloppy annotation, ambiguity, missing information, and subjectivity. Supported by the perspectivist paradigm introduced in Cabitza *et al.* (2023), we decided to experimentally use residual disagreement to reflect individual viewpoints that may arise in the interpretation of online hate detection in a multiparty setting. We specifically investigate the application of the Weirdness Index (Ahmad *et al.*, 1999). In its original formulation, the W-index is used to extract domain-specific terms by comparing the relative frequencies of words in a domain-specific corpus vs. a generic corpus. The index was later applied to annotated corpora in order to rank the words according to their association to a specific human-provided label (Basile, 2020). We further adapt the method to automatically compute the association between each word and the disagreement between a pair of annotators. Given a pair of annotators a, b, the dataset is divided in two parts: $A_{a,b}$, i.e. the set of messages on which a and b agree on a specific label, and $D_{a,b}$, i.e. the set of messages on which a and b disagree. The Agreement Weirdness (AW) index for a word w is therefore defined as:

$$AW(w, a, b) = \sigma\left(\frac{P(w|D_{a,b})}{P(w|A_{a,b})}\right)$$

where $P(w|D)$ and $P(w|A)$ are the relative frequencies of w in D and A respectively, and σ is the standard logistic function. In essence, $AW(w)$ will be a number in $[0, 1]$ close to 1 if it occurs more often in texts on which a and b disagree, and close to 0 if it occurs more often in texts on which a and b agree.

In order to explore the results of the AW-index analysis, we introduce an ad-hoc visualization method where the level of disagreement associated with a word is correlated to its distance from fixed points. In each figure, the three blue dots represent the three annotators. Starting from the center of the triangle, each word is moved toward the line connecting two annotators based on $AW(w, a, b)$, which quantifies the degree of pairwise disagreement between annotators a and b for the word w. As a consequence, we may observe three main patterns:

– a words stays close to the center, if its disagreement levels are balanced across all three annotators;

– a word is close to an edge, if its disagreement is observed between a specific pair of annotators only. We call this **bilateral** disagreement;

– a word is close to a corner, if its disagreement is observed between a specific annotator and both the others, but not between the other two. We call this **multilateral** disagreement.

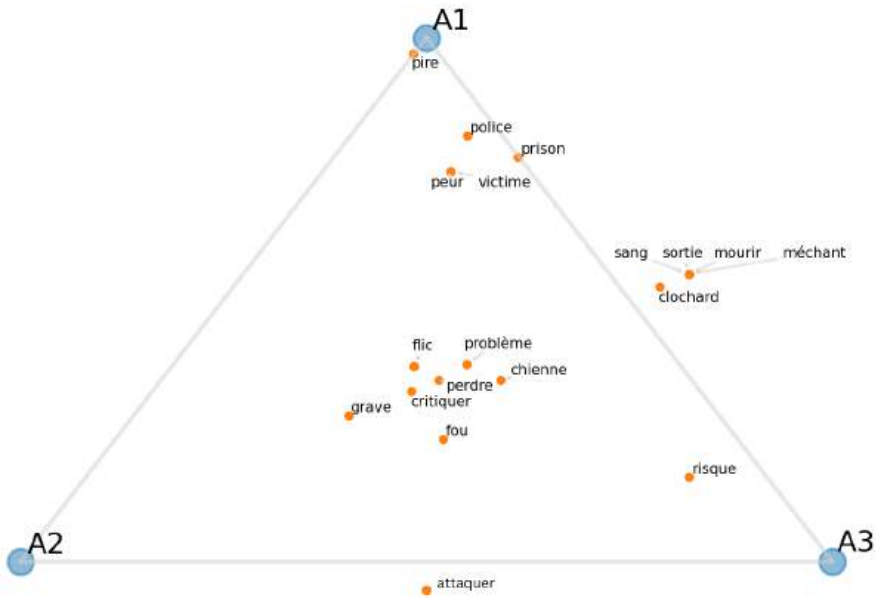

Figure 1. *A sample of visualization obtained for the emotion fear on the topic of obesity.*

Figure 1 presents a generated visualization using the AW-index on the full *CyberAgressionAdo-Large* dataset. Here, we can observe a bilateral disagreement

among the annotators A1 and A3 concerning utterances containing the reported words, as well as a multilateral disagreement for A1. The analysis of the visualizations obtained for the labels HATE and VERBAL_ABUSE, considering all the emotions, confirms the benefits of such a visualization method in unveiling factors of divergence of opinions, all the while facilitating the discovery of meaningful patterns and causal relationships among the systematic disagreements [3]. For instance, repeated bilateral disagreements among annotators could unveil divergence of opinions from a community perspective, while multilateral ones would refer to divergences influenced by individual perspectives. Moreover, this visualization method facilitates tailored research and the exploration of various analytical perspectives. For instance, when paired with a word affect lexicon such as *NRC VAD* (Mohammad, 2018), it highlights the potential interplay between the affective connotations of words and their interpretation in conveying hate.

In conclusion, observations reported in annotators' disagreements confirm that capturing pragmatic depictive social dynamics and interactions shaping conversations is achievable through the incorporation of annotation layers. Additionally, we believe that providing scenarios to annotators has influenced their interpretations, a factor that warrants further study to fully understand its impact. However, it remains undeniable that preserving annotations provided by different individuals is necessary in this context to access multiple potential interpretations of conversational data. This diversity of annotations allows for a comprehensive understanding of real-world scenarios and human values, thereby empowering the development of NLP systems to more accurately reflect and respect the intricacies of human communication and interaction. This is particularly crucial for addressing tasks related to online hate detection.

6. Analysis of cyberbullying practices

In this section, we present statistical evidence of cyberbullying practices observed in the annotated scenarios. The reported observations are based on frequent patterns identified at the instance level (i.e., a single message) or at the implicature level (i.e., a message and its subsequent reply). The patterns presented at the instance level are derived from observations that consider each annotator's perspective individually. In contrast, the observations reported at the implicature level are based on frequent patterns identified collectively among all the annotators.

In detail, Table 6 presents the most prevalent patterns observed in cyberbullying practices by analyzing individual author utterances (instances). Multiple recurrent cyberbullying behaviors are identified, which coincide with the roles of involvement concerning the type of hate expressed, the role of the individual(s) targeted, as well as the authors' intentions behind the posted message. Across all annotators, both bullies and their bystanders tend to target victims and their bystanders with the intention

3. The visualizations are available here: https://github.com/aollagnier/CyberAgression-Large/viz/.

ROLE	HATE	TARGET	INTENTION	FREQ. (%)
bully	OAG	victim	ATK	34.0
	NAG	-	OTH	19.5
	OAG	victim_support	ATK	13.2
bully_support	OAG	victim	ATK	29.9
	NAG	-	OTH	19.8
	CAG	victim	ATK	11.7
conciliator	NAG	-	OTH	30.9
	NAG	-	CR	24.1
victim	NAG	-	OTH	26.9
	NAG	-	DFN	17.3
	NAG	-	CNS	13.2
victim_support	NAG	-	OTH	22.9
	OAG	bully	DFN	17.0
	OAG	bully_support	DFN	15.0

(a) *Annotator 1*

ROLE	HATE	TARGET	INTENTION	FREQ. (%)
bully	OAG	victim	ATK	34.3
	NAG	-	OTH	20.1
	OAG	victim_support	ATK	13.4
	CAG	victim	ATK	11.4
bully_support	OAG	victim	ATK	29.1
	NAG	-	OTH	21.1
	OAG	victim_support	ATK	10.9
conciliator	NAG	-	OTH	29.7
	NAG	-	CR	22.8
victim	NAG	-	OTH	29.5
	NAG	-	DFN	15.6
	NAG	-	CNS	13.7
victim_support	NAG	-	OTH	24.5
	OAG	bully	DFN	16.9
	OAG	bully_support	DFN	14.0

(b) *Annotator 2*

ROLE	HATE	TARGET	INTENTION	FREQ. (%)
bully	OAG	victim	ATK	34.7
	NAG	-	OTH	20.6
	OAG	victim_support	ATK	13.4
	CAG	victim	ATK	10.8
bully_support	OAG	victim	ATK	28.8
	NAG	-	OTH	21.3
	OAG	victim_support	ATK	10.7
conciliator	NAG	-	OTH	31.1
	NAG	-	CR	22.1
victim	NAG	-	OTH	30.7
	NAG	-	DFN	16.0
	NAG	-	CNS	12.8
victim_support	NAG	-	OTH	24.0
	OAG	bully	DFN	16.8
	OAG	bully_support	DFN	14.5

(c) *Annotator 3*

Table 6. *The patterns of cyberbullying practices observed for each annotator at the instance level (i.e., a single message). The percentages indicate the frequency of each pattern relative to all messages sent by the corresponding authors across all scenarios annotated by the same annotator.*

of deliberately harming (ATK) them. The primary observed intention typically falls within a proactive aggression scheme characterized by repeated attacks intended to deliberately inflict harm or escalate the level of violence. Conversely, the intentions of victims and their supporters are predominantly characterized by OTH, which typically corresponds to neutral utterances (messages not conveying explicit or implicit harm). According to annotators' observations, neutral utterances often involve participants engaging in arguments, potentially impacting conflicts. DFN and CNS are also among the primary discursive devices used, representing behaviors aligned with a reactive aggression scheme describing impulsive aggressive responses to provocation. Conciliators are mainly non-aggressive, employing neutral utterances (29.7%-31.1%) or actively striving to resolve conflicts and de-escalate situations, with between 22.1%-24.1% of their messages dedicated to these objectives. While Annotators 2 and 3 noted the presence of covertly aggressive (CAG) messages directed at victims (10.8%-11.4%), it appears that the use of figurative devices is less common in this setting compared to other social media platforms (Ocampo *et al.*, 2023).

(source ↪ reply)	HATE	TARGET	INTENTION
bully	OAG	victim	ATK
↪ victim_support	OAG	bully	DFN
victim_support	OAG	bully	DFN
↪ victim	OAG	bully	DFN
bully_support	OAG	victim	ATK
↪ bully	OAG	victim	ATK
victim	OAG	bully	DFN
↪ victim_support	OAG	bully	DFN

Table 7. *Patterns of cyber-aggressions observed at the implicature level (i.e., one message and the subsequent reply), common to all annotators.*

Table 7 offers a comprehensive overview of cyberbullying practices observed across pairs of utterances (implicatures), taking into account all annotators' perspectives. In this context, "source" refers to the initial message, while "reply" denotes the immediate subsequent message. These pairs denote an implicature relationship as they comprise messages generated within the same context, with the "reply" often reliant on the preceding message for context and meaning. It's noteworthy that each recurring pattern involves distinct roles, shedding light on the intricate dynamics of cyberbullying situations. Additionally, these patterns consistently emerge among all annotators across all scenarios (with a support measure of 1.0), providing generalizable and reliable insights essential for studying this complex behavioral phenomenon. In detail, bystanders of the victim frequently intervene in bullying episodes (ATK setting) by directly assisting victims against the bullies. Victims and their bystanders tend to support each other against the bullies, while the bullies and their bystanders unite with the aim of jointly attacking the victims.

Overall, the observations derived from these tables offer an initial depiction of cyberbullying practices in this specific multiparty context, providing valuable insights into the complex nature of cyberbullying phenomena. Firstly, despite variances in annotators' perceptions, common practices being topic-agnostic emerge and recur in each scenario. Secondly, non-aggressive exchanges are prevalent and should be analyzed, as they can contribute to either the escalation or de-escalation of situations. A recent study presented in Kaliampos *et al.* (2022) confirms this finding by examining potential behaviors of bystanders in bullying episodes. It reports that neutral utterances by victims' supporters can aim to de-escalate tension, seek clarification, maintain normalcy, or subtly intervene without provoking further hostility. Lastly, while the proactive-reactive aggression scheme has been extensively studied in the operationalization of cyberbullying, it appears that peer support schemes play a crucial role in the unfolding of events (Cowie, 2014). Under the umbrella of peer support, activities such as befriending, peer counseling, conflict resolution, or mediation, as well as interventions in bullying situations, are included. These activities should be con-

sidered with differing intentions, depending on whether peer support is offered by the bullies or the victims.

7. Conclusion

In this paper, we introduce the *CyberAgressionAdo-Large* dataset, which applies a hierarchical, fine-grained tagset designed for annotating bullying narrative events in multi-party chat conversations. Currently, the dataset comprises 36 conversations in French, mimicking online aggression commonly observed among teenagers on private instant messaging platforms. Our data collection efforts are ongoing, with additional sessions planned in French high schools over the coming months to expand both the size and diversity of the dataset. Given that participants in the role-playing game have the freedom to influence their group's storyline, it is crucial to conduct more scenarios and gather additional data from schools to ensure comprehensive coverage of real-world bullying practices. Furthermore, we intend to enhance the existing tagset by incorporating labels that facilitate computational modeling of multi-party dialogues. These enhancements aim to support tasks such as identifying participant roles, managing initiative and turn-taking, and analyzing discourse relations, which are essential for detecting online hate and related phenomena effectively within this context.

8. Ethics statement

NLP research focusing on online aggression and harassment detection inevitably raises ethical considerations. In our work, we place significant emphasis on the importance of ensuring that students involved are fully informed, that the data collected replicate naturalistic interactions, and our support for an annotation methodology promoting diverse opinions and perspectives.

Firstly, all students under 18 participated with parental consent, receiving comprehensive explanations about research objectives, data usage, and associated risks. Transparency was paramount as both parents and students were informed about AI's potential benefits in detecting hostile online messages. Prior to participation, students underwent education on cyber aggression and AI to foster informed consent. Our research protocol underwent rigorous review and approval by each participating school, adhering to European ethical standards and university guidelines. Throughout the study, we maintained strict confidentiality, anonymity, and respect for participants' autonomy. To ensure a positive experience, we provided support during role-playing sessions and conducted post-session feedback and training on cyber aggression's impact on victims and perpetrators.

Secondly, the validity of our data collection process was validated by a sociologist and an expert in education sciences. The scenario designs were based on real experiences shared by young people, ensuring authenticity and relevance to actual online interactions. The spontaneous nature of multi-party chats minimized scripted

responses, aligning with research that shows role-plays provide a more genuine portrayal of natural language use compared to methods such as interviews, questionnaires, human-machine interactions, or reconstructing conversations from threads.

Finally, despite the lack of comprehensive sociodemographic information about the annotators provided by the company, our work underscores the importance of acknowledging and incorporating annotator subjectivity in NLP applications. Indeed, diverse annotator viewpoints can be utilized to mitigate biases and reflect real-world human values. Moreover, our corpus serves as a foundation for exploring perspectivist computational approaches to address subjective tasks in conversational data.

In conclusion, by addressing these ethical concerns and promoting diversity, the NLP community can significantly advance in combating online hate across diverse digital environments, relying on more effective, fairer, and transparent NLP models.

Acknowledgements

This work is funded under the IDEX UCA OTESIA "L'intelligence artificielle au service de la prévention de la cyberviolence, du cyberharcèlement et de la haine en ligne", and by the UCA Academy 1 project with the reference number C870A021 – D103 – ACAD1_FIN_17_20Y. It has also been supported by the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

9. References

Ahmad K., Gillam L., Tostevin L. *et al.*, "University of Surrey Participation in TREC 8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER)", p. 1-8, 1999.

Alderson P., Morrow V., *The ethics of research with children and young people: A practical handbook*, Sage, 2011.

Alhashmi A. A., KM A. K., Eid A. A., Mansouri W. A., Othmen S., Miled A. B., Darem A. A., "TAXONOMY OF CYBERBULLYING: AN EXPLORATION OF THE DIGITAL MENACE", *Journal of Intelligent Systems and Applied Data Science*, 2023.

Alkomah F., Ma X., "A Literature Review of Textual Hate Speech Detection Methods and Datasets", *Inf.*, vol. 13, nº 6, p. 273, 2022.

Baider F., "Pragmatics lost?: Overview, synthesis and proposition in defining online hate speech", *Pragmatics and Society*, vol. 11, nº 2, p. 196-218, 2020.

Basile V., "Domain Adaptation for Text Classification with Weird Embeddings", *in* J. Monti, F. Dell'Orletta, F. Tamburini (eds), *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, vol. 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

Bauman S., *Cyberbullying: What counselors need to know*, John Wiley & Sons, 2014.

Blaya C., Audrin C., "Toward an Understanding of the Characteristics of Secondary School Cyberhate Perpetrators", *Frontiers in Education*, vol. 4, p. 46, 2019.

Bucchianeri M. M., Eisenberg M. E., Wall M. M., Piran N., Neumark-Sztainer D., "Multiple types of harassment: Associations with emotional well-being and unhealthy behaviors in adolescents", *Journal of Adolescent Health*, vol. 54, n° 6, p. 724-729, 2014.

Cabitza F., Campagner A., Basile V., "Toward a Perspectivist Turn in Ground Truthing for Predictive Computing", *in* B. Williams, Y. Chen, J. Neville (eds), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, AAAI Press, p. 6860-6868, 2023.

Cecillon N., Labatut V., Dufour R., Linarès G., "WAC: A Corpus of Wikipedia Conversations for Online Abuse Detection", *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, European Language Resources Association, p. 1382-1390, 2020.

Cowie H., "Understanding the role of bystanders and peer support in school bullying.", *International journal of emotional education*, vol. 6, n° 1, p. 26-32, 2014.

Curry A. C., Abercrombie G., Rieser V., "ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Detection in Conversational AI", *in* M. Moens, X. Huang, L. Specia, S. W. Yih (eds), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Association for Computational Linguistics, p. 7388-7403, 2021.

Fillies J., Peikert S., Paschke A., "Hateful Messages: A Conversational Data Set of Hate Speech produced by Adolescents on Discord", *CoRR*, 2023.

Fortuna P., Soler Company J., Wanner L., "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets", p. 6786-6794, 2020.

Gamal D., Alfonse M., Jiménez-Zafra S. M., Aref M., "Intelligent Multi-Lingual Cyber-Hate Detection in Online Social Networks: Taxonomy, Approaches, Datasets, and Open Challenges", *Big Data Cogn. Comput.*, vol. 7, n° 2, p. 58, 2023.

Ganesh A., Palmer M., Kann K., "A Survey of Challenges and Methods in the Computational Modeling of Multi-Party Dialog", *in* Y.-N. Chen, A. Rastogi (eds), *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, Association for Computational Linguistics, Toronto, Canada, p. 140-154, July, 2023.

Hada R., Sudhir S., Mishra P., Yannakoudakis H., Mohammad S. M., Shutova E., "Ruddit: Norms of Offensiveness for English Reddit Comments", *in* C. Zong, F. Xia, W. Li, R. Navigli (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Association for Computational Linguistics, p. 2700-2717, 2021.

Kahle L., Peguero A. A., "Bodies and bullying: The interaction of gender, race, ethnicity, weight, and inequality with school victimization", *Victims & Offenders*, vol. 12, n° 2, p. 323-345, 2017.

Kaliampos G., Katsigiannis K., Fantzikou X., "Aggression and bullying: a literature review examining their relationship and effective anti-bullying practice in schools", *International Journal of Educational Innovation and Research*, vol. 1, n° 2, p. 89–98, Jul., 2022.

Kasper G., "Data collection in pragmatics research", *University of Hawai'i Working Papers in English as a Second Language 18 (1)*, 1999.

Kumar R., Ratan S., Singh S., Nandi E., Devi L. N., Bhagat A., Dawer Y., Lahiri B., Bansal A., Ojha A. K., "The ComMA Dataset V0.2: Annotating Aggression and Bias in Multilingual Social Media Discourse", *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, European Language Resources Association, p. 4149-4161, 2022.

Lin Z., Wang Z., Tong Y., Wang Y., Guo Y., Wang Y., Shang J., "ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation", *in* H. Bouamor, J. Pino, K. Bali (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Association for Computational Linguistics, p. 4694-4702, 2023.

Llorent V. J., Ortega-Ruiz R., Zych I., "Bullying and cyberbullying in minorities: Are they more vulnerable than the majority group?", *Frontiers in psychology*, vol. 7, p. 1507, 2016.

Madukwe K. J., Gao X., Xue B., "In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets", *in* S. Akiwowo, B. Vidgen, V. Prabhakaran, Z. Waseem (eds), *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020*, Association for Computational Linguistics, p. 150-161, 2020.

Mohammad S. M., "Word Affect Intensities", *in* N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, European Language Resources Association (ELRA), 2018.

Ocampo N., Sviridova E., Cabrio E., Villata S., "An In-depth Analysis of Implicit and Subtle Hate Speech Messages", *in* A. Vlachos, I. Augenstein (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Association for Computational Linguistics, p. 1989-2005, 2023.

Ollagnier A., "CyberAgressionAdo-v2: Leveraging Pragmatic-Level Information to Decipher Online Hate in French Multiparty Chats", *in* N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (eds), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, ELRA and ICCL, p. 4287-4298, 2024.

Ollagnier A., Cabrio E., Villata S., "Harnessing Bullying Traces to Enhance Bullying Participant Role Identification in Multi-Party Chats", *in* M. Franklin, S. A. Chun (eds), *Proceedings of the Thirty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2023, Clearwater Beach, FL, USA, May 14-17, 2023*, AAAI Press, 2023a.

Ollagnier A., Cabrio E., Villata S., Blaya C., "CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game", *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, European Language Resources Association, p. 867-875, 2022.

Ollagnier A., Cabrio E., Villata S., Tonelli S., "BiRDy: Bullying Role Detection in Multi-Party Chats", *in* B. Williams, Y. Chen, J. Neville (eds), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, AAAI Press, p. 16464-16466, 2023b.

Papegnies E., Labatut V., Dufour R., Linarès G., "Impact of Content Features for Automatic Online Abuse Detection", *in* A. F. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Budapest, Hungary, April 17-23, 2017, Revised Selected Papers, Part II*, vol. 10762 of *Lecture Notes in Computer Science*, Springer, p. 404-419, 2017.

Puhl R. M., Wall M. M., Chen C., Austin S. B., Eisenberg M. E., Neumark-Sztainer D., "Experiences of weight teasing in adolescence and weight-related outcomes in adulthood: A 15-year longitudinal study", *Preventive medicine*, 2017.

Räsänen P., Hawdon J., Holkeri E., Keipi T., Näsi M., Oksanen A., "Targets of online hate: Examining determinants of victimization among young Finnish Facebook users", *Violence and victims*, vol. 31, n° 4, p. 708-725, 2016.

Sandri M., Leonardelli E., Tonelli S., Jezek E., "Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks", *in* A. Vlachos, I. Augenstein (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Association for Computational Linguistics, p. 2420-2433, 2023.

Sprugnoli R., Menini S., Tonelli S., Oncini F., Piras E., "Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying", *in* D. Fiser, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, J. Wernimont (eds), *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*, Association for Computational Linguistics, p. 51-59, 2018.

Tokunaga R. S., "Following you home from school: A critical review and synthesis of research on cyberbullying victimization", *Computers in human behavior*, vol. 26, n° 3, p. 277-287, 2010.

Tran G. Q., "The naturalized role-play: An innovative methodology in cross-cultural and interlanguage pragmatics research", vol. 5, Reflections on English Language Teaching, p. 1-24, 2006.

Tufa W. T., Markov I., Vossen P., "The Constant in HATE: Analyzing Toxicity in Reddit across Topics and Languages", *CoRR*, 2024.

Van Amsterdam N., Knoppers A., Claringbould I., Jongmans M., "A picture is worth a thousand words: Constructing (non-)athletic bodies", *Journal of Youth Studies*, vol. 15, n° 3, p. 293-309, 2012.

Vidgen B., Nguyen D., Margetts H. Z., Rossini P. G. C., Tromble R., "Introducing CAD: the Contextual Abuse Dataset", *in* K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (eds), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Association for Computational Linguistics, p. 2289-2303, 2021.

Watts L. K., Wagner J., Velasquez B., Behrens P. I., "Cyberbullying in higher education: A literature review", *Comp. in Human Behavior*, 2017.

Comparaison de méthodes pour la détection du discours des *incels* sur Reddit

Camille Demers* — **Dominic Forest***

* *Université de Montréal, École de bibliothéconomie et des sciences de l'information*

RÉSUMÉ. *Les* incels *(célibataires involontaires) regroupent typiquement des hommes se trouvant dans l'incapacité de former des relations amoureuses ou intimes et partageant par conséquent des opinions négatives à l'endroit des femmes. Compte tenu de la gravité des attaques commises par des individus incels et de leur propension à se radicaliser sous l'effet de chambres d'écho, il s'avère plus que nécessaire de détecter le discours de ces communautés virtuelles. Cette étude compare la performance de différents systèmes de détection du discours incel utilisant une approche d'apprentissage par sacs de communautés. Les expérimentations menées permettent de comparer l'efficacité de diverses représentations vectorielles pour entraîner différents algorithmes d'apprentissage supervisé à détecter le discours incel dans un corpus de commentaires provenant de Reddit. Nos modèles les plus performants obtiennent une mesure-F globale variant entre 82,35 % en phase d'apprentissage et 79,70 % en phase de test.*

MOTS-CLÉS : *détection de propos misogynes,* incel*, classification supervisée, comparaison de méthodes, Reddit.*

TITLE. *Comparison of methods for detecting incel speech on Reddit*

ABSTRACT. *Incels (involuntary celibates) typically bring together men that are unable to form romantic or intimate relationships, and therefore share negative opinions about women. Given the seriousness of attacks committed by incel individuals as well as their propensity to radicalize under the effect of echo chambers, it is more than necessary to detect the discourse of these virtual communities. This study compares the performance of various incel speech detection systems using a bag-of-communities learning approach. The experiments carried out compare the effectiveness of various vector representations for training supervised learning algorithms to detect incel speech in a corpus of comments from Reddit. Our best-performing models achieve a macro F-score ranging from 82,35% in the learning phase to 79,70% in the test phase.*

KEYWORDS: *online misogyny detection, incel, supervised classification, comparison of methods, Reddit.*

1. Introduction

> Incel, /ˈɪn.sel/
> Nom masculin : « Membre d'une communauté virtuelle composée principalement de jeunes hommes ne se sentant pas attirant envers les femmes et partageant régulièrement des opinions négatives à leur endroit.»
>
> (Hornby, 2020)

Le 7 novembre 2017, suivant l'adoption d'une nouvelle politique interdisant toute forme de contenu prônant ou incitant à la violence sur sa plateforme, le réseau social Reddit bannit le forum *r/Incels*, regroupant alors plus de 40 000 membres (Hauser, 2017). Un an plus tard, le 15 octobre 2018, le forum en ligne *Incels.me* est suspendu du domaine .me pour violation de sa politique en matière d'abus et de promotion de la violence et des discours de haine (.ME, 2018). Malgré la fermeture de ces espaces numériques, les *incels*, ou « célibataires involontaires », désignent encore aujourd'hui un ensemble de communautés virtuelles formées majoritairement d'hommes partageant une incapacité à lier des relations avec les femmes et alimentant par conséquent une subculture caractérisée par la propagation d'idéologies aux tendances misogynes et antiféministes (Halpin, 2022 ; Meier et Sharp, 2024). Selon le Réseau de sensibilisation à la radicalisation de l'Union européenne (RSR, 2021) :

> Les *Incels* sont persuadés que leur incapacité à avoir des relations sexuelles est due à des facteurs génétiques, des processus évolutivement prédéterminés de sélection du partenaire, ainsi qu'aux structures sociales. Ils pensent que les femmes ne les trouvent pas séduisants et qu'elles ne s'intéressent qu'aux beaux « mâles alpha » (également appelés « chads »). Fréquemment mentionnée chez les incels, la « règle 80/20 » signifie que les 20 % des hommes les plus attirants ont monopolisé 80 % des femmes.

Bien que les espaces de discussion des *incels* se voient fréquemment censurés par les plateformes qui les hébergent, les activités de ces communautés ont franchi les frontières du numérique à un certain nombre de reprises au cours des dernières années, ayant donné lieu à des actes de violence extrême de la part d'individus s'identifiant comme *incels*. L'un des événements les plus médiatisés à cet égard fut la tuerie d'Isla Vista perpétrée en mai 2014, où Elliot Rodger a abattu six personnes après avoir annoncé son plan de représailles envers les femmes dans une vidéo publiée sur sa chaîne YouTube quelques heures avant l'événement : *« You forced me to suffer all my life, and now I'll make you all suffer »* (Associated Press, 2014). Cet événement a contribué par la suite à motiver l'attaque au camion-bélier ayant eu lieu en avril 2018 à Toronto, où Alek Minassian a happé dix personnes à bord d'un véhicule de location après avoir inauguré l'événement sur sa page Facebook, en saluant au passage son prédécesseur : *« The Incel Rebellion has already begun ! We will overthrow all the Chads and Stacys ! All hail the Supreme Gentleman Elliot Rodger ! »* (Nadeau, 2022). Plus récemment, en

juin 2023, Geovanny Villalba-Aleman s'est infiltré dans la salle d'un cours en études de genre à l'université de Waterloo et y a poignardé trois personnes (Shetty, 2023). Bien que l'auteur de cet acte ne se soit pas identifié comme un *incel*, son geste a fait l'objet de célébrations au sein de ces communautés (Halpin *et al.*, 2024).

Depuis la mise à jour de sa politique d'utilisation en 2017, la plateforme Reddit bannit régulièrement des forums de discussion occupés par des *incels*, cependant la fermeture de ces espaces de discussion ne semble pas véritablement arrêter la progression de ces communautés. Ainsi, après que le forum *r/Incels* a été supprimé en 2017, seuls quelques jours ont suffit pour que ses utilisateurs migrent vers de nouveaux espaces, notamment *r/Braincels* (Ribeiro *et al.*, 2021), qui a à son tour été banni en 2019, suivi de *r/AskAnIncel*, banni en 2020. Plus d'actions sont donc requises pour modérer la radicalisation des propos émis au sein des forums de discussion des *incels* et pour éviter la perpétration de nouveaux actes de violence par ces individus.

Différents projets de loi ont été proposés au cours des dernières années pour encadrer les pratiques d'utilisation des médias sociaux et pour favoriser la mise en place d'environnements numériques qui soient plus sécuritaires. En février 2024, le gouvernement du Canada a notamment présenté le projet de loi C-63 sur les préjudices en ligne (gouvernement du Canada, 2024), visant plus particulièrement sept types de contenu préjudiciable dont le contenu fomentant la haine et le contenu incitant à la violence. Ce projet de loi prévoit la mise en place de mesures de protection des publics ainsi que la responsabilisation des plateformes numériques à l'égard de ces contenus, notamment en obligeant la mise en œuvre de mesures pour identifier les contenus préjudiciables et en rendant ces contenus inaccessibles au public dans des délais raisonnables (gouvernement du Canada, 2024 ; Benmoussa *et al.*, 2024).

Étant donné la difficulté à modérer les propos des communautés *incels* en raison de l'anonymat de leurs membres et de leur propension à se radicaliser sous l'effet de chambres d'écho, le développement d'outils permettant de détecter le discours des *incels* constitue une piste d'action concrète visant à mettre en œuvre les mesures proposées par de tels projets de loi. Cet article s'inscrit dans cette perspective. Il rend compte d'expérimentations visant à évaluer différents systèmes de détection automatique permettant d'extraire les spécificités du discours des *incels* sur Reddit en le comparant aux autres formes de discours émises sur cette plateforme. Les retombées associées au développement de ce type d'outils visent à améliorer les fonctionnalités de modération déjà en place pour signaler aux utilisateurs de Reddit la sensibilité des contenus associés au discours véhiculé au sein des espaces de discussion des *incels*[1]. De telles fonctionnalités ont déjà fait leur place au sein des plateformes du groupe Meta [2], sous la forme d'écrans d'avertissement présentés à l'utilisateur avant d'afficher le contenu

1. https://www.reddit.com/r/help/comments/aayoxb/what_is_a_quarantined_subreddit/
2. https://transparency.meta.com/enforcement/taking-action/context-on-sensitive-misleading-content/

sensible. Ce type d'outils permet d'ailleurs de suivre les migrations des communautés produisant ces contenus afin d'en faciliter la suppression.

La suite de l'article est organisée de la manière suivante. Nous dressons d'abord un portrait de travaux portant sur la détection de discours misogynes et sexistes ainsi que sur l'analyse du discours *incel* à proprement parler. La section « Méthodologie » décrit les étapes de constitution d'un corpus de commentaires issus de forums de discussion *incels* et non-*incels* sur Reddit ainsi que le paramétrage de différents modèles de classification visant à détecter ce type de discours. La section « Résultats et discussion » compare la performance des modèles en phase d'apprentissage et de test, puis analyse les meilleurs paramètres de détection. Des recommandations liées aux développements futurs de ce type d'approche sont formulées en conclusion.

2. État de la question

Un important nombre de travaux en traitement automatique des langues, en lexicométrie et en fouille de textes ont récemment visé à caractériser et à identifier des discours sexistes et misogynes sur les réseaux sociaux. Ces travaux incluent ceux réalisés dans le cadre de plusieurs campagnes d'évaluation en traitement automatique des langues, notamment Evalita 2018 et Evalita 2020, *Automatic Misogyny Identification* (Fersini, 2018 ; Fersini *et al.*, 2020), SemEval-2022, *Multimedia Automatic Misogyny Identification* (Fersini *et al.*, 2022) et SemEval-2023, *Explainable Detection of Online Sexism* (Kirk *et al.*, 2023), ou encore les campagnes EXIST, *sEXism Identification in Social neTworks* (Rodríguez-Sanchez *et al.*, 2021 ; Rodríguez-Sánchez *et al.*, 2022 ; Plaza *et al.*, 2023 ; Plaza *et al.*, 2024).

Pour sa part, l'analyse du discours des *incels* a fait l'objet de travaux situés à l'intersection des disciplines de la communication, de l'étude des médias sociaux et des études de genre. Peu de travaux se sont cependant intéressés spécifiquement à détecter le discours des *incels* parmi d'autres formes de discours. Outre les travaux de Gemelli et Minnema (2024) portant sur la constitution et la description d'un corpus de propos *incels* en italien à l'aide de FrameNet, l'annotation de corpus misogynes pose généralement d'importants défis (Sheppard *et al.*, 2024). À cet égard, les travaux liés à la détection automatique de sexisme et de misogynie présentent néanmoins un intérêt pour le présent contexte, cette tâche bénéficiant d'une communauté de pratique plutôt bien établie. Plusieurs de ces travaux se sont penchés non seulement sur la détection, mais également la catégorisation de diverses formes de sexisme et de misogynie.

En l'occurrence, pour répondre à la tâche *Automatic Misogyny Identification* (AMI) de la campagne *Evalita 2018* (Fersini, 2018), Saha *et al.* (2018) ont développé un modèle d'apprentissage automatique permettant de détecter la présence de propos sexistes dans des publications provenant de Twitter et de catégoriser le caractère actif ou passif de la cible des propos. Leur modèle a été le plus performant en utilisant conjointement des méthodes de *sentence embedding* et de vecteurs pondérés par TF-IDF, avec une performance de 70,4 % d'exactitude. Frenda *et al.* (2019) ont réutilisé

le corpus de commentaires misogynes de la compétition AMI de 2018 ainsi qu'un corpus supplémentaire de tweets sexistes pour extraire automatiquement des analogies et des distinctions entre les propos misogynes et sexistes publiés sur Twitter. Ils ont entre autres employé des mesures de similarité lexicale pour comparer la richesse du vocabulaire des deux corpus. Ils ont également développé un système de détection des tweets sexistes et misogynes en utilisant une approche de machines à vecteurs de support dont les traits discriminants ont été pondérés selon leur valeur de TF-IDF, et ont obtenu une performance de 76,05 % d'exactitude.

Dans une étude longitudinale portant sur l'évolution des communautés de la manosphère sur le Web, Ribeiro *et al.* (2021) ont constitué un corpus s'échelonnant sur une période de 14 ans issu de nombreux espaces de discussion de la plateforme Reddit (appelés *subreddits*). Leurs travaux ont permis d'étudier les migrations et intersections des utilisateurs à travers ces plateformes en employant des mesures de similarité sémantique (l'indice de Jaccard et le coefficient de chevauchement). Ceux-ci ont également pu caractériser le niveau de toxicité des propos au sein de ces communautés en développant un système reposant sur un réseau neuronal convolutif (CNN).

Dans un article récent, Morales-Castro *et al.* (2023) ont cherché à détecter automatiquement les propos misogynes en extrayant des informations subjectives de textes non structurés à l'aide du jeu de données d'évaluation de la campagne Evalita. Pour ce faire, ils ont comparé les performances de plusieurs méthodes d'apprentissage supervisé dont les machines à vecteurs de support (SVM), le classifieur bayésien naïf, l'algorithme de régression logistique, les arbres de décisions et l'algorithme des K plus proches voisins (KNN). Les auteurs ont ensuite sélectionné les trois systèmes présentant la plus grande précision et les ont combinés en un métaclassificateur basé sur la régression logistique, atteignant une précision de 81,8 %.

Dans le même ordre d'idées, les travaux de Muti *et al.* (2024) ont abordé la détection des propos mysogynes en ligne en déployant une technique basée sur le raisonnement argumentatif à l'aide de grands modèles de langues. Les résultats obtenus sur un corpus de textes en anglais et en italien sont encourageants (certaines configurations permettent d'obtenir des mesures de rappel de 91,3 %, bien que leur approche basée sur le raisonnement se heurte à d'importantes limites.

En ce qui concerne spécifiquement le discours des *incels*, les travaux réalisés par Jaki *et al.* (2019) ont cherché à développer un système de détection des propos à caractère misogyne, homophobe et/ou raciste parmi les commentaires publiés sur le forum Incels.me. Ceux-ci ont permis de comparer un CNN et un perceptron entraînés à partir de n-grammes de différentes tailles de caractères et de mots. Les deux approches ont permis de détecter les propos misogynes, homophobes et/ou racistes parmi les commentaires issus du forum *incel* avec 95 % d'exactitude.

Pelzer *et al.* (2021) ont pour leur part développé un modèle de détection automatique du niveau de toxicité de trois des plus importants forums Incels connus en 2021, soit Incels.is, Lookisms.net et Looksmax.org. Ils ont utilisé une approche d'apprentissage par transfert (*transfert-learning*) basée sur le modèle de langue BERT. Leurs

analyses ont démontré que les propos de ces trois forums présentaient un taux de toxicité significativement supérieur à celui d'un corpus contrôle provenant de Reddit.

Une approche lexicométrique a été employée par Gothard *et al.* (2021) pour caractériser et analyser les spécificités des patrons de langage propres à trois forums incels sur la plateforme Reddit, soit *r/Braincels*, *r/Incels* et *r/Shortcels*. Ces travaux ont également comparé la richesse lexicale de *r/Braincels* à celui d'un corpus contrôle issu de différents *subreddits*, en mobilisant des mesures de statistiques lexicales. Les lexiques des deux corpus ont été comparés à l'aide d'une analyse rang-rang (*rank-rank*) des formes les plus fréquentes. Cette approche a démontré que le vocabulaire des commentaires issus de *subreddits Incels* a tendance à être moins riche que celui d'autres communautés de la plateforme.

Plus récemment, Yoder *et al.* (2023) ont utilisé des mesures de statistiques textuelles (fréquences, distribution, etc.) et une analyse de réseau sur un ensemble de données comprenant 6 248 234 commentaires postés sur le forum *incels.is* pour évaluer la construction d'une identité de groupe à travers le discours des incels. Parallèlement à l'utilisation de méthodes lexicales classiques, ils ont entraîné un modèle word2vec à partir de ces commentaires afin d'évaluer la diversité du vocabulaire associé à l'identité des incels. Leurs résultats suggèrent une forte importance accordée à l'apparence physique comme déterminante de la valeur individuelle des êtres humains ainsi qu'une prévalence de la question du genre au centre des discussions entre individus incels. Ces chercheurs ont également extrait des termes identitaires fortement péjoratifs à l'égard des femmes, ce qui corrobore les travaux mentionnés ci-dessus. Ceux-ci ont identifié une prévalence de 30 % de ces termes d'identité dans le corpus utilisé, et suggèrent que tout classificateur visant à détecter les discours haineux ou misogynes devrait inclure ces termes comme caractéristiques discriminantes.

Les travaux menés par Hajarian et Khanbabaloo (2021) sont les seuls à s'être spécifiquement intéressés à détecter les utilisateurs *incels* sur Facebook et sur Twitter. Pour ce faire, ces chercheurs ont combiné une méthode d'analyse de sentiments à un système de détection de propos injurieux qu'ils ont appliqué à un corpus de commentaires issus de ces deux plateformes. Leurs travaux ont été en mesure de détecter le discours *incel* dans 78.8 % des cas.

Finalement, dans une étude récente et très importante, Arango *et al.* (2022) ont souligné le contraste entre d'une part l'incapacité, malgré des investissements colossaux, des principaux réseaux sociaux à détecter automatiquement les contenus haineux et, d'autre part, les résultats de recherche, générés principalement par des chercheurs issus du secteur académique, indiquant que les approches de classification supervisée permettent d'atteindre des performances très appréciables. Selon ces chercheurs, l'écart entre les performances des systèmes issus de la recherche et ceux des concepteurs des réseaux sociaux s'explique par des problèmes d'ordre méthodologique et par un biais dans la conception des ensembles de données employés pour valider ces systèmes. Par conséquent, les performances des systèmes de pointe seraient, selon ces auteurs, considérablement surestimées. Ainsi, ceux-ci ont ré-évalué les résultats de certains travaux documentés dans la littérature et ont révélé une importante baisse

de performance de ces systèmes, passant dans certains cas de performances de plus de 90 % à des performances qui seraient plutôt de l'ordre de 50 % à 80 %, sur des ensembles de données plus représentatifs de la réalité des réseaux sociaux.

La recherche dont nous rendons compte dans le présent article vise à comparer rigoureusement différentes techniques de détection automatique en faisant varier divers paramètres afin d'en favoriser l'application à plus large échelle.

3. Méthodologie

3.1. *Constitution de corpus*

Les données servant à l'apprentissage des modèles proviennent de jeux de données existants en langue anglaise. Celles-ci représentent d'une part des propos représentatifs du discours *incel* (classe « *incels* ») et d'autre part des données représentatives des pratiques langagières propres à la plateforme Reddit, mais sans qu'une thématique particulière n'y soit associée (classe « neutres » ou « non-*incels* »). Cette dernière classe est également constituée de commentaires provenant de Reddit afin de limiter l'introduction de biais reflétant les spécificités langagières propres à cette plateforme plutôt que du discours d'intérêt à détecter (Tranchese et Sugiura, 2021).

L'annotation des données servant à entraîner les modèles de classification repose sur une approche appelée « sac de communautés » (*Bag-of-Communities* [BoC]) (Chandrasekharan *et al.*, 2017). Cette approche consiste à identifier une communauté entière comme étant haineuse ou toxique plutôt que d'annoter chacun des contenus publiés par ses utilisateurs individuellement. L'un des avantages de cette approche est qu'elle permet de pallier l'absence de données d'évaluation de même que le recours à un travail d'annotation manuelle auprès d'experts de domaine (Chandrasekharan *et al.*, 2017 ; Pelzer *et al.*, 2021). Cette méthode offre donc la possibilité d'exploiter les données de communautés sources pour classer les discours haineux dans une communauté cible, en se basant sur l'hypothèse que la communauté cible présente des similarités linguistiques avec les communautés sources (Muralikumar *et al.*, 2023).

Un enjeu potentiellement associé à l'annotation basée sur les sacs de communautés réside dans la configuration des données d'entraînement, laquelle rend difficile de distinguer si les propos détectés constituent véritablement des propos haineux ou s'ils reflètent uniquement des similitudes entre les communautés concernées (Berglind *et al.*, 2019). Or dans un contexte où les communautés sélectionnées sont reconnues pour la toxicité de leurs propos, cette approche pourrait offrir le potentiel d'améliorer l'adaptabilité des modèles de détection de discours haineux entre des communautés issues de différentes plateformes (p. ex. Reddit, Facebook ou X), ou encore de sous-communautés distinctes au sein d'une même plateforme (par exemple les *subreddits* sur Reddit) (Almerekhi *et al.*, 2020). Nous pensons donc que les sacs de communautés présentent un intérêt particulier pour la détection du discours des *incels*. D'une part, ces communautés partagent une vision du monde et une identité commune se manifestant d'une manière similaire dans les différents espaces de discussion où elles sont

actives ; d'autre part, la migration constante des utilisateurs vers de nouveaux espaces faisant suite au banissement de *subreddits incels* implique l'existence d'un alignement entre les espaces de discussion de ces communautés.

Le corpus constitué dans le cadre de cet article regroupe donc un ensemble de commentaires issus de *subreddits* reconnus comme des espaces de discussion occupés par des *incels*, par opposition à un ensemble de commentaires issus de *subreddits* non dédiés à ces communautés. Chaque commentaire est étiqueté comme étant *incel* ou non-*incel* en fonction du *subreddit* au sein duquel celui-ci a été publié.

3.1.1. *Données* incels

Deux jeux de données ont été mobilisés pour constituer la classe *incels*. Nous avons d'abord identifié un corpus rendu disponible par Ribeiro *et al.* (2020)[3] pour identifier un ensemble de *subreddits* caractéristiques du discours *incel*. Ce jeu de données regroupe des espaces de discussion de l'androsphère (*manosphere*) sur le Web, un ensemble de communautés aux revendications masculinistes incluant les communautés *incels*. Ce corpus compte plus de 28,8 millions de publications provenant de différents forums en ligne, dont 56 forums de discussion Reddit (*subreddits*). Dans le cadre de leurs travaux, les auteurs ont catégorisé chacun de ces 56 *subreddits* en fonction de sa communauté d'appartenance au sein de la manosphère (p. ex. *Incels*, *Men Get Their Own Way [MGTOW], Pick Up Artists [PUA], The Red Pill [TRP], Men's Rights Activist [MRA]*). Au regard de cette analyse, 23 des 56 *subreddits* ont été identifiés comme étant spécifiquement *incels* (p. ex. *r/Braincels, r/ForeverUnwanted, r/AskAnIncel*)[4]. Nous avons donc retenu ces 23 *subreddits* comme point de départ pour constituer les données d'apprentissage de la classe *incels*. Comme ce corpus couvre une période temporelle s'arrêtant en 2019, nous avons récupéré les données provenant des *subreddits incels* à partir des archives du projet PushShift (Baumgartner *et al.*, 2020), disponibles sur la plateforme The-Eye[5], et ce afin de représenter une période temporelle couvrant les 10 dernières années (janvier 2014 à décembre 2023). Nous avons choisi de ne retenir que les commentaires et publications émis depuis 2014 afin de refléter le jargon actuel de ces communautés ainsi que de tenir compte d'éventuelles évolutions dans ses pratiques discursives.

Parmi les 23 *subreddits* identifiés par Ribeiro *et al.* (2021) comme étant *incels*, nous en avons retenu 9 figurant parmi les 40 000 *subreddits* les plus populaires de Reddit à travers la totalité de son historique[6] : *r/AskAnIncel*, *r/BlackPillScience*, *r/Braincels*, *r/ForeverAlone*, *r/ForeverAloneDating*, *r/ForeverUnwanted*, *r/Incels*,

3. Ces données sont disponibles à l'adresse suivante : `https://zenodo.org/records/4007913`

4. La catégorisation effectuée par Ribeiro *et al.* (2021) est disponible à l'adresse suivante : `https://github.com/idramalab/manosphere_analysis/blob/master/data/subreddit_descriptions.csv`

5. `https://the-eye.eu/redarcs/`

6. Les données constituant la classe « *incels* » ont été récupérées à l'adresse suivante : `https://academictorrents.com/details/56aa49f9653ba545f48df2e33679f014d2829c10`

r/IncelSelfies, *r/IncelsWithoutHate*. Les publications ont été sélectionnées aléatoirement au sein de ces 9 *subreddits incels* pour une période s'échelonnant de 2014 à 2023, en maintenant égale la proportion de publications par année. Le tableau 1 illustre quelques exemples de données issues des *subreddits incels*.

Subreddit	**Commentaire**
r/Braincels	Women are the root. The reason we express our emotions through anger 90% of the time is because that's what women find attractive. Femoids, fuck you. You did this. fuck you !
r/ForeverUnwanted	Abandon all hope, gentlemen. Any woman you ever meet and fall in love with will be able to find some other dude to replace you quicker than all fuck.
r/Incels	Who gives a fuck. Bitch gave her word to be wife for life. No wonder nobody respects females. they can't hold up their end of such a sacred agreement. 80% of divorces initiated by females. Almost always for bullshit reasons too.

TABLEAU 1. *Exemples de commentaires issus de* subreddits incels

.

3.1.2. *Données non*-incels

Pour constituer la classe de données non-*incels*, nous avons extrait un échantillon de commentaires à partir des archives du projet PushShift (Baumgartner *et al.*, 2020) (via la plateforme The-Eye), lesquelles couvrent la totalité des publications et commentaires émis sur Reddit entre 2005 et 2023[7]. Nous avons extrait de ces données un échantillon de commentaires publiés entre 2014 et 2023 dans différents *subreddits* non reconnus comme *incels* (p. ex. *r/OffMyChest, r/Electronic_Cigarette, r/FinalFantasy*, etc.). Étant donné le volume important de cette archive, nous avons extrait, pour chaque année allant de 2014 à 2023, l'ensemble des publications et commentaires sur une période de 1 mois sélectionné aléatoirement (p. ex. 2014 = avril, 2015 = mai, etc.). Ces publications sont issues de 13 828 *subreddits* distincts. Le tableau 2 illustre quelques exemples de données provenant des *subreddits* non-*incels*.

3.1.3. *Partitionnement des données d'apprentissage et de test*

Le partitionnement des données a été effectué selon un ratio classique deux tiers un tiers : 40 000 commentaires ont été utilisés pour la phase d'apprentissage, tandis que 20 000 commentaires inédits ont été utilisés pour évaluer la performance des modèles paramétrés, c'est-à-dire tester leur capacité de généralisation à détecter les propos *incels* dans des commentaires inédits provenant de Reddit.

Pour les données d'apprentissage, nous avons constitué différents jeux de données en faisant varier le ratio de commentaires *incels* contenus dans ceux-ci. Neuf corpus

7. Les données constituant la classe « neutres » ont été récupérées à l'adresse suivante : `https://academictorrents.com/details/9c263fc85366c1ef8f5bb9da0203f4c8c8db75f4`

Subreddit	Commentaire
r/WhoWouldWin	He's on par with a lot of movie showings of Superman. When Superman shows up in long running series, like comics, he tends to gradually become significantly more powerful. This is true for most characters like him, as well.
r/SpinnCoffee	I ordered my Spinn in January. According to my order, I am also batch 6. I feel like June to January is a long batch period. I'm a bit worried now about when I'll actually get mine.
r/OldHagFashion	Years ago, a group of us had a "fun vest night." I wore this vest, made by my mom (apparently from a kit), when I was a kid. Showed up at my friend's door...he was wearing the exact same vest. It's a good one.

TABLEAU 2. *Exemples de commentaires issus de* subreddits *non*-incels

d'apprentissage ont ainsi été créés, contenant de 10 % à 90 % de commentaires *incels*. Cette stratégie a été employée dans le but de pallier le problème de déséquilibre des classes auxquelles font typiquement face les tâches de détection automatique (Ling et Sheng, 2010), où la tâche à détecter se trouve sous-représentée dans la réalité, rendant plus difficile l'extraction de ses spécificités par un modèle de classification. Bien que la proportion réelle du discours *incel* sur Reddit demeure inconnue, cette approche vise à évaluer l'effet d'une surreprésentation des données *incels* dans les données d'apprentissage sur les performances de détection résultantes. Cette approche s'oppose aux stratégies traditionnelles de suréchantillonnage, qui consistent à dédoubler des exemplaires des catégories de la classe à détecter dans les données d'apprentissage, lesquelles seraient à la source des problèmes méthodologiques associés à la surestimation des performances de certains systèmes selon Arango *et al.* (2022). Notre approche vise plutôt à faire varier la proportion de commentaires uniques associés à chaque classe dans les données d'apprentissage dans le but de contrer ce type de biais méthodologique. Ainsi, chaque corpus d'apprentissage totalise 40 000 documents, avec un ratio variable de données *incels* et non-*incels*. Pour chaque proportion testée, les données ont été sélectionnées au moyen d'une technique d'échantillonnage aléatoire stratifiée visant à maintenir égale la proportion de publications par année. Le tableau 3 illustre les caractéristiques des corpus d'apprentissage, comptant en moyenne 1 208 385 occurrences de mots (*tokens*) et 57 955 formes uniques (*types*).

Le corpus dédié à l'évaluation des modèles totalise pour sa part 20 000 commentaires, dont la proportion de commentaires *incels* a été fixée à 10 %. Comme la proportion réelle des propos *incels* sur Reddit est inconnue, ce ratio a été retenu en fonction des résultats obtenus par Hajarian et Khanbabaloo (2021), lesquels ont rapporté des proportions de propos *incels* de l'ordre de 9,1 % sur Facebook et de 8,5 % sur Twitter. Afin de tenir compte de ces proportions, le corpus d'évaluation des modèles compte donc 2 000 commentaires *incels* et 18 000 commentaires neutres, tel qu'illustré dans le tableau 4.

% *Incels*	Commentaires *incels*	Commentaires non-*incels*	Total	Occurrences de mots (tokens)	Mots uniques (types)
10	4 000	36 000	40 000	1 087 228	66 037
20	8 000	32 000	40 000	1 117 519	63 960
30	12 000	28 000	40 000	1 156 716	63 543
40	16 000	24 000	40 000	1 167 474	60 345
50	20 000	20 000	40 000	1 214 857	59 124
60	24 000	16 000	40 000	1 238 734	56 267
70	28 000	12 000	40 000	1 259 718	53 694
80	32 000	8 000	40 000	1 302 462	50 446
90	36 000	4 000	40 000	1 330 756	48 180

TABLEAU 3. *Variation de la proportion des commentaires* incels *et non-*incels *dans les données d'apprentissage, avec occurrences et types de mots*

% *Incels*	Commentaires *incels*	Commentaires non-*incels*	Total	Occurrences de mots (tokens)	Mots uniques (types)
10	2 000	18 000	20 000	461 414	39 554

TABLEAU 4. *Proportion des commentaires* incels *et non-*incels *dans les données d'évaluation (données de test)*

3.2. *Prétraitements*

Différents filtrages ont été appliqués de manière à supprimer les commentaires non pertinents. Les publications vides (« *[removed]* », « *[deleted]* ») ou ne contenant qu'un seul caractère ont été retirées, de même que les publications de robots modérateurs (auteur *« AutoModerator »*). L'ensemble du texte des commentaires a été minusculisé. Des patrons d'expressions régulières ont été employés pour retirer les URL ainsi que certains artefacts issus de l'API de Reddit tels que des entités HTML (p. ex. « *>* »). Les commentaires ont ensuite été segmentés en mots (*tokenisés*) avec la fonction *word_tokenize* de la librairie Python NLTK (Bird *et al.*, 2009). Les mots fonctionnels ont été filtrés au moyen d'un antidictionnaire de l'anglais, de même que les expressions contenant des chiffres ou des caractères spéciaux.

3.3. *Phase d'apprentissage*

Pour entraîner les modèles de classification, nous avons comparé trois approches permettant de représenter numériquement les commentaires provenant de Reddit : (1) des vecteurs basés sur une pondération TF-IDF des termes du lexique ; (2) des vecteurs basés sur le modèle de plongement lexical *Continuuous Bag-of-Words* (CBOW) (Mikolov *et al.*, 2013) ; (3) des vecteurs basés sur le modèle de plongement de phrases

Sentence Transformers (SBERT) (Reimers et Gurevych, 2019). Le choix de ces trois types de représentations a pour objectif d'évaluer la capacité de modèles de classification à exploiter différentes caractéristiques textuelles pour la tâche de détection de cette étude. Cette approche vise en l'occurrence à comparer des méthodes classiques issues de la statistique lexicale (TF-IDF) à des méthodes plus récentes exploitant les relations de dépendance contextuelle entre les mots pour produire des représentations sémantiques denses (Word2Vec, SBERT). Ces trois modèles ont été comparés compte tenu de l'existence de résultats contradictoires concernant l'emploi de plongements lexicaux pour des tâches de classification textuelle, notamment Word2Vec (Abubakar *et al.*, 2022 ; Truşcă, 2019), BERT ou SBERT (Jamshidian, 2023), en comparaison avec des vecteurs TF-IDF. Nos travaux cherchent donc à mettre en perspective ces résultats dans le contexte de la détection du discours *incel*.

Le *Term Frequency-Inverse Document Frequency* (TF-IDF) est une mesure de pondération statistique permettant de calculer un score de spécificité pour chaque terme d'un corpus en fonction de sa fréquence relative dans chaque document par rapport à sa proportion inverse sur l'ensemble des documents du corpus (IDF) (Ramos, 2003). Une motivation à employer cette approche dans le cadre d'une tâche de classification supervisée réside dans le fait que les termes présentant une forte fréquence dans un nombre restreint de documents obtiendront un score TF-IDF plus élevé pour ceux-ci, reflétant la pertinence de ces termes pour représenter ce groupe de documents. Les vecteurs TF-IDF ont été générés avec la fonction *Tfidfvectorizer* de la librairie Python Scikit-learn (Pedregosa *et al.*, 2011).

Le modèle de plongement lexical *Continuous Bag-of-Words* est pour sa part basé sur une architecture neuronale simple permettant de prédire la probabilité conditionnelle d'occurrence associée à chacun des mots d'un document compte tenu d'une fenêtre contextuelle donnée, laquelle est typiquement constituée des 5 mots entourant le mot à prédire (Azmy *et al.*, 2018 ; Mikolov *et al.*, 2013). Ce modèle permet de générer une représentation vectorielle de longueur fixe pour chaque mot d'un document, puis ces représentations de mots peuvent ensuite être combinées au moyen d'une fonction d'agrégation (par exemple la somme ou la moyenne des éléments des vecteurs) pour obtenir un seul vecteur par document. Les vecteurs CBOW ont été générés au moyen du modèle Word2Vec disponible dans la librairie Python Gensim, avec une fenêtre contextuelle de 5 mots. Pour chaque commentaire, un seul vecteur a ensuite été généré par l'agrégation des vecteurs de mots en utilisant la moyenne comme fonction d'agrégation.

Le modèle de plongement de phrases *Sentence Transformers* (SBERT) est un réseau neuronal basé sur l'architecture Transformer permettant de générer des représentations vectorielles de phrases ou de textes courts (Reimers et Gurevych, 2019). Ce modèle constitue une extension du réseau neuronal préentraîné BERT spécifiquement conçue pour la comparaison de paires de phrases. Il utilise une architecture bi-encodeur employant deux instances parallèles de BERT pour traiter indépendamment des paires de phrases, le rendant plus efficace que BERT pour des tâches nécessitant une comparaison de similarité textuelle (Reimers et Gurevych, 2019). Dans le

cadre de cette étude, les vecteurs SBERT ont été générés au moyen du modèle *all-MiniLM-L6-v2* de la librairie Python Sentence Transformers, qui permet d'encoder les commentaires issus de Reddit en un vecteur de longueur fixe de 384 dimensions.

Nous avons fait varier le nombre de dimensions des vecteurs TF-IDF et CBOW afin d'évaluer l'effet de ce paramètre sur la performance de détection résultante. Nous avons employé des valeurs allant de 1 000 à 15 000 éléments pour les modèles basés sur une pondération TF-IDF étant donné la haute dimensionnalité des vecteurs typiquement associés à cette approche. En contrepartie, l'une des caractéristiques des vecteurs générés au moyen d'approches par plongement comme CBOW et SBERT réside dans leur capacité à représenter l'information de manière dense, et ce avec une dimensionnalité relativement restreinte. Nous avons donc retenu des valeurs allant de 100 à 500 éléments pour les plongements associés aux modèles CBOW. Pour leur part, comme les plongements SBERT sont issus d'un modèle de langue préentraîné générant des vecteurs dont la dimension est fixée à 384 éléments (*all-MiniLM-L6-v2*), nous n'avons pas fait varier ce paramètre pour ce type de représentations.

3.4. *Classification*

Différents algorithmes de classification ont été évalués pour détecter la présence de propos *incels* dans les commentaires issus de Reddit. Pour ce faire, nous avons sélectionné quatre algorithmes implémentés dans la librairie Python Scikit-learn (Pedregosa *et al.*, 2011) : (1) la régression logistique (LR) ; (2) les machines à vecteurs de support (SVM) ; les forêts aléatoires (RF) ; (4) la classification bayésienne naïve multinomiale (MNB). Le choix de ces méthodes repose sur leur capacité à traiter des données dans les formats de représentation que nous faisons varier dans nos expérimentations, mais aussi sur le fait que plusieurs d'entre elles ont été fréquemment employées dans des tâches de détection du cyberharcèlement et, plus spécifiquement, des propos misogynes. Nous avons évalué la pertinence de ces méthodes sur des ensembles de données textuelles représentées numériquement à l'aide des trois méthodes de vectorisation mentionnées dans la section précédente (TF-IDF, CBOW, SBERT). D'un point de vue technique, ces approches présentent plusieurs avantages dans le contexte de nos travaux : elles ne nécessitent pas de grands ensembles de données (compte tenu de la taille relativement petite de notre corpus – 60 000 commentaires totalisant moins de 2 millions de mots) pour générer des résultats de qualité et sont relativement peu couteuses computationnellement.

La régression logistique (LR) est une approche de classification probabiliste permettant de déterminer l'importance de chaque attribut d'un document en termes d'un poids relatif (coefficient de régression) contribuant à une fonction de décision, ce poids pouvant être positif ou négatif. La classification consiste à appliquer une fonction logistique à la somme des poids associés aux attributs d'un document afin d'obtenir une probabilité que ce document appartienne à une classe d'intérêt, ce dernier étant classifié comme tel au-delà d'un certain seuil (Jurafsky et Martin, 2019).

Les machines à vecteurs de support (SVM) sont des modèles de classification binaire reposant sur l'identification d'un hyperplan séparant deux classes de données en maximisant la distance (marge) entre cet hyperplan et tout point de données dans un espace vectoriel (Manning *et al.*, 2008). L'hyperplan est identifié à partir de données d'apprentissage et peut ensuite être appliqué à de nouvelles données dans un même espace vectoriel.

Les forêts aléatoires (RF) sont issues d'approches d'apprentissage ensemblistes, c'est-à-dire effectuant la tâche de classification en combinant les décisions d'un ensemble de classifieurs à travers un processus de vote (Pal, 2005). Dans une forêt aléatoire, les classifieurs employés reposent sur un apprentissage par arbre de décision, où chaque arbre traite un sous-ensemble d'attributs relatifs à l'élément à classer, lesquels sont échantillonnés aléatoirement. Chaque arbre vote ensuite pour la classe sous laquelle l'élément devrait être classé en fonction de ces attributs (Breiman, 1999).

Les classifieurs bayésiens naïfs (NB) sont des algorithmes probabilistes modélisant la probabilité qu'un élément appartienne à une classe donnée en fonction d'un ensemble de caractéristiques considérées comme indépendantes les unes des autres (Feldman et Sanger, 2006). Des variantes de ces classifieurs dépendent du choix de modélisation des probabilités conditionnelles selon différentes distributions (Manning *et al.*, 2008). Nous avons employé un classifieur bayésien naïf basé sur une distribution multinomiale (MNB) pour les vecteurs générés par pondération TF-IDF. Nous n'avons pas testé ce classifieur pour les modèles CBOW et SBERT, comme ils génèrent des représentations vectorielles dont les éléments sont des valeurs continues.

Pour chacun des classifieurs ci-dessus, nous avons testé l'ensemble des configurations associées aux paramètres mentionnés dans les sections précédentes : ratio de données *incels* dans les données d'apprentissage, technique de vectorisation utilisée, nombre de dimensions des vecteurs, algorithme de classification employé. La figure 1 illustre l'ensemble des configurations testées. Chacun des modèles a fait l'objet d'une validation croisée à 5 plis. Les paramètres optimaux pour chaque modèle ont été identifiés au moyen d'une recherche en grille (*GridSearchCV*).

Les expérimentations ont été réalisées avec Python 3.11 sous le système d'exploitation Windows 11, sur une machine dotée de 16 Go de mémoire vive et d'un processeur à 2,9 GHz et 8 cœurs. L'entraînement des modèles utilisant SentenceTransformers ont été réalisés en utilisant une unité de calcul graphique (GPU) NVIDIA GeForce RTX 3070. Les scripts utilisés pour prétraiter les données, entraîner les modèles et générer les résultats sont disponibles à l'adresse suivante : `https://github.com/CamilleDemers/incels-detection-reddit`.

Les 20 configurations ayant généré les meilleures performances en phase d'apprentissage ont été évaluées sur un ensemble de test composé de 20 000 documents inédits. Ces performances sont présentées dans la section suivante.

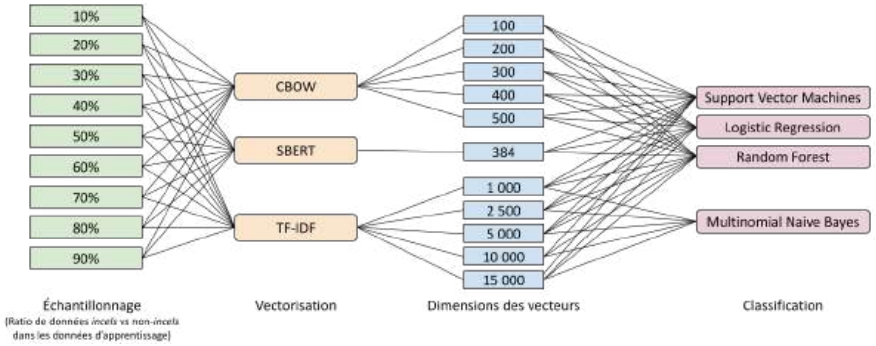

FIGURE 1. *Configurations testées pour l'entraînement des modèles de classification*

4. Résultats et discussion

4.1. *Performances de classification*

4.1.1. *Phase d'apprentissage*

Les résultats sont évalués au regard de l'exactitude prédictive, la précision et le rappel. La performance globale est évaluée au moyen de la mesure-F (macro f1). Cette mesure reflète d'ailleurs plus adéquatement la performance que l'exactitude prédictive dans un contexte de déséquilibre des classes (Zhao et Chen, 2014). Les résultats sont obtenus en calculant la moyenne arithmétique des métriques associées à chacun des plis de la validation croisée réalisée sur les données d'apprentissage.

Les tableaux 5a, 5b et 5c illustrent les 5 meilleures configurations en phase d'apprentissage pour les trois modèles de vectorisation sélectionnés. Globalement, les modèles de classification basés sur des vecteurs SBERT performent le mieux (mesure-F maximale = 84,60), suivis des vecteurs TF-IDF (mesure-F maximale = 81,98) et des vecteurs CBOW (mesure-F maximale = 78,58) (SBERT > TF-IDF > CBOW).

Pour les vecteurs TF-IDF, le classifieur bayésien naïf multinomial entraîné sur un corpus contenant 40 % de données *incels* avec des vecteurs de 15 000 dimensions obtient la meilleure performance de détection (mesure-F = 81,98). Autrement, un ratio plus élevé de données *incels* dans les données d'apprentissage donne lieu à de meilleures performances prédictives, pour un ratio allant jusqu'à 60 % de données *incels*. En ce qui concerne l'algorithme de détection employé, la classification bayésienne naïve multinomiale ainsi que la régression logistique obtiennent les meilleures performances. Finalement, les modèles retenant 10 000 à 15 000 dimensions performent mieux que ceux retenant un nombre plus faible de traits discriminants.

% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F
40	MNB	TF-IDF	15 000	82,71	82,01	81,95	**81,98**
40	MNB	TF-IDF	10 000	82,69	82,05	81,76	81,89
50	LR	TF-IDF	15 000	81,77	82,07	81,77	81,73
50	LR	TF-IDF	10 000	81,59	81,86	81,59	81,55
60	LR	TF-IDF	15 000	82,24	81,48	81,62	81,54

(a) *Performances de classification des 5 meilleurs modèles basés sur des vecteurs TF-IDF, triés par mesure-F. MNB = Multinomial Naive Bayes, LR = Logistic Regression*

% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F
50	RF	CBOW	500	78,59	78,61	78,59	**78,58**
50	RF	CBOW	200	78,55	78,56	78,55	78,55
50	RF	CBOW	450	78,50	78,51	78,50	78,49
50	RF	CBOW	350	78,46	78,48	78,46	78,46
50	RF	CBOW	100	78,43	78,44	78,43	78,43

(b) *Performances de classification des 5 meilleurs modèles basés sur des vecteurs CBOW (Word2Vec), triés par mesure-F. RF = Random Forest*

% Incels	Algorithme	Vecteurs	Dim.	Exactit.	Précision	Rappel	Mesure-F
40	SVM	SBERT	384	85,38	83,95	78,47	**84,60**
50	SVM	SBERT	384	84,50	85,32	83,34	84,50
40	LR	SBERT	384	85,26	83,64	78,51	84,47
50	LR	SBERT	384	84,37	85,14	83,29	84,37
60	SVM	SBERT	384	84,82	87,01	87,81	84,15

(c) *Performances de classification des 5 meilleurs modèles basés sur des plongements SBERT, triés par mesure-F. SVM = Support Vector Machine, LR = Logistic Regression*

TABLEAU 5. *Performances de classification associées aux trois approches de vectorisation testées, en phase d'apprentissage. Pour chaque modèle, la mesure-F maximale est indiquée en gras.*

Les modèles basés sur les vecteurs CBOW illustrent des tendances singulières. Ces modèles performent en effet systématiquement mieux en mobilisant des corpus d'apprentissage contenant 50 % de données *incels* et en exploitant un classifieur basé sur des forêts aléatoires (RF). Ces résultats ne permettent cependant pas de déterminer si la dimension des vecteurs est liée à la performance de détection pour les modèles basés sur des vecteurs CBOW. Bien que les meilleures performances soient associées à des vecteurs de haute dimension (500 dimensions), des vecteurs de plus faible dimension (100) se retrouvent également parmi les 5 meilleures configurations.

Les modèles associés aux plongements SBERT donnent pour leur part de meilleures performances en mobilisant des classifieurs basés sur les machines à vecteurs de support (SVM) et sur la régression logistique (LR). Ces modèles obtiennent par ailleurs de meilleurs résultats lorsqu'entraînés sur des corpus d'apprentissage contenant 40, 50 ou 60 % de données *incels*, ce qui est similaire aux modèles basés sur une pondération TF-IDF du lexique.

Autrement, l'écart entre les résultats relatifs aux vecteurs TF-IDF et CBOW semble cohérent avec les travaux de Wang *et al.* (2017), lesquels ont observé de meilleures performances de classification pour des textes courts avec des vecteurs TF-IDF plutôt qu'avec les modèles Word2Vec et Doc2Vec, en utilisant trois des classifieurs mobilisés dans cette étude (LR, SVM et MNB). Similairement, les travaux de Truşcă (2019) suggèrent que les vecteurs TD-IDF surpassent les vecteurs Word2Vec pour des problèmes linéairement séparables. En contrepartie, les différences de performance observées pour les vecteurs SBERT et TF-IDF font contraste avec les travaux de Jamshidian (2023), lesquels ont obtenu une meilleure performance de classification avec des vecteurs TF-IDF plutôt que des plongements SBERT pour entraîner des modèles d'analyse de sentiment basés sur des SVM.

4.1.2. *Phase de test*

Pour la phase de test, les 20 meilleurs modèles en phase d'apprentissage ont été retenus et évalués sur 20 000 commentaires inédits. L'ensemble de ces modèles sont basés sur des vecteurs TF-IDF ou des plongements SBERT, ceux-ci ayant illustré de meilleures performances d'apprentissage. Autrement, les meilleurs modèles sont issus des quatre algorithmes de détection testés : (1) la classification bayésienne naïve multinomiale ; (2) la régression logistique ; (3) les machines à vecteurs de support et (4) les forêts aléatoires. Pour les modèles TF-IDF, ces configurations sont associées à des vecteurs de 10 000 et de 15 000 dimensions et à des corpus d'entraînement contenant de 40 à 50 % de commentaires *incels*. Pour les modèles SBERT, les meilleures performances sont associées à des corpus d'entraînement contenant de 20 à 80 % de commentaires *incels*.

Les tableaux 6a et 6b présentent les métriques de performance des 5 meilleures configurations en phase d'apprentissage et de test pour chaque classe individuelle. Ces résultats indiquent une meilleure performance de détection pour la classe non-*incel*, qui atteint une mesure-F maximale de 96,13 avec une exactitude de 92,87. À l'opposé, la classe de données *incel* est l'objet de résultats nettement inférieurs, avec

une mesure-F maximale de 63,35. Cet écart entre les résultats des deux classes pourrait en partie s'expliquer par le débalancement des données dans le corpus d'évaluation (10 % *incel* / 90 % non-*incel*), lequel favorise la classe non-*incel*. Il est également possible que l'approche de sac de communautés employée pour l'annotation des données d'évaluation sous-estime la capacité réelle des modèles à extraire les caractéristiques permettant de détecter le discours *incel* parmi les autres formes de discours sur Reddit. En effet, la présence d'erreurs d'annotation est inhérente à cette approche puisqu'elle considère uniquement l'espace de discussion où un commentaire a été publié (le *subreddit*) pour lui attribuer une catégorie. Une évaluation sur un corpus annoté manuellement permettrait de mesurer plus justement la performance des modèles.

Le tableau 6c présente les performances des 5 modèles en phase de test. Le modèle le plus performant repose sur une régression logistique avec des vecteurs SBERT et un corpus d'apprentissage contenant 20 % de données *incels* (mesure-F = 79,70). Globalement, les meilleurs modèles sont associés aux ratios de 20 % et de 30 % de données *incels* en apprentissage (mesure-F = 79,70, 79,65 et 78,15), tandis que les pires sont ceux entraînés sur des corpus contenant 80 % de données *incels* (mesure-F = 43,76, 56,71 et 56,82). Cette tendance suggère un suréchantillonnage trop élevé de la classe *incel* en phase d'apprentissage, résultant en une proportion plus élevée de prédictions faussement positives pour la classe *incels* en phase de test. À titre indicatif, les modèles TF-IDF les plus performants arrivent aux dixième, onzième et douzième rangs parmi tous les modèles évalués en phase de test. Il s'agit des modèles suivants : classification bayésienne naïve multinomiale avec 10 000 traits discriminants, entraînée sur un corpus contenant 40 % de données *incels* (mesure-F = 71,67) ; régression logistique avec 15 000 traits discriminants et 50 % de données *incels* en apprentissage (mesure-F = 71,44) ; classification bayésienne naïve multinomiale, 40 % de données *incels* en apprentissage et 10 000 traits discriminants (mesure-F = 71,39).

4.2. *Analyse des traits prédictifs de chaque classe*

Malgré la meilleure performance de classification des modèles basés sur des vecteurs SBERT pour la détection du discours *incel*, les caractéristiques extraites par ces modèles ne permettent pas de rendre compte des propriétés textuelles contribuant au processus de décision de ces modèles. En contrepartie, les modèles issus de vecteurs basés sur une pondération TF-IDF du lexique offrent de manière générale une plus grande interprétabilité quant au processus de classification résultant. Puisque la performance de ces derniers modèles demeure comparable à celle basée sur des vecteurs SBERT, nous avons mené une analyse supplémentaire permettant d'examiner les termes revêtant une plus grande importance pour la frontière de décision des modèles de classification basés sur des vecteurs TF-IDF.

Le tableau 7 présente les caractéristiques les plus prédictives pour chaque classe. Ces caractéristiques ont été analysées à partir du modèle de régression logistique basé sur des vecteurs TF-IDF ayant obtenu les meilleures performances en phase de test (50 % de données *incels* en apprentissage, 15 000 traits discriminants).

incel											
				Apprentissage				**Test (= 10 % *Incels*)**			
% Incels	**Algorithme**	**Vecteurs**	**Dim.**	**Exactit.**	**Précision**	**Rappel**	**Mesure-F**	**Exactit.**	**Précision**	**Rappel**	**Mesure-F**
20	LR	SBERT	384	89,61	79,42	64,88	71,41	92,69	63,39	63,55	**63,47**
20	SVM	SBERT	384	89,53	79,35	64,38	71,08	92,72	63,75	62,95	63,35
30	SVM	SBERT	384	87,01	82,14	72,46	**77,00**	91,09	54,16	70,70	61,33
30	LR	SBERT	384	86,80	82,18	71,50	76,47	90,86	53,25	70,45	60,65
30	LR	TF-IDF	15000	84,91	85,54	59,83	70,40	91,99	59,84	60,50	60,17

(a) *Performances de détection de la classe de données* incel *pour les 5 meilleurs modèles de classification en phase de test*

non-*incel*											
				Apprentissage				**Test (= 10 % *Incels*)**			
% Incels	**Algorithme**	**Vecteurs**	**Dim.**	**Exactit.**	**Précision**	**Rappel**	**Mesure-F**	**Exactit.**	**Précision**	**Rappel**	**Mesure-F**
20	RF	SBERT	384	88,15	88,44	98,00	92,97	92,87	93,93	98,44	**96,13**
20	SVM	SBERT	384	89,53	91,50	95,81	93,60	92,72	95,89	96,02	95,96
20	LR	SBERT	384	89,61	91,60	95,80	**93,65**	92,69	95,95	95,92	95,94
30	RF	SBERT	384	84,53	84,76	94,96	89,57	92,16	95,05	96,30	95,67
30	SVM	SBERT	384	87,01	88,76	93,25	90,95	91,09	96,63	93,35	94,96

(b) *Performances de détection de la classe de données non-*incel *pour les 5 meilleurs modèles de classification en phase de test*

incel* + non-*incel											
				Apprentissage				**Test (= 10 % *incels*)**			
% Incels	**Algorithme**	**Vecteurs**	**Dim.**	**Exactit.**	**Précision**	**Rappel**	**Mesure-F**	**Exactit.**	**Précision**	**Rappel**	**Mesure-F**
20	LR	SBERT	384	89,51	79,22	64,50	82,35	92,69	79,67	79,74	**79,70**
20	SVM	SBERT	384	89,63	79,80	64,50	82,50	92,72	79,82	79,49	79,65
30	SVM	SBERT	384	87,05	82,50	72,15	83,98	91,09	75,39	82,02	78,15
30	LR	SBERT	384	86,89	81,95	72,21	83,82	90,86	74,92	81,79	77,74
40	SVM	SBERT	384	85,38	83,95	78,47	**84,60**	88,71	71,67	83,39	75,55

(c) *Performances macro (moyennes arithmétiques des métriques de chaque classe) pour les 5 meilleurs modèles de classification en phase de test*

Tableau 6. *Métriques de performance des 5 meilleurs modèles de classification en phase de test, pour les classes* incel ***(a)*** *et non-*incel ***(b)*** *respectivement. Les performances macro (moyenne pour les deux classes) sont présentées en (c). Les données sont triées par mesure-F en phase de test, les performances en phase d'apprentissage sont indiquées à titre indicatif. Les valeurs maximales (mesure-F) sont indiquées en gras. MNB = Multinomial Naive Bayes, LR = Logistic Regression, SVM = Support Vector Machine*

L'importance relative de chaque trait pour la fonction de décision du modèle est représentée par son coefficient de régression logistique. Ainsi, les traits associés à un plus haut coefficient pour une classe donnée contribuent plus fortement à la probabilité qu'un commentaire soit catégorisé dans cette classe. Ces coefficients sont accessibles par un attribut du modèle de régression logistique de *scikit-learn* (`model.coef`).

incel		**non-*incel***	
Trait	**Coefficient**	**Trait**	**Coefficient**
incel	6,5982	team	3,5702
chad	6,5319	kink	3,0854
woman	5,7303	player	2,8360
ugly	5,7011	host	2,7403
incels	5,5302	character	2,5176
normies	4,6894	use	2,4960
alone	4,6703	appreciated	2,4724
virgin	4,5697	trade	2,4536
loneliness	4,5553	season	2,4162
relationship	4,4181	item	2,4016
attractive	4,3775	using	2,3349
normie	4,2995	system	2,2578
social	4,2354	issue	2,2470
life	4,2053	server	2,2433
cope	4,0723	advance	2,2372
personality	4,0205	running	2,2040
dating	3,9777	horny	2,1822
girl	3,9139	version	2,1594
hobby	3,8762	suggestion	2,1390
lonely	3,7593	killed	2,1345

TABLEAU 7. *Traits prédicitfs des classes* incel *et non*-incel

Les termes prédictifs de la classe *incel* obtiennent de manière générale de plus hauts coefficients de régression que ceux associés à la classe de données non-*incel*. Malgré les faibles métriques de performance observées pour la classe *incel* individuellement, ces coefficients suggèrent l'existence d'un vocabulaire fortement relié à cette classe pour la fonction de décision du modèle. Ces termes incluent un ensemble d'expressions dénotant la subculture *incel* ou témoignant de la vision du monde qui y est associée (p. ex. *incel* = 6,5982, *chad* = 6,5319, *normies* = 4,6894). Un certain nombre de ces termes font référence aux femmes (p. ex. *women* = 5,7303, *girl* = 3,9138), tandis que d'autres relèvent de la solitude associée au célibat involontaire (p. ex. *alone* = 4,6703, *loneliness* = 4,5553, *lonely* = 3,7593). Certains traits reflètent finalement l'importance accordée à l'apparence physique dans ces communautés (p. ex. *ugly* = 5,7011, *attractive* = 4,3775) ou dénotent d'aspects relatifs aux relations intimes (p. ex. *virgin* = 4,5697, *relationship* = 4,4181, *dating* = 3,9777).

Ces traits sont représentatifs de la sous-culture *incel* dans une perspective générale. Il convient ici de rappeler que nous considérons dans cette étude l'ensemble des *subreddits incels* comme formant une communauté homogène, ce qui est attesté par le fait que plusieurs utilisateurs sont fréquemment abonnés aux mêmes subreddits (Ribeiro *et al.*, 2021). Une analyse plus fine pourrait mettre en lumière les thématiques de discussion de *subreddits* spécifiques, par exemple *r/gymscels*, dédié au fitness, ou *r/shortcels*, dédié aux difficultés relationnelles associées à la grandeur physique chez les *incels*.

En contrepartie, plusieurs des traits présentant les plus hauts coefficients pour la classe de données non-*incels* relèvent davantage de l'usage général de la langue sur Reddit sans évoquer de thématique particulière (p. ex. *use* = 2,4960, *appreciated* = 2,4724, *item* = 2,4016, *suggestion* = 2,1390). D'autres de ces termes pourraient relever de thématiques populaires sur la plateforme telles que le sport (p. ex. *team* = 3,5702, *player* = 2,8360, *trade* = 2,4536, *season* = 2,4162) ou la sexualité (p. ex. *kink* = 3,0854, *horny* = 2,1822). Néanmoins, les valeurs plus faibles associées aux coefficients de cette classe par rapport à la classe *incel* suggèrent que ces caractéristiques sont moins discriminantes pour la fonction de décision.

5. Conclusion

Cette étude évalué différents modèles de classification supervisée pour détecter le discours des *incels* sur la plateforme Reddit. Les expérimentations menées ont fait varier le ratio de la classe à détecter dans les données d'apprentissage, l'approche de vectorisation employée, le nombre de traits discriminants retenus par les modèles ainsi que l'algorithme de classification utilisé. Nos résultats indiquent de meilleures performances avec 20 % de données *incel* dans les données d'apprentissage pour des vecteurs SBERT et un modèle de régression logistique. Le système le plus performant obtient une mesure-F globale de 82,35. Les résultats obtenus sur les données test, lesquelles contiennent un ratio de 10 % de données *incel*, sont toutefois légèrement inférieurs (mesure-F globale de 79,70). Cela est possiblement explicable par le débalancement des classes à prédire, mais également par la manière dont les données sont annotées (en assignant une catégorie à des *subreddits* entiers) lors de la constitution des corpus. Un examen plus approfondi des erreurs de classification de nos modèles permettrait de faire la lumière sur ces résultats. Les performances obtenues sont cependant encourageantes. Notre démarche comparant rigoureusement différents modèles répond aux critiques d'ordre méthodologique parfois soulevées dans ce type de tâches. En nous inspirant des travaux de Klein et Golbeck (2024), nous entendons poursuivre cette recherche en explorant l'évolution des patrons langagiers les plus pertinents pour la détection de ce type de discours, en évaluant les performances d'autres méthodes de classification adaptées aux particularités de notre corpus.

6. Remerciements

Les auteurs remercient les évaluateurs dont les commentaires ont grandement contribué à la qualité de cet article. Nous remercions également Isabelle Fontaine et Audrée Frappier pour leurs contributions aux phases antérieures du projet.

7. Bibliographie

Abubakar H. D., Umar M., Bakale M. A., « Sentiment classification : Review of text vectorization methods : Bag of words, Tf-Idf, Word2vec and Doc2vec », *SLU Journal of Science and Technology*, vol. 4, n^{o} 1 & 2, p. 27-33, 2022.

Almerekhi H., Jansen S. b. B. J., Kwak c.-s. b. H., « Investigating toxicity across multiple Reddit communities, users, and moderators », *Companion proceedings of the web conference 2020*, p. 294-298, 2020.

Arango A., Pérez J., Poblete B., « Hate speech detection is not as easy as you may think : A closer look at model validation (extended version) », *Information Systems*, vol. 105, p. 1-11, 2022.

Associated Press, « Le suspect d'une tuerie en Californie est le fils d'un réalisateur de Hollywood », *Radio-Canada*, May, 2014.

Azmy W. M., Moulahi B., Bringay S., Azé J., Servajean M., « Lirmm@ deft-2018–modèle de classification de la vectorisation des documents », *Actes de DEFT, Rennes, France*, 2018.

Baumgartner J., Zannettou S., Keegan B., Squire M., Blackburn J., « The Pushshift Reddit Dataset », *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, p. 830-839, May, 2020.

Benmoussa M., Chénier I., Keenan-Pelletier M., Mason R., Robichaud M., Tanguay L., Valiquet D., Walker J., « Résumé législatif du projet de loi C-63 : Loi édictant la Loi sur les préjudices en ligne, modifiant le Code criminel, la Loi canadienne sur les droits de la personne et la Loi concernant la déclaration obligatoire de la pornographie juvénile sur Internet par les personnes qui fournissent des services Internet et apportant des modifications corrélatives et connexes à d'autres lois », *Bibliothèque du Parlement du Canada*, 2024.

Berglind T., Pelzer B., Kaati L., « Levels of hate in online environments », *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, p. 842-847, 2019.

Bird S., Klein E., Loper E., *Natural language processing with Python : analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.", 2009.

Breiman L., « Random Forests—Random Features [Technical Report 567] », 1999.

Chandrasekharan E., Samory M., Srinivasan A., Gilbert E., « The Bag of Communities : Identifying Abusive Behavior Online with Preexisting Internet Data », *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, Association for Computing Machinery, New York, NY, USA, p. 3175-3187, May, 2017.

Feldman R., Sanger J., *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, Cambridge, 2006.

Fersini E., Gasparini F., Rizzi G., Saibene A., Chulvi B., Rosso P., Lees A., Sorensen J., « SemEval-2022 Task 5 : Multimedia Automatic Misogyny Identification », *in* G. Emerson,

N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (eds), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Association for Computational Linguistics, Seattle, United States, p. 533-549, July, 2022.

Fersini E., Nozza D., Rosso P., « AMI @ EVALITA2020 : Automatic Misogyny Identification », *in* V. Basile, D. Croce, M. D. Maro, L. C. Passaro (eds), *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, vol. 2765 of *CEUR Workshop Proceedings*, CEUR, Online event, December, December, 2020.

Fersini Elisabetta e. a., *EVALITA Evaluation of NLP and Speech Tools for Italian*, Accademia University Press, chapter Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI), p. 59-66, 2018.

Frenda S., Ghanem B., Montes-y Gómez M., Rosso P., « Online Hate Speech against Women : Automatic Identification of Misogyny and Sexism on Twitter », *Journal of Intelligent & Fuzzy Systems*, vol. 36, nº 5, p. 4743-4752, January, 2019.

Gemelli S., Minnema G., « Manosphrames : exploring an Italian incel community through the lens of NLP and Frame Semantics », *in* P. Sommerauer, T. Caselli, M. Nissim, L. Remijnse, P. Vossen (eds), *Proceedings of the First Workshop on Reference, Framing, and Perspective @ LREC-COLING 2024*, ELRA and ICCL, Torino, Italia, p. 28-39, May, 2024.

Gothard K., Dewhurst D. R., Minot J. R., Adams J. L., Danforth C. M., Dodds P. S., « The incel lexicon : Deciphering the emergent cryptolect of a global misogynistic community », *arXiv [cs]*, May, 2021.

gouvernement du Canada, « Loi édictant la Loi sur les préjudices en ligne, modifiant le Code criminel, la Loi canadienne sur les droits de la personne et la Loi concernant la déclaration obligatoire de la pornographie juvénile sur Internet par les personnes qui fournissent des services Internet et apportant des modifications corrélatives et connexes à d'autres lois », , Projet de loi no C-63 (dépôt et 1re lecture – 26 février 2024), 1e sess., 44e légis., 2024.

Hajarian M., Khanbabaloo Z., « Toward Stopping Incel Rebellion : Detecting Incels in Social Media Using Sentiment Analysis », *2021 7th International Conference on Web Research (ICWR)*, p. 169-174, May, 2021.

Halpin M., « Weaponized Subordination : How Incels Discredit Themselves to Degrade Women », *Gender & Society*, vol. 36, nº 6, p. 813-837, December, 2022.

Halpin M., Preston K., Lockyer D., Maguire F., « A solider and a victim : Masculinity, violence, and incels celebration of December 6th », *Canadian Review of Sociology/Revue canadienne de sociologie*, vol. 61, p. cars.12460, January, 2024.

Hauser C., « Reddit Bans 'Incel' Group for Inciting Violence Against Women », *The New York Times*, November, 2017.

Hornby A. S., « Incel », *Oxford Advanced Learner's Dictionary*, 2020.

Jaki S., Smedt T. D., Gwóźdź M., Panchal R., Rossa A., Pauw G. D., « Online hatred of women in the Incels.me forum : Linguistic analysis and automatic detection », *Journal of Language Aggression and Conflict*, vol. 7, nº 2, p. 240-268, November, 2019.

Jamshidian M., « Evaluation of Text Transformers for Classifying Sentiment of Reviews by Using TF-IDF, BERT (word embedding), SBERT (sentence embedding) with Support Vector Machine Evaluation », 2023.

Jurafsky D., Martin J. H., *Speech and language processing*, Stanford Univ, 2019.

Kirk H., Yin W., Vidgen B., Röttger P., « SemEval-2023 Task 10 : Explainable Detection of Online Sexism », *in* A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (eds), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Association for Computational Linguistics, Toronto, Canada, p. 2193-2210, July, 2023.

Klein E., Golbeck J., « A Lexicon for Studying Radicalization in Incel Communities », *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, Association for Computing Machinery, New York, NY, USA, p. 262–267, 2024.

Ling C. X., Sheng V. S., « Class Imbalance Problem », *in* C. Sammut, G. I. Webb (eds), *Encyclopedia of Machine Learning*, Springer, Boston, MA, p. 171-171, 2010.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Higher Education from Cambridge University Press, July, 2008.

.ME, « The suspension of incels.me », *.ME blog*, November, 2018.

Meier M. L., Sharp K., « Death to Chad and Stacy : Incels and anti-fandom as group identity », *International Journal of Cultural Studies*, vol. 27, p. 349-367, January, 2024.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient Estimation of Word Representations in Vector Space », *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

Morales-Castro J. C., Hernandez-Rayas A., Ruiz-Pinales J., Guzmán-Cabrera R., « Automatic Identification of Misogynistic Sentiments on Social Networks », *Journal of Social Researches*, vol. 9, n° 23, p. 10-18, 2023.

Muralikumar M. D., Yang Y. S., McDonald D. W., « A human-centered evaluation of a toxicity detection api : Testing transferability and unpacking latent attributes », *ACM Transactions on Social Computing*, vol. 6, n° 1-2, p. 1-38, 2023.

Muti A., Ruggeri F., Khatib K. A., Barrón-Cedeño A., Caselli T., « Language is Scary when Over-Analyzed : Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts », *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, p. 21091-21107, November, 2024.

Nadeau J.-P., « Attaque au camion-bélier : Alek Minassian condamné à 25 ans minimum », *Radio-Canada*, June, 2022.

Pal M., « Random forest classifier for remote sensing classification », *International journal of remote sensing*, vol. 26, n° 1, p. 217-222, 2005.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., « Scikit-learn : Machine Learning in Python », *Journal of Machine Learning Research*, vol. 12, p. 2825-2830, 2011.

Pelzer B., Kaati L., Cohen K., Fernquist J., « Toxic language in online incel communities », *SN Social Sciences*, vol. 1, n° 8, p. 213, August, 2021.

Plaza L., Carrillo-de Albornoz J., Amigó E., Gonzalo J., Morante R., Rosso P., Spina D., Chulvi B., Maeso A., Ruiz V., « EXIST 2024 : sEXism Identification in Social neTworks and Memes », *in* N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (eds), *Advances in Information Retrieval*, Springer, Cham, p. 498-504, 2024.

Plaza L., Carrillo-de Albornoz J., Morante R., Amigó E., Gonzalo J., Spina D., Rosso P., « Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Charac-

terization », *in* A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (eds), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer, Cham, p. 316-342, 2023.

Ramos J., « Using tf-idf to determine word relevance in document queries », *Proceedings of the first instructional conference on machine learning*, vol. 242, Citeseer, p. 29-48, 2003.

Reimers N., Gurevych I., « Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks », *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 11, 2019.

Ribeiro M. H., Blackburn J., Bradlyn B., Cristofaro E. D., Stringhini G., Long S., Greenberg S., Zannettou S., « Dataset for : The Evolution of the Manosphere Across the Web », *Zenodo*, August, 2020.

Ribeiro M. H., Blackburn J., Bradlyn B., Cristofaro E. D., Stringhini G., Long S., Greenberg S., Zannettou S., « The Evolution of the Manosphere across the Web », *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, p. 196-207, May, 2021.

Rodríguez-Sanchez F. J., de Albornoz J. C., Plaza L., Gonzalo J., Rosso P., Comet M., Donoso T., « Overview of EXIST 2021 : sEXism Identification in Social neTworks », *Proces. del Leng. Natural*, vol. 67, p. 195-207, 2021.

Rodríguez-Sánchez F., Carrillo-de Albornoz J., Plaza L., Mendieta-Aragón A., Marco-Remón G., Makeienko M., Plaza M., Gonzalo J., Spina D., Rosso P., « Overview of EXIST 2022 : sEXism Identification in Social neTworks », *Procesamiento de Lenguaje Natural*, vol. 69, p. 229-240, 09, 2022.

RSR R. d. S. l. R., Le phénomène incel : exploration des problèmes internes et externes touchant les célibataires involontaires, Technical report, Commission européenne, 2021.

Saha P., Mathew B., Goyal P., Mukherjee A., « Hateminers : Detecting Hate speech against Women », *arXiv :1812.06700*, December, 2018.

Sheppard B., Richter A., Cohen A., Smith E., Kneese T., Pelletier C., Baldini I., Dong Y., « Biasly : An Expert-Annotated Dataset for Subtle Misogyny Detection and Mitigation », *in* L.-W. Ku, A. Martins, V. Srikumar (eds), *Findings of the Association for Computational Linguistics : ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand, p. 427-452, August, 2024.

Shetty A., « Accused in University of Waterloo stabbings charged with attempted murder, added to other counts », *CBC News*, 2023.

Tranchese A., Sugiura L., « "I Don't Hate All Women, Just Those Stuck-Up Bitches" : How Incels and Mainstream Pornography Speak the Same Extreme Language of Misogyny », *Violence Against Women*, vol. 27, n° 14, p. 2709-2734, November, 2021.

Truşcă M. M., « Efficiency of SVM classifier with Word2Vec and Doc2Vec models », *Proceedings of the International Conference on Applied Statistics*, p. 496-503, 2019.

Wang Y., Zhou Z., Jin S., Liu D., Lu M., « Comparisons and selections of features and classifiers for short text classification », *Iop conference series : Materials science and engineering*, vol. 261, IOP Publishing, 2017.

Yoder M., Perry C., Brown D., Carley K., Pruden M., « Identity Construction in a Misogynist Incels Forum », *The 7th Workshop on Online Abuse and Harms (WOAH)*, Association for Computational Linguistics, Toronto, Canada, p. 1-13, 2023.

Zhao Y., Chen Y., *Data Mining Applications with R*, Elsevier, 2014.

Automated Speech Act Classification in Offensive German Language Tweets

Melina Plakidis[1,2] — Elena Leitner[1] — Georg Rehm[1,2]

[1] *DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany*
[2] *Humboldt-Universität zu Berlin, Dorotheenstraße 24, 10117 Berlin, Germany*

ABSTRACT. *One under-researched avenue for hate speech and offensive language detection is the integration of knowledge related to speech acts. In previous work, we investigated whether the distribution of speech acts differs across offensive and non-offensive language. Our findings revealed supporting evidence. In the present article, we fine-tune several BERT models and LLMs on the German Speech Acts Dataset. Our goals are two-fold: we want to contribute relevant research results to speech act theory by developing and providing models that detect and classify speech acts in documents or other types of discourse such as tweets. We hope that detected speech acts can be used in a beneficial way as additional features in the detection of hate speech. Our best-performing model achieves a macro-averaged F_1-score of 68.68%.*

RÉSUMÉ. *En matière de détection de discours de haine et de langage offensant, l'intégration des connaissances sur les actes de langage représente une voie de recherche encore peu explorée. Dans nos précédents travaux, nous avons analysé si la répartition des actes de langage variait selon que les propos étaient injurieux ou non. Les résultats que nous avons obtenus ont confirmé cette hypothèse. Dans le présent article, pour affiner plusieurs modèles BERT et LLM, nous avons utilisé le jeu de données des actes de langage allemands. Nous poursuivons un double objectif. Nous souhaitons fournir des résultats pertinents à la théorie des actes de langage en développant et en mettant à disposition des modèles capables de mettre en œuvre la détection et la classification d'actes de langage dans des documents ou d'autres types de propos, des tweets notamment. Nous espérons que les actes de langage détectés pourront servir de caractéristiques supplémentaires et bénéficier à la détection des discours de haine. Notre modèle le plus performant atteint un score F_1 macro-moyenné de 68,68 %.*

KEYWORDS: *hate speech detection, offensive language, speech acts.*

MOTS-CLÉS : *détection de discours haineux, langage offensant, actes de langage.*

1. Introduction

Hate speech and the use of offensive language have become a pervasive phenomenon online. Wiegand *et al.* (2018) define offensive language as "hurtful, derogatory or obscene comments made by one person to another person". A study conducted by Bilewicz and Soral (2020) shows that increased exposure to hate speech can lead to desensitisation and thus decrease people's ability to identify hate speech. Furthermore, encountering derogatory language targeting immigrants and minority groups can contribute to political radicalisation (Bilewicz and Soral, 2020). The vast amount of newly created daily posts, messages and other types of content makes the manual handling of offensive language impossible. Automatic processes are needed, but even with the recent emphasis on hate speech detection (Poletto *et al.*, 2021), there are still various challenges when it comes to detecting hate speech automatically. [1]

In previous work (Plakidis and Rehm, 2022), we took a closer look at pragmatic properties of offensive language, i. e., by combining the field of *speech act theory* with *hate speech detection*, aiming to enrich text data with pragmatic characteristics and exploring possible differences between offensive and non-offensive language. We created a dataset of offensive and non-offensive German tweets and annotated them for coarse- and fine-grained speech acts. Our findings suggest a difference in the distribution of speech acts between offensive and non-offensive tweets as well as between different offensiveness categories. A similar observation made by previous studies also shows that speech acts vary depending on the discussed topic (Zhang *et al.*, 2011; Vosoughi and Roy, 2016; Laurenti *et al.*, 2022).

Building on our previous work, in this article we experiment with state-of-the-art encoder and decoder models to train speech act classifiers on our German Speech Act Dataset and investigate which models are better suited considering different levels of speech acts. For encoders, we implement various fine-tuning strategies such as default, hyperparameter search and few-shot classification to improve performance of selected models. For decoders, we focus on the parameter-efficient fine-tuning method instead of exploring different prompting approaches such as zero-shot or few-shot. This method allows to optimise performance of a model by retraining with specific data (Wang *et al.*, 2024). In addition, we provide an error analysis to identify the core issues of our best-performing classifiers.

The remainder of this article is structured as follows. Section 2 presents related work and Section 3 describes the dataset. Section 4 introduces our experiments on training a speech act classifier, providing information on the approach as well as on the evaluation. Section 5 reports on our results. Finally, Section 6 concludes the article.

1. In the wider field of research, a variety of similar terms are used such as "abusive" (Nobata *et al.*, 2016), "toxic" (Risch *et al.*, 2021) or "offensive" (Wiegand *et al.*, 2018; Zampieri *et al.*, 2019) language. We use *hate speech* and *offensive language* synonymously.

2. Related Work

The research dedicated to speech acts used in hate speech is still limited. Nevertheless, there are some works dealing with the combination of speech acts and hate speech which we will present in the following.

Oktaviani and Nur (2022) analyse a twitter account using Searle's speech act theory. They use an exploratory, qualitative approach and assign a hate speech label as well as a speech act label for each tweet. They observe the occurrence of *assertives*, *directives* and *expressives*, stating that *directives* appear most often in the data. Nevertheless, it is not clear which speech act classes occur most often in which hate speech category and how both categories relate to each other. Similarly to the study by Oktaviani and Nur (2022), Mubarok *et al.* (2024) also analyse comments of a selected social media account. They find 11 *directives*, 15 *expressives* and five *assertive* speech acts in a small sample of 31 abusive comments. In a study by Dhayef and Ali (2020), seven newspaper article extracts are selected from a Rwandan newspaper which are expected to contain racial hate speech. They examine them using Searle's five speech act classes (Searle, 1979) and additionally include Searle's distinction between direct and indirect speech acts. They present a qualitative as well as quantitative analysis and put forward three hypotheses for which their results seem to provide confirming evidence. First, they expect the excerpts to contain a high quantity of *directives*. Second, they assume that the excerpts will contain more indirect than direct speech acts and third, they estimate that direct *assertives* and indirect *expressives* are the most dominant speech acts in the excerpts. However, Dhayef and Ali (2020) provide only seven short extracts for their pragmatic analysis and they do not state what constitutes an utterance or how they intend to segment the extracts. A more recent study by Ollagnier (2024) introduces the dataset CyberAgressionAdo-V2 on cyberbullying in French multiparty chats, which, inter alia, is annotated with pragmatic aspects. These pragmatic aspects are located on the discursive level which comprises eight distinct categories such as *gaslighting*, *defend*, and *attack*. Similar to speech acts, these categories denote the intention that the user attempts to convey with his message. In addition, Ollagnier (2024) also considers the context in which these messages occur. The annotations are not restricted to aggressive messages, but comprise all messages regardless of their level of aggressiveness. The findings show that most of the time, bullies and their supporters intentionally send messages that attack the victims, while victims and their supporters predominantly issue either neutral or defensive messages.

Several attempts have been made to classify speech acts automatically and their annotation taxonomies have often been influenced to a great extent by Austin (1962) and Searle (1979). Compagno *et al.* (2018) represent one of these works. Their hierarchically structured speech act taxonomy is based on Searle's five classes which they applied to a Reddit corpus dealing with autoimmune diseases. In total, their fine-grained classification consists of 17 speech acts. Other approaches influenced by Searle (1979) and Austin (1962) include Vosoughi and Roy (2016) and Zhang *et al.* (2011). Zhang *et al.* (2011), for instance, aim to classify tweets into one of five speech act classes: *statement*, *question*, *comment*, *suggestion* and *miscellaneous*. They achieve an F_1-

score of almost 70.00% on average using a Support Vector Machine classifier with a linear kernel in addition to word- and character-based features. Similarly, Vosoughi and Roy (2016) classify tweets into one of six categories: *assertion*, *recommendation*, *expression*, *question*, *request*, and, again, *miscellaneous*. With the use of semantic and syntactic features in combination with a Logistic Regression classifier, they manage to achieve an average F_1-score of 70.00%. Another approach with the aim of annotating speech acts automatically, currently however semi-automatically, is presented by Weisser (2018). He introduces the Dialogue Annotation and Research Tool (DART), which is publicly available. Its current version 3.0 [2] classifies dialogue using various features including syntactic categories and speech acts. The proposed speech act tags in the latest version of the DART taxonomy [3] result in a total of 162 speech act tags. In a further study by Laurenti *et al.* (2022), French tweets posted during crisis events were annotated for speech acts on both tweet level and a more fine-grained segment level. Their speech act annotations on the level of tweets comprise five classes, namely *assertives*, *jussives*, *subjectives*, *interrogatives* and *other*. Additionally, their segment level annotations consist of eight speech act classes. Their findings indicate a correlation between urgent messages during crisis events and higher occurrence of *proper assertions* (assertions not relying on a third-party source). Additionally, they observe a higher occurrence of *subjective* speech acts in non-urgent tweets. Their best-performing model for tweet-level annotations with four classes is CamemBERT (Martin *et al.*, 2020) with focal loss (Lin *et al.*, 2020) and extra-features and achieves an F_1-score of 73.55%. Building on the work by Laurenti *et al.* (2022), Benamara *et al.* (2024) further extend the dataset by Laurenti *et al.* (2022) to about 13,000 French tweets. Their experiments show that FlauBERT (Le *et al.*, 2019) pre-trained on crisis domain tweets (Kozlowski *et al.*, 2020) with focal loss and additional features is the best performing model (F_1: 67.37%) for predicting the five speech act classes on tweet level. Their best-performing classifier for the eight fine-grained speech act classes is FlauBERT base with cross-entropy loss in a multi-label setting achieves an F_1 of 87.80%.

3. Dataset

Our German Speech Act Dataset (Plakidis and Rehm, 2022) comprises 600 tweets of the dataset created for task two of the 2019 GermEval Shared Task on the Identification of Offensive Language (Struß *et al.*, 2019). We chose Twitter (now: X) as the main source of data because it is the most frequently used platform in the field of hate speech detection (Poletto *et al.*, 2021). The 600 tweets were selected with the aim to analyse whether the speech act distribution differs across different offensive language classes. For each of the six offensive language classes established by Struß *et al.* (2019), i. e., *implicit*, *explicit*, *profanity*, *insult*, *abuse* and *other*, we randomly selected 100 tweets.

2. `http://martinweisser.org/publications/DART_manual_v3.0.pdf`.
3. `http://martinweisser.org/DART_scheme.html`.

According to Struß *et al.* (2019), these classes can be described as follows. In contrast to being *explicitly* offensive, offensive language counts as being *implicit* when the reader needs to infer that the tweet is offensive, as the offense is only implied. Moreover, implicit offensive language also entails using figurative language (e. g., sarcasm or irony). Tweets are labeled as *profanity* if they consist of profane words like swearwords but lack abusive language as well as insults. If they also contain abusive language or insults, they either belong to the class *insult* or *abuse*. While the class *insult* only contains offensive language targeting individuals, the class *abuse* contains tweets that target group representatives, assigning them universally negative traits.

We extended the dataset by adding speech act annotations which we included for both fine-grained as well as coarse-grained speech acts. In contrast to Struß *et al.* (2019), these annotations relate to the sentence level and not to the tweet level. Thus, the unit for a speech act is the sentence. [4] However, Twitter users often do not use punctuation properly in their tweets. During the annotation process, in order to clarify how to segment tweets into sentences, rules had to be established which are specified in our previous work (Plakidis and Rehm, 2022).

Table 1 shows the results of our speech act annotations, revealing distinct differences between offensive and non-offensive language in terms of speech act usage. Offensive language generally features more *expressives* and fewer *assertives* than non-offensive language. As tweets consisting of implicit offensive language tend to lack emotional expression, thus increasing the use of *assertives* and decreasing the use of *expressives*, this difference is most pronounced when comparing explicitly offensive with implicitly offensive tweets. The results indicate that offensive and non-offensive language differ in how speech acts are distributed.

In the following, we present our speech act annotation scheme which is based on Compagno *et al.* (2018) and Searle (1979). [5] In addition, we also provide information on the inter-annotator agreement in Section 3.3. The German Speech Act Dataset is publicly available under a CC-BY-4.0 license and can be accessed at GitHub [6].

3.1. *Coarse-Grained Speech Act Level*

The coarse-grained speech act level includes six classes: *assertive*, *directive*, *expressive*, *commissive*, *unsure* and *other*. Quoting Searle (1979), the first four speech acts can be defined as follows: "We tell people how things are (Assertives), we try

4. Exceptional cases include user mentions, hashtags and emojis. The difference between tweet and sentence level is best illustrated in examples from the dataset in Sections 3.1 and 3.2.

5. Building upon Weisser (2018), our annotation scheme contains a syntactical level and a speech act level. In this article, we focus on the speech act level exclusively.

6. The most recent Version 1.1 of the dataset contains several bugfixes: `https://github.com/MelinaPl/speech-act-analysis/blob/main/version_1-1_changes.md`.

7. The class *accept* has been constructed and does not represent a real instance because the category did not occur in the data at all.

Table 1. *Frequency of coarse-grained and fine-grained speech acts in offensive language categories. Note that speech acts were annotated on sentence level, while offensive language categories were annotated on tweet level.*

	Offensive		**Other**		**Implicit**		**Explicit**		**Abuse**		**Profanity**		**Insult**		**Total**	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Assertive	557	34.3	126	37.7	116	41.6	85	28.9	118	32.2	111	33.8	127	35.5	683	34.9
Assert	473	29.1	117	35.0	97	34.8	73	24.8	99	27.0	93	28.4	111	31.0	590	30.1
Sustain	11	0.7	2	0.6	2	0.7	0	0.0	5	1.4	1	0.3	3	0.8	13	0.7
Guess	26	1.6	1	0.3	9	3.2	2	0.7	3	0.8	7	2.1	5	1.4	27	1.4
Predict	32	2.0	2	0.6	6	2.2	8	2.7	6	1.6	5	1.5	7	2.0	34	1.7
Agree	11	0.7	2	0.6	2	0.7	1	0.3	4	1.1	4	1.2	0	0.0	13	0.7
Disagree	4	0.2	2	0.6	0	0.0	1	0.3	1	0.3	1	0.3	1	0.3	6	0.3
Expressive	353	21.7	47	14.1	44	15.8	76	25.9	78	21.3	73	22.3	82	22.9	400	20.4
Rejoice	14	0.9	3	0.9	1	0.4	6	2.0	1	0.3	4	1.2	2	0.6	17	0.9
Complain	240	14.8	17	5.1	37	13.3	55	18.7	39	10.7	46	14.0	63	17.6	257	13.1
Wish	10	0.6	1	0.4	0	0.0	3	1.0	3	0.8	4	1.2	0	0.0	11	0.6
Apologize	0	0.0	1	0.3	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.0
Thank	4	0.2	4	1.2	0	0.0	0	0.0	1	0.3	2	0.6	1	0.3	8	0.4
expressEmoji	85	5.2	21	6.3	6	2.2	12	4.1	34	9.3	17	5.2	16	4.5	106	5.4
Commissive	17	1.0	3	0.9	0	0.0	3	1.0	1	0.3	12	3.7	1	0.3	20	1.0
Engage	11	0.7	2	0.6	0	0.0	0	0.0	0	0.0	11	3.4	0	0.0	13	0.7
Accept[7]	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Refuse	1	0.0	0	0.0	0	0.0	1	0.3	0	0.0	0	0.0	0	0.0	1	0.0
Threat	5	0.3	1	0.3	0	0.0	2	0.7	1	0.3	1	0.3	1	0.3	6	0.3
Directive	524	32.2	109	32.6	99	35.5	100	34.0	131	35.8	85	25.9	109	30.4	633	32.3
Request	130	8.0	33	9.9	23	8.2	23	7.8	36	9.8	24	7.3	24	6.7	163	8.3
Require	66	4.1	12	3.6	7	2.5	17	5.8	13	3.6	13	4.0	16	4.5	78	4.0
Suggest	15	0.9	1	0.3	2	0.7	1	0.3	5	1.4	3	0.9	4	1.1	16	0.8
Greet	1	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.3	1	0.0
Address	312	19.2	63	18.9	67	24.0	59	20.1	77	21.0	45	13.7	64	17.9	375	19.1
Unsure	113	7.0	37	11.1	18	6.5	15	5.1	30	8.2	35	10.7	15	4.2	150	7.7
Other	61	3.8	12	3.6	2	0.7	15	5.1	8	2.2	12	3.7	24	6.7	73	3.7
Total	1,625	100.0	334	100.0	279	100.0	294	100.0	366	100.0	328	100.0	358	100.0	1,959	100.0

to get them to do things (Directives), we commit ourselves to doing things (Commissives), [and] we express our feelings and attitudes (Expressives)" (p. viii). The categories *assertive*, *directive* and *expressive* are shown in Examples (1, 2) and *commissive* in (3). The category *unsure* is used in cases where an utterance in a tweet cannot be classified due to missing or insufficient context as in Example (4). Finally, the category *other* in (2) is used for all speech acts not represented in this annotation scheme. The examples below reflect coarse- and fine-grained labels which are separated using "|".

(1) *[@Alexplantsatree @griechenwoos2 @Die_Gruenen]*$_{directive|address}$ *[Schon die*
@Alexplantsatree @griechenwoos2 @Die_Gruenen already the
Worter "schmutzige Technologien" implizieren, dass der
words dirty technologies imply that the
Automobilbau eine Technologie ist, die entsorgt werden
automobile.manufacturing a technology is that disposed.of be
musse.]$_{assertive|assert}$ *[Nur leider ist die Elektromobilitat keine adaquate*
must now unfortunately is the electric.mobility no adequate

Alternative zum Auto mit Verbrennungsmotor und Kernenergie wurde aus
alternative to cars with combustion.engines and nuclear.energy was of
reinem Opportunismus aufgegeben.]$_{expressive\backslash complain}$
pure opportunism up.give
'@Alexplantsatree @griechenwoos2 @Die_Gruenen Already the words "dirty technologies" imply that automobile manufacturing is a technology that must be disposed of. Unfortunately, electric mobility is not an adequate alternative to cars with combustion engines and nuclear energy has been abandoned out of pure opportunism.'

(2) *[2/2]*$_{other}$ *[sollten wir nicht in Berlin und Brüssel stehen und die Banditen*
2/2 should we not in Berlin and Brussels stand and the bandits
aus ihren Ämtern jagen?]$_{directive\backslash request}$ *[Mit Schimp und Schande, geteert*
from their offices chase with disgrace and shame tarred
und gefedert?]$_{directive\backslash request}$ *[Suizid?]*$_{directive\backslash request}$ *[Unfassbar.]*$_{expressive\backslash complain}$
and feathered suicide unbelievable.
2/2 Shouldn't we stand in Berlin and Brussels and chase the bandits from their offices? With disgrace and shame, tarred and feathered? Suicide? Unbelievable.'

(3) *[@_denk_mal_]*$_{directive\backslash address}$ *[ES WIRD ZEIT, DIESE KRANKE ZU*
@_denk_mal_ it will time this sick.person to
WARNEN!]$_{commissive\backslash threat}$
warn
'@_denk_mal_ It will be time to warn this sick person!'

(4) *[@Snakecleaver @Metalwilli]*$_{directive\backslash address}$ *[OK.....!]*$_{unsure}$
@Snakecleaver @Metalwilli okay
'@Snakecleaver @Metalwilli OK....!'

3.2. *Fine-Grained Speech Act Level*

The fine-grained speech act level consists of 23 speech acts. We modified the taxonomy by Compagno *et al.* (2018) by adding the categories *predict*, *expressEmoji*, *threat*, *address* and *unsure* as well as by moving *greet* to *directives* and maintaining the distinction between the two classes *request* and *require*. Each fine-grained speech act has a corresponding coarse-grained speech act. However, the categories *unsure* and *other* remain the same on both levels. Several examples are shown in (5-9) and in the previous subsection in (1-4). [8]

(5)

8. Additional examples can be found in our repository: `https://github.com/MelinaPl/speech-act-analysis/blob/main/README.md`.

[Er geht mir ziemlich auf den Keks, aber wegen Vorstehendem habe
he goes me quite on the biscuit but because.of before.standing have
ich ihn noch nicht einfach geblockt!]$_{sustain}$
I him yet not simply blocked
'*He really gets on my nerves but because of the preceding I haven't blocked him yet.*'

(6) *[ich kotzt das so an,]*$_{complain}$ *[fragt die deutschen Staatsbürger,]*$_{require}$
me throw.up that so of ask the German citizens
[schätze 80% sind gegen den Migrationspackt]$_{guess}$ *[#Maischberger]*$_{other}$
estimate 80% are against the migration.pact #Maischberger
'I'm so sick of it, ask the German citizens, I estimate that 80% are against the migration pact.'

(7) *... [ich werde ihnen auch in den Hintern Kriechen so bald ich bei der*
... I will you also in the butt creep as soon.as I by the
Merkel raus bin.]$_{predict}$ *[Ich biete Ihnen gute Zusammenarbeit an..]*$_{engage}$ *...*
Merkel out am I offer you good cooperation on..
'I will kiss their asses as soon as I leave Merkel. I offer you good cooperation.'

Assertive speech acts comprise statements that *assert* something, statements sustained with arguments (*sustain* in (5)) as well as weaker forms of assertions (*guess* in (6) or *predict* in (7)). In addition, assertive speech acts can also be used to signal agreement (*agree*) or disagreement (*disagree*) with something or someone.

Expressive speech acts include statements about positive (*rejoice*) or negative (*complain* in (1, 2, 4, 6)) attitude towards someone or something, serve by wishing for something (*wish* in (8)), apologising to someone for something (*apologize*) or thanking someone (*thank*). Additionally, *expressEmoji* is used for an emoji or series of emojis.

Directive speech acts either *require* or *request* someone to do something, provide a suggestion about something (*suggest* in (9)) or are used to greet (*greet*) or address someone (*address* in (1, 2, 4)).

Commissive speech acts are utterances that either *engage* oneself to do something (7), *accept* or *refuse* something based on a previous utterance or are used to threaten someone (*threat* as in (3)).

(8) *[Schönen Freitag.]*$_{wish}$
beautiful Friday
'Have a nice Friday.'

(9) *[Die linke, deutsch/islamische #Bundesregierung kann den #korantreuen*
the left German/Islamic #federal.government can the #Koran.faithful
#Moslems #IS #Hamas doch gleich den Schlüssel zu Deutschland
#Muslims #IS #Hamas still immediately the key to Germany
überreichen.]$_{suggest}$
over.give
'The leftist, German/Islamic #federalgoverment may as well hand the #Koranfaithful #Muslims #IS #Hamas the key to Germany.'

3.3. *Inter-Annotator Agreement*

As our original dataset had only been annotated by one annotator, we decided to extend it by including two more annotators. Two of the annotators are authors of this paper, while the third annotator is a Master student with a background in linguistics. Currently, 200 of the 600 tweets have been annotated by two annotators (100 tweets by each of the two additional annotators). We pre-segment the tweets so that the annotators only have to choose the correct speech act labels. To compute the agreement between two annotators, we choose Cohen's κ (Cohen, 1968), resulting in an average κ score of 0.69 for coarse-grained speech acts and a κ of 0.66 for fine-grained speech acts. The values for both granularities indicate a substantial agreement. Similar values were achieved by Laurenti *et al.* (2022), who report a Cohen's κ of 0.62. Furthermore, Compagno *et al.* (2018) report a moderate to substantial agreement for all annotators with values between 0.57 and 0.87 for five coarse-grained speech act classes and between 0.48 and 0.73 for 18 fine-grained speech act classes. However, it should be noted that the authors computed the inter-annotator agreement using Fleiss' κ (Fleiss, 1971).

One of the greatest difficulties during the annotation process was distinguishing between *assertives* and *expressives* as it is often challenging to specify whether an utterance merely describes reality (= *assertive*) or expresses the speaker's feelings or attitude towards something (= *expressive*). Sometimes, both can be true at the same time, resulting in a rather subjective choice by the annotator. Similar observations were also made by Laurenti *et al.* (2022), who report issues distinguishing between *assertives* and *subjectives*, the latter class being comparable to *expressives*, and by Compagno *et al.* (2018), who report that their annotation results point to a continuity between *assertives* and *expressives*.

4. Experiments

The following section presents the experiments.[9] First, we describe which annotations are used for the training process and how we modify the classes with sparse

9. In addition to training several speech act classifiers, we also conducted an initial experiment to fine-tune an offensive language classifier with and without speech acts. These results have

data. Second, we provide information on the evaluation method and metric. Third, we present the selected models and fine-tuning strategies.

4.1. *Training Data*

For the experiments, we use both sets of annotations in the German Speech Act Dataset. The first version consists of data annotated for six coarse-grained speech act classes. The second version consists of fine-grained classes that have been modified. Due to rather sparse occurrences of a few fine-grained classes, only classes with ten or more instances are included. *Disagree*, *apologize*, *thank* and *greet* were combined in the new *excluded* class. As for the coarse-grained speech act *commissive*, we decided not to divide it into fine-grained classes due to sparse occurrences of its fine-grained classes. Thus, the number of fine-grained speech acts was reduced from 23 to 17.

4.2. *Evaluation Method and Metric*

Due to the dataset size and label distribution, we apply 5-fold cross-validation, a best-practice evaluation method. Sentences were shuffled and stratified in order to preserve the percentage of samples for each class in each fold and split. A train split contains 1,567 sentences (80%) and a validation split 392 sentences (20%). For the coarse- and fine-grained labels, we created individual splits. The mean number of instances across the 5-fold splits can be found in Table 2. As evaluation metrics, we use precision, recall and macro F_1, which calculate the unweighted mean between all labels.

4.3. *Models*

For training, we use state-of-the-art encoder and decoder models developed for German. As encoders, we selected three pre-trained BERT (Devlin *et al.*, 2018) models as many previous text classification experiments make use of Transformer-based architectures (Risch *et al.*, 2021) which have been shown to be more effective than other approaches (Struß *et al.*, 2019). We use the base versions of the cased [10] and uncased [11] pre-trained Digitale Bibliothek Münchener Digitalisierungszentrum (DBMDZ) BERT models. The two models were trained on Wikipedia, the EU Bookshop

not shown improvements when including speech act features: F_1 of 67.99% without speech acts and F_1 of 65.31% with speech acts. However, as offensive language classification is conducted on the tweet level and not on the sentence level, the experiments were carried out on a much smaller dataset. We plan to replicate this experiment on a significantly increased dataset in the future.

10. `https://huggingface.co/dbmdz/bert-base-german-cased`.

11. `https://huggingface.co/dbmdz/bert-base-german-uncased`.

	Train	Val	#
Assertive	546	137	683
Expressive	320	80	400
Commissive	16	4	20
Directive	506	127	633
Unsure	120	30	150
Other	59	14	73
Total	1,567	392	1,959

(a) *Coarse-grained annotations.*

	Train	Val	#
Assert	472	118	590
Sustain	10	3	13
Guess	22	5	27
Predict	27	7	34
Agree	10	3	13
Rejoice	14	3	17
Complain	206	51	257
Wish	9	2	11
Expressemoji	85	21	106
Commissive	16	4	20
Request	130	33	163
Require	62	16	78
Suggest	13	3	16
Address	300	75	375
Unsure	120	30	150
Other	58	15	73
Excluded	13	3	16
Total	1,567	392	1,959

(b) *Fine-grained annotations.*

Table 2. *Mean number of speech acts in 5-fold splits.*

corpus, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl. Additionally, we use Deepset's base version of the German BERT model called GBERT [12] (Chan *et al.*, 2020).

As decoders, we utilise Gemini 1.5 Flash [13] from Google AI, multilingual Llama 3.2 [14] from Meta (3B) as well as German Llama3 [15] (8B). Gemini is a transformer decoder model with 2M+ context and multimodal capabilities trained on a variety of multimodal and multilingual data (Gemini Team *et al.*, 2024). This model achieved good results as a text classifier (Wang *et al.*, 2024). Llama 3.2 (Llama Team *et al.*, 2024) is an auto-regressive language model trained on up to 9 trillion tokens from publicly available online data. German Llama3 is developed by the open-source research collective Disco Research that concentrates on the German language. This model is based on Meta's Llama3-8B and was pretrained on 65 billion tokens.

12. `https://huggingface.co/deepset/gbert-base`.
13. `https://ai.google.dev/gemini-api/docs/models/gemini`.
14. `https://huggingface.co/meta-llama/Llama-3.2-3B`.
15. `https://huggingface.co/DiscoResearch/Llama3-German-8B`.

4.4. *Fine-Tuning Strategy*

For each encoder model and granularity level (coarse-grained and fine-grained), we apply three methods: (i) default, (ii) bestrun with hyperparameter search, (iii) few-shot classification. The default models were fine-tuned with default hyperparameters which were the same for each model and granularity. We also performed a hyperparameter search on the first train and validation split using Ray Tune [16] (Liaw *et al.*, 2018). The goal was to maximise the macro F_1-score during 30 trials. After finding the best hyperparameters, we trained and evaluated a bestrun model on 5-folds. For few-shot classification, we used Fastfit. [17] This method utilises an approach that integrates batch contrastive learning and a token-level similarity score which provides accurate classification of semantically similar classes (Yehudai and Bandel, 2024).

For the decoder models, instead of prompting strategies such as few-shot prompting, we apply a supervised fine-tuning strategy to improve the model's performance as a speech act classifier. Each decoder model was fine-tuned in the same manner as encoder models, each on a full train set from 5-folds and evaluated on a corresponding validation set. For Gemini, we leverage the fine-tuning procedure with suggested hyperparameters described in its model tuning card. For both pretrained base Llama models, we performed parameter efficient fine-tuning (PEFT) (Mangrulkar *et al.*, 2022) using quantized low-rank adaptation (QLoRA) (Dettmers *et al.*, 2023).

Detailed information on the hyperparameters and results for each model, granularity and fine-tuning strategy (including results per class) as well as training scripts can be found in our GitHub Repository. [18]

5. Results

The following section presents the results of the experiments as well as a brief error analysis.

5.1. *Performance*

Table 3 presents the mean results during 5-fold cross-validation for the encoder models GBERT, $\text{BERT}^{\text{cased}}_{\text{german}}$ and $\text{BERT}^{\text{uncased}}_{\text{german}}$ as well as for the decoder models Gemini 1.5 Flash, Llama 3.2 and German Llama 3 across the two granularities and the different fine-tuning strategies. Regarding the encoder models, we can see that the results are improving based on the fine-tuning strategy, i. e., hyperparameter search is better than the default, and few-shot classification is better than hyperparameter search. The only exceptions are GBERT and $\text{BERT}^{\text{cased}}_{\text{german}}$ trained on coarse-grained labels; here,

16. `https://docs.ray.io/en/latest/tune/index.html`.
17. `https://github.com/IBM/fastfit`.
18. `https://github.com/elenanereiss/German-Speech-Act-Classification`.

	coarse-grained labels			*fine-grained labels*		
encoder	**GBERT**	**BERT**$^{\text{cased}}_{\text{german}}$	**BERT**$^{\text{uncased}}_{\text{german}}$	**GBERT**	**BERT**$^{\text{cased}}_{\text{german}}$	**BERT**$^{\text{uncased}}_{\text{german}}$
	Default					
precision	68.81	67.04	70.44	55.80	57.82	55.91
recall	65.62	64.33	66.21	48.39	50.70	49.88
F_1*-score*	66.51	65.05	<u>67.76</u>	50.14	<u>52.55</u>	51.84
	Hyperparameter search – Bestrun					
precision	69.44	70.19	65.80	63.18	58.74	57.20
recall	68.76	67.11	64.27	54.15	51.68	51.39
F_1*-score*	**68.68**	67.96	64.47	<u>56.37</u>	53.48	52.72
	Few-shot classification – Fastfit					
precision	73.97	70.07	72.39	67.68	63.58	62.46
recall	66.25	65.03	66.06	53.11	52.48	53.13
F_1*-score*	<u>68.45</u>	66.39	68.15	**57.04**	55.29	55.80
decoder	**Gemini**$^{\text{1.5 Flash}}$	**Llama3.2**$^{\text{3B}}$	**Llama3**$^{\text{8B}}_{\text{german}}$	**Gemini**$^{\text{1.5 Flash}}$	**Llama3.2**$^{\text{3B}}$	**Llama3**$^{\text{8B}}_{\text{german}}$
precision	33.42	64.97	62.69	45.06	39.68	40.62
recall	31.07	64.05	62.59	32.93	41.80	42.97
F_1*-score*	28.96	<u>62.56</u>	61.41	34.26	39.51	<u>39.88</u>

Table 3. *Mean precision, recall and macro F_1-score during 5-fold cross-validation. The best F_1-score in each setting is underlined, the best overall F_1-score is typeset in bold.*

hyperparameter search provides the best results. Overall, the best performing model on coarse-grained labels is bestrun GBERT with 68.68 macro F_1-score after hyperparameter search. The results for few-shot classification are almost similar and differ by 0.23 points. Regarding the fine-grained labels, few-shot classification with Fastfit has a clear advantage regarding macro F_1-score. Compared to the default, Fastfit achieves 3-7 points more on macro F_1-score; compared to bestrun with hyperparameter search, it achieves 0.7-3 points more. The best results are achieved by GBERT with 57.04 macro F_1-score.

Concerning the results of the decoder-based models, we observe that Gemini 1.5 Flash achieves exceptionally low scores in both settings (macro F_1-score of 28.96 and 34.26, respectively). While Llama 3.2 achieves the best F_1-score with 62.56 in the coarse-grained setting, and Llama 3 achieves the best F_1-score with 39.88 in the fine-

grained setting. All encoder-based models still outperform the two Llama models. We observe a particularly large difference concerning the results on the fine-grained speech acts: even the default models with default hyperparameters achieve macro F_1-scores that are at least 10 points better.

	precision	*recall*	*F_1-score*	*support*
Assertive	73.65	82.77	77.81	137
Expressive	71.25	63.25	66.79	80
Commissive	57.33	60.00	58.10	4
Directive	93.83	89.29	91.45	127
Unsure	31.46	28.67	29.31	30
Other	89.10	88.57	88.60	14
macro F_1-score	69.44	68.76	68.68	392

Table 4. *Mean results per class during 5-fold cross-validation for the best performing model GBERT on the coarse-grained labels after hyperparameter search.*

Table 4 presents the mean results of the best-performing model GBERT after hyperparameter search for each coarse-grained speech act class. The class *directive* achieves the best macro F_1-score (91.45) while the class *unsure* achieves the lowest macro F_1-score (29.31). This indicates that the class either was not well defined during the annotation or that it is difficult to predict whether the surrounding context is sufficient enough for a valid interpretation. An interesting finding is that the *commissive* class does not achieve the lowest macro F_1-score, although it is the class with the lowest number of instances during training and validation.

Finally, Table 5 shows the mean results of GBERT (in the few-shot classification setting with Fastfit) for each fine-grained speech act class. The best-performing classes are *address* and *expressEmoji* with a macro F_1-score of 99.60 and 98.56, respectively. This should come as no surprise as these are the most well-defined classes that are almost exclusively used whenever emojis or mentions are involved. The class *request* has the third best macro F_1-score (87.84), closely followed by *other* with an F_1-score of 86.13. As *request* is mostly used for questions, the presence of a question mark is most probably the best indicator of the class, leading to the good F_1-score. Similarly, the class *other* is most often used for uses of hashtags. Thus, the presence of a hashtag is a very likely sign to classify the utterance as an instance of *other*. The two worst-performing classes are *rejoice* and *other* with a macro F_1-score of 20.00 and 21.46, respectively.

5.2. *Error Analysis*

We conducted an error analysis with the aim of improving our understanding of the models' performance. For our analysis, we choose the best-performing models for each data version. As we evaluated each model in a 5-fold cross-validation manner,

	precision	*recall*	*F_1-score*	*support*
Assert	67.97	76.95	72.16	118
Sustain	50.00	20.00	28.00	3
Guess	55.71	36.00	40.00	5
Predict	60.00	42.86	48.21	7
Agree	63.33	33.33	41.33	3
Rejoice	40.00	13.33	20.00	3
Complain	50.37	56.47	53.17	51
Wish	60.00	40.00	46.67	2
expressEmoji	99.09	98.10	98.56	21
Commissive	83.33	65.00	71.90	4
Request	93.15	83.64	87.84	33
Require	62.53	50.00	54.61	16
Suggest	76.67	33.33	44.67	3
Address	100.00	99.20	99.60	75
Unsure	23.71	20.00	21.46	30
Other	84.63	88.00	86.13	15
Excluded	80.00	46.67	55.43	3
macro F_1-score	67.68	53.11	57.04	392

Table 5. *Mean results per class during 5-fold cross-validation for best performing model GBERT on the fine-grained labels in few-shot classification with Fastfit.*

we selected a fold where the results of the model were closest to the mean results on all folds. We create confusion matrices for the two models to illustrate common errors. Figure 1 shows the results of GBERT which was fine-tuned on the coarse-grained data version consisting of six classes and Figure 2 shows the confusion matrix for GBERT which was fine-tuned on the fine-grained data version. It should be noted that the two confusion matrices only show incorrectly predicted labels, while all correctly predicted labels were removed.

The confusion matrix for the coarse-grained model shows that most instances of errors consist either of *assertives* that were classified as *expressives* or *expressives* classified as *assertives*. This finding corroborates our observations made during the annotation of the data, where the greatest challenge was to distinguish between *assertives* an *expressives*. An example can be seen in Table 6, ex. (1). While the declarative sentence structure might appear like a mere statement on the surface, the utterance is actually used to express a negative attitude which is not recognised by the classifier. Furthermore, the confusion matrix illustrates that the class *unsure* often leads to misclassifications with *assertives*. This class was originally created to address uncertainties due to missing or unclear contexts. As the nature of this class heavily relies on the surrounding context of the utterance which is not available to the classifier during prediction, the classifier cannot accurately predict the class *unsure* which is reflected in the high error rate. Thus, the classifier tends to incorrectly classify instances which

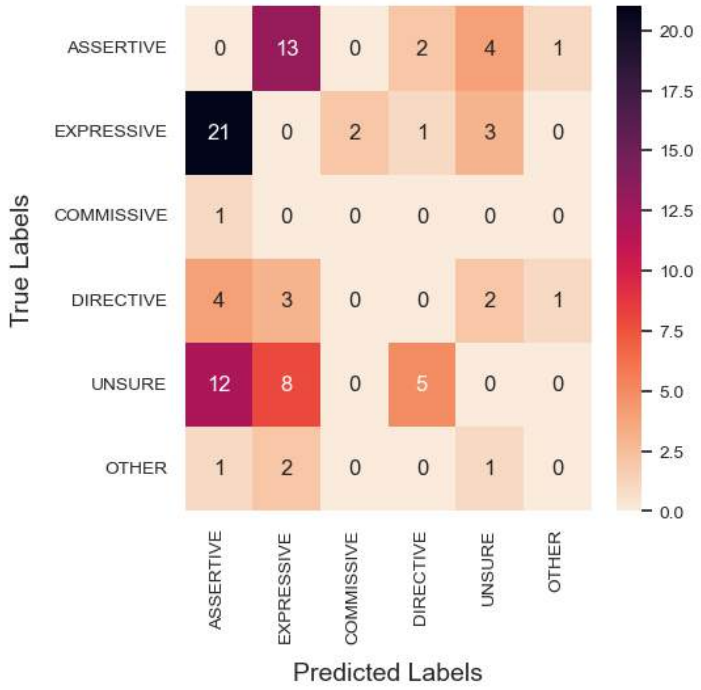

Figure 1. *Confusion matrix for all incorrectly classified instances of the best performing coarse-grained classifier GBERT after hyperparameter search to illustrate common errors.*

are usually shorter and do not provide much context by themselves, as can be seen in Examples (2), (3) and (4) in Table 6.

Figure 2 shows similar observations. The class *unsure* leads to several errors, repeatedly, while the classes *assert* (a subclass of *assertives*) and *complain* (a subclass of an *expressive*) are frequently confused with each other. In example (5), for instance, one could argue that the utterance is both describing the world and expressing an attitude, simultaneously, leading to misclassifications. In example (6) in Table 6, the expression is clearly used to express a negative feeling of the speaker. Nevertheless, the form of a declarative sentence, which is the sentence structure most often used for *assertives* (Plakidis and Rehm, 2022), might have led the classifier to incorrectly classify it as an instance of the *assert* class.

	Text	*Correct label*	*Predicted label*
	coarse-grained labels		
1	*Wir leben in einem Irrenhaus.* we live in a madhouse 'We live in a madhouse.'	Expressive	Assertive
2	*Ein Witz.* a joke 'A joke.'	Expressive	Unsure
3	*Wie kannst du!* how can you 'How could you!'	Unsure	Directive
	fine-grained labels		
4	*Ja!* yes 'Yes!'	Agree	Unsure
5	*Kein Wunder, dass Bewegungen wie z.B. AfD usw. so viel Zulauf haben.* no wonder that movements like e.g. AfD etc. so much support have 'No wonder movements like the AfD, etc., have so much support.'	Assert	Complain
6	*Ich habe eine Scheißangst.* I have a shit.fear 'I'm scared as hell.'	Complain	Assert

Table 6. *Examples of misclassifications for coarse- and fine-grained speech act classification.*

6. Conclusion

Our results demonstrate that encoder-based models outperform decoder-based models in the task of speech act classification. The best performing classifier is GBERT in both settings, achieving a macro F_1-score of 68.68 for coarse-grained classification and a macro F_1-score of 57.04 for fine-grained classification.

Our results show that there is still room for improvement regarding the automated detection of speech acts, which could involve a new annotation scheme with more precise guidelines in order to diminish potentially overlapping classes and vagueness concerning definitions. During the annotation process, we observed that some examples in our data fit multiple classes at the same time, especially with regard to the distinction between *assertives* and *expressives*, which renders the task of speech act annotation rather subjective. This is also reflected in the error analysis of this paper which shows that the distinction between *assertives* and *expressives* is a frequent error.

For future work, we thus intend to revise our annotation scheme and annotate a larger, more balanced dataset, enabling us to train improved speech act classifiers as well as an offensive language classifier to investigate whether the inclusion of speech acts improves the detection of offensive language on a larger dataset. In addition, we plan to release a curated version of the annotations.

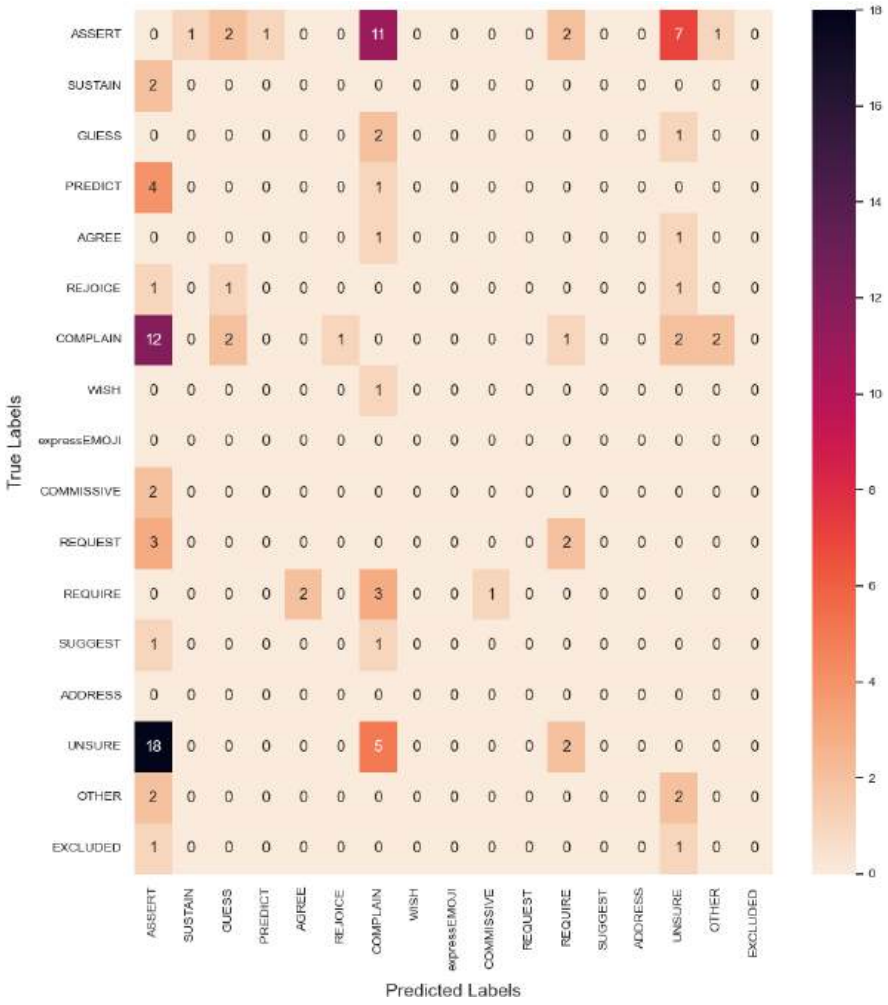

Figure 2. *Confusion matrix for all incorrectly classified instances of the best performing fine-grained classifier GBERT to illustrate common errors.*

7. References

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

Benamara F., Mari A., Meunier R., Moriceau V., Moudjari L., Tinarrage V., "Digging Communicative Intentions: The Case of Crises Events", *Dialogue Discourse*, 2024.

Bilewicz M., Soral W., "Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization", *Political Psychology*, vol. 41, p. 3-33, 2020.

Chan B., Schweter S., Möller T., "German's Next Language Model", *arXiv preprint arXiv:2010.10906*, 2020.

Cohen J., "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.", *Psychological Bulletin*, vol. 70, p. 213-220, 1968.

Compagno D., Epure E., Deneckere R., Salinesi C., "Exploring Digital Conversation Corpora with Process Mining", *Corpus Pragmatics*, vol. 2, p. 193-215, 2018.

Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L., "QLORA: efficient finetuning of quantized LLMs", *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

Devlin J., Chang M.-W., Lee K., Toutanova K., "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*, 2018.

Dhayef Q., Ali A., "A Pragmatic Study of Racial Hate Speech", *Journal of Tikrit University for Humanities*, vol. 27, n° 8, p. 24-1, 2020.

Fleiss J. L., "Measuring nominal scale agreement among many raters.", *Psychological bulletin*, vol. 76, n° 5, p. 378, 1971.

Gemini Team, Georgiev P., Lei V. I., Burnell R., Bai L., Gulati A., Tanzer G., Vincent D., Pan Z., Wang S., Mariooryad S., et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context", *arXiv*, 2024.

Kozlowski D., Lannelongue E., Saudemont F., Benamara F., Mari A., Moriceau V., Boumadane A., "A three-level classification of French tweets in ecological crises", *Information Processing & Management*, vol. 57, n° 5, p. 102284, 2020.

Laurenti E., Bourgon N., Benamara F., Mari A., Moriceau V., Courgeon C., "Give me your Intentions, I'll Predict Our Actions: A Two-level Classification of Speech Acts for Crisis Management in Social Media", *13th Conference on Language Resources and Evaluation (LREC 2022)*, p. 4333-4343, 2022.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., "Flaubert: Unsupervised language model pre-training for french", *arXiv preprint arXiv:1912.05372*, 2019.

Liaw R., Liang E., Nishihara R., Moritz P., Gonzalez J. E., Stoica I., "Tune: A Research Platform for Distributed Model Selection and Training", *arXiv preprint arXiv:1807.05118*, 2018.

Lin T.-Y., Goyal P., Girshick R., He K., Dollár P., "Focal Loss for Dense Object Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, n° 2, p. 318-327, 2020.

Llama Team, Grattafiori A., Dubey A., Jauhri A., Pandey A., Kadian A., Al-Dahle A., Letman A., Mathur A., Schelten A., Vaughan A., et al., "The Llama 3 Herd of Models", *arXiv*, 2024.

Mangrulkar S., Gugger S., Debut L., Belkada Y., Paul S., Bossan B., "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods", 2022.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D., Sagot B., "CamemBERT: a Tasty French Language Model", *in* D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 7203-7219, July, 2020.

Mubarok Y., Sudana D., Yanti D., Aisyah A. D., Af'idah A. N. *et al.*, "Abusive Comments (Hate Speech) on Indonesian Social Media: A Forensic Linguistics Approach", *Theory and Practice in Language Studies*, vol. 14, nº 5, p. 1440-1449, 2024.

Nobata C., Tetreault J., Thomas A., Mehdad Y., Chang Y., "Abusive Language Detection in Online User Content", *Proceedings of the 25th International Conference on World Wide Web*, p. 145-153, 2016.

Oktaviani A. D., Nur O. S., "Illocutionary Speech Acts and Types of Hate Speech in Comments on @Indraakenz's Twitter Account", *International Journal of Science and Applied Science: Conference Series*, 2022.

Ollagnier A., "CyberAgressionAdo-v2: Leveraging Pragmatic-Level Information to Decipher Online Hate in French Multiparty Chats", *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024.

Plakidis M., Rehm G., "A Dataset of Offensive German Language Tweets Annotated for Speech Acts", *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 4799-4807, 2022.

Poletto F., Basile V., Sanguinetti M., Bosco C., Patti V., "Resources and benchmark corpora for hate speech detection: a systematic review", *Language Resources and Evaluation*, vol. 55, p. 477-523, 2021.

Risch J., Stoll A., Wilms L., Wiegand M., "Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments", *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, Association for Computational Linguistics, Duesseldorf, Germany, p. 1-12, 2021.

Searle J. R., *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge University Press, Cambridge, 1979.

Struß J. M., Siegel M., Ruppenhofer J., Wiegand M., Klenner M., "Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language", *German Society for Computational Linguistics. Proceedings of the 15th Conference on Natural Language Processing (KONVENS) 2019*, Nürnberg/Erlangen, p. 354-365, 2019.

Vosoughi S., Roy D., "Tweet Acts: A Speech Act Classifier for Twitter", *Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, Cologne, Germany, 2016.

Wang Z., Pang Y., Lin Y., Zhu X., "Adaptable and Reliable Text Classification using Large Language Models", *arXiv*, 2024.

Weisser M., *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond*, John Benjamins Publishing Company, Amsterdam/Philadelphia, 2018.

Wiegand M., Siegel M., Ruppenhofer J., "Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language", *Proceedings of GermEval 2018 Workshop (GermEval)*, 2018.

Yehudai A., Bandel E., "When LLMs are Unfit Use FastFit: Fast and Effective Text Classification with Many Classes", *ArXiv*, 2024.

Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R., "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)", *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, p. 75-86, 2019.

Zhang R., Gao D., Li W., "What Are Tweeters Doing: Recognizing Speech Acts in Twitter", *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Thibault BAÑERAS-ROUX : thibault.baneras.roux@gmail.com

Titre : Analyse et compréhension de l'évaluation des systèmes de reconnaissance automatique de la parole : vers des métriques intégrant la perception humaine

Mots-clés : reconnaissance automatique de la parole, métriques d'évaluation, perception humaine, sémantique.

Title: *Analysis and Understanding of the Evaluation of Automatic Speech Recognition Systems: Towards Metrics Integrating Human Perception*

Keywords: *automatic speech recognition, evaluation metrics, human perception, semantics.*

Thèse de doctorat en informatique, LS2N, UFR Sciences et techniques, Nantes Université, sous la direction de M. Richard Dufour (Pr, Nantes Université), Mme Jane Wottawa (MC, Le Mans Université) et M. Mickael Rouvier (MC HDR, Avignon Université). Thèse soutenue le 17/01/2025.

Jury : M. Richard Dufour (Pr, Nantes Université, directeur), Mme Irina Illina (MC HDR, université de Lorraine, rapporteuse), M. Cyril Grouin (IR HDR, université Paris-Saclay, rapporteur), M. Benjamin Lecouteux (Pr, université Grenoble Alpes, président), Mme Béatrice Daille (Pr, Nantes Université, examinatrice), Mme Martine Adda-Decker (DR, CNRS, examinatrice), Mme Jane Wottawa (MC, Le Mans Université, co-encadrante), M. Mickael Rouvier (MC HDR, Avignon Université, co-encadrant).

Résumé : *Le taux d'erreur de mots reste la métrique principale pour évaluer les systèmes de reconnaissance automatique de la parole (RAP), mais il ne reflète pas toujours la perception humaine. Cette thèse propose des métriques alternatives pour*

évaluer non seulement l'orthographe, mais aussi la grammaire, la sémantique et la phonétique. À travers le corpus HATS, annoté par 143 francophones, nous avons mesuré la corrélation entre ces métriques et les choix humains. Les résultats montrent que SemDist, basée sur les représentations sémantiques de BERT, est la plus pertinente, tandis que le taux d'erreur de mots se révèle peu performant. Une analyse des hyperparamètres des systèmes de RAP révèle que chaque métrique évalue des aspects distincts, soulignant l'importance d'une évaluation multimétrique. Enfin, pour rendre les métriques sémantiques plus compréhensibles, nous avons développé la méthode minED, qui identifie la gravité des erreurs et améliore l'interprétation des scores, offrant des outils précieux pour évaluer et perfectionner les systèmes RAP.

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-04931310

Johanna CORDOVA : johanna.cordova@inalco.fr

Titre : Le quechua dans les outils numériques, un défi pour le TAL ? Développement de ressources linguistiques et numériques pour le quechua ancashino

Mots-clés : TAL, quechua, langue agglutinante, langue minoritaire, analyse morphologique, corpus aligné, corpus arboré, *Universal Dependencies*, OCR.

Title: *Quechua in Digital Tools: A Challenge for NLP? Development of Linguistic and Digital Resources for Ancash Quechua*

Keywords: *NLP, Quechua, agglutinative language, low-resourced language, morphological analysis, parallel corpora, treebank, Universal Dependencies, OCR.*

Thèse de doctorat en traitement automatique des langues, école doctorale Langues, littératures et sociétés du monde, institut national des langues et civilisations orientales, sous la direction de M. César Itier (Pr, institut national des langues et civilisations orientales) et M. Damien Nouvel (MC, institut national des langues et civilisations orientales). Thèse soutenue le 18/12/2024.

Jury : M. César Itier (Pr, institut national des langues et civilisations orientales, codirecteur), M. Damien Nouvel (MC, institut national des langues et civilisations orientales, codirecteur), Mme Capucine Boidin (Pr, université Sorbonne-Nouvelle, institut des hautes études de l'Amérique latine, IHEAL, centre de recherche et de documentation sur les Amériques, CREDA, présidente), Mme Kata Gábor (MC, institut national des langues et civilisations orientales, ERTIM, examinatrice), Mme Katharina Haude (CR, CNRS, Structure et dynamique des langues, SeDyl, UMR 8202, examinatrice), M. Elwin Huaman (*senior researcher*, Semantic Technology Institute Innsbruck, université d'Innsbruck, Autriche, examinateur), M. Sylvain Kahane (Pr, université Paris Nanterre, rapporteur), M. Matthias Urban (CR, CNRS, Dynamique du langage, DDL, rapporteur).

Résumé : *Les langues quechuas constituent l'une des familles linguistiques amérindiennes comptant le plus grand nombre de locuteurs natifs. Au Pérou, selon le recensement de 2017, 13,9 % de la population a le quechua pour première langue et environ 20 % le parle. Pourtant, cette macrolangue est presque totalement absente des usages numériques. En traitement automatique des langues (TAL), c'est une langue peu dotée, avec une forte disparité de ressources selon la variété de quechua considérée. L'objectif de cette thèse est de développer un ensemble d'outils fondamentaux pour le traitement automatique d'une variété du quechua central, le quechua ancashino, parlé par environ 400 000 personnes, et en danger d'extinction d'après la classification de l'UNESCO. Ce processus comporte trois étapes : la première est la numérisation des ressources disponibles dans cette variété (dictionnaires, corpus écrits), qui permet de constituer le plus grand corpus bilingue quechua-espagnol aligné par phrases disponible à ce jour pour le traitement automatique (6 120 phrases pour 38 500 tokens en quechua). La seconde étape, indispensable pour une langue agglutinante comme le quechua, est l'implémentation d'un analyseur morphologique permettant d'identifier, à partir d'une forme de surface, les morphes et leurs morphèmes correspondants, ainsi que leur étiquette suivant le formalisme de Leipzig. Outre l'analyse au niveau du mot, nous proposons une méthode basée sur les CRF pour sélectionner les annotations correctes en fonction du contexte phrastique lorsque plusieurs analyses sont possibles. Nous terminons par l'élaboration d'un corpus arboré pour l'analyse en morphosyntaxe, suivant le formalisme* Universal Dependencies. *Ces développements s'accompagnent d'une description approfondie de la morphologie du quechua ancashino, largement sous-étudiée. Les ressources développées sont distribuées sous licence libre pour des travaux ultérieurs en TAL, mais aussi adaptées pour une utilisation grand public à travers des applications telles qu'un moteur de recherche permettant d'interroger l'ensemble des dictionnaires ou un outil d'analyse morphologique en ligne. Dans un contexte global de valorisation des langues originaires et alors que d'ambitieuses politiques liées aux droits linguistiques sont en cours de déploiement dans les pays de l'aire andine, nous faisons l'hypothèse que la présence du quechua dans les technologies constitue un important levier pour renforcer sa pratique et faciliter son enseignement.*

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-04989396

Nicolas GUTEHRLÉ : nicolas.gutehrle@univ-fcomte.fr

Titre : Extraction d'informations appliquée aux documents non structurés pour la valorisation de périodiques historiques : application au patrimoine de la région Bourgogne Franche-Comté en France

Mots-clés : extraction d'information, gestion des connaissances, annotation sémantique, interfaces de recherche, exploitation et exploration des documents, humanités numériques.

Title: *Information Extraction from Unstructured Documents for the Valorisation of Historical Periodicals: Application to the Heritage of the Bourgogne Franche-Comté Region in France*

Keywords: *information extraction, knowledge management, semantic annotation, search interfaces, exploration and exploitation of documents, digital humanities.*

Thèse de doctorat en sciences du langage, mention traitement automatique des langues, centre de recherches interdisciplinaires et transculturelles, CRIT, UR 3224, UFR Sciences du langage, de l'homme et de la société, université de Franche-Comté, Besançon, sous la direction de Mme Iana Atanassova (MC HDR, université de Franche-Comté, centre de recherches interdisciplinaires et transculturelles, CRIT, UR 3224, Besançon). Thèse soutenue le 21/06/2024.

Jury : Mme Iana Atanassova (MC HDR, université de Franche-Comté, centre de recherches interdisciplinaires et transculturelles, CRIT, UR 3224, Besançon, directrice), M. Antoine Doucet (Pr, La Rochelle Université, laboratoire Informatique, image et interaction, L3i, président), M. Bruno Bachimont (Pr, université de technologie de Compiègne, Connaissance, organisation et systèmes techniques, Costech, rapporteur), M. Pavel Pecina (MC, Charles University, Institute of Formal and Applied Linguistics, ÚFAL, Prague, République tchèque, rapporteur), M. Jean-Charles Lamirel (MC, université de Strasbourg, laboratoire lorrain de recherche en informatique et ses applications, LORIA, SYNALP, Nancy, examinateur), M. Mohand Boughanem (Pr, université Paul Sabatier, institut de recherche en informatique de Toulouse, IRIT, examinateur), M. Adam Jatowt (Pr, Universität Innsbruck, Data Science Group, Autriche, examinateur).

Résumé : *Ces dernières années, les bibliothèques et archives ont entrepris de nombreuses campagnes de numérisation afin d'élargir l'accès du public à leurs collections d'archives. Cependant, le défi de promouvoir le contenu des collections et de rendre ces ressources accessibles reste entier. La numérisation produit souvent un contenu non structuré dans lequel il est difficile de naviguer, tandis que les interfaces qui s'appuient sur des requêtes basées sur des mots-clés pour accéder aux documents d'archives peuvent fournir aux utilisateurs des résultats non pertinents. Afin d'exploiter le potentiel des «* Big Data of the Past *», notion introduite par Kaplan et di Lenardo en 2017, il est essentiel de développer des méthodes et des cadres pour structurer le contenu textuel des documents, dans le but d'en améliorer l'exploration et l'exploitation. Dans ce contexte, la présente thèse de doctorat aborde le problème du traitement des documents historiques numérisés, en se concentrant sur l'extraction des entités nommées et des relations afin de créer des interfaces pour l'exploitation efficace des données textuelles historiques. Premièrement, nous proposons une nouvelle méthode pour déterminer la structure logique des journaux historiques en utilisant une approche à base de règles. Deuxièmement, nous présentons une méthode pour extraire les entités et les relations concernant les personnes et les lieux mentionnés dans les textes. Notre approche s'intitule* Extensible, Lightweight and Interpretable Joint

Extraction of Relations and Entities *(ELIJERE). Elle est basée sur des ressources linguistiques obtenues par supervision distante. Enfin, nous proposons un cadre général pour l'étude de l'expression d'informations spatiales dans les documents, et un autre cadre pour l'application des méthodes de* TimeLine Summarisation *à des collections de documents. Nous montrons comment ces méthodes peuvent être appliquées pour produire des interfaces sémantiquement riches, telles que des frises chronologiques et des cartes, qui permettent au grand public une lecture proche ou distante de ces collections.*

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-04719778

Thi Phuong Hang Le : hangtp.le@gmail.com

Titre : Architectures et techniques d'entraînement pour la traduction parole-texte multilingue

Mots-clés : multilinguisme, traduction automatique, traduction automatique de la parole, *transformer*, traduction parole-texte, traduction.

Title: *Model Architectures and Training Techniques for Multilingual Speech-to-Text Translation*

Keywords: *multilingualism, automatic speech recognition, ASR, speech-to-text translation, ST, transformer, translation.*

Thèse de doctorat en mathématiques et informatique, laboratoire d'informatique de Grenoble, LIG, école doctorale Mathématiques, sciences et technologies de l'information, université Grenoble Alpes, sous la direction de M. Didier Schwab (Pr, université Grenoble Alpes) et M. Benjamin Lecouteux (Pr, université Grenoble Alpes). Thèse soutenue le 25/03/2024.

Jury : M. Didier Schwab (Pr, université Grenoble Alpes, codirecteur), M. Benjamin Lecouteux (Pr, université Grenoble Alpes, codirecteur), Mme Caroline Rossi (Pr, université Grenoble Alpes, présidente), M. Frédéric Béchet (Pr, université Aix-Marseille, rapporteur), M. François Yvon (DR, CNRS, rapporteur), M. Laurent Besacier (*principal scientist*, HDR, NaverLabs Europe, examinateur), M. Juan Pino (*research scientist*, Meta AI Research, examinateur).

Résumé : *Speech-to-text translation (ST) consists in translating a speech audio input in one language into a text output in another language. This task is highly challenging due to its multimodal and multilingual nature (the former involves both speech and text modalities, while the latter involves more than one language). In this thesis, we make three major contributions spanning two primary research areas of ST, namely model architectures and training techniques.*

First, in terms of model architectures, we introduce the dual-decoder transformer, a new model architecture that jointly performs automatic speech recognition (ASR) and multilingual ST. Our model consists of two decoders, each responsible for one task (ASR or ST), that can interact with each other through a novel dual-attention mechanism. This design allows the decoders to specialize in their respective tasks while being helpful to each other. We propose two variants, called the parallel and cross dual-decoder transformers, corresponding to two different levels of dependencies between the decoders. The proposed model also generalizes existing approaches using two independent or weakly tied decoders. Experiments on standard benchmarks show that our models outperform previous work in terms of translation performance under both bilingual and multilingual settings.

Second, in terms of training techniques, we propose a parameter-efficient fine-tuning approach based on adapter modules. We show that language-specific adapters can enable a fully trained multilingual ST model to be further specialized in each language pair. With these adapter modules, one can efficiently obtain a single multilingual ST system that outperforms the original multilingual model as well as multiple bilingual systems while maintaining a low storage cost and simplicity in deployment. In addition, we show that adapters can also be used to connect available pre-trained models such as an ASR model and a multilingual denoising auto-encoder to form strong multilingual ST systems.

Finally, as a second contribution in training techniques, we propose an effective supervised pre-training method to address the so-called speech-text modality gap, a well-known major challenge in ST. Our method combines the connectionist temporal classification loss and optimal transport in a Siamese-like model. This model is composed of two encoders, one for acoustic inputs and the other for textual inputs, which are trained such that they produce representations that are close to each other in the Wasserstein space. Extensive experiments on standard benchmarks show that our pre-training method applied to the vanilla encoder-decoder transformer achieves state-of-the-art performance under the no-external-data setting, and performs on par with recent strong multi-task learning systems trained with external data. Finally, our method can also be applied on top of these multi-task systems, leading to further improvements for these models.

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-04726713

Anna PAPPA : apappa@univ-paris8.fr

Titre : Contributions à la création de corpus et de modèles d'apprentissage profond pour les données textuelles multilingues

Mots-clés : création de corpus, corpus multilingues, annotation sémantique, analyse textuelle, apprentissage profond, méthodes hybrides.

Title: *Contributions to Corpus Creation and Deep Learning Models for Multilingual Textual Data*

Keywords: *corpus creation, multilingual corpora, semantic annotation, textual analysis, deep learning, hybrid methods.*

Habilitation à diriger des recherches en informatique, laboratoire d'intelligence artificielle et sémantique des données, LIASD, UFR des sciences et des technologies du numérique, département Programmation et informatique fondamentale, université Paris 8. Habilitation soutenue le 22/04/2024.

Jury : M. Nicolas Jouandeau (Pr, université Paris 8, rapporteur), M. Christian Boitet (Pr émérite, université Grenoble Alpes, rapporteur), M. Benoît Crabbé (Pr, université Paris Cité, rapporteur), M. Max Silberztein (Pr, université de Franche-Comté, examinateur), Mme Tita Kyriacopoulou (Pr, université Gustave Eiffel, examinatrice), M. Tristan Cazenave (Pr, université Paris Dauphine, examinateur), M. Jean-Jacques Bourdin (Pr, université Paris 8, examinateur).

Résumé : *Cette habilitation à diriger des recherches se situe à l'intersection de l'informatique et de la linguistique, synthétisant près d'une décennie de travaux qui explorent cette interface multidisciplinaire. Elle est structurée autour de trois axes principaux, chacun apportant des contributions au domaine du traitement automatique des langues (TAL).*

Le premier axe présente la création de corpus multilingues et thématiques, spécifiquement dédiés à l'analyse des opinions et des aspects. Les méthodologies et les outils développés visent à réduire le bruit et les irrégularités dans les données, en se concentrant sur des avis d'utilisateurs provenant de diverses plateformes en ligne. Ces corpus multilingues sans bruit servent de base solide pour les expérimentations subséquentes et permettent une analyse plus précise et pertinente des sentiments exprimés dans différentes langues.

*Le deuxième axe explore l'analyse de sentiment par le biais de modèles hybrides, combinant des réseaux neuronaux convolutifs (CNN) et récurrents (RNN) tels que les LSTM (*Long Short-Term Memory*). Ces modèles atteignent une précision élévée, oscillant entre 90 % et 100 % sur divers corpus multilingues non annotés. L'utilisation de réseaux neuronaux convolutifs hiérarchiques (ConvNet) et de réseaux neuronaux récurrents permet de relever les défis de la prédiction de la polarité et de la classification thématique, améliorant ainsi les performances et la robustesse des systèmes de TAL.*

*Le troisième axe aborde l'annotation d'aspects en utilisant une architecture combinée BiLSTM-CNN-CRF (*Bidirectional Long Short-Term Memory-Convolutional Neural Network-Conditional Random Field*), enrichie par des techniques d'apprentissage profond telles que l'apprentissage actif et l'apprentissage par transfert. Ces méthodes se sont avérées particulièrement efficaces pour améliorer les performances des modèles dans des contextes de rareté de données ou de langues peu représentées, en*

permettant une annotation plus précise et fiable des aspects dans des textes multilingues.

En synthèse, cette habilitation constitue une contribution pertinente à la fusion des méthodologies informatiques et linguistiques. Elle exploite des jeux de données sur mesure et des architectures de modèles hybrides pour résoudre des défis complexes en annotation sémantique et en analyse multilingue. Ce travail ouvre également la voie à des recherches futures pour affiner ces méthodologies dans des scénarios encore plus exigeants ou moins étudiés, et propose des perspectives prometteuses pour le développement de nouveaux outils et techniques en TAL. Les résultats obtenus démontrent l'importance et l'efficacité de l'approche multidisciplinaire dans le traitement et l'analyse des données textuelles.

URL où le mémoire peut être téléchargé :
https://univ-paris8.hal.science/tel-04595140