Progress in the Modeling of Violent Messages in Spanish Social Networks

Beatriz Botella-Gil

Dept. of Software and Computing Systems, University of Alicante, Apdo. de Correos 99, E-03080, Alicante, Spain

Abstract

Society advances loaded with new and highly accessible knowledge, that is published in the virtual world. It is a reality that ICTs have brought many benefits for our lives but we also see how year after year the use of violence in platforms increase. The application of PLN is fundamental in this type of research given the large volume of existing data, which facilitates a breakthrough in the investigation of the detection of this type of messageThe doctoral work focuses on the detection of violent messages in the social network twitter from different perspectives.

Keywords

Natural Language Processing, Annotation Guideline, Dataset Annotation, Detection of Violent Messages

1. Introduction

The Internet has become an indispensable part of our lives, being used in practically all of society's daily activities. Nowadays it is possible to have immediate contact with any person in the world through an electronic device. Society is moving forward with new and very accessible knowledge that is published in the virtual world. Personal relationships have also been affected, not only in the private sphere, but also in the workplace.

According to We are Social[1], almost 44 million people in Spain spend more than 6 hours a day on the Internet and around 41 million Spaniards are users of social networks. It is a reality that ICTs have brought many benefits to our lives, but also, thanks to the possibility of being an anonymous user and the absence of observing face to face the damage that our words can generate, problems still to be solved are created [2]. In particular, many researchers call this type of violent action as hate speech, an offensive behavior through language towards people or groups and whose detection is being a problem for researchers, since, it is possible that violence is not used explicitly in a discourse, but as a single word or even implicitly through the use of emoticons [3], or by using humor, irony, sarcasm [4, 5] or stereotypes [6].

Given the number of users present in social networks, it is impossible to manually control the comments that are registered and their intention, creating impunity for people who use these networks in order to do harm. The identification of violent messages and control of hate speech on the Internet has been approached from different points of view, being essential the use of Natural Language Processing (NLP) to develop computational systems that help to interpret and process human language quickly and effectively.

Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA, 28 September 2023, Jaén, Spain.

beatriz.botella@dlsi.ua.es (B. Botella-Gil)

One barrier we encountered right at the start of the study is the collection of messages in social networks, since, as Bruns points out[7], the restriction of access to social network data makes it difficult to analyze issues of great importance such as abusive language, harassment, hate speech or disinformation campaigns. That is why in the present research the social network Twitter will be used, where as Ott [8] defines: "Twitter discourse is disrespectful because its register is informal, and because it depersonalizes social interactions". This research aims to provide solutions to the existing problems in the detection of violent messages in social networks in a fast, automatic and effective way.

2. Background and Related Work

Many studies have been carried out on the analysis of violent messages in social networks and media. In particular, much research can be found focused on discovering the characteristics of human behavior that promote the emission of such messages, as well as those that focus on discovering the characteristics of the messages themselves through PLN techniques.

2.1. Language and behavior study

There is a wealth of research on human behavior in the face of violent messages and the language used. As McMenamin said [9], "hate speech is studied according to how it is defined, how it is interpreted, and what are the best practices to deal with it". That is why we find works such as Salado [10], who based their research on a syntactic analysis of language, and discovered that there are different linguistic elements to take into account that are present in the forms of violent speech such as, the linguistic category, the lexicon used or how the words are placed. Plaza-Del-Arco et al. [11] carried out a study of the implicit and explicit linguistic phenomena of offensive language. Others such as Gitari [12] focused on something as specific as the creation of a list of verbs that can be indicators of violent messages. On the other hand, there are works that focus on the roles present in these acts, such as, for example, Nielsen, who, through interviews and a study of the participants, observed the consequences for the victim, his or her harm and the possibility of crime in the messages.

2.2. PLN applied to the detection of violent messages

The application of PLN is fundamental in this type of research given the large volume of existing data, which facilitates a breakthrough in the investigation of the detection of this type of messages, thanks to the following techniques:

- **Keyword-based classifiers**
  Part of the research in this field has focused on the development of lists of insults that help automatic detection. In this sense, lexicons and dictionaries have been developed in order to observe whether the presence of these terms determines the violence in the message [13]. Although such lists have aided detection, they have fallen short of being the sole tool for determining violence. Violent language is constantly evolving, language varies from place to place varies depending on where it occurs and there may be terms that in some geographic areas are insults and in others are not [14].

- **Machine learning**
  Most of the work related to the detection of violent messages addresses this problem using classical machine learning (ML) algorithms. Works such as Xu et al. [15] and Dadvar et al. [16] have used support vector machines (SVM) in their research and obtained satisfactory results, proving to be very effective with large training samples. SVM is not the only classical algorithm used in research in this field; works such as [17], used other algorithms, showing in their results that the one that offered the best performance is logistic regression, followed by Naive Bayes and SVMs.
  Most ML-based classifiers use traditional text representations such as bag-of-words (BOW), n-grams, term frequency (TF), among others. In Burnap and Williams [18] all of the above techniques are used. This research compares the results obtained individually by the classifiers with the use of a set of classifiers (ensemble) that integrates them all, demonstrating greater accuracy in the latter. Sentiment analysis is another of the most widely used tools in this field. With it we can extract the polarity of the message and use this indicator along with other tasks to determine more accurately whether we are facing a violent or non-violent message [19].
  Corpus development, have an important role in offensive language research when ML techniques are applied. In recent years we have observed a large volume of work by PLN researchers to generate these resources [20, 21, 22, 23, 24, 25]. These authors created English-language resources, with SOLID [25] being the resource containing more than nine million English-language tweets tagged in a semi-supervised manner.
  On the other hand, HurtLex [26] is a multilingual lexicon of hate speech spanning multiple languages and hatebase3 [1] is a collaborative repository of hate speech that is also multilingual. The main drawback of these resources is their paucity of English terms, and those that are present have been compiled using a semi-automatic translation from another language, neglecting the importance of cultural and linguistic factors in each country. However, despite the fact that Spanish is one of the most widely spoken languages in the world, we found a shortage of resources in this language to carry out the task of detection. There are resources in Spanish for offensive words such as Plaza-Del-Arco et al. [11] for misogynistic and xenophobic terms; and Share [27] that label them as offensive and not. After the study carried out on the literature, it is considered necessary to elaborate another corpus to collect more characteristics present in violent messages, which can help in the explicability and detail of the detection.
- **Deep learning**
  Within AI there are other more complex techniques that have also been used in this task. We refer to deep learning (DL), as is the case of the research by Arcila-Calderón et al. [17] that after using ML tools and neural networks, the latter improved the evaluation metrics against the models generated with traditional ML algorithms. To the same end Badjatiya et al. [28] uses DL models to train different word embeddings validating that, using these representations, obtains better results than traditional representations such as term frequency - inverse document frequency (TF-IDF) or BoW.
  Models based on *transformer* architecture, such as BERT, RoBERTa and ALBERT, show the best state-of-the-art results in the detection of violent messages in tasks recognized

[1] urlhttps://hatebase.org/

as OffensEval or HatEval [29]. In Sarkar et al. [29] fine-tuning (*fine-tuning*) to BERT is performed using SOLID, the largest English offensive language identification corpus, improving the results obtained with BERT in the tasks mentioned above.

In Song et al. [30] an ensemble of classifiers (*ensemble*) based on RoBERTa and BERT is used which obtains the best results in the shared task "SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense"[2] which includes a subtask of detecting offensive messages. This work consists of fine tuning these models to create a classifier and clustering them into a set of classifiers based on *stacking*.

3. Main Hypothesis and Objectives

After literature reviews, future work was found to improve the detection of violent messages in social networks, more specifically focused not only on the language itself, but also on determining the existence of patterns of behavior and user profiles. This problem is being studied taking as references data such as user role, type of violence, message form (implicit/explicit), isolated message vs. thread of messages, number of followers and activity in the social network (number of # or mentions of the user).

3.1. Objectives

In our research the following objectives will be pursued:

1. Expanding the corpus VIL [31].
2. Define behavioral patterns for violent identification.
3. Achieve a system that is able to give a user violence score, both to know if the user is violent and to know if is receiving violence.
4. Identify the phases of the violence process (beginning, in process, completed).
5. Creation of a program for automatic detection of violence in social networks.

3.2. Hypothesis

For the study of violence detection through messages on social networks with the help of NLP, the following hypotheses are pursued:

1. Is it possible to detect violence on social networks through language analysis?. Certain words and phrases could serve as indicators that increase the likelihood of a message being considered violent. Therefore, it is relevant to investigate how language can be used offensively in online communication, using linguistic cues as potential markers of violence in messages.
2. Do patterns exist that aid in the detection of violence on social networks?. The study of the characteristics surrounding the issue of violence on social networks, including user behavior and their use of language, could provide essential insights that bring us closer to identifying effective techniques to address this problem.

[2] http://bit.ly/3J8uHOX

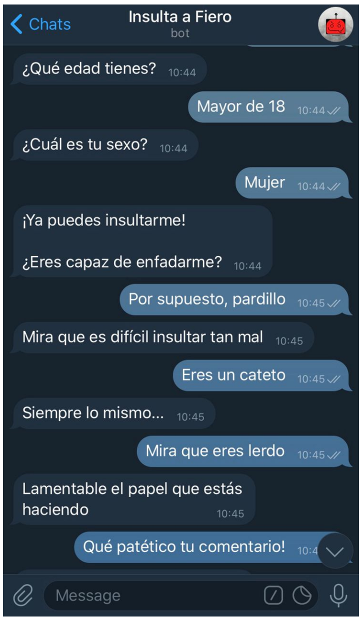
3. Is it possible to detect violent messages using NLP tools?. We believe that ML models trained with labeled data significantly enhance the ability to detect violent messages compared to rule-based approaches alone.

4. Experiments

We have conducted 3 experiments during this thesis, a FIERO chatbot, an annotation guide and a Corpus of violent messages through the social network Twitter.

4.1. Fiero

Fiero, a virtual assistant that maintains a conversation with the user encouraging him to express expletives through Telegram. This application is popularly known in the environment of the use of digital tools by the population. The collected dialogue will be used to generate linguistic resources that can be used in automatic artificial intelligence systems to combat social problems such as cyberbullying or hate speech [32]. In Figure 1, you can observe what a conversation with the chatbot looks like to collect insults from users.



¿Qué edad tienes? 10:44

Mayor de 18 10:44

¿Cuál es tu sexo? 10:44

Mujer 10:44

¡Ya puedes insultarme!

¿Eres capaz de enfadarme? 10:44

Por supuesto, pardillo 10:45

Mira que es difícil insultar tan mal 10:45

Eres un cateto 10:45

Siempre lo mismo... 10:45

Mira que eres lerdo 10:45

Lamentable el papel que estás haciendo 10:45

Qué patético tu comentario! 10:4

Figure 1: Fiero user interface.

4.2. Annotation Guide

Pursuing the goal of creating a resource to aid in the detection of violent messages, we decided to generate a fine-grained annotation guide for messages, with a certain degree of semantic complexity, which not only marks whether a message is violent or not, but also certain important elements regarding the content of the message. In this annotation guide shown in figure 2 we collect information about: Insults, Violent vs No-Violent, Level of violence, Role and Type of Violence.

4.3. VIL Corpus

Having studied the literature on the task and the importance of the application of PLN and ML and DL techniques, and because these techniques are fed by training data, we conclude the need to create a resource in Spanish that can be used in the effective detection of violent messages, with a level of detail that goes beyond simple binary detection, marking features, which we detail in our annotation guide, such as the degree, the role or type of violence, since we consider that if detection in the messages that can help future explainability in the decisions taken [31]. In the figure 3 you can see an example of how the labeled tweets would look in our corpus.

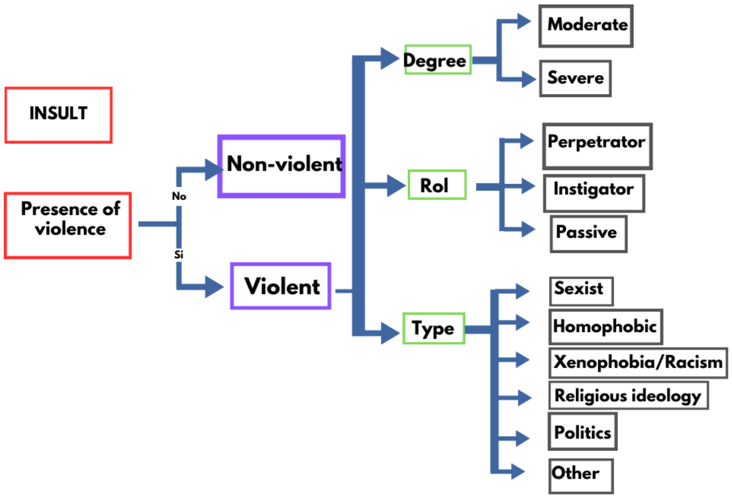


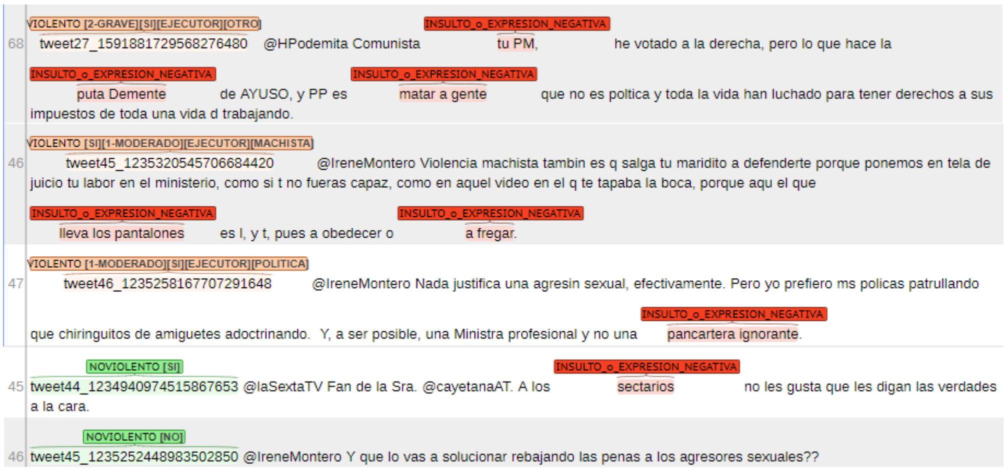
Figure 2: Annotation guide

5. Research issues to discuss

Following our research and the creation of an annotation guide and corpus presented in the previous sections, we have questions that could be addressed in the future:

1. Is possible to detect violent messages through Machine Learning to curb the problem of the use of violence in social networks? Due to the ambiguity and personal subjectivity in how users understand violence, we may encounter difficulties in reaching an agreement on what constitutes a violent message and what does not.

2. Different phases of violence can be defined in order to know what level of severity we find or in what phase of violence we are in? We could determine the severity level of violence based on the message content to take appropriate actions with the violent user, imposing different consequences based on severity.
3. An automatic alert could be created to warn us that violence is being generated? A prompt response in a case of violence would be beneficial for the virtual community.


VIOLENTO [2-GRAVE][SI][EJECUTOR][OTRO]
INSULTO_o_EXPRESION_NEGATIVA
INSULTO_o_EXPRESION_NEGATIVA
INSULTO_o_EXPRESION_NEGATIVA

tweet27_1591881729568276480 @HPodemita Comunista tu PM, he votado a la derecha, pero lo que hace la puta Demente de AYUSO, y PP es matar a gente que no es poltica y toda la vida han luchado para tener derechos a sus impuestos de toda una vida d trabajando.

VIOLENTO [SI][1-MODERADO][EJECUTOR][MACHISTA]
INSULTO_o_EXPRESION_NEGATIVA
INSULTO_o_EXPRESION_NEGATIVA

tweet45_1235320545706684420 @IreneMontero Violencia machista tambin es q salga tu maridito a defenderte porque ponemos en tela de juicio tu labor en el ministerio, como si t no fueras capaz, como en aquel video en el q te tapaba la boca, porque aqu el que lleva los pantalones es l, y t, pues a obedecer o a fregar.

VIOLENTO [1-MODERADO][SI][EJECUTOR][POLITICA]
INSULTO_o_EXPRESION_NEGATIVA

tweet46_1235258167707291648 @IreneMontero Nada justifica una agresin sexual, efectivamente. Pero yo prefiero ms policas patrullando que chiringuitos de amiguetes adoctrinando. Y, a ser posible, una Ministra profesional y no una pancartera ignorante.

tweet44_1234940974515867653 @laSextaTV Fan de la Sra. @cayetanaAT. A los sectarios no les gusta que les digan las verdades a la cara.

tweet45_1235252448983502850 @IreneMontero Y que lo vas a solucionar rebajando las penas a los agresores sexuales??

Figure 3: Examples of tweets annotated in the VIL corpus.

6. Acknowledgments

This research work is part of the Spanish research project Cleartext «TED2021-130707B-I00>. This work is funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/ PRTR. I would also like to thank my thesis directors Patricio Martinez-Barco and Estela Saquete Boro for their help and collaboration with the project Coolang (proyecto de I+D+i «PID2021-122263OB-C22», funded por MCIN/ AEI/10.13039/501100011033/ and "FEDER Una manera de hacer Europa").

References

[1] WeAreSocial, Hootsuite, DIGITAL REPORT ESPAÑA 2022, 2022. URL: https://encr.pw/8avSe.

[2] J. Flores Fernandez, Guía rápida para la prevención del acoso por medio de las nuevas tecnologías, 2008. URL: https://www.pantallasamigas.net/ciberbullying-guia-rapida.

[3] L. Alonso, V. J. Vázquez, Sobre la libertad de expresión y el discurso del odio: Textos críticos, Athenaica ediciones universitarias, 2017.

[4] S. Frenda, V. Patti, P. Rosso, Killing me softly: Creative and cognitive aspects of implicitness in abusive language online, Natural Language Engineering (2022) 1–22.

[5] S. Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, P. Rosso, The unbearable hurtfulness of sarcasm, Expert Systems with Applications 193 (2022) 116398.

[6] J. Sánchez-Junquera, P. Rosso, M. Montes, B. Chulvi, et al., Masking and bert-based models for stereotype identication, Procesamiento del Lenguaje Natural 67 (2021) 83–94.

[7] A. Bruns, After the 'apicalypse': Social media platforms and their fight against critical scholarly research, Information, Communication & Society 22 (2019) 1544–1566.

[8] B. L. Ott, The age of twitter: Donald j. trump and the politics of debasement, Critical studies in media communication 34 (2017) 59–68.

[9] G. R. McMenamin, Introducción a la lingüística forense: un libro de curso, Press at California State University, Fresno, 2017.

[10] M. R. Salado, Análisis lingüístico del discurso de odio en redes sociales, VISUAL REVIEW. International Visual Culture Review/Revista Internacional de Cultura Visual 9 (2022) 1–11.

[11] F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Detecting misogyny and xenophobia in spanish tweets using language technologies, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–19.

[12] N. D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering 10 (2015) 215–230.

[13] S. O. Sood, E. F. Churchill, J. Antin, Automatic identification of personal insults on social news sites, Journal of the American Society for Information Science and Technology 63 (2012) 270–285.

[14] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[15] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmore, Learning from bullying traces in social media, in: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2012, pp. 656–666.

[16] M. Dadvar, D. Trieschnigg, R. Ordelman, F. d. Jong, Improving cyberbullying detection with user context, in: European Conference on Information Retrieval, Springer, 2013, pp. 693–696.

[17] C. Arcila-Calderón, J. J. Amores, P. Sánchez-Holgado, D. Blanco-Herrero, Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on twitter in spanish, Multimodal technologies and interaction 5 (2021) 63.

[18] P. Burnap, M. L. Williams, Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making (2014).

[19] R. Martins, M. Gomes, J. J. Almeida, P. Novais, P. Henriques, Hate speech classification in social media using emotional analysis, Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018 (2018) 61–66. doi:10.1109/BRACIS.2018.00019.

[20] M. Wiegand, J. Ruppenhofer, A. Schmidt, C. Greenberg, Inducing a lexicon of abusive words–a feature-based approach, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Volume 1 (Long Papers), 2018, pp. 1046–1056.

[21] J. Qian, M. ElSherief, E. Belding, W. Y. Wang, Learning to decipher hate symbols, arXiv preprint arXiv:1904.02418 (2019).

[22] A. Olteanu, C. Castillo, J. Boy, K. Varshney, The effect of extremist violence on hateful speech online, in: Proceedings of the international AAAI conference on web and social media, volume 12, 2018.

[23] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[24] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.

[25] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, P. Nakov, A large-scale semi-supervised dataset for offensive language identification, arXiv preprint arXiv:2004.14454 (2020).

[26] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, CEUR-WS, 2018, pp. 1–6.

[27] F. M. Plaza-del Arco, A. B. P. Portillo, P. L. Úbeda, B. Gil, M.-T. Martín-Valdivia, Share: A lexicon of harmful expressions by spanish speakers, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1307–1316.

[28] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[29] D. Sarkar, M. Zampieri, T. Ranasinghe, A. Ororbia, Fbert: A neural transformer for identifying offensive content, arXiv preprint arXiv:2109.05074 (2021).

[30] B. Song, C. Pan, S. Wang, Z. Luo, Deepblueai at semeval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), 2021, pp. 1130–1134.

[31] B. Botella, R. Sepúlveda-Torres, P. M. Barco, E. Saquete, Violencia identificada en el lenguaje (vil). creación de recurso para mensajes violentos, Procesamiento del Lenguaje Natural 70 (2023) 187–198.

[32] B. B. Gil, F. M. P. del Arco, A. B. P. Portillo, Y. Gutiérrez, Fiero: Asistente virtual para la captación de insultos, volume 2968 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2968/paper8.pdf.