Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 159-164 recibido 15-12-2023 revisado 20-01-2024 aceptado 27-01-2024

# Detecting offensive language by integrating multiple linguistic phenomena

## *Detección del lenguaje ofensivo mediante la integración de diferentes fenómenos lingüísticos*

**Flor Miriam Plaza-del-Arco**
[1]Bocconi University, Milan, Italy
flor.plaza@unibocconi.it

**Abstract:** This is a summary of the Ph.D. thesis conducted by Flor Miriam Plaza del Arco at the University of Jaén under the supervision of Ph.D. M. Teresa Martín Valdivia and Ph.D. L. Alfonso Ureña López. The thesis defense took place in Jaén on January 30, 2023, with the doctoral committee comprising Ph.D. Mariona Taulé Delor from the University of Barcelona, Ph.D. José Camacho Collados from the University of Cardiff, and Ph.D. Eugenio Martínez Cámara from the University of Granada. Notably, the thesis was awarded the distinction of Summa Cum Laude and received international recognition.
**Keywords:** Natural language processing, Spanish corpora, offensive language detection, hate speech, multitask learning, linguistic phenomena

**Resumen:** Este es un resumen de la tesis doctoral realizada por Flor Miriam Plaza del Arco en la Universidad de Jaén, bajo la supervisión de la Dra. M. Teresa Martín Valdivia y el Dr. L. Alfonso Ureña López. La defensa de la tesis tuvo lugar en Jaén el 30 de enero de 2023 y la comisión de doctorado estuvo formada por la Dra. Mariona Taulé Delor de la Universidad de Barcelona, el Dr. José Camacho Collados de la Universidad de Cardiff y el Dr. Eugenio Martínez Cámara de la Universidad de Granada. Cabe destacar que la tesis obtuvo la calificación de Summa Cum Laude y la mención internacional.
**Palabras clave:** Procesamiento del Lenguaje Natural, corpus en español, detección del lenguaje ofensivo, discurso de odio, aprendizaje multitarea, fenómenos lingüísticos

### *1 Introduction*

One of the characteristics that distinguish humans from other living beings is the ability to communicate systematically and understandably, i.e. through language. Language is defined as a sophisticated system of both phonetic and written symbols that allows two or more individuals to communicate ideas, thoughts, sentiments, attitudes, and different situations. Since the emergence of Web 2.0, users were no longer limited to face-to-face communication but rather used online platforms to interact. This interaction has resulted in an increasing amount of textual data being available on the Web. Natural Language Processing, a tract of Artificial Intelligence and Linguistics, arises for the development of computational systems to interpret human language and thus enable human-computer interaction. Giving computers this skill offers a plethora of benefits, including the potential to moderate harmful conduct on social media.

**This doctoral thesis focuses on both the creation of linguistic resources and the development of NLP-based techniques to aid in the automatic detection of offensive language on the Web**. On the one hand, for the development of these techniques, data labeled are essential to learning the language patterns characteristic of this behavior; however, the available resources are mainly focused on English, leaving aside other languages such as Spanish with very scarce or non-existent resources of this nature. Therefore, a fundamental part of this doctoral thesis is focused on the generation of these resources for Spanish. On the

ISSN 1135-5948 DOI 10.26342/2024-72-13 

other hand, for the implementation of automatic systems based on NLP, one of the main contributions has been the integration of different linguistic phenomena that might be involved in the expression of offensiveness in computational systems. In particular, we developed a Multitask Learning (MTL) method based on Transfer Learning (TL). We believe that this methodology plays an important role in their application to the detection of more specific problems in our society, such as Hate Speech (HS), misogyny, or sexism, that have been addressed in the frame of this doctoral thesis. As a result, it should be mentioned that this thesis has both a social and technological dimension to contribute to society's improvement.

## 1.1 Motivation

Social media have grown into the primary means of communicating between people, allowing users to have conversations, share opinions, and create content. The rise in digital social connections has led to the dissemination of harmful communication, which is sometimes aided by the anonymity afforded by these platforms (Aguilera-Carnerero and Azeez, 2016). As a consequence, offensive language and one of its most damaging forms, HS, tends to proliferate swiftly and is difficult to regulate. For instance, according to a Spanish report in 2020 on the evolution of hate crimes in Spain[1], threats, insults, and discrimination are counted as the most repeated criminal acts, with the Internet (45%) and social media (22.8%) as the most widely used media to commit these actions. Similarly, a recent survey on hate crimes in Spain 2021[2] shows that 41.65% of the participants, out of a total of 437, have been victims of hate crimes on more than one occasion in the last 5 years. On the one hand, they have received offensive comments on more than 10 occasions. On the other hand, more than 50% of them have received offenses or threats through social networks or the Internet. Finally, more than 70% of the respondents have received discriminatory treatment on one or more occasions in the last 5 years.

In this regard, inaction against offensive language allows for the further reinforcement of prejudices and stereotypes, while this type of hostile communication may lead to negative psychological effects among online users, causing anxiety, harassment, and, in extreme cases, suicide (Hinduja and Patchin, 2010). As a result, this scenario has motivated interested stakeholders (governments, online communities, and social media platforms) to look for efficient solutions to prevent Internet hostility. One strategy used to tackle this problem is through legislation, by implementing laws and policies. For instance, since 2013 the Council of Europe has sponsored the "No Hate Speech" movement[3] seeking to mobilize young people to combat HS and promote human rights online. In May 2016, the European Commission reached an agreement with Facebook, Microsoft, Twitter, and YouTube to implement the "Code of Conduct on countering illegal HS online"[4]. From 2018 to 2020, platforms such as Instagram, Snapchat, and TikTok adopted the Code. One of the initial and most common approaches to hatred intervention adopted by social media platforms is content moderation. This approach is based on the suspension of user accounts and the removal of hate messages while attempting to balance the right to freedom of expression.

Although these approaches have the clear advantage of analyzing the context and accurately identifying this behavior, still these strategies do not seem to achieve the desired effect because they involve an intense, time-consuming, and costly procedure that limits its scalability and quick solutions. At the same time, hate content is continuously growing and adapting, making it harder to identify (Davidson et al., 2017). As a result of these challenges, an alternative and preferable option is to rely on NLP-based methods to automatically detect this type of harmful online communication. Advances in NLP can be used to detect offensive content online thus decreasing the time and effort in fighting this problem. Offensive language detection and analysis has become a major area of research in NLP. However, existing NLP-based methods face several drawbacks. Firstly, detecting offensive content is challenging for machines (Zampieri et al., 2019; Wiegand, Ruppenhofer, and Kleinbauer, 2019; Poletto et al., 2020), since this type of language presents a subjective nature as well as social and cultural implications. Though recent

[1] https://shorturl.at/hlnAX
[2] https://shorturl.at/mpxLR
[3] https://shorturl.at/DQ345
[4] https://shorturl.at/kvHOT

approaches of sequence-to-sequence models (Zampieri et al., 2020; Tontodimamma et al., 2021) have achieved good performance in detecting this type of content, most of them have not considered linguistic phenomena that may occur in the expression of offensive language such as those of an implicit nature such as sarcasm and irony (Chauhan et al., 2020; Wiegand, Ruppenhofer, and Eder, 2021). Secondly, since most of the available corpora contain messages from the Twitter platform, automatic systems have specialized in learning the language style and register used by the users on this platform, making cross-domain transfer difficult when using such systems on other platforms. Thirdly, so far most of the research to solve this problem has been focused on English (Fortuna and Nunes, 2018), leaving other languages such as Spanish in second place, although combating this type of behavior is a global concern.

These challenges motivate this doctoral thesis to explore methods for accurately detecting offensive language on the Web using NLP techniques to aid in this process. **This thesis relies on advanced methods in NLP such as deep learning to tackle this issue**. First, it faces the problem of limited training data, especially in Spanish, generating appropriate resources to combat offensive textual content. These resources will also help to solve the limitation of the systems specialized in Twitter since messages from other social platforms such as YouTube and Instagram are considered. Secondly, it introduces different linguistic phenomena that could be involved in the expression of offensiveness and could help in the detection of this content. Then, a novel method is proposed where these identified phenomena are integrated for the detection of offensive language, using state-of-the-art techniques based on transfer learning. Finally, this novel method is applied for the detection of different offensive language scenarios (HS, sexism, toxicity), analyzing which specific linguistic phenomena are beneficial in each of them.

## 1.2 Hypotheses

This thesis studies the problem of automatically detecting offensive textual language with deep learning techniques for NLP. The main hypothesis of this thesis is the following: **Advanced NLP methods based on deep learning, in particular transfer learning, aid in the detection of offensive textual language. We subdivide this hypothesis into three hypotheses:**

**Hypothesis 1 (H1)** The subjective nature of offensive language can have strong cultural, demographic, and social implications, and therefore language-specific resources and models are required.

**Hypothesis 2 (H2)** Transfer learning models leveraging linguistic phenomena related to offensive language expression outperform those that do not integrate this information in offensive language detection tasks.

**Hypothesis 3 (H3)** Incorporating specific linguistic phenomena into transfer learning methodologies can enhance the detection of various offensive scenarios. Offensive language detection encompasses a range of scenarios, such as identifying sexist content, hate speech, or toxic language.

## *2 Thesis outline*

This thesis is structured into 8 chapters, outlined as follows:

- **Chapter 2** includes an overview of the background information that is significant for understanding the content of this thesis. We review traditional ML and Neural Network (NN) based methods for offensive language research in NLP. We furthermore provide a compilation of different existing resources labeled with offensiveness. Then, we present the research challenges and opportunities based on the previous research approaches reviewed.

- **Chapter 3** introduces our preliminary research in the thesis, focusing mainly on traditional ML approaches to address HS detection, including misogyny and xenophobia. In addition, we present the first experiments with monolingual and multilingual pre-trained language models based on Transformers.

- **Chapter 4** describes the different corpora and lexicons we generate during the thesis for the research on offensive language and emotion analysis. Specifically, three corpora and three lexicons, mainly focused on Spanish, are presented.

- **Chapter 5** introduces our contribution to addressing offensive language detection. We propose a novel approach that uses the MTL paradigm to combine different phenomena inextricably related to the expression of offensive language. This approach aims to benefit from shared knowledge across tasks to improve the detection of offensive language. We identify some linguistic phenomena that might be involved in the expression of offensive language and present initial experiments.
- **Chapter 6** focuses on the evaluation of the proposed MTL learning approach in different offensive language scenarios studying the integration of the linguistic phenomena defined in Chapter 5. We show the success of our MTL methodology by comparing its performance with previous state-of-the-art approaches that do not consider this useful information.
- **Chapter 7** presents two different shared tasks organized in the framework of this doctoral thesis to promote the research on emotion analysis and offensive language detection in Spanish. The task descriptions, the corpora and evaluation measures used as well as the participants and results achieved are described.
- **Chapter 8** finally summarizes our conclusions where we present the main findings of this doctoral thesis and suggest future research directions within offensive language research.

## *3 Main contributions*

The research conducted in this doctoral thesis has resulted in several contributions that support the hypothesis outlined in Section 1.2.

**Contributions to support H1:**

**Contribution 1** The generation of different linguistic resources for offensive language research and emotion analysis focused mainly on Spanish (Plaza-del-Arco et al., 2020; Plaza-del-Arco et al., 2021; Plaza-del-Arco et al., 2022).

**Contribution 2** We have developed our annotation scheme for each of the resources generated.

**Contribution 3** Using the resources generated, we have organized different shared tasks in the IberLEF evaluation campaign to promote offensive language research in Spanish (Plaza-del-Arco et al., 2021a; Plaza-del-Arco et al., 2021).

**Contributions to support H2:**

**Contribution 4** We have identified different linguistic phenomena that might be involved in the expression of the offense.

**Contribution 5** We have proposed the main methodology conducted in this doctoral thesis which follows an MTL paradigm and relies on integrating the selected linguistic phenomena in a comprehensive computational system for detecting offensive language more accurately (Plaza-del-Arco et al., 2021; Plaza-del-Arco et al., 2022).

**Contributions to support H3:**

**Contribution 6** We have applied the proposed approach to different scenarios involved in offensive language research including sexism, hate speech, and toxicity.

**Contribution 7** We have analyzed which linguistic phenomena benefit the most in each scenario through extensive experiments. We have provided a valuable discussion with the primary findings for each scenario (Plaza-del-Arco et al., 2021c; Plaza-del-Arco et al., 2021b; Plaza-del-Arco et al., 2022; Plaza-del-Arco et al., 2022)

**Contribution 8** The superior performance of our proposed approach over the previous state-of-the-art approaches.

## *Acknowledgments*

This research has been partially supported by the scholarship (FPI-PRE2019-089310) from the Ministry of Science, Innovation, and Universities, the LIVING-LANG project (RTI2018-094653-B-C21), and RED-ES project (TIN2015-65136-C2-1-R) from the Spanish Government.

## *References*

Aguilera-Carnerero, C. and A. H. Azeez. 2016. 'Islamonausea, not Islamophobia': The many faces of cyber hate speech. *Journal of Arab & Muslim media research*, 9(1):21–40.

Chauhan, D. S., D. S R, A. Ekbal, and P. Bhattacharyya. 2020. Sentiment and

Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online, July. Association for Computational Linguistics.

Davidson, T., D. Warmsley, M. W. Macy, and I. Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

Fortuna, P. and S. Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4), jul.

Hinduja, S. and J. W. Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221.

Plaza-del-Arco, F., M. Casavantes, H. Escalante, M. T. Martín-Valdivia, A. Montejo-Ráez, M. Montes-y-Gómez, H. Jarquín-Vásquez, and L. Villaseñor-Pineda. 2021. Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67.

Plaza-del-Arco, F. M., S. Halat, S. Padó, and R. Klinger. 2022. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. In *ACM SIGIR Special Interest Group on Information Retrieval TCS Research*.

Plaza-del-Arco, F. M., S. M. Jiménez-Zafra, A. Montejo-Ráez, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021a. Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67:155–161.

Plaza-del-Arco, F. M., M. D. Molina-González, U.-L. L. Alfonso, and M. V. M. Teresa. 2021b. Sinai at iberlef-2021 detoxis task: Exploring features as tasks in a multi-task learning approach to detecting toxic comments. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, volume 21, pages 580–590.

Plaza-del-Arco, F. M., M. D. Molina-González, L. López, and M. Martín-Valdivia. 2021c. Sexism identification in social networks using a multi-task learning system. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*.

Plaza-del-Arco, F. M., M. D. Molina-González, L. López, and M. Martín-Valdivia. 2022. Exploring the use of different linguistic phenomena for sexism identification in social networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022)*, volume 2943, pages 491–499.

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M.-T. Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965.

Plaza-del-Arco, F. M., A. Montejo-Ráez, L. A. Ureña-López, and M.-T. Martín-Valdivia. 2021. OffendES: A New Corpus in Spanish for Offensive Language Research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108, Held Online, September. INCOMA Ltd.

Plaza-del-Arco, F., A. B. Parras Portillo, P. López-Úbeda, B. Botella-Gil, and M. T. Martín-Valdivia. 2022. SHARE: A Lexicon of Harmful Expressions by Spanish Speakers. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1307–1316, Marseille, France, June. European Language Resources Association.

Plaza-del-Arco, F., C. Strapparava, L. A. Ureña-López, and M. T. Martín-Valdivia. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages

1492–1498, Marseille, France, May. European Language Resources Association.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Tontodimamma, A., E. Nissi, A. Sarra, and L. Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179, January.

Wiegand, M., J. Ruppenhofer, and E. Eder. 2021. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online, June. Association for Computational Linguistics.

Wiegand, M., J. Ruppenhofer, and T. Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zampieri, M., P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), December. International Committee for Computational Linguistics.