Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 239-242 recibido 16-12-2022 revisado 01-02-2023 aceptado 11-02-2023

Sarcasm and Implicitness in Abusive Language Detection: A Multilingual Perspective

Sarcasmo e implícidad en el reconocimiento automático del lenguaje abusivo en una perspectiva multilingüe

Simona Frenda[1,2]
[1]PRHLT Research Center, Universitat Politècnica de València, Spain
[2]Dipartimento di Informatica, Università degli Studi di Torino, Italy
simona.frenda@unito.it

Abstract: PhD thesis in Computer Science focused on Natural Language Processing, written by Simona Frenda under the supervision of Prof. Viviana Patti and Prof. Paolo Rosso. This thesis was developed in a co-tutelle program between the PRHLT Research Center of the Universitat Politècnica de València (Spain) and the Computer Science Department of the University of Turin (Italy). In this work, we analysed, linguistically and computationally, the characteristics of the implicit abusive language, especially when it is masked as sarcastic. The thesis defence was held in Torino on June 6th, 2022. The doctoral committee was composed by: Prof. Liviu Petrisor Dinu (University of Bucharest, Romania), Prof. Els Lefever (Ghent University, Belgium) and Prof. Elena Cabrio (Université Côte d'Azur, France). An international mention was achieved, and the work was graded as excellent and awarded Cum Laude.
Keywords: Natural Language Processing, Computational Linguistics, Abusive Language Detection, Irony Detection, Stance Detection.

Resumen: Tesis doctoral en Informática con tema en Procesamiento del Lenguaje Natural realizada por Simona Frenda y dirigida por la Profa. Viviana Patti y el Prof. Paolo Rosso en el marco de un convenio de cotutela entre el PRHLT Research Center de la Universitat Politècnica de València (España) y el Departamento de Informática de la Universidad de Turín (Italia). En esta tesis se analiza a nivel lingüístico y computacional las características del lenguaje abusivo implícito, especialmente cuando está disfrazado como sarcástico. La defensa de la tesis fue en Turín el 6 de junio de 2022 ante un tribunal compuesto por el Prof. Liviu Petrisor Dinu (Universidad de Bucarest, Rumania), la Profa. Els Lefever (Universidad de Ghent, Bélgica), y la Profa. Elena Cabrio (Universidad de Côte d'Azur, Francia). Se obtuvo la mención internacional y una calificación de sobresaliente cum laude.
Palabras clave: Procesamiento del lenguaje natural, Lingüística computacional, Detección del lenguaje abusivo, Detección de la ironía, Detección de la stance.

1 Introduction

The possibility to monitor hateful content online on the basis of what people write is becoming an important topic for several actors such as governments, ICT companies, and NGO's operators conducting active campaigns in response to the worrying rise of online abuse and hate speech. Hand in hand, abusive language detection turns into a task of growing interest in Natural Language Processing (NLP), especially when applied to the recognition of various forms of hatred in social media posts. Abusive language is a broad umbrella term which is commonly used for denoting different kinds of hostile user-generated contents that intimidate or incite to violence and hatred, targeting many vulnerable groups in social platforms (Poletto et al., 2021). Such hateful contents are pervasive nowadays and can also be detected even in other kinds of texts, such as online newspapers. The importance of understanding and automatically detecting abusive language is due to the observation of real manifestations

ISSN 1135-5948 DOI 10.26342/2023-70-21

of violent acts connected to negative behaviours online in its various forms, such as cyberbullying, racism, sexism, or homophobia. Various approaches have been proposed in the last years to support the identification and monitoring of these phenomena, but unfortunately, they are far from solving the problem due to the inner complexity of abusive language, and to the difficulties to detecting its implicit forms (Wiegand, Ruppenhofer, and Eder, 2021).

In our doctoral investigation, we have studied the issues related to automatic identification of abusive language online, investigating various forms of hostility against women, immigrants and cultural minority communities in languages such as Italian, English, and Spanish. The analysis of the results of different methods of classification of hateful and non-hateful messages revealed important challenges that lie principally on the implicitness of some manifestations of abusive language expressed through the use of figurative devices (i.e., irony and sarcasm) (Frenda et al., 2022), recall of inner ideologies (i.e., sexist ideology) (Frenda et al., 2019a) or cognitive schemas (i.e., stereotypes) (Frenda, Patti, and Rosso, 2022), and expression of unfavourable stance (Frenda et al., 2019b).

To face these challenges, we have proposed distinct solutions applicable also to different textual genres. We observed that, in particular, cognitive (i.e., stereotypes) and creative aspects (i.e., sarcasm) of abusive language are harder to infer automatically from texts. Sarcasm, for instance, is a recurrent element in this kind of texts, and tends to affect the accuracy of the systems of recognition (Frenda, 2018). Indeed, for its peculiarities, sarcasm is apt to disguise hurtful messages, especially in short and informal texts such as the ones posted on Twitter. Its ironic sharpness and its echoic function of recalling a meaning that is the opposite or an extension of the literal one, make sarcasm appropriate to lower tones without losing the hurtfulness of the message. Moreover, funny messages are more likely to be accepted and shared by the community, making the abuse viral. Therefore, our hypothesis is that **information about the presence of sarcasm could help to improve the detection of hateful messages, even when they are camouflaged as sarcastic**. To verify it, we elaborated specific research questions:

RQ1 How to make abusive language detection systems sensitive to implicit manifestations of hate?
RQ2 What is the role played by sarcasm in hateful messages online?
RQ3 Could the awareness of the presence of sarcasm increase the performance of abusive language detection systems?

Focusing on these questions, 1) we investigated the characteristics of implicit manifestations of hate speech and examined, in terms of performance, the techniques that could help systems to infer them, such as the use of lexical resources, specific models to capture semantic relations, and the use of transfer learning techniques combined with linguistic features; 2) we analysed the role of ironic language in hateful texts, observing the multilingual characteristics of irony and especially of sarcasm, validating, with experiments of classification, these traits in terms of features; 3) we evaluated the benefits of ironic awareness in hate speech detection exploiting computational techniques that make systems aware of ironic language, such as the multi-task learning approach. This technique enables systems of abusive language detection to acquire specific knowledge about ironic language. Finally, we measured the significance of the obtained results in comparison to existing approaches and baseline models.

The corpora used in our experiments have been exploited as benchmark datasets within the EVALITA evaluation campaign for NLP tools for Italian, contributing to creating a new state of the art for these tasks in Italian: IronITA 2018 (Cignarella et al., 2018) and HaSpeeDe 2020 (Sanguinetti et al., 2020). Moreover, the multidisciplinary and multilingual frame of our analyses allowed us to reflect on the boundaries between dimensions and topical focuses that often overlap in computational approaches to detect abusive language and related phenomena.

2 Thesis Overview

The work presented in the thesis has been organized in 7 chapters grouped in 3 principal parts.

I part: Abusive Language Detection Chapter 1. The first chapter is the introductory section, where we described the social problems related to the new technologies, introducing the issue of the *abusive language* and the difficulties to detect it automatically.

Chapter 2 In the second chapter, we defined the concept of *abusive language*, looking at the juridical and linguistic theories. Moreover, we resumed the state of the art from a computational perspective, focusing especially on the open challenge of implicit abusive language detection.

Chapter 3 In the third chapter, we reported the linguistic, statistical and computational analysis performed on benchmark datasets to individuate the characteristics of the explicit and implicit manifestations of hate speech. Additionally, we described the linguistic resources created manually, and the designed approaches that make systems able to infer indirect abusive messages such as negative stereotypes (**RQ1**). Finally, we presented the second edition of the HaSpeeDe[1] shared task organized at EVALITA 2020 on hate speech and stereotypes detection in Italian tweets and news headlines.

II part: Irony and Sarcasm Detection

Chapter 4 In the fourth chapter, we defined what is *ironic language*, looking at the linguistic theories stretching from pragmatic to cognitive studies. In addition, we introduced the state of the art on irony and sarcasm detection, focusing especially on studies that analysed the peculiarities of sarcasm.

Chapter 5 In the fifth chapter, we proposed statistical and computational analysis to individuate the characteristics of irony and sarcasm. We observed linguistic traits of irony from a mono and multilingual perspective, and emotional and aggressive language involved in the expression of irony and sarcasm, especially when the topic of the text regards controversial issues such as the integration of cultural minorities (**RQ2**). In this chapter, we described also our experience as organizers of the IronITA[2] shared task at EVALITA 2018 on irony and sarcasm detection.

III part: Abusive and Ironic Language

Chapter 6 Taking into account the findings emerged from previous chapters, in the sixth one, we proposed a new computational approach that exploits the simultaneous learning from abusive and ironic language to detect hate speech in Italian tweets and news headlines. The results showed an improvement of the performance (33 % Δ), especially in hate speech detection in tweets, evaluated as significant (below a cut-off of 0,05) for all the metrics by means of a bootstrap sampling significance test (**RQ3**).

Chapter 7 In the last chapter, we reported the obtained results and the observations emerged from our analyses. We individuated the remaining challenges that we plan to address in further works, and we summarized the contributions to the NLP community in terms of findings, methodologies, resources, and publications.

3 Conclusions and Contributions

Various scholars in linguistics stress the mutual relation between language and society, composed of the speakers of that language. Actually, with *words* we are not just speaking, but we do things, things that could help people or things that could marginalize or hurt people. Our investigation aimed at contributing to the comprehension of how abuses, such as misogyny and racism, are expressed directly and indirectly, and how they could be recognized by machines.

The corpora-based analysis, the statistical tests and computational experiments on various benchmark datasets showed that abusive language towards women and immigrants involves important social biases that appear to be pervasive even in discussions that involve these targets. Another recurrent element is the presence of irony in these messages, used to lessen the social cost of their meaning. To make the systems of abusive language detection aware of stereotypes or prejudices, we experimented various approaches, discovering that especially lexica-based features are very useful even in the systems with neural architectures. Approaching abusive language detection as a classification problem, we noticed that one of the points that remained unsolved was related to the presence of ironic devices. Irony, in fact, is used to mask the purpose of haters to insult specific vulnerable targets. Ironic texts have been found to be aggressive, above all when the sarcastic form of irony is employed; proving, therefore, some arguments in favour of linguistic and pragmatic theories (Bowes and Katz, 2011).

Considering that, we designed a new approach of detection, exploiting the presence of irony in manual annotated texts. We designed a system that fine-tune Italian language models simultaneously on the tasks of hateful and ironic language recognition in a multi-task framework. We compared its re-

[1] http://di.unito.it/haspeede2020

[2] http://di.unito.it/ironita2018

sults with the one obtained with the previous approach that combines general knowledge, coming from language models, and linguistic information, provided by means of specific features. We discovered that the awareness of sarcasm helps the system to retrieve correctly hate speech in social media texts, such as tweets; and that linguistic features make the system sensible to stereotypes in both tweets and news headlines. Finally, our research questions encouraged also the investigation about irony and its manifestations in various contexts. Therefore, our analyses contributed also to the more theoretical and linguistic discussion on: 1) the peculiarities of sarcasm compared to other forms of irony, and 2) mono and multilingual characteristics of irony. Sarcasm, defined in literature as a sharp form of irony with the intent of scorning a victim, proved to be characterized by: hurtful language, explicit contradictions marked with adverbial locutions, semantic and polarity shifts, and false assertions and euphemistic forms. The computational experiments carried out on irony detection revealed, instead, that negative emotions are involved in the expression of irony, regardless of the language, the context, and the genre.

Although some important issues in Abusive Language detection have been addressed in this work, other challenges remain open for further investigation, such as: the misclassification of texts containing swear words used with non-abusive intent (surprise, friendly nicknames); and the processing of texts only at message level leaving unexplored the contextual information that could help to give a more informed perspective to interpret them as abuses or not.

Acknowledgments

The work in this thesis was partially supported by various financial projects. Among them: the Spanish research project SomEMBED funded by Ministerio de Economía y Sostenibilidad, the NII International Internship Program funded by JSPS KAKENHI, the Italian project M.EMO.RAI funded by RAI - Radiotelevisione Italiana Spa, the Italian project IhatePrejudice funded by Compagnia di San Paolo, and the European project “STERHEOTYPES” funded by Compagnia di San Paolo Foundation, Volkswagen Stiftung and Carlsberg Fondation.

References

Bowes, A. and A. Katz. 2011. When sarcasm stings. *Discourse Processes: A Multidisciplinary Journal*, 48(4):215–236.

Cignarella, A. T., S. Frenda, V. Basile, C. Bosco, V. Patti, and P. Rosso. 2018. Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA). In *EVALITA 2018*, volume 2263. CEUR-WS.

Frenda, S. 2018. The role of sarcasm in hate speech: A multilingual perspective. In *Proceedings of Doctoral Symposium at SEPLN 2018*. CEUR-WS.

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

Frenda, S., B. Ghanem, M. Montes-y Gómez, and P. Rosso. 2019a. Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Frenda, S., K. Noriko, V. Patti, P. Rosso, et al. 2019b. Stance or insults? In *Ninth International Workshop on Evaluating Information Access*, pages 15–22. National Institute of Informatics.

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

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

Sanguinetti, M., G. Comandini, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. 2020. Haspeede 2@ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In *EVALITA 2020*. CEUR.

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 NAACL: Human Language Technologies*, pages 576–587.