The role of sarcasm in hate speech. A multilingual perspective

La función del sarcasmo en los discursos de odio. 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: The importance of the detection of aggressiveness in social media is due to real effects of violence provoked by negative behavior online. For this reason, hate speech online is a real problem in modern society and the necessity of control of user-generated contents has become one of the priorities for governments, social media platforms and Internet companies. Current methodologies are far from solving this problem. Indeed, several aggressive comments are also disguised as sarcastic. In this perspective, this research proposal wants to investigate the role played by creative linguistic devices, especially sarcasm, in hate speech in multilingual context.
Keywords: Hate speech, social media, aggressiveness, misogyny, sarcasm

Resumen: La importancia del reconocimiento de la agresividad en las redes sociales es debido al hecho que esas conductas negativas se traducen en violencias en la vida real también. Por esa razón los discursos de odio online son un problema real en nuestra sociedad y la necesidad del control de los contenidos generados por usuarios se ha convertido en una de las prioridades de gobiernos, de las redes sociales y de empresas de Internet. Las metodologías corrientes están lejos de resolver este problema. De hecho gran parte de los comentarios agresivos son disfrazados como sarcasticos. En esta perspectiva, esta propuesta de investigación propone de estudiar la función de las figuras retóricas, con particular atención al sarcasmo, en los discursos de odio en un contexto multilingüe.
Palabras clave: Discursos de odio, redes sociales, agressividad, misoginia, sarcasmo

1 Introduction

The web facilitates the large resonance of hate speech, inciting racism, misogyny or xenophobia also in the real world. Actually, it is common that misbehaviours online are traduced in physical attacks, such as rapes or bulling. For instance, Fulper et al. (2014) demonstrated the existence of a correlation between the number of rapes and the amount of misogynistic tweets per state in USA, suggesting the fact that social media can be used as a social sensor of violence.

In addition, the persistence and diffusion of misogynistic or offensive content can hurt and distress psychologically the victims, causing sometime their suicide, such as the case of the teenager Amanda Todd in 2012[1]. In order to contrast the origin of these hate events and to monitor the uncontrolled flow of users texts, several initiatives have been taken in the last years. An example is the campaign *No Hate Speech Movement*[2] of the Council of Europe for human rights online.

The growing interest of NLP (Natural Language Processing) research community is demonstrated by the proposal of national and international workshops (such as ALW 2018[3]) or campaigns of evaluation fostering the research in this issue in various languages, such as EvalIta 2018[4], IberEval 2018[5] and SemEval 2019[6]. These initiatives allow to share

[1] https://www.theguardian.com/commentisfree/2012/oct/26/amanda-todd-suicide-social-media-sexualisation
[2] https://www.coe.int/en/web/no-hate-campaign
[3] https://sites.google.com/view/alw2018
[4] http://www.evalita.it/2018
[5] https://sites.google.com/view/ibereval-2018
[6] http://alt.qcri.org/semeval2019/index.

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Moreno, I. (eds.) *Proceedings of the Doctoral Symposium of the XXXIV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, p. 13–17 Sevilla, Spain, September 19th 2018.

information and results exploring the different topics regarding the hate speech online. As well as, the organizers of these competitions provide resources such as annotated datasets that are very costly to obtain.

The fact that the majority of data are collected from Twitter or Facebook supports the analysis of the computer-mediated communication. As well as, the context of short text incites the creativity of authors who use figurative devices to express their opinion. One of the most used figures of speech to manifest negative opinions is the sarcasm. In fact, it is used to disguise and, at the same time, to reinforce the negative thinking, such as:

i) *Un pensiero di ringraziamento ogni mattina va sempre ai comunisti che ce li hanno portati fino a casa musulmani rom e delinquenti grazie*[7].

The ironic sharpness of the sarcasm seems to be appropriated to express contempt and to offend individuals subtly. In order to study this correlation between sarcasm and hate speech, we proposed the shared task IronITA[8] at Evalita 2018 that asks participants to recognize ironic and sarcastic tweets in a dataset containing also offensive messages addressed, especially, immigrants (Sanguinetti et al., 2018).

Moreover, we participated in two tasks proposed at IberEval 2018 about hate speech: aggressiveness detection in Mexican Spanish tweets (MEX-A3T)[9] organized by Álvarez-Carmona et al. (2018) and identification of misogynistic English and Spanish tweets (AMI)[10] organized by Fersini, Anzovino, and Rosso (2018). As a confirmation of our intuition, the systems proposed for these tasks show some difficulties to classify the sarcastic abusive tweets. Indeed, sarcasm, independently from the differences between languages, disguises the real intention of the message which is with difficulty recognized by machine. In line with these early experiments, IronITA could be a good step of analysis.

The rest of the paper is structured as follows. Section 2 introduces the literature that inspired our investigation. Section 3 describes our participation in IberEval tasks with the used approach and obtained results. In Section 4 we analyze the presence of sarcasm in analyzed aggressive and offensive texts. Finally, in Section 5 and 6 we draw our research proposal and the future work.

2 Related work

The literature about hate speech detection includes different issues, such as: cyberbullying, misogyny, nastiness and aggressiveness. The most commercial methods, currently, rely on the use of blacklists. However, filtering the messages in this way does not provide a sufficient remedy because it falls short when the meaning is more subtle or altered by sarcasm. Actually, some authors, such as Justo et al. (2014) and Nobata et al. (2016), underline the fact that sarcasm makes the interpretation of the message difficult, generally requiring world knowledge. Also Smokey, one of the first systems, implemented by Spertus (1997), uses syntactic and semantic rules with lexicons to recognize flames.

In this context, the research is oriented at investigating deeply the language using classical (Samghabadi et al., 2017) and deep learning methods (Del Vigna et al., 2017). Differently from Mehdad and Tetreault (2016) and Gambäck and Sikdar (2017), for MEX-A3T task in Frenda and Banerjee (2018) we applied an experimental technique that combines linguistic features and Convolutional Neural Network (CNN).

For the first time, Anzovino, Fersini, and Rosso (2018) propose a classical machine learning approach to identify misogyny in English, comparing different classifiers. Taking into account this previous work and the psychological studies about sexism (Ford and Boxer, 2011), in Frenda and Ghanem (2018) we combined sentiment and stylistic information with specific lexicons involving several aspects of misogyny online.

In the following section we report how we addressed the identification of aggressiveness and misogyny in Twitter, the experiments carried out and the results obtained.

php?id=tasks

[7] *Each morning, I would like to thank communists who bring home musulmans, roms and delinquents thanks.* Tweet from IronITA corpus.

[8] http://di.unito.it/ironita18

[9] https://mexa3t.wixsite.com/home/aggressive-detection-track

[10] https://amiibereval2018.wordpress.com/

3 *Hate speech, aggressiveness and misogyny*

Considering our motivations, our early experiments focus mainly on hate speech detection. For this purpose, we participated at two tasks at IberEval 2018 respectively about aggressiveness and misogyny detection.

3.1 Aggressiveness detection

The first task aims to classify aggressive and non-aggressive tweets in Mexican Spanish. We applied a deep learning approach incorporating into CNN architecture a set of linguistic features (DL+FE) concerning: proper characteristics of a tweet, such as emoticons, abbreviations and slang words; stylistic information, such as the length of tweets, the use of the punctuation and the uppercase characters; bags of words weighted with tf-idf; emotive traits of the aggressiveness; and derogatory adjectives and vulgar expressions typical of Mexican culture.

By means of Information Gain, we noticed that anger and disgust are the principal emotions that incite the aggressive behaviour. We compared this system with a simple CNN architecture (DL) in order to understand the contribution of features to deep learning approach. The measure used for the competition is F-score for positive class (i.e. aggressive class). Despite the novel approach, the results obtained are low and the features seem not to help deep learning, as showed in Table 1.

	Prec.	*Rec.*	*F-pos*	*Rank*
DL	0.34	0.34	**0.34**	9
DL+FE	0.27	0.38	0.31	10

Table 1: Results for aggressiveness detection

Therefore, in order to understand what are the difficulties of DL+FE, we carried out the error analysis. We mainly noticed that there are several humorous cases, especially sarcastic (see Section 4), which are misclassified.

3.2 Automatic misogyny identification

The second task proposes to identify misogyny in two collection of English and Spanish tweets. In the case a tweet is classified as misogynistic (Task A), we need to distinguish (Task B) if the target is an individual or not (*Tar.*) and identify the type of misogyny, according to the following classes (*Cat.*): stereotype and objectification, dominance, derailing, sexual harassment and threats of violence, and discredit. This subdivision of misogyny allows us to explore the different aspects of misogyny and compare them in two different languages. Moreover, the data are not geolocalized. Therefore, in order to gather the linguistic variations and consider the various traits of misogyny, we proposed an approach based on stylistic features captured by means of the character n-grams, sentiment and affective information, and on a set of lexicons concerning: sexuality, profanity, femininity, human body and stereotypes. In addition, we considered slangs, abbreviations and hashtags.

By means of Information Gain, we discovered some differences between the two languages: sexual language is more used in English misogynistic tweets, whereas profanities or vulgarities are more used in Spanish ones. For this task, we applied Support Vector Machine (SVM) and majority voting technique. To evaluate the Task A the organizers used Accuracy measure and for Task B the average Macro-F1 measure. In Table 2 and Table 3 we report the promising results obtained with better runs for both languages.

	Approach	*Acc*	*Rank*
En	Ensemble	**0.87**	2
Sp	Ensemble	**0.81**	3

Table 2: Results for Task A of misogyny detection

	Approach	*F1*	*Cat.*	*Tar.*	*Rank*
En	SVM	**0.44**	0.29	0.59	1
Sp	Ensemble	**0.44**	0.33	0.55	2

Table 3: Results for Task B of misogyny detection

4 *Sarcasm*

In *Traité des tropes* (1729) Dumarsais has defined the sarcasm as an *ironie faite avec ai-*

greur et emportement[11], that is a kind of aggressive and sharp irony addressed a target to hurt or criticize him without to exclude the possibility to amuse. This statement is corroborate by our analyses on English, Spanish, Mexican and Italian hate speech corpora. As said above, we carried out the error analysis for both tasks.

In the first competition we noticed that our approach fails in the classification of sarcastic aggressive utterances, such as:

ii) *@USUARIO #LOS40MeetAndGreet 9 . Por q es una mamá luchona que cuida a su bendiciòn*[12].

Actually, the sarcasm is a type of figurative devices that modifies the perception of message, hindering the correct detection of hate speech by automatic systems. We found, in fact, the same difficulty for the recognition of misogynistic tweets in both languages, such as:

iii) *¿Cuál es la peor desgracia para una mujer? Parir un varón, porque después de tener un cerebro dentro durante 9 meses, van y se lo sacan*[13];

iv) *What's the difference between a blonde and a washing machine? A washing machine won't follow you around all day after you drop a load in it.*

In virtual as in real life, sexist jokes are very common. In general, they are considered innocent by the majority of people. However, Ford and Boxer (2011) reveal that sexist jokes are experienced by women as sexual harassment as well as offences. Moreover, Ford, Wentzel, and Lorion (2001) investigate on the effects of exposure to sexist jokes and they underline that a continue exposition can also modify the perception of sexism as norm and not as misbehavior.

5 Research Proposal

These early observations suggest the necessity to address the use of figures of speech such as sarcasm, in order to accurate, in a multilingual perspective, the automated methods to flag abusive language.

For this purpose, we propose an accurate analysis of different kinds of hate speech online especially in Italian, English and Spanish, taking into account also the geographical linguistic variations. We focus in particular on short texts such as tweets, posts or comments, exploring the informal language.

Considering the previous observations, we propose approaching the hate speech detection issues taking into account the figurative dimension of language and especially of abusive language. Moreover, it is necessary to examine the appropriateness of various computational techniques to solve this problem. In this line, we want to examine the contribution of the linguistic features to deep learning approaches by comparison with the performances of classical techniques. Finally, the multilingual context allows to discover the typical aspects of hate speech in order to recognize it independently from the languages.

Indeed, the scope of this investigation is to propose a methodology for monitoring correctly the user-generated contents allowing the system to work as sensor of the violence, also in real world.

6 Future work

Our research aims to explore the several dimensions of hate speech considering, above all, the use of figurative devices that hinder the automatic processes of recognition. In order to investigate the remarks observed in these first experiments, as future work, we would like to participate in HaSpeeDe[14] and AMI[15] at Evalita 2018 for Italian.

In addition, similar tasks are proposed at SemEval 2019 concerning: multilingual hate speech against immigrants and women (HatEval)[16], and the identification and categorization of offensive language in social media (OffensEval)[17]. Analyzing different kinds of abusive language allows to understand the boundaries between them and their singular aspects. Finally, multilingual context gives us the opportunity to delineate the differ-

[11] *"type of irony done with sharpness and a fit of anger"*

[12] *@User #LOS40MeetAndGreet 9 . Because she is a fighter mother who takes care of her kid.*

[13] *What's the worst disgrace for a woman? Giving birth to boy, because after she has got a brain into her for 9 months, it is taken out*

[14] `http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html\#`

[15] `https://amievalita2018.wordpress.com/`

[16] `https://competitions.codalab.org/competitions/19935`

[17] `https://competitions.codalab.org/competitions/20011`

ences and analogies between the various languages, inferring general characteristics of hate speech online.

References

Álvarez-Carmona, M. Á., E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes. 2018. Overview of mex-3at at ibereval: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September*.

Anzovino, M., E. Fersini, and P. Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Del Vigna, F., A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of ITASEC17*.

Fersini, E., M. Anzovino, and P. Rosso. 2018. Overview of the task on automatic misogyny identification at ibereval. In *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September*.

Ford, T. E. and C. F. Boxer. 2011. Sexist humor in the workplace: A case of subtle harassment. In *Insidious Workplace Behavior*. Routledge, pages 203–234.

Ford, T. E., E. R. Wentzel, and J. Lorion. 2001. Effects of exposure to sexist humor on perceptions of normative tolerance of sexism. *European Journal of Social Psychology*, 31(6):677–691.

Frenda, S. and S. Banerjee. 2018. Deep analysis in aggressive mexican tweets. In *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September*.

Frenda, S. and B. Ghanem. 2018. Exploration of misogyny in spanish and english tweets. In *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September*.

Fulper, R., G. L. Ciampaglia, E. Ferrara, Y. Ahn, A. Flammini, F. Menczer, B. Lewis, and K. Rowe. 2014. Misogynistic language on twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on ChASM*.

Gambäck, B. and U. K. Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Justo, R., T. Corcoran, S. M. Lukin, M. Walker, and M. I. Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133.

Mehdad, Y. and J. Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Samghabadi, N. S., S. Maharjan, A. Sprague, R. Diaz-Sprague, and T. Solorio. 2017. Detecting nastiness in social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 63–72.

Sanguinetti, M., F. Poletto, C. Bosco, V. Patti, and M. Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. LREC.

Spertus, E. 1997. Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*, pages 1058–1065.