Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 132–136
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Automatic Detection of Offensive Language in Social Media: Defining Linguistic Criteria to build a Mexican Spanish Dataset

María José Díaz-Torres*, Paulina Alejandra Morán-Méndez*
Luis Villaseñor-Pineda†‡, Manuel Montes-y-Gómez†
Juan Aguilera†, Luis Meneses-Lerín‡
*Facultad de Lenguas, Universidad de las Américas Puebla, México,
{maria.diazto, paulina.moranmz}@udlap.mx
†Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico,
{villasen, mmontesg, jaguilera}@inaoep.mx
‡Centre de Recherche en Linguistique Française GRAMMATICA, Université d'Artois, France,
luis_meneses_lerin@yahoo.fr

Abstract

Phenomena such as bullying, homophobia, sexism and racism have transcended to social networks, motivating the development of tools for their automatic detection. The challenge becomes greater when speakers make use of popular sayings, colloquial expressions and idioms which may contain vulgar, profane or rude words, but not always have the intention to offend; a situation often found in the Mexican Spanish variant. Under these circumstances, the identification of the offense goes beyond the lexical and syntactic elements of the message. This first work aims to define the main linguistic features of aggressive, offensive and vulgar language in social networks in order to establish linguistic-based criteria to facilitate the identification of abusive language. For this purpose, a Mexican Spanish Twitter corpus was compiled and analyzed. The dataset included words that, despite being rude, need to be considered in context to determine they are part of an offense. Based on the analysis of this corpus, linguistic criteria were defined to determine whether a message is offensive. To simplify the application of these criteria, an easy-to-follow diagram was designed. The paper presents an example of the use of the diagram, as well as the basic statistics of the corpus.

Keywords: aggressiveness detection, corpus annotation, text classification, Spanish

1. Introduction

As of today, social media platforms such as Facebook, Twitter and YouTube have facilitated and encouraged interpersonal communication. Through them, people interact and share their opinions through posts, messages and comments online. Unfortunately, since these platforms guarantee to some extent the freedom of expression of their users, they can and often use these means to attack or offend other persons. This situation leads to safety issues: online aggression and abuse not only create mental and psychological health problems for the victims but have also been proved to cause self-harm and even suicide (Kumar et al., 2018).

Some of the major challenges for detecting abusive language in social networks are the speed and volume of online communication. Every second, approximately 6,000 tweets are published, which is equivalent to more than 500 million tweets per day[1], making manual monitoring impossible. The previous scenario has motivated the development of methods for the automatic detection of abusive messages. Current methods are of two main kinds: supervised (Burnap and Williams, 2016; Plaza-del Arco et al., 2019) which require labeled data for learning a classification model, and, unsupervised (Gitari et al., 2015; Wiegand et al., 2018; Guzmán-Falcón, 2018), which detect hostile messages by searching for words in a given lexicon of profane words. Both kinds of approaches have their own advantages and disadvantages. In particular, the creation of supervised learning methods for offensive language detection requires of large, accurate, manually annotated resources. Nevertheless, most corpora available are in English (Pamungkas and Patti, 2019), which greatly hinders this task in low-resource languages. Annotation criteria for this type of datasets have only seldom been detailed (Ousidhoum et al., 2019), and, moreover, the labeling of offensive and non-offensive messages is commonly a costly and highly subjective task due to several socio-cultural and domain dependent issues. A greater challenge is posed by the richness of colloquial expressions and vulgar language that characterizes communication in social networks, since the identification of offenses goes beyond the lexical and syntactic elements of the message, and requires the annotator to understand the context beyond individual terms. With this motivation, through the present research we sought to define the main linguistic features that characterize abusive language manifested in social networks. As a first step, our work departs from the fact that the language used in social networks is abundant in colloquial expressions, commonly composed of rude or profane words, but they are not used to offend. Hence, the interest of this work is the definition of an annotation scheme with enough elements to discriminate these situations. To this end, we defined the concepts of offensive, aggressive and vulgar language, based on Austin's Speech Acts theory (Austin, 1962), with the aim of establishing criteria to facilitate their identification and thus define an accurate, fine-grained and linguistic-based annotation scheme.

[1]Internet Live Stats, 2019 - www.internetlivestats.com/twitter-statistics

2. Related Work

The task of automatically detecting aggressive content aimed at individuals or communities has recently been studied in different academic forums. However, most of them focus on the English language (Álvarez-Carmona et al.,

2018). In 2017, the 1st Workshop on Abusive Language Online (ALW1) was organized, where different approaches were presented for the detection of abusive language in social networks, focusing particularly on written communications in English and German (Waseem et al., 2017a). Subsequently, more workshops of the same court emerged, but due to the lack of consensus on a definition for "offensive language", the scope of the task was narrowed to more specific and identifiable behaviors. This was the case of the recent First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) (Kumar et al., 2018). In this workshop, the phenomena of online aggression such as trolling and cyberbullying were discussed.

By the same token, issues such as racism (Tulkens et al., 2016), sexism (Lee et al., 2010), and bullying (Samghabadi et al., 2017) have been studied in this line of research. Along the definitions proposed for these abusive behaviors we can find certain patterns, such as the presence of curse words, discriminatory vocabulary, derogatory adjectives and the explicit mention of others; manifested through names, pronouns, and user tags (Waseem et al., 2017b).

With respect to the efforts made for Mexican Spanish, the last two years, the evaluation forum "Authorship and Aggressiveness Analysis in Twitter: a case study in Mexican Spanish" (MEX-A3T) has been held. This forum -which took place within the IberEval 2018 (Álvarez-Carmona et al., 2018) and IberLEF 2019 (Aragón et al., 2019) conferences- evaluated an aggressiveness detection task in Mexican Spanish tweets. The results confirmed the complexity of this task, and the need for well-defined criteria to differentiate offensive, aggressive and vulgar language. Therefore, the goal of the present research was to establish criteria to facilitate the identification of offensive language and thus define a detailed, linguistic-based annotation scheme.

3. Data Collection

To collect data, we considered Twitter as the source media since it is open and its anonymity allows people to write judgments or assessments about other people, including offenses or aggressions. The interest of this first work is the definition of criteria to distinguish the offense or the aggression when using the same vocabulary. That is, it is necessary to collect messages that, despite using the same words (*i.e.* rude words), it is the context that determines whether a word is used to offend, or is part of a colloquial expression that is not intended to offend. To build the corpus, we collected tweets from August to November of 2017. We used some rude words and controversial hashtags to narrow the search. We collected a set of 143 terms that served as seeds for extracting the tweets, which includede words classified as vulgar and non-colloquial in the *Diccionario de Mexicanismos de la Academia Mexicana de la Lengua*, as well as words and hashtags identified by the *Instituto Nacional de las Mujeres* as related to violence and sexual harassment against women on Twitter (Guzmán-Falcón, 2018). Table 1 shows examples of these seed words.

To ensure their origin, the tweets were collected considering their geolocation. We considered Mexico City as the center and extracted all tweets that were within a radius of 500km. Finally, nearly 10,500 tweets in Mexican Spanish were collected and analyzed to define the annotation scheme.

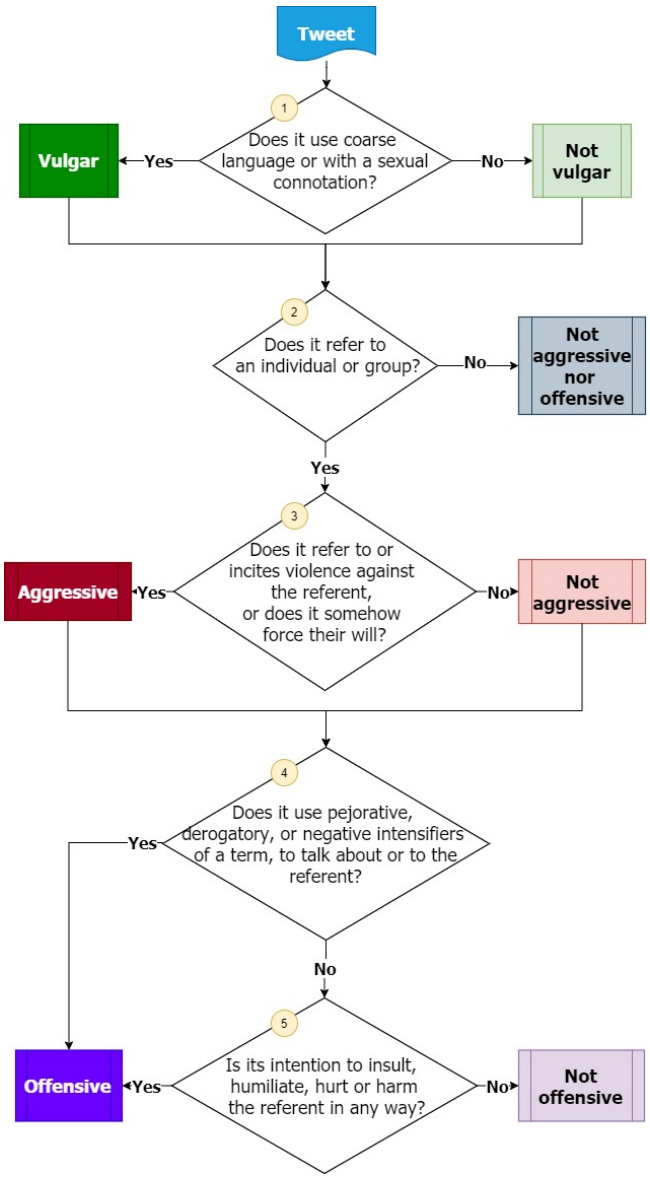

Figure 1: Annotation flowchart for abusive language categorization

4. Annotation Scheme

The creation of the annotation scheme and the annotation task itself were part of an incremental and complementary process. Two linguists from our research team studied the abusive language phenomenon through the literature and analyzed the collected tweets, to arrive to a typology that identified the defining characteristics of vulgar, aggressive and offensive language. Then, the linguists wrote the annotation diagram and used it to classify the corpus. For the purpose of creating said linguistic-based annotation scheme, first, it was necessary to arrive at a definition for the concepts of offensive, aggressive, vulgar language. Having a conceptualization of each term is a critical task, since it allows to establish linguistic criteria for the iden-

Table 1: Sample of the vocabulary applied for the recovery of tweets.

Spanish	English Translation
luchona	*hard-working woman (single mother; derogatory)*
pendejo(a)	*asshole (masc./fem.)*
prieto(a)	*dark-skinned (masc./fem.; derogatory)*
vergazos	*strong blow (vulgar)*
golfas	*whores*
puta	*slut*
lameculos	*ass kisser*

tification and classification of these linguistic phenomena. Once the theoretical framework on these linguistic manifestations was outlined, we looked for the lexical and semantic elements representative of the aggressive, offensive or vulgar messages.

4.1. Offensive, Aggressive and Vulgar Language

In order to identify the most characteristic features of aggressive, offensive and vulgar language, we first studied the definitions formulated in several academic forums and workshops. Among the proposed conceptualizations, recurrent linguistic characteristics can be found: the presence of rudeness, discriminatory vocabulary, derogatory adjectives and the mention of others, which is manifested through names, pronouns, and user tags (Waseem et al., 2017b). Beyond these lexical and syntactic elements, the pragmatic aspect of the messages is crucial to qualify them as aggressive, offensive or vulgar. According to the Speech Acts theory (Austin, 1962), the production of a statement performs three types of actions or acts at the same time: the locutionary act, the linguistic expression itself, its syntactic structure and the literal meaning semantic; the illocutionary act, the force or intention of the expression provided by the speaker; and the perlocutionary act, the consequence or effect of the statement on the interlocutor. The second act is the one that interests the detection of abusive language, since the illocutionary force of a message is its underlying purpose, which could go from asking a question, an invitation, a reminder, to a warning, a promise, or a threat, among many others. This wide range of intentions is delineated in the classification of illocutionary speech acts by (Searle, 1976). It is important to emphasize that the illocutionary force of a speech act always depends on the context of the expression (Fromkin et al., 2011), and since tweets provide very little context other than the linguistic expression itself, the annotators must rely on their sociopragmatic knowledge of the language to identify the illocutionary force of the message. That is the reason why linguistic variation must be taken into account for the definition of these concepts. Linguistic variation is the intrinsic characteristic of all languages that refers to the systematic differences in pronunciation, vocabulary and grammar of different social and regional groups of speakers of a language (Holmes and Wilson, 2017). This is a relevant phenomenon for any natural language processing task, and in the case of abusive language detection it should be considered not only because of the distinctive lexical and syntactic characteristics of the dialect, but also because these patterns convey social meanings (Wardhaugh, 2011), which would affect the way of expressing aggressiveness.

After revising the literature on the subject and analyzing the definitions of other related linguistic manifestations such as hate speech, cyberbullying, and racism, an offensive, aggressive and vulgar language typology was reached:

- Offensive language: aims at insulting or humiliating a group or individual, usually using derogatory or derogatory terms. An example from the corpus is: *No es que estés gorda, lo gordo se quita. Es tu cara de caballo.* This tweet humiliates a woman, makes fun of her body and compares her to an animal.
- Aggressive language: seeks to harm or hurt a group or individual by referring to or inciting violence. An example from the corpus is: *pero estas gorda... aprovecha tu fin pendeja que el lunes te violo.* This tweet involves insults and a rape threat.
- Vulgar language: it involves profanity, with sexual connotation and sometimes double entendre, but may or may not refer to an individual or collective. An example from the corpus is: *Martes con de M de Mamando onvre se arreglan las cosas... creo... eso dicen..* This tweet uses obscene vocabulary and is sexually explicit.

5. Diagram Description

Our annotation scheme was designed as a flowchart, for the purpose of supporting abusive language categorization into aggressive, offensive and vulgar in a clear, visual way. It was devised with the goal to be easy to read and useful for annotators without strong linguistics knowledge, to account for the diversity of backgrounds in the field of natural language processing. The typology portrays each concept as a non-exclusive quality of the message or tweet. This way, the tool allows for a better characterization of the texts when considering the possibility of a tweet belonging to one, two or even all classes, which represents more accurately the nature of these messages in social networks. The flowchart presents questions regarding the form and function of the message, about the presence of insults, derogatory, or sexually-charged vocabulary, but most prominently it is concerned on the illocutionary force of the message; that is, the intention and target of the tweet. As shown in Fig. 1, the labeling process begins with the selection of a tweet, and the first question that asks if the tweet uses coarse language or with a sexual connotation. If the answer is yes, this indicates the message is vulgar, otherwise it is not. Following, the annotator is asked whether the tweet refers to an individual or to a group of people, or not. This question serves to make an early discard of aggressiveness and offensiveness, since these classes, unlike vulgar language, require of a target to qualify as such. If the message does not have a specific referent, the labeling process ends there. On the contrary, if the answer is positive, then the next question concerns aggressiveness, and asks if the tweet incites violence or tries to force the will of its referent. Finally, to determine if the message is offensive, the diagram

Table 2: Examples showing the use of the proposed scheme. The number in parentheses refers to the question in the annotation flowchart.

Message	Vulgar?	Aggressive?	Offensive?
Lo más rico de coger no es lo que tú sientes; sino ver al cabrón retorcerse de placer... #Bottom #Sex #Coger *The best part about sex is not the feeling you get, but watching the man shiver of pleasure... #Bottom #Sex #Fuck*	Yes (1)	No (3)	No (5)
@USUARIO Estoy hasta la puta madre jajajajajaja *@USER I've fucking had it hahahahaha*	Yes (1)	No (3)	No (5)
Vrg que feas botas *Holy fuck those are some ugly boots*	Yes (1)	No (2)	No (3)
Lloran cuando las golpean, ah pero en la calle andan de golfas :) #MujerGolpeadaHombreFeliz *They cry when they're beaten, oh but they're out whoring on the street :) #BeatenWomanHappyMan*	No (1)	Yes (3)	Yes (4)
Tu no por qué eres MACHORRA!! *Not you because you're a BUTCH!!*	No (1)	No (3)	Yes (4)
Te recomiendo que te vayas comprando tus Tampax joto agachón!!! *I recommend you buy tampons bitch boy!!!*	No (1)	No (3)	Yes (5)
Ya me tienes hasta la madre pendejo. Al chile el martes el Richi y yo te vamos a partir la madre. *I'm fucking sick of you asshole. I swear on Tuesday Richi and I are going to fuck you up.*	Yes (1)	Yes (3)	Yes (5)

directs the annotator to observe if the tweet uses pejorative, derogatory or negative intensifiers of a term to refer to its target; if the tweet seeks to humiliate or insult its referent. Be any of these questions answered affirmatively, the tweet shall be labeled as offensive.

It should be noted that each of these classifications, vulgar, aggressive, and offensive, are non-exclusive qualities of the tweet. That is the reason why the flowchart continues after every decision, with the exception of the message having no referent. Table 2 shows examples that correspond to each of the categories.

6. Towards automatic detection of abusive language

This research work generated two digital linguistic resources: a linguistic annotation scheme for the classification of offensive, aggressive and vulgar language; and a corpus of offensive language in Mexican Spanish. As it was previously explained, the scheme was designed based on an abusive language typology, which served to annotate the dataset. This obtained a Kappa coefficient of inter-evaluator agreement of 0.91, which means that as a result we had a consistent annotation when making use of the proposed scheme while annotating the corpus with both of the evaluators. Clearly, the high level of agreement is because they labelled the corpus at the time of analysis. A second exercise with new annotators is needed to confirm the applicability of the proposed scheme.

Table 3 shows the general characteristics of this corpus: the distribution of the messages in the offensive and non-offensive classes, as well as the size of their vocabularies. Using this corpus, a first classification exercise was carried out. To do this, a traditional method for text classification was applied[2]. The objective of this exercise was to observe the strong overlap between both classes. As mentioned in previous sections, the collection of messages was done with a single set of seed words. Consequently, the common vocabulary between the two classes is high. However, although many of the messages in the non-offensive class use the same rude words, they are not considered offenses or aggressions.

Table 3: Corpus' distribution.

Class	Tweets	Vocabulary	Tweet size
Non-offensive	7,460	13,696	16.1±5.9
Offensive	3,015	7,365	16.3±5.8
Total	**10,475**	**17,067**	**16.1±5.9**

Table 4 shows the results obtained. As it can be seen, the non-offensive class achieves greater F1-measure, an effect expected by the imbalance in the classes. On the other hand, as expected, the classifier does not correctly discriminate between the two classes, because this simple representation (*i.e.* unigrams) does not consider the entire context.

Table 4: Offensive detection results, Acc=0.77±0.06 (stratified 10-fold cross validation).

Class	Precision	Recall	F1-measure
Non-offensive	0.83±0.05	0.86±0.06	0.84±0.04
Offensive	0.63±0.13	0.56±0.18	0.58±0.14

[2]A unigram based representation with frequency weights; frequency threshold >= 10; SVM classifier (linear kernel, C = 1).

7. Conclusions

This research work focuses on the annotation process of corpora for the detection of abusive language. The proposed annotation scheme provides specific criteria to identify aggressive, offensive and vulgar language based on its linguistic characteristics and intent of the message. This initial scheme took special care to include in the analysis messages that, despite the use of rude words, are not considered offensive. On the other hand, the collected corpus of abusive language is representative of the variant of Mexican Spanish, encouraging the creation of more resources in our language and giving visibility to one of its many dialects. Our contribution encourages the emergence of proposals for automatic methods that will be able to obtain better results thanks to a more accurate dataset, consistent with the reality of this online language phenomenon. Lastly, it should be noted that the diagram will be made available, and our corpus will be made available through the MEX-A3T 2020 forum[3]. Any future participant in the forum will have access to the dataset presented in this work.

8. Acknowledgements

We would like to thank CONACyT for partially supporting this work under grants CB-2015-01-257383 and the Thematic Networks program (Language Technologies Thematic Network).

9. Bibliographical References

Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., and Rico-Sulayes, A. (2018). Overview of mex-a3t at ibereval 2018: 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*, volume 6.

Aragón, M. E., Álvarez-Carmona, M. Á., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., and Moctezuma, D. (2019). Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain*.

Austin, J. (1962). How to do things with words, 2nd edn, jo urmson and m. *Sbasa (eds)*.

Burnap, P. and Williams, M. L. (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Fromkin, V., Rodman, R., and Hyams, V. (2011). An introduction to language, 9e. *Boston, MA: Wadsworth, Cengage Learning*.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Guzmán-Falcón, E. (2018). *Detección de lenguaje ofensivo en Twitter basada en expansión automática de lexicones*. Tesis de maestría en ciencias computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica.

Holmes, J. and Wilson, N. (2017). *An introduction to sociolinguistics*. Routledge.

Kumar, R., Ojha, A. K., Zampieri, M., and Malmasi, S. (2018). Proceedings of the first workshop on trolling, aggression and cyberbullying (trac-2018). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

Lee, T. L., Fiske, S. T., Glick, P., and Chen, Z. (2010). Ambivalent sexism in close relationships:(hostile) power and (benevolent) romance shape relationship ideals. *Sex Roles*, 62(7-8):583–601.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Pamungkas, E. W. and Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370.

Plaza-del Arco, F. M., Molina-González, M. D., Martín-Valdivia, M. T., and Lopez, L. A. U. (2019). Sinai at semeval-2019 task 6: Incorporating lexicon knowledge into svm learning to identify and categorize offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 735–738.

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

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, 5(1):1–23.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016). The automated detection of racist discourse in dutch social media. *Computational Linguistics in the Netherlands Journal*, 6:3–20.

Wardhaugh, R. (2011). *An introduction to sociolinguistics*, volume 28. John Wiley & Sons.

Waseem, Z., Chung, W. H. K., Hovy, D., and Tetreault, J. (2017a). Proceedings of the first workshop on abusive language online. In *Proceedings of the First Workshop on Abusive Language Online*.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017b). Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018). 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)*, pages 1046–1056.

[3] sites.google.com/view/mex-a3t/