Deep analysis in aggressive Mexican tweets

Simona Frenda[1,2] and Somnath Banerjee[3]

[1] University of Turin, Italy
[2] Universitat Politècnica de València, Spain
sfrenda@unito.it
[3] Jadavpur University, India
sb.cse.ju@gmail.com

Abstract. The importance of the detection of aggressiveness in social media is due to real effects of violence provoked by negative behavior online. Indeed, this kind of legal cases are increasing in the last years. For this reason, the necessity of controlling user-generated contents has become one of the priorities for many Internet companies, although current methodologies are far from solving this problem. Therefore, in this work we propose an innovative approach that combines deep learning framework with linguistic features specific for this issue. This approach has been evaluated and compared with other ones in the framework of the MEX-A3T shared task at IberEval on aggressiveness analysis in Spanish Mexican tweets. In spite of our novel approach, we obtained low results.

Keywords: Aggressiveness Detection · Deep Learning · Linguistic Analysis.

1 Introduction

The opinions expressed online by users are usually uncontrolled and this lack of control facilitates and supports negative online behaviors such as cyberbullying, racism, sexism and any form of hate. In the last few years, governments, social media platforms, Internet companies and communities of citizens are spending a growing amount of efforts to monitor and contrast such forms of online aggressive behaviors and attitudes, with the main aim of limiting it. An example of governmental dedication about this subject is the campaign No Hate Speech Movement of the Council of Europe for human rights online. On the academic side, the research interest about this issue is increasing and the approach is naturally interdisciplinary. Especially in the natural language processing (NLP) field, the attention is supported by international and national workshops or campaigns of evaluation like the competition proposed in the framework of IberEval 2018 by the organizers of MEX-A3T[4] [1] on the aggressiveness analysis in Twitter. This track proposes to detect the aggressiveness on Mexican Spanish tweets providing texts containing offensive messages that disparage or humiliate specific target. In this paper we present our participation in this task proposing a new approach

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

that combines deep learning with linguistic features.
The remainder of the paper is organized as follows. In Section 2 we describe synthetically the previous approaches used until today. In Section 3 we present our proposal followed by the results obtained in the competition (Section 4). Finally, in Section 5 we draw some conclusions.

2 Related Work

Currently, commercial and simple methods to deal with the automatic detection of negative online behaviors rely on the use of blacklists , essentially composed with slurs and swear words. However, filtering the messages in this way does not provide a sufficient remedy because it falls short when user-generated content is more subtle. Therefore, the research challenges in this field are oriented at investigating deeply all dimensions of language and also the communication on the Web, to envision deeper and more sophisticated solutions exploiting surface features ([6], [12]), syntactic features [4], semantic and conceptual features, polarity information [11], word-embedding techniques ([13], [16]), world knowledge information from ontologies [8], or proposing profile-based approach [9]. Some authors focus also on the extraction of meta-information from social platforms about users (like gender) and on their social activity (like history or geolocalization of posts) as predictive features [7]. In addition, some scholars take advantage of the connection between sentiment analysis and hate speech, benefiting from sentiment lexicons [18] or using a multi-step approach that combines sentiment or subjectivity classifiers with systems of hate speech detection ([8], [10]). This relation is due to the fact that hate speech expressions mostly exhibit a negative polarity, although the polarity intensity depends above all on cultural factors. Indeed, the aggressiveness involves different aspects of the user/author of the message that are difficult to define. So, taking into account the literature, we analyzed linguistically the data and we tried to understand what are the characteristics of aggressive tweets in the context of the Mexican culture and also the emotions that arouse this behavior.

3 Methodology

The common approach to detect aggressiveness online is formulating a prediction task, and in particular MEX-A3T organizers proposed a classification task with the aim to distinguish aggressive tweet from the non-aggressive [1]. Considering the complexity of this task, we needed to analyze the provided data in order to contemplate the different factors involved: linguistic characteristics proper of a tweet (like shortness or informal language), emotive traits of the aggressiveness and cultural aspects considering the fact that the provided data are geolocalized in Mexico. Therefore, we propose in this paper an innovative approach that incorporates linguistic features into deep learning architecture.
In the next subsections we describe the set of features provided along with the training data and deep learning architecture operations.

3.1 Linguistic Features

The linguistic features employed aim to cover all the above aspects about the aggressiveness in the context of a tweet. *Textual features* As textual features we take into account the polarity (positive/negative/neutral) of emoticons[5], used especially for giving contextual information to readers.

Style and writing density We consider also stylistic traits of authors, such as: the use of specific abbreviations used in Mexican tweets (*hdp*, *alv*), the number of characters per sentence and word, the use of some elements of punctuation (question, exclamation marks and sequences of dots) and the uppercase characters, inspecting if the user writes all in uppercase or just some letters.

Bag of words In order to understand the importance of some words respect to others, we extract trigrams of words weighted with tf-idf.

Lists of aggressive words Considering the fact that the aggressive text aims to offend, attack, humiliate and hurt an individual or collective target, we created two lists containing specifically derogatory adjectives and vulgar expressions like profanities and insults (*chinga a tu madre*, *vete a la verga*).

Syntactic patterns Another factor involved in aggressive texts is the target, implicit or explicit, to whom the insults or profanities are addressed. Therefore, we examined the syntactic combinations of target explicited with mention (@usuario) or proper name with derogatory adjectives and vulgar expressions.

Affective features Finally, as said above, we take into account the emotions concerning aggressiveness and we observed that anger and disgust are the principal emotions that provoke this kind of behavior. For this feature, we used Spanish Emotion Lexicon (SEL) provided by [17] and [15], considering the words with a higher Probability Factor of Affective use in Spanish language. In addition, we increased it with slang words usually used in social networks [14], taking into consideration also the cases of synonymy.

In order to allow our architecture to process these features, we preprocessed the data deleting symbols and urls that can hinder the process of extraction of features and pos-tagging the texts using FreeLing [5].

3.2 Deep Learning Framework

In this section, we describe the deep learning (DL) framework for detecting the aggressive tweets. The proposed model is inspired by the deep architecture proposed in [3]. They combined the feature engineering with DL and increased the classification accuracy for the code-mixed question classification task [2]. To understand the effectiveness of combining feature engineering with the DL framework, we have experimented with two setups: one with feature engineering and another without it. Therefore, we have proposed two models: Model-1 is based on the DL framework with feature engineering and Model-2 is based on DL

[5] The annotated list of emoticons used for this work is provided by the Unicode Consortium: http://www.unicode.org/

framework without feature engineering. The deep learning framework is based on Convolutional Neural Network (CNN).

Embedding layer: Instead of using any pre-trained word embedding scheme, we have built a vocabulary table which is learned from the training data. The embedding layer works as a lookup table which maps vocabulary word indices into low-dimensional vector representations. As the aggressive tweets are of variable length, we used the zero-padding (i.e., the missing part replaced by zeros) to maintain the input vector to a fixed size L.

Features: For Model-1, we integrated the features in the embeddings. We derived a feature set as described in Section 3.1. We combined these features with DL in Model-1. However, we did not combine the features with DL framework in Model-2.

Convolutional layer: Let $t_i \in \mathbb{R}^k$ be the k-dimensional vector corresponding to the i-th word in the tweet. A tweet is represented as $t_{1:n} = t_1 \oplus t_2 \oplus ... \oplus t_n$, where, the tweet contains the words $t_1, t_2, \ldots, t_n$ and $\oplus$ is the concatenation operator.
Also, let $tf_{1:m} = tf_1 \oplus tf_2 \oplus ... \oplus tf_m$ be the feature set for the tweet $t_{1:n}$. After combining the feature set $tf_{1:m}$ with the vector representation of the tweet $t_{1:n}$, the resulting vector is $l_{1:m+n} = tf_{1:m} \oplus t_{1:n}$. Therefore, $l_{1:m+n} = l_1 \oplus l_2 \oplus ... \oplus l_{m+n}$, where either $l_i \in tf_{1:m}$ or $l_i \in t_{1:n}$.
Let $l_{i:i+j}$ refer to the concatenation of $l_i, l_{i+1}, \ldots, l_{i+j}$. In the convolution operation, the filter $w \in \mathbb{R}^{hk}$ is applied to a window of h words to produce new features such as feature s_i is generated from a window of words $l_{i:i+h1}$ by $s_i = f(w.l_{i:i+h-1} + b)$, where, $b \in \mathbb{R}$ is a bias term and f is a non-linear function. A feature map $s = [s_1, s_2, \ldots, s_{nh+1}]$ (where, $s \in \mathbb{R}^{n-h+1}$) is produced by applying the aforesaid filter to each possible window of h words (i.e., $\{l_{1:h}, l_{2:h+1}, \ldots, l_{nh+1:n}\}$) in the tweet. The max-pooling operation is applied to the feature map s to obtain the maximum value $s' = \max\{s\}$ for the particular filter. The objective of the max pooling is to capture the most important feature with the highest value for each feature map. However, the proposed architecture uses multiple filters with varying window sizes to obtain multiple features. Then, these features are passed to the next layer, i.e., a fully-connected layer.

Fully-connected layer: The fully-connected layer is also known as the dense layer. The max-pooling operation selects the best features from each convolutional kernel. Thus, all the resulting features which are selected from the max-pooling are combined in the fully-connected layer. The output of fully connected layer is passed to the output layer.

Output layer: The final layer (i.e., the output layer) is made of 2 neurons as the given tweets are of 2 target classes (i.e., aggressive and non aggressive). The output layer uses 'softmax' as the nonlinear activation function.

4 Results

In the framework of the evaluation campaign, we have submitted two runs: the DL based on CNN with feature engineering (DLF+FE) and a simple DL frame-

work based on CNN (DLF). In order to evaluate the performance of the systems in the competition, the organizers use the F-measure of aggressiveness class. In Table 1, we report the scores obtained along with our position in the ranking for the aggressive tweets prediction. In spite of the novelty of our approach, the results are low and the feature engineering does not outperform the deep learning based model.

Table 1. Results

	Overall prediction				**Aggressiveness prediction**			
	Accuracy	*F-score*	*Precision*	*Recall*	*Precision*	*Recall*	*F-score*	Rank
DLF	0.6702	0.5585	0.5585	0.5585	0.3363	0.3367	**0.3365**	9
DLF+FE	0.5865	0.5094	0.5149	0.5183	0.2676	0.3827	**0.3150**	10

5 Discussion and Conclusions

In this work, we investigate the automatic detection of aggressive texts by incorporating linguistic features into deep learning architecture. Considering the low results, we carry out error analysis that reveals that our systems mainly fail to classify tweets with orthographic errors and sarcastic or ironic utterances, such as: "*USUARIO #LOS40MeetAndGreet 9. Por q es una mamá luchona que cuida a su bendición*"; "*Quiero hablar con el que inventó el hecho de "levantarse temprano" Que xxxxx estaba pensando*". Therefore, taking into account these observations, we will investigate the use of humorous devices to express negative opinions. Moreover, as future work, in order to make deeper analysis about the impact of feature engineering on deep learning approach, we would like to propose this approach in similar research issues.

Acknowledgement

The work of Simona Frenda was partially funded by the Spanish MINECO under the research project SomEMBED (TIN2015-71147-C2-1-P).

References

1. Álvarez-Carmona, Miguel Á and Guzmán-Falcón, Estefanía and Montes-y-Gómez, Manuel and Escalante, Hugo Jair and Villaseñor-Pineda, Luis and Reyes-Meza, Verónica and Rico-Sulayes, Antonio: Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September (2018)
2. Banerjee, Somnath, Sudip Kumar Naskar, Paolo Rosso, and Sivaji Bandyopadhyay: The First Cross-Script Code-Mixed Question Answering Corpus. In MultiLingMine@ ECIR, pp. 56-65. (2016)
3. Banerjee, Somnath, Sudip Naskar, Paolo Rosso, and Sivaji Bandyopadhyay: Code mixed cross script factoid question classification-A deep learning approach. Journal of Intelligent & Fuzzy Systems 34, no. **5**: 2959-2969. (2018)

4. Burnap, Pete, and Matthew L. Williams: Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Science 5, no. **1**: 11. (2016)
5. Carreras, Xavier, Isaac Chao, Lluis Padr, and Muntsa Padr: FreeLing: An Open-Source Suite of Language Analyzers. In LREC, pp. 239-242. (2004)
6. Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu: Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pp. 71-80. IEEE. (2012)
7. Dadvar, Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong: Improving cyberbullying detection with user context. In European Conference on Information Retrieval, pp. 693-696. Springer, Berlin, Heidelberg. (2013)
8. Dinakar, Karthik, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS) 2, no. **3**: 18. (2012)
9. Escalante, Hugo Jair, Esa Villatoro-Tello, Sara E. Garza, A. Pastor Lpez-Monroy, Manuel Montes-y-Gmez, and Luis Villaseor-Pineda: Early detection of deception and aggressiveness using profile-based representations. Expert Systems with Applications **89**: 99-111. (2017)
10. Gitari, Njagi Dennis, Zhang Zuping, Hanyurwimfura Damien, and Jun Long: A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering 10, no. **4**: 215-230. (2015)
11. Justo, Raquel, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Ins Torres: Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Systems **69**: 124-133. (2014)
12. Mehdad, Yashar, and Joel Tetreault: Do Characters Abuse More Than Words?. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 299-303. (2016)
13. Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang: Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web, pp. 145-153. International World Wide Web Conferences Steering Committee. (2016)
14. Posadas-Durn, Juan-Pablo, Ilia Markov, Helena Gmez-Adorno, Grigori Sidorov, Ildar Batyrshin, Alexander Gelbukh, and Obdulia Pichardo-Lagunas: Syntactic n-grams as features for the author profiling task. Working Notes Papers of the CLEF. (2015)
15. Daz Rangel, Ismael, Grigori Sidorov, and Sergio Surez Guerra: Creacin y evaluacin de un diccionario marcado con emociones y ponderado para el espaol. Onomazein 1, no. **29** (2014).
16. Samghabadi, Niloofar Safi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and Thamar Solorio: Detecting Nastiness in Social Media. In Proceedings of the First Workshop on Abusive Language Online, pp. 63-72. (2017)
17. Sidorov Grigori, Miranda-Jimnez Sabino, Viveros-Jimnez Francisco, Gelbukh Alexander, Castro-Snchez No, Velsquez Francisco, Daz-Rangel Ismael, Surez-Guerra Sergio, Trevio Alejandro and Gordon Juan: Empirical study of opinion mining in Spanish tweets. LNAI 7629, pp. 1-14 (2012)
18. Van Hee, Cynthia, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Vronique Hoste: Detection and fine-grained classification of cyberbullying events. In International Conference Recent Advances in Natural Language Processing (RANLP), pp. 672-680. (2015)