Aggressive Analysis in Twitter using a Combination of Models

Gretel Liz De la Peña Sarracén and Paolo Rosso

PRHLT Research Center
Universitat Politècnica de València, Spain
gredela@posgrado.upv.es
prosso@dsic.upv.es

Abstract. This paper describes the system we developed for the task on Aggressive detection in Authorship and aggressiveness analysis in Twitter (MEX-A3T)The task focuses on the detection of aggressive comments in tweets that come from Mexican users. We have analyzed three kinds of models and the proposed system is a combination of them. The first model is based on Convolutional Neuronal Networks whose outputs feed a LSTM Neural Network. The second one uses the pre-trained Universal Sentence Encoder for encoding sentences into embedding vectors. Finally, the third one consists in a simple Multi-layer Perceptron. The final results show that our model achieves good results.

Keywords: Convolutional Neural Network, LSTM Model, Universal Sentence Encoder, Multi-layer Perceptron, Aggressive Detection Track, Twitter

1 Introduction

Nowadays, the use of social networks is increasing rapidly. Among them, Twitter stands out as a broadcast medium of information. Many users use this social media as one of the main sources for obtaining news. However, many of those users are attacked by tweets with aggressive messages.

This phenomenon constitutes a problem that affects different groups of people, due to harassment towards immigrants, women or for instance, sexist comments [6]. Therefore, some treatment that controls this situation is essential. Different researches have been done in this regard. Some approaches use traditional classifiers such as Naive Bayes and Linear SVM [15, 13, 9]. Others use models based on Deep Learning with architectures such as LSTM and Convolutional Neural Networks (CNN) [2, 7, 12]. Several international competitions have also been organized to motivate the creation of systems for the detection of this type

of messages. Such as the Workshop on Trolling, Aggression and Cyberbullying [10], that included a shared task on aggression identification; the tasks on Automatic Misogyny Identification (AMI) [4] at IberEval 2018 and EVALITA 2018 [5], the Workshop on Abusive Language [14] and the task on Autohorship and Aggressiveness Analysis in Twitter task (MEX-A3T) [11] proposed at IberEval 2018. This year the second edition of MEX-A3T [1] has been launched. Its aim is to further improve the research in autohorship and aggressiveness analysis tasks and to push the computational processing of the Mexican tweets.

In this work, we propose a system formed by the combination of three strategies. Each of them analyzes the tweet to be classified in a different way. The first one is based on Convolutional Neural Networks whose outputs feed a LSTM Neural Network. The second one uses the pre-trained Universal Sentence Encoder for encoding sentences into embedding vectors. The third one consists of a simple Multi-layer Perceptron which gets the TF-IDF representation of the tweet. Then, the strategies are combined in order to build a system that takes into account each of the analysis and predicts whether a given tweet is aggressive or not.

The rest of the paper is organized as follows. Section 2 describes our system. Experimental results are then discussed in Section 3. Finally, we present our conclusions with a summary of our findings in Section 4.

2 System

2.1 Preprocessing

The first step for the development of the system is the preprocessing of the texts. In this phase different characteristics, typically present in the tweets, and that possibly do not have discriminatory semantic information, are normalized. In this way, the numbers are replaced by the *_num* tag, dates by the *_date* tag, and all the links by the *_url* tag. In addition, user mentions, identified by the first character @, are replaced by *_user*. The hashtags were not processed to avoid losing information that they may contain.

2.2 Method

We propose a system that consists of a combination of different strategies as Figure 1 shows. The first one is a deep learning model (CNN-LSTM) at the word level. For each tweet, CNN-LSTM receives as input the word embbedings, which are processed by a CNN for obtaining a sequence of vectors. These vectors can be seen as the representation of n-grams according with the size of the kernel. In the next section, the details are discussed. Then, the vectors feed a LSTM model for obtaining a prediction. The second model (USE-MLP) takes as input a vector for a tweet. This vector is obtained with the pre-trained Universal Sentence Encoder based on the transformer architecture. Then, a Multi-layer Perceptron is used to get a prediction. Finally, a similar model to the previous

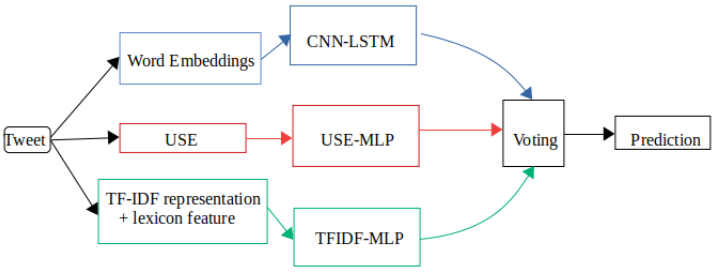

Fig. 1. General structure of the system

one is used in the third one (TFIDF-MLP). The difference is in the input of the Multi-layer Perceptron. In this case, the vector is the TF-IDF representation of the tweet. In addition, a new component is concatenated to the vector according to a linguistic feature based on a lexicon of obscene and vulgar phrases in the Mexican Spanish. Then, the final prediction is obtained by majority of votes, given the prediction of each model. In each case, cross entropy is used as the loss function.

2.3 Convolutional Neural Network and LSTM Model

In this first model, as was mentioned before, the tweets are represented with a sequence of word embeddings. For this, the Word2vec MEX-A3T model provided by the organizers of the competition is used. This has been trained with the MEX-A3T corpus containing 500,000 tokens. The size of the embeddings is 200. The objective of this model is to process bigrams present within a tweet in a sequential manner. The approximation used to obtain the sequence of bigrams

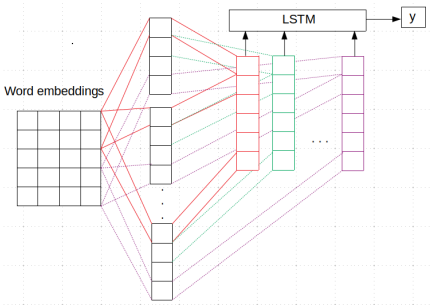

Fig. 2. CNN-LSTM Model

vectors is shown in the figure 2. Where 150 filters of 2x200 are used, 2 correspond

with the size of the bigrams and 200 correspond with that of the embeddings. The result is a column matrix with depth 150, so that the i-th component taken in depth, can be seen as a high level representation of the i-th bigram. Then, each of these vectors is the input at each time step of the LSTM Recurrent Neural Network which can process them sequentially. Finally, a Softmax layer is used to obtain the prediction.

2.4 Universal Sentences Encoder Model

The second model takes advantage of the pre-trained Universal Sentence Encoder [3] to get the prediction for a tweet. It takes variable length text as input and as outputs a 512-dimensional vector. We have used the encoder architectures based on the transformer architecture trained for Spanish. Two dense layers with a Relu function are used to process the vector and finally the prediction is obtained with a Softmax at the end.

2.5 Multi-layer Perceptron model

A problem that frequently occurs with the approaches based on deep learning is the lack of data to train the models. To solve this problem, a model based on a traditional approach has been included in the system. For this, each tweet has been represented as a TF-IDF vector. Additionally, a linguistic feature has been incorporated into the vector. Basically, this feature corresponds to the number of aggressive phrases contained in the tweet. The identification of these phrases is based on the study carried out in the work [8], where the authors propose a methodology for the detection of obscene and vulgar phrases in Mexican tweets. Then, the prediction for a tweet is obtained by a Multi-layer Perceptron of three layers whose input is the correspondent vector.

3 Results

The Training set has been divided in the experiments, separating 30% for the Validation set. The results of the F-measure of the aggressive class on that Validation set were 0.64 for USE-MLP, 0.68 for TFIDF-MLP and 0.65 for CNN-LSTM, while for the combination of the three models 0.68 was obtained. As can be seen, the best results were achieved with the simplest model, in the same way as in the Test set as shown below.

Table 1 shows the results on the Test set for different variants and the result of the best system in the competition (best). The run1 corresponds to the combination of the commented models. On the other hand, run2 and run3 are systems that only take into account the CNN-LSTM and TFIDF-MLP models respectively. Our best result is obtained with the simplest model, which reaches the third position in the competition with a value very close to the first two in the F-measure of the aggressive class (F1), and in both class (F(P,R)). Our particular results show that the lack of data can affect the models based on deep

Table 1. Performance on the testing set

System	run1	run2	run3	run4 best
F1	0.4635	0.4405	0.4749	0.4796
F(P,R)	0.6205	0.5920	0.6349	0.6464

learning, with which in this case worse results were obtained. In addition, other problem that may affect the performance of the deep learning based system is the fact that rare or misspelled words can not be represented with the embeddings. This can badly condition the training, since important information may be lost.

4 Conclusion and Future work

We proposed a combination of three different models for the MEX-A3T task on aggressive detection in Twitter. The first one uses a CNN whose outputs feeds to a LSTM model. The second model analyzes the input at the full text level with the Universal Sentences Encoder. The third model is the simplest one that takes a TF-IDF representation of the text, and obtains the prediction with a Multi-layer Perceptron. The best results have been obtained with this last model, instead of the system which combines all the three models. This can be for the lack of data to train deep learning models, or for the problem of rare words that can not be represented with the embeddings. Thus, for future works, it is important dealing with these problems to improve the performance of the system.

Acknowledgments. The work of the second author was partially funded by the the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

1. Aragón, Mario Ezra and Álvarez-Carmona, Miguel Á and Montes-y-Gómez, Manuel and Escalante, Hugo Jair and Villaseñor-Pineda, Luis and Moctezuma, Daniela. Overview of MEX-3AT at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, September. (2019).
2. Badjatiya, Pinkesh and Gupta, Shashank and Gupta, Manish and Varma, Vasudeva. Deep Learning for Hate Speech Detection in Tweets. Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee. (2017).

3. Cer, Daniel and Yang, Yinfei and Kong, Sheng-yi and Hua, Nan and Limtiaco, Nicole and John, Rhomni St and Constant, Noah and Guajardo-Cespedes, Mario and Yuan, Steve and Tar, Chris and others. Universal Sentence Encoder. arXiv preprint arXiv:1803.11175. (2018).
4. Elisabetta Fersini, Maria Anzovino, and Paolo Rosso. Overview of the Task on Automatic Misogyny Identification at Ibereval 2018. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). CEUR Workshop Proceedings. CEUR-WS. org, Seville, Spain. 2150. (2018).
5. Elisabetta Fersini, Debora Nozza, and Paolo Rosso. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA18), Turin, Italy. CEUR. org. 2263. (2018).
6. Frenda, Simona and Ghanem, Bilal and Montes-y-Gómez, Manuel and Rosso, Paolo. Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. Journal of Intelligent Fuzzy Systems. 36.5. pp. 4743-4752. (2019).
7. Gambäck, Björn and Sikdar, Utpal Kumar. Using Convolutional Neural Networks to Classify Hate-Speech. Proceedings of the First Workshop on Abusive Language Online. (2017).
8. Guzmán, Estefania and Beltrán, Beatriz and Tovar, Mireya and Vázquez, Andrés and Martínez, Rodolfo. Clasificación de Frases Obscenas o Vulgares dentro de Tweets. Research in Computing Science. 85, pp. 65–74. (2014).
9. Gómez-Adorno, Helena and Bel-Enguix, Gemma and Sierra, Gerardo and Sánchez, Octavio and Quezada, Daniela. A Machine Learning Approach for Detecting Aggressive Tweets in Spanish. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings. 2150, pp. 102–107. (2018).
10. Ritesh Kumar, Atul Kr Ojha, Marcos Zampieri, and Shervin Malmasi. Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). (2018).
11. Miguel Àlvarez-Carmona, Estefanìa Guzmàn-Falcòn, Manuel Montes-y Gòmez, Hugo Jair Escalante, Luis Villasenor-Pineda, Verònica Reyes-Meza, and Antonio Rico-Sulayes. 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, 6. (2018).
12. Nikhil, Nishant and Pahwa, Ramit and Nirala, Mehul Kumar and Khilnani, Rohan. LSTMs with Attention for Aggression Detection. arXiv preprint arXiv:1807.06151. (2018).
13. Waseem, Zeerak and Hovy, Dirk. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proceedings of the NAACL Student Research Workshop. (2016).
14. Waseem, Zeerak and Kyong Chung, Wendy Hui and Hovy, Dirk and Tetreault, Joel. Proceedings of the First Workshop on Abusive Language Online. In Proceedings of the First Workshop on Abusive Language Online. (2017).
15. Xiang, Guang and Fan, Bin and Wang, Ling and Hong, Jason and Rose, Carolyn. Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM. (2012).