

Hate Speech Detection on Twitter

Notebook for PAN at CLEF 2021

Carolina Martín-del-Campo-Rodríguez, Grigori Sidorov and Ildar Batyrshin

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Juan de Dios Bátiz Avenue, Mexico City, 07738, Mexico

Abstract

With the use of social networks, the automatic detection of hate speech has become of great importance to prevent people, being protected by anonymity, from feeling free to discriminate against different groups. This document describes two approaches taken to detect hate speech by author: the first based on the individual processing of tweets by the author, which establishes a threshold of hate tweets to identify hate speech; the second based in the concatenation of tweets by author for processing.

Keywords

Hate speech, Twitter, SVM, Deep Neural Network

1. Introduction

The use of social networks has been increasing in recent years and the presence of hate speech has increased in the same way; so the study of this topic has gained attention. In [1] different types of hate speech are pointed out. But, even with these categories, the task of hate speech detection is not easy, even for human, because this is intrinsically associated to relationships between groups, and also rely on language nuances [2]. Hence, this task becomes even more difficult for computers, even though various approaches have been taken to try to solve this problem.

In [3] authors focus in detection of Misogynistic Language on Twitter. They created five sub-categories to detect the phenomena: Discredit, Stereotype and Objectification, Sexual Harassment and Threats of Violence, Dominance, and Derailing. For the final model they used a token n-grams representation and SVM for classification.

In this paper we describe our approach to detect hate speech authors on Twitter applying Support Vector Machine and Deep Neural Networks, participating in the PAN 2021 [4] task "Profiling Hate Speech Spreaders on Twitter" [5].

1.1. Corpus Description


The corpus for the development phase was divided into two languages (English and Spanish), each language was made up of 200 authors, with 200 tweets each. The labels were at the author

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ cm.del.cr@gmail.com (C. Martín-del-Campo-Rodríguez); sidorov@cic.ipn.mx (G. Sidorov); batyr1@gmail.com (I. Batyrshin)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

level, that is, there were no tags for each tweet. For the test phase, for each language there were 100 authors with 200 tweets each.

Links, user mentions, and hashtags were substitute, for all tweets, with #URL#, #USER# and #HASHTAG#, respectively.

2. Data Preprocessing

The pre-processing steps for both languages were the same. Links, user mentions, hashtags, and retweets were removed. All non-English and non-Spanish characters, respectively, were removed. The occurrence of more than two consecutive characters (letters) was replaced by only two characters. All punctuation was remove and all numbers were substituted with 0. Emojis (emoticons) were decoded into their text equivalents. Then all the text was lowercase.

3. Methodology

Different approaches were followed for each language. For English, a manual labeling of the tweets was carried out, so each tweet was considered individually for the training. For Spanish, the original author labeling was used, concatenating all the tweets by author separated by a special token. Each approach is detailed in the next subsection.

3.1. English

A manual labeling of tweets from 45 authors was made by us, based in the original labeling: 23 authors labeled as hate speech and 22 authors as non-hate speech. From these tweets, all tweets labeled as hate speech were taken, 405 tweets; from the rest of the tweets, the same number of tweets was randomly selected in order to have a balanced training corpus. GloVe was used for the embeddings.

A deep neural network was used for the classification¹, the architecture is defined in figure 1. The activation function used for the inner dense layers was relu, for the last dense layer sigmoid was used. To avoid overfitting, two Dropouts 0.8 and 0.5 were used, as can be seen in Figure 1. For the inner dense layers, a kernel regularizer l2 with a value of 0.0001 was set. For the last Dense layer a l2 activity regularizer was set with the same value. The model was configured with binary crossentropy for the loss, an Adam optimizer and using accuracy as metric.

A 10-Fold Cross-Validation was used in the training. Once the training was finished, the rest of the tweets in the corpus were classified. Next, the sum of Hate Speech Tweets (HST) by author was performed. Taking into account the original labeling (by author) and the HST number, a threshold was defined i.e., if an author had more than a certain HST number, this author was classified as a hate speech spreaders. The threshold was set to 30.

¹Using tensorflow 2

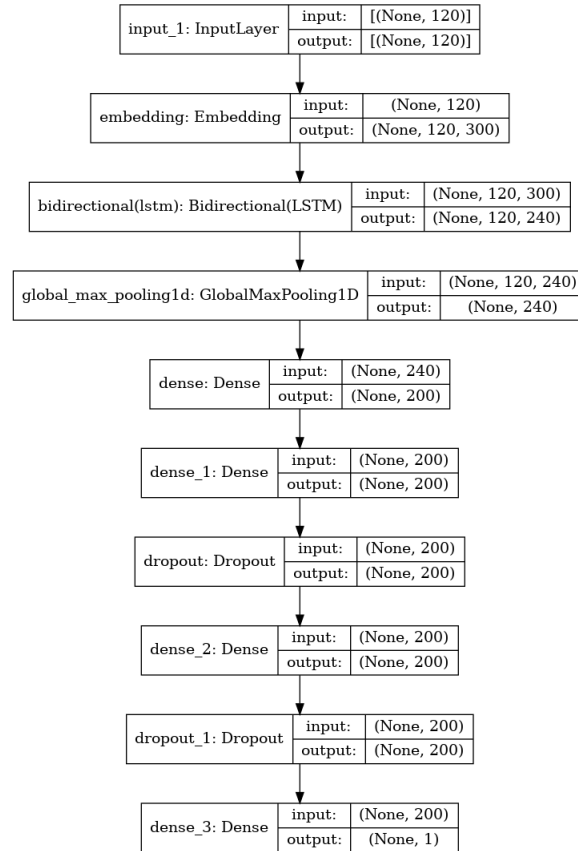


Figure 1: Specifications of the Neural Network model for English

3.2. Spanish

All tweet per author were concatenated, using the token <EOT> as separation. For classification a SVM was used². A linear kernel was used and a max number of iterations of 2000 was set (all the other parameters were left as default). For the training, 10-Fold Cross-Validation was used. Tokenization was perform with CountVectorizer, using the default values.

4. Results

Accuracy was used as evaluation metric according to the specifications of the organizer of PAN 2021 and TIRA [6]. For the training dataset the accuracy for English was $80.11\% \pm 3.41$, for Spanish 79.12 ± 1.23 , having around 79% of accuracy. For the testing phase, the result obtained for English was **65%** and for Spanish was **77%**, getting an average of **71%**.

²Using scikit-learn

5. Conclusions

Using the manual tweets labeling approach not only flawed the criteria for detecting hate speech but also makes it impossible to recognize the full context by author thus, hate spreaders whose individual tweets do not represent hate speech, but in a general way constantly attack a specific group, cannot be identified with this approach. On the other hand, concatenate all the tweets allows to analyze the complete context by author, but makes processing difficult (200 tweets per author). The results obtained shows that the concatenate approach is better generalizing.

As future work, a more in-depth analysis using deep neural network with tweets concatenation and SVM with labeling of tweets is proposed. In the same way, using the concatenation approach analyzes how the use of mini batches of tweets by author is carried out, maintaining the author's labels, so that part of the context is preserved and processing is facilitated.

Acknowledgments

The work was done with support of the Government of Mexico via CONACYT, SNI and the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico

References

- [1] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, in: LREC 2020, 2020.
- [2] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, *ACM Computing Surveys (CSUR)* 51 (2018) 1 – 30.
- [3] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: M. Silberstein, F. Atigui, E. Kornysheva, E. Métais, F. Meziane (Eds.), *Natural Language Processing and Information Systems*, Springer International Publishing, Cham, 2018, pp. 57–64.
- [4] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection, in: *12th International Conference of the CLEF Association (CLEF 2021)*, Springer, 2021.
- [5] F. Rangel, G. L. D. L. P. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: *CLEF 2021 Labs and Workshops, Notebook Papers*, CEUR-WS.org, 2021.
- [6] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), *Information Retrieval Evaluation in a Changing World, The Information Retrieval Series*, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.