

Profiling Hate Speech Spreaders using Characters and Words N-grams

(Notebook for PAN at CLEF 2021)

Daniel Jacob Espinosa, Grigori Sidorov

Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico

Abstract

With the increase interactions in social networks, it is important to take care of the health of information and relationships between the users. One of the big problems today are hate speech within them, this type of comments as well as the users who share them can be very dangerous for the integrity of society. In this occasion we show a solution based on N-grams of characters and words for the task of "Profiling Hate Speech Spreaders on Twitter", as classifier we use SVM Support Vector Machines (libSVM) for English and Spanish corpus.

1. Introduction

Today the use of social networks are useful for many things, from connections between friends to online classes, according to reports from Twitter [1] in this period of pandemic there was an increase in its audience of 30% worldwide . Due to the importance of health and communication within social networks, Facebook has improved its detection by 59 % its hate detection algorithm as of 2020. Where there are no large amounts of data to train the algorithm [2]. Despite the fact that these algorithms are relatively advanced, human verifiers are still needed for their correct classification.

Unfortunately, there are various investigations that mention how social networks are particularly effective for sharing feelings of hatred, creating "hate chambers" where users express themselves freely where they can go very far to the point of having incitements to annoy someone or carry out crimes [3]. This research shows how to create a community that supports political parties with anti-refugee and anti-immigration ideas in Facebook Germany without any kind of rules of conduct. The research indicates that many members of the community show extreme right-wing behaviors which are related to municipalities where attacks on the refugee community are registered.

Although this problem is not new, it has reached levels which can threaten the life of any person. Massimo's research shows a combination of the use of bots with hate speech in social networks showing how these feelings are easier to share via social networks [5]. Although bots

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

✉ espinosagonzalezdaniel@gmail.com (D. Y. Espinosa); sidorov@cic.ipn.mx (G. Sidorov)

🌐 <http://www.cic.ipn.mx/~sidorov/> (G. Sidorov)



© 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)

Table 1

Training Corpus of "Profiling Hate Speech Spreaders on Twitter" [4]

Language	English	Spanish	Total
Files	200	200	400

are not real users, they can be programmed to share hate speech creating problems in human user communities.

Due to these inconveniences, this year PAN [4] shared a corpus for the task called "Profiling Hate Speech Spreaders on Twitter" where we must obtain the best precision of an algorithm that performs said classification. This time we decided to do a combination of word and character N-grams to perform the classification, this task there are two languages to classify: English and Spanish.

2. Corpus

The corpus provided by PAN consists of a set of Tweets which are messages published on the social network Twitter. In these messages, photos, videos or GIFs can be attached, and they can also be responded to by other users by creating a reply tweet[4].

The training corpus for this task are: 400 files[4], where 200 are in English and 200 are in Spanish, show Table 1. All these files are tweets extracted from different Twitter users in xml format. Each file is from a specific user and contains 200 tweets where links, user mentions and hashtags were previously tagged. The emoticons remain in the texts; These tweets are 140 characters long.

3. Methodology

3.1. Pre-processing steps

For the data preprocessing layer we carry out the following steps:

Digits For the digits we decided to remove it since for the extraction of characteristics we do not see them useful.

Emoticons In the emoticons, we assigned a specific label for each different emotion and these were added to the bag of words.

Punctuation marks For punctuation marks due to the bag of words we use, we decided to have them removed.

Other symbols For the symbols that are not within the reference Standard ASCII (American Standard Code for Information Interchange) they were not considered important to be taken as

Table 2
Features of N-grams

Spanish	English
2 Char-Ngrams	2 Char-Ngrams
3 Char-Ngrams	3 Char-Ngrams
5 Char-Ngrams	5 Char-Ngrams
1 Words-Ngrams	1 Words-Ngrams
2 Words-Ngrams	3 Words-Ngrams

characteristics and therefore were removed.

3.2. Features

To solve this task, a structure of N-grams shown in Table 2 were selected. These N-grams are character and word N-grams and were selected in this way for their precision obtained by testing the training set.

It is important to say that the data sets for each language should not be mixed, so that these procedures are similar for each language but with different N-gram structures.

Let us remember that each file corresponds to the tweets of a specific user, then the structures of Table 2 were applied for each file, after these N-grams are selected, now we must model these structures in frequency matrices [6], to That is why we will use a bag of words [6] where we will put all the N-grams, after this we will count the frequency of repetition of these structures for each user. In this way we create a matrix with the occurrences of the N-grams within the bag of words. Now this model is replicated for each of the files in the corpus, this is better known as the **Term-Document Matrix** [7].

After having all those matrices, we create a matrix with N dimensions, where each new dimension is is a user, these new matrices that are added to the N-dimensional matrix must be organized and place their columns in relation to the N-grams of the previous one, if there is no column because that N-gram did not come out of the bag of words is given a 0.

When we complete the entire matrix of all the files for each language, we proceed to remove the columns with 2 or less repetitions, we do this because they are minimum values, which we can save processes when we use a classifier. We will call this matrix, already organized the Vector Space Model [7].

3.3. Vector Space Model for Texts

The Vector space model helps us to find a hyperplane [6] which to try to separate or look for a similarity between dimensions, we call this training since we intend to create a classification learning model. As a classifier we decided to use SVM since with this model it gave us the best performance to perform this classification, we can see the results in Table 3.

Table 3

Testing with different algorithms classification

Algorithm	Spanish data	English data
J48	69.00	68.72
NaiveBayes	74.55	72.99
RandomForest	77.45	76.44
Multi-layer Perceptron	83.03	83.44
SVM	83.09	81.90
LinearSVM	87.86	86.38

4. Experiments

One of the main reasons we decided to use N-gram structures was that we had previously worked with tweets; In previous works we participated in PAN 2019 [8] and PAN 2020 [9]. These results are shown in the CLEF congresses which are held every year and show community evaluation experiments with different laboratories and discussions around the world [10].

For 2019 we participate with the work with the name "Bots and Gender Profiling using Character Bigrams" [11] and for 2020 with the work called "Profiling Fake News Spreaders using Characters and Words N-grams" [12]. With the implementations carried out, we obtained an accuracy greater than 80% in both investigations.

To arrive at the solution of the use of the SVM as a classifier, we carried out several experiments to obtain the best result in precision. One of our implementations in this research was to use a neural network where the precision obtained was 64% for the corpus in English and 67% for the language in Spanish. An important characteristic in this type of tasks organized by PAN uses a TIRA platform [13] that serves to run your programs and these can be evaluated by the committee, so that the virtual computers that share you do not have many resources to perform large data handling or simultaneous processing, so you need to perform code optimizations so that you do not lack resources to run your algorithms.

5. Conclusions

It is important to find a solution to these problems since they are experiments with real data from the corpus and that can have a direct impact on society, we consider that currently these types of problems can be better solved by applying sentiment analysis with deep networks, since they have shown to work with better precision performed in this research. One problem with this type of implementation is that we need large amounts of data to train the model, although there is still the possibility of errors with the use of sarcasm or ironic teasing on social media.

We previously used hashtags as a new feature that helps improve accuracy, so we believe these methods would improve performance by adding hashtags as well as considering add the 280 character length that Twitter currently allows for a tweet. Even with this, we believe that it is an excellent task to solve because it represents a current problem and it is a challenge to train a model with so little data.

In the future we would like this type of methodologies and technologies to be implemented directly in the most used social networks since these types of problems are really a social danger that can be classified as terrorist weapons for a nation.

References

- [1] J. Vives, El coronavirus dispara el número de usuarios de Twitter, <https://www.lavanguardia.com/tecnologia/20200324/4882705311/coronavirus-dispara-numero-usuarios-twitter.html>, 2020. [Online].
- [2] Time, Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch, <https://time.com/5739688/facebook-hate-speech-languages/>, 2019. [Online].
- [3] K. Müller, C. Schwarz, Fanning the flames of hate: Social media and hate crime, *SSRN Electronic Journal* (2017). doi:10.2139/ssrn.3082972.
- [4] R. Francisco, L. D. L. P. S. Gretel, C. BERTa, F. Elisabetta, R. Paolo, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: A. J. M. M. F. P. Guglielmo Faggioli, Nicola Ferro (Ed.), *CLEF 2021 Labs and Workshops, Notebook Papers*, CEUR-WS.org, 2021.
- [5] M. D. D. Massimo Stella, Emilio Ferrarab, Bots increase exposure to negative and inflammatory content in online social systems, *PNSA 115* (2018) 12435–12440. URL: <https://www.pnas.org/content/115/49/12435>.
- [6] G. Sidorov, *Syntactic n-grams in Computational Linguistics*, SpringerBriefs in Computer Science, Springer, 2019. doi:<https://doi.org/10.1007/978-3-030-14771-6>.
- [7] G. Sidorov, Formalization in computational linguistics, in: *Syntactic n-grams in Computational Linguistics*, SpringerBriefs in Computer Science, Springer, 2016.
- [8] F. Rangel, P. Rosso, *CLEF 2019 Labs and Workshops, Notebook Papers*, in: C. L., F. N., M. H, L. D. (Eds.), *Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling*, CEUR Workshop Proceedings, 2019.
- [9] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter, in: *CLEF, 2020*.
- [10] 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.
- [11] D. Espinosa, H. Gómez-Adorno, G. Sidorov, Bots and Gender Profiling using Character Bigrams, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), *CLEF 2019 Labs and Workshops, Notebook Papers*, CEUR-WS.org, 2019. URL: <http://ceur-ws.org/Vol-2380/>.
- [12] D. Espinosa, H. Gómez-Adorno, G. Sidorov, Profiling Fake News Spreaders using Characters and Words N-grams—Notebook for PAN at CLEF 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), *CLEF 2020 Labs and Workshops, Notebook Papers*, CEUR-WS.org, 2020. URL: <http://ceur-ws.org/Vol-2696/>.
- [13] 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.