Profiling Fake News Spreaders using Characters and Words N-grams

Notebook for PAN at CLEF 2020

Daniel Yacob Espinosa[1], Helena Gómez-Adorno[2], and Grigori Sidorov[1]

[1] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico City, Mexico
espinosagonzalezdaniel@gmail.com, sidorov@cic.ipn.mx
[2] Universidad Nacional Autónoma de México,
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Mexico City, Mexico
helena.gomez@iimas.unam.mx

Abstract With the use of social networks as mass media; The spread of fake news becomes an investigative problem. This article describes our approach to the PAN 2020 task on "Profiling Fake News Spreaders on Twitter" [7].
The objective is to distinguish which users share fake news. Our approach includes a data cleaning part and feature extraction using N-grams of characters and words. The experiments were carried out with different N-gram structures depending on the languages: English and Spanish. We experimented machine learning algorithm Support Vector Machines (libSVM).

1 Introduction

Currently, social networks are one of the most important means of communication. Twitter has become an extremely active social network where users can give their opinion on any topic. With so much data and information exposed, millions of users are unable to validate the veracity of the information shared in this social network. So users can be deceived by lies told by word of mouth; in this case from user to user, this type of information is better known as "fake news".

Fake news aim to disqualify or create controversy about issues important to society [4]. Fake news are currently difficult to detect due to the use of bots for their replication and distribution. Bots are the main source of distribution of these type of messages. Bots are programs that pretend to have human behavior within social networks [2]. Through bots, the fake news has a wide distribution and reach to real users, which damages the real news ecosystem on social networks. Bots can distribute fake news with a wide user reach which becomes a serious problem on many daily activities [1].

Table 1. Corpus of "Profiling Fake News Spreaders Twitter" [7]

Language	Fake news Profiles	Real news Profiles	**Total**
Spanish	150	150	300
English	150	150	300

What makes it a problem is that this type of news is not filtered or fully verified that it is true. It is simply posted and with the interactions of users, it becomes a widely mentioned fake news that many people believe to be true [5]. Currently this problem is still under investigation due to the complexity of the issue. A possible solution to this problem is to identify the users' profiles which are propagators of false notifications and not to trust their replicated information. In order to address this kind of solutions, PAN organizes the "Profiling Fake News Spreaders on Twitter" task, which aims to find users who share fake news using only a sample of their shared tweets [7].

Previously we participated in PAN 19 [8], with the task "Bots and Gender Profiling" where we showed a solution formed by N-gram structures, in particular we used character bi-grams; so for this task we decided to use the same methodology; using N.grams; but increasing the number of characters and adding the n-grams of words.

2 Corpus

For the assigned task, the corpus given by PAN 20 [7] consists of 600 files (300 in Spanish and 300 in English) where each file represents a user; so that each user contains 100 tweets, in Table 1 shows the corpus configuration in detail.
A particularity of all the tweets in this corpus is that they have 140 characters, in addition; all the urls, links, hashtags and usermentions are masked, this means that one of these characteristics was assigned a label.

3 Methodology

3.1 Pre-processing steps

To perform the experiments, we performed a series of steps to pre-process the data:

Digits On the part of the digits we decided to remove them since we consider them not important for our methodology.

Emoticons Many of the tweets contain emoticons and by doing some experiments we decided to do something interesting with them: each emoticon with the same symbol, we decided to assign it a unique label; it means that if inside the tweet it contained the potato chips emoticon then a unique label is assigned for potato chips; with this we increase a little value of the percentage of accuraccy in the classification.

Punctuation marks The punctuation marks for our methodology were also not relevant, so we decided to remove them as well.

Table 2. Features of Ngrams

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

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 characteristics and therefore were removed.

3.2 Features

Since we preprocess the data, we need to find a pattern that helps us differentiate the users who share fake news. We can find that pattern using N-grams. N-grams are structures formed from the selected data, which we can organize by the repetition frequency. With this, it is possible to create a matrix which abstractly describes a representation form of said file, this is called a vector space model, let's not forget that each file represents a user.

Because we have 2 languages to classify, each one has different N-gram configurations. We selected the configuration of Table 2 where we show the structures of the N-grams assigned for each language.

3.3 Vector Space Model for Texts

Now that we have the characteristics of the tweets represented in a **Term-Document Matrix** [9], it is necessary to order them so that each column is in the same order for each dimension. When this our ordered matrix is commonly called vector space model where an abstraction of objects is represented so that they can be classified by a classification algorithm.

4 Experiments

Mainly the N-gram structures were considered due to our methodology used in the PAN19 task "Bots and Gender Profiles" [8], where good results were obtained with a single characteristic: **Character bi-grams** [3]. Using this same methodology, we propose the use of word and character N-grams to solve this task.

Table 3. Testing with SVM

Algorithm	Spanish data	English data
SVM	80.99	81.33
LinearSVM	**81.83**	**86.28**

After these experiments we made the comparisons between two classifiers with the vector space model. In Table 3 we show the results of the classifiers with which we carried out the experiments.

Due to our methodology, it is possible to test various configurations since we have the vector space model organized and filled with the frequencies of the N-grams of the tweets. Regarding the experiments, we tested different classifiers, selecting **LinearSVM** and which is the one that gives us the best performance. All the results shown in Table 3 were evaluated with an accuracy score.

An interesting aspect was assigning a label for each different emoji, when we assign the labels for the emojis these characteristics are not part of the N-gram structures within the characteristics matrix; that is, they were assigned directly with the assigned label. This method gave us an improvement in precision of **6.0**% for each language

All these experiments were taken care of by TIRA's technological resources [6], where it is necessary to upload the programs for their execution. Within this virtual machine, the scores of the researchers who are within the competition are organized, identifying the best precision of all the experiments.

5 Conclusions

We propose this solution for the task "Profiling Fake News Spreaders on Twitter" [7], using class and word N-gram structures, we use **LinearSVM** like classifier.We decided to use these functions and methods because we had previously worked with tweets; and although the task was not the same, the use of this methodology gave very interesting scores. So we understand the differences between the tasks, but we consider it important to note if they can be solved in a similar way as the data from the same social network.

Something interesting was previously we used the hashtags and usermentions to obtain a greater precision in the task which this time could not be used because they had already been tagged with the corpus; which means that the task solution can be used in another social network where these characteristics do not exist. Another aspect that we would like to test is the use of the 280 characters of a tweet; This corpus was limited with the use of the 140 but we would like to know if we could improve the performance of our method using the 280 that could has a tweet. We also noticed that some of the characteristics of social networks such as hashtags or emojis, provide an improvement in accuracy. We consider that for future experiments they can be included and see how these converge for the solution of a task.

This task is really interesting due to the use of social networks nowadays; As long as social networks have the same or greater interest than today, they will continue to serve as mass media where it will be essential to create a virtual environment as healthy as possible.

References

1. Bakshy, E., Hofman, J., Mason, W., Watts, D.: Everyone's an influencer: Quantifying influence on twitter. pp. 65–74 (01 2011). https://doi.org/10.1145/1935826.1935845
2. Cai, C., li, L., Zengi, D.: Behavior enhanced deep bot detection in social media. pp. 128–130 (07 2017). https://doi.org/10.1109/ISI.2017.8004887
3. Espinosa, D., Gómez-Adorno, H., Sidorov, G.: Bots and Gender Profiling using Character Bigrams. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019), http://ceur-ws.org/Vol-2380/
4. Ghanem, B., Rosso, P., Rangel, F.: An Emotional Analysis of False Information in Social Media and News Articles. ACM Transactions on Internet Technology (TOIT) **20**(2), 1–18 (2020)
5. Popat Kashyap, Mukherjee Subhabrata, Y.A., Gerhard, W.: Declare: Debunking fake news and false claims using evidence-aware deep learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1003, https://www.aclweb.org/anthology/D18-1003
6. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World. Springer (Sep 2019)
7. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
8. Rangel, F., R.P.: CLEF 2019 Labs and Workshops, Notebook Papers. In: Cappellato L., Ferro N., M.H.L.D. (ed.) Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. CEUR Workshop Proceedings (2019)
9. Sidorov, G.: Formalization in computational linguistics. In: Syntactic n-grams in Computational Linguistics. SpringerBriefs in Computer Science, Springer (2016)