Bots and Gender Profiling using Character Bigrams
Notebook for PAN at CLEF 2019

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 This paper describes our approach to tackle the Author Profiling task at PAN 2019. The objective is to distinguish between bot and human users and for human users it is also necessary to detect their gender. We are given only Twitter messages in two languages (Spanish and English). Our preprocessing stage includes data cleaning as well as the extraction of features using character *bi*-grams. We experimented with several feature representations and machine learning algorithms (Support Vector Machines (SVM) from libSVM). For both languages we use the same methods of feature extraction and classification.

1 Introduction

Thanks to artificial intelligence, learning using computer is possible, because with each interaction in technology, it can learn more from us to give us more comfort in some tasks or to provide us with solutions, which are more according to our tastes or interests. Actually with the help of artificial intelligence, what we want to do is to model the human intelligence [11].

Currently the use of artificial intelligence to make predictions is very involved in most streaming services or social networks, to mention some internet services. They are constantly learning about users to give them the best service according to their interests, for the streaming services we can consider artificial intelligence to predict what a user may like and in this way invite him to continue using the services. On the other hand, social networks are also used to show news, pages, forums, friends or simply to meet new people. In this new generation of web 2.0, social networks are a great double-edged sword, since both companies and users with a more direct interaction [4] are what can be called horizontal communication. Thanks to this, companies, agencies, and some ministrations can interact more directly with users so that users can give their opinion

about a product or service, now imagine that many people have similar opinions and these are shared on social networks, since that we interact with the comments of others we can have empathy or perhaps disgust and express it in the same way [2].

One of the main reasons to study bots is the impact they generate on social networks through opinions [3], then it is tried to explore the text that they generate to detect if it is a bot. We have to realize the importance of social networks today and the use of technology for these, they can warn of a catastrophe situation in some part of the world to creation of "Trending topics" about trends in the world of Fashion. Unfortunately, so deeply penetrated social networks there are companies and governments that benefit from this creating bots and using them to spread false news and thus create doubts, discontents, generate uncertainties to much of the community interested in these issues [6].

Then PAN workshop is organized every year since 2011 with aim of promoting research on authorship analysis which includes authorship attribution, author profiling, and plagiarism detection, among others [7]. In this year campaign, the organizers included a subtask of automatic bot detection. The aim is to discriminate between real users and bots based only on text messages posted in Twitter [1].

2 Corpus

The task proposed by PAN is to predict if a user is a bot or not, if it is a human user then it is also necessary to predict the gender of the user. The released dataset contains two languages: Spanish and English. It is important to mention that each user is represented by 100 tweets, which will be analyzed and separated depending on the language. The dataset contains only tweets in which each file corresponds to a user. This dataset was mainly used for training a system, which was tested with other datasets for PAN evaluations.

3 Methodology

3.1 Preprocessing steps

Having only the tweets of the users it is necessary to do a preprocessing considering some features that the tweets can have:

Digits For the part of the digits we decided to remove them since we considered that they were not necessary for text feature representation.

URLs Since the links are resource identifiers in this case are Internet pages are not necessary for the bi-gram structures either.

@Mentions Mentions refer to other Twitter users with whom they interact in the message; they are important to quote on Twitter but in our case they will not be necessary so we will also eliminate them.

Emoticons There are messages that contain emoticons but for the structure that we use they are not necessary, however, we consider them not to be helpful for the classification.

Considering the data, a procedure for standardization is necessary:

Punctuation marks For the case of our selection of characteristics, we will not need to use punctuation marks. We extracted them to have our data as clean as possible.

3.2 Features

First for the preprocessing of the data we removed punctuation marks, since in the experiment we will not use them as a feature, we also removed the references to other Twitter users as well as links, numbers and emojis contained in the messages as well as characters that are not inside of the Standard ASCII (American Standard Code for Information Interchange).

Since we have the data somehow clean, it is necessary to eliminate the spaces between the words by the following procedures.

The main idea of the extraction of features is to obtain particular features of the object so we can then compare those features with others and consider some patterns that have in common. So one way in which we can obtain these characteristics or features is with the use of character n-grams [5]. With the use of traditional character n-grams we discovered that we had a good performance for solving the problem, but the best results for both languages (Spanish and English) were with the formation of character bi-grams [9]. When the bi-grams are generated, if there are equal bi-grams then they will be added in a counter of the frequency of that word where in this case is the character bi-gram, if a new character bi-gram comes out then it will be a new feature where the frequency it will be 1 since it is the first time it appears, and so on until the analysis of each user is complete, where we do not forget that each file corresponds to a user.

3.3 Vector Space Model for Texts

Now we have the characteristics obtaining the frequencies of characterbi-gram per user, we need a method in which we can organize the data of all the documents with respect to their characteristics; because of this we created a vector space model.

Table 1. Example of Term-Document Matrix with character bi-gram

Matrix	**Doc1**	**Doc1**	**Doc3**
bi-gram1	3	6	0
bi-gram2	0	3	2
bi-gram3	5	0	1

The main idea of using vector space model is to represent the characteristics of each object with its corresponding object but in an organized manner where the objects can be compared later [10].

We proceed to organize our data in a table **Term-Document Matrix** [11] where for the part of the columns we have the document and in the part of the rows is a description of the character bi-gram, in this way the content of the table will be the difference of the character bi-gram in the analyzed file. If a character bi-gram is not found in the document, the value of the box must be 0.

Having this matrix you have all the documents with character bi-grams in an orderly way and can be analyzed in a much more efficient way.

3.4 Experiments

Thanks to the structure of the organized matrix, they were tested with several classifiers and evaluated the accuracy to know which could be the best classifier for this task. All the n-grams tested were of character since with them we had much better accuracy than with other structures to obtain characteristics.

After having the results of the classification between humans and bot, we only use humans for the classification of gender using same methods of extraction of characteristics and the same classification models.

Table 2. Evaluation results in terms of **the classification accuracy between humans and bots** on the PAN Author Profiling 2019 test corpus and classification method.

	Spanish			**English**		
classification method	**1-gram**	**2-grams**	**3-grams**	**1-gram**	**2-grams**	**3-grams**
J48 Split 70%	62.00	67.82	82.03	63.44	65.54	71.29
NaiveBayes Split 70%	65.11	71.11	83.30	66.48	69.25	72.69
RandomForest Split 70%	91.77	74.44	85.11	83.21	85.55	83.24
RandomForest 20-Fold	92.22	92.7	90.6	92.79	92.80	90.31
SVM CrossValidation-10	91.86	**92.38**	90.76	92.42	**92.86**	90.41

Table 3. Evaluation results in terms of **the classification accuracy of gender** on the PAN Author Profiling 2019 test corpus and classification method.

	Spanish			**English**		
classification method	**1-gram**	**2-grams**	**3-grams**	**1-gram**	**2-grams**	**3-grams**
J48 Split 70%	53.22	57.1	56.31	54.44	57.02	55.90
NaiveBayes Split 70%	65.11	71.11	83.33	62.60	66.08	69.91
RandomForest Split 70%	71.77	74.44	85.111	66.08	78.86	75.65
RandomForest 20-Fold	76.35	75.53	77.42	73.22	76.82	76.31
SVM CrossValidation-10	76.13	**80.72**	72.40	75.22	**83.37**	78.39

Comparing the results, it was decided to use the SVM algorithm with cross validation of 10 boxes (SKlearn SVM) as the classifier for the best classifier among the comparisons with the other classifiers.

4 Conclusions

In this paper, we present an approach to get the solution for the Task "Bots and Gender Profiling" of PAN at CLEF 2019. Our final system for the classification between bots and humans followed by classifying users who are human by extracting characteristics from the tweets and placing them in a structure formed by character bi-grams. In this way, a term-document matrix is formed in which the entire data set is ordered to pass through a classification process. With respect to the tests carried out, we decided to use Support Vector Machine as a classifier with cross validation with 10 boxes for training the model and later use it with the PAN tests. We realized that for the Spanish and English languages it did not differ much in the value of the accuracy for the classifications, so we used the same method for the extraction of characteristics: as well as to determine between human users and bots and the gender classification in the human users. In the same way we use the same classifier for both languages [8].

For future work due to the good performance between the classification of humans and bots we would like to try with the different characteristics that social networks allow to introduce in messages (for example, using the 250 characters that Twitter allows in each Tweet); perhaps we can find more efficient ways to classify between humans and bots using natural language processing techniques.

References

1. Daelemans, W., Kestemont, M., Manjavancas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., Zangerle, E.: Overview of PAN 2019: Author Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D., Heinatz, G., Cappellato, L., Ferro, N. (eds.) Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Springer (Sep 2019)
2. Emilio Ferrara, Onur Varol, F.M.A.F.: Detection of promoted social media campaigns. In: The 10th International AAAI Conference on Web and Social Media. pp. 563–566. SpringerBriefs in Computer Science, ICWSM (2016)
3. Zakaria el Hjouji, D. Scott Hunter, N.G.d.M.T.Z.: The impact of bots on opinions in social networks. In: arXiv preprint arXiv:1810.12398 (2018)
4. Linda S. L. Lai, E.T.: Groups formation and operations in the web 2.0 environment and social networks. In: Group Decision and Negotiation. p. 387–402. Springer (2008)
5. Manning, C.D., Schütze, H.: Statistical estimation: n-gram models over sparse data. In: Foundations of Statistical Natural Language Processing. MIT Press, MIT (1999)
6. Martin Potthast, Tim Gollub, M.W.B.S.: TIRA Integrated Research Architecture. In: Nicola Ferro, C.P. (ed.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)
7. 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)

8. Rangel, F., R.P.F.M.: Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics. In: A Low Dimensionality Representation for Language Variety Identification, pp. 156–169. Springer-Verlag (2018)
9. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies. NAACL-HLT '15, Association for Computational Linguistics (2015)
10. Sidorov, G.: N-gramas sintácticos y su uso en la lingüística computacional. In: Vectores de investigación, 6(6). pp. 1–15. SpringerBriefs in Computer Science, Springer (2013)
11. Sidorov, G.: Formalization in computational linguistics. In: Syntactic n-grams in Computational Linguistics. SpringerBriefs in Computer Science, Springer (2016)