Using N-grams to detect Fake News Spreaders on Twitter

Notebook for PAN at CLEF 2020

Juan Pizarro[0000−0001−9598−1929]

jpizarrom@gmail.com

Abstract. This paper synthesizes our participation in the CLEF conference 2020 regarding the Profiling Fake News Spreaders on Twitter task, organized at the PAN lab on digital text forensics and stylometry. The models that we suggested obtained one of the two best results, based on an average accuracy of 0.7775 –0.7350 for English and 0.8200 for Spanish. In summary, we propose a Support Vector Machine (SVM) classifier with character and word n-gram features to determine whether the author of a Twitter feed is keen to be a spreader of fake news.

Keywords: Author Profiling · Fake News · Twitter · Spanish · English.

1 Introduction

In the past few years, social media has been changing how people communicate and interact. Currently, we use these platforms daily for a variety of purposes –searching for information, buying products, reaching out to bank representatives, or as a marketing and commercialization channel. For example, in the first quarter of 2019, Twitter reported an average of 330 million monthly active users [22].

As well as social media platforms have reached popularity, they have become tools that directly influence the perception of events, people, or products. Certain people and organizations have been reaching this goal through the spread of fake news, rumors, and misinformation.

This paper presents our participation in the Author Profiling task at PAN. That work aims to identify users in two categories: faker user (fake news spreader) and legitimate user (real news spreader) [4] [17]. Our method follows the one presented in 2019 [15], focused on comparing n-grams of chars and words as features, and an SVM as a classifier. However, this year we intend to explore the use of different preprocessing strategies for a specific classifier.

2 Related Work

Author Profiling distinguishes between classes of authors by studying shared language among people. Therefore, through this perspective, it is possible to address fake news by studying stylistic deviations of users that tend to spread them [4].

As follows, a brief timeline regarding Author Profiling research achievements. In 2002, the use of function words and part of speech tagging served to identify the author's gender on a corpus consisting of 920 labeled documents [9]. Additionally, uni-grams and bi-grams, Naive Bayes, maximum entropy classification, and support vector machines helped to classify the sentiment around movie-data [13]. In 2004, a support vector machine and Naive Bayes were used to determine if movie reviews were positive or negative [12].

In 2006, an accuracy of 0.80 was obtained in the task of gender identification in a corpus of 85.000 blogs using style and content words [19]. During the 2010 U.S. midterm elections, primitive social bots played the part of supporters of some candidates and attacked their opponents [11] [10] [23]. In 2014, the problem of identifying bots on all of Twitter was studied, and 19 of the 25 top features they use were identified as sentiment-related [7]. A grid search was used to find the best hyper-parameters for each of the classifiers. In 2016, social bots were found generating a large amount of content, possibly distorting online conversations. They noted that bots tweeting about Donald Trump generated the most positive tweets [3] [23].

More recently, in 2018, it was reported a case of political manipulation on social media that used sentiment analysis [20]. Finally, in 2019, the use of a variety of different semantic and stylistic features, and a neural recurrent model helped to detect fake news on Twitter's accounts [8]. Regarding this discovery, word embeddings and style features served to profile fake news in different accounts. On the contrary, information such as hashtags was not useful.

Related to Author Profiling task at PAN in CLEF 2017, the gender identification task obtained an accuracy of 0.8233 for English and 0.8321 in Spanish [1]. Besides, an SVM classifier was trainded with combination of characters n-grams and TF-IDF. For the Author Profiling task in PAN at CLEF 2018, the result showed an accuracy of 0.8221 for English and 0.82 for Spanish [6]. In this case, they used char and word n-grams as features and an SVM as a classifier. By applying similar strategies, the method presented in [21] obtained 0.8121 for English and 0.8005 for Spanish. In 2019, different classifiers were evaluated using characteristics similar to those applied in previous years. N-grams of chars and words were used, and the best results were obtained using also SVM [15].

3 Our Method

The following section presents our method.

Table 1. Preprocessing options for Profiling Fake News Spreaders on Twitter

Name	Description
pres-case	whether to maintain letter case or downcase for everything except for emoticons
red-len	whether to replace repeated character sequences
rpl-dgt	whether to replace numbers by xxdgt
demojify	whether to replace emojis by word representations
rpl-anon	whether to replace anonymized tags *#URL#* by *xxurl* *#USER#* by *xxusr* *#HASHTAG#* by *xxhst*

3.1 Preprocessing

The corpus considers the text of tweets whose authors qualify as fake news spreaders or legitimate users. At first, each author's tweet groups together on a long chain. Subsequently, a preprocessing strategy applies to them.

A preprocessing strategy consists of a stack of text transformations that creates a complete list of preprocessing options. For example, emoji[1] replaces emojis, *xxdgt* instead of numbers, and lower case for specific cases. Table 1 shows a complete list of the preprocessing options.

3.2 Features

N-grams of characters and words are generated with different n-gram orders. Each document, composed by each set of tweets per author, portraits the use of Term Frequency – Inverse Document Frequency (TF-IDF). Afterward, the n-grams vectors group together to obtain one feature vector per author.

3.3 Classifiers

Linear Support Vector Machine operates as a classifier due to the positive results obtained in the previous years.

4 Experimental Work

This section presents the dataset and the experimental setup. To accomplish our method goals, we choose to use nltk[5], sklearn[14], and hyperopt[2] to implement it.

[1] Emoji for Python https://github.com/carpedm20/emoji/

Table 2. Profiling Fake News Spreaders on Twitter Corpus

Lang	Author		Tweets	
	Fake News Spreader	***legitimate user***	***Fake News Spreader***	***legitimate user***
en	150	150	15000	15000
es	150	150	15000	15000

Table 3. Feature representation hyperparamenters for the Profiling Fake News Spreaders on Twitter task

N-gram type	Param	Values
word	ngram_range	$(1,2),(1,3),(2,3)$
word	max_df	$0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0$
word	min_df	$0.0001, 0.001, 0.01, 0.02, 0.03,$ $0.04, 0.05, 0.06, 0.1, 1, 2, 5$
char	ngram_range	(1, 3),(1, 5),(2, 5),(3, 5),(1, 6),(2, 6)
char	max_df	$0.6, 0.7, 0.75, 0.8,$ $0.85, 0.9, 0.95, 1.0$
char	min_df	$0.0001, 0.001, 0.01, 0.02, 0.03,$ $0.04, 0.05, 0.06, 0.1, 1, 2, 5$

4.1 Datasets

In Table 2 can be seen that the corpus consists of the text of 60.000 tweets whose authors qualify as fake news spreaders or legitimate users. The 100 tweets of each author balance in terms of the types previously described and their language –30.000 tweets are in English and 30.000 in Spanish.

4.2 Hyper-parameter Tuning

The hyper-parameters used for the feature representation are shown in Table 3, for the preprocessor strategy are shown in Table 4 and for the classifier are shown in Table 5.

5 Results

A TIRA [16] account is given in order to evaluate our models and get the results in a hidden test set. Additionally, task organizers permit to evaluate models in an early bird software submission phase that enables the configuration environment subjected to verified and provides an early approximation of the model's accuracy.

During the training phase, the 5-fold cross-validation strategy supported the choice of the best model. Nevertheless, as shown in Table 6, the results of our best models in the early bird software submission phase differ from the training.

Table 4. Preprocessing strategy in Profiling Fake News Spreaders on Twitter 2020 (T=True, F=False)

Strategy	***pres-case***	***red-len***	***rpl-dgt***	***demojify***	***rpl-anon***
v0_0	T	T	F	F	F
v0_1	F	T	F	F	F
v1_0	T	T	F	T	F
v1_1	F	T	F	T	F
v2_0	T	T	F	F	T
v2_0_1	T	T	T	F	T
v2_1	F	T	T	F	T
v2_1_1	F	T	T	T	T
v3_0	T	T	F	T	T
v3_0_1	T	T	T	T	T
v3_1	F	T	F	T	T
v3_1_1	F	T	T	T	T

Table 5. SVM hyper-parameters for the Profiling Fake News Spreaders on Twitter task

Param	**Values**
C	loguniform(log(1e-5), log(1e5))
loss	[hinge, squared_hinge]
tol	loguniform(log(1e-5), log(1e-2))
intercept_scaling	loguniform(log(1e-1), log(1e1))
class_weight	[None, balanced]
max_iter	2000

Table 6. Results in the Profiling Fake News Spreaders on Twitter Datasets

Model	Dataset	Lang		Average
		en	***es***	
5-fold cv	Train	0.7833	0.7900	0.7891
10-fold cv	Train	0.7500	0.8267	0.7883
Our model	Early bird	0.7300	0.8200	0.775
Our model	Test [18]	0.7350	0.8200	0.7775

For that reason, the experiments evaluated during the training changed to the 10-fold cross-validation method.

For English, our model uses the v2_1_1 preprocessing strategy –for both char and word n-grams. We apply a variety of replacement strategies, such as exchanging digits for custom tags, emojis for words, anonymous tags for custom tags without the character #, and downcasting letters. As well, we use char n-grams of orders between 1 and 6, and unigrams and bigrams for word n-grams.

For Spanish, our model uses the v3_1 preprocessing strategy –for both char and word n-grams. However, the difference regarding the English's model is that custom tags do not replace digits in Spanish. Furthermore, we use chars n-grams of orders between 2 and 6, and unigrams, bigrams, and trigrams for words n-grams.

The models that we suggested obtained one of the two best results on the Profiling Fake News Spreaders on Twitter task at the 8th Author Profiling Task at PAN 2020 [18], based on an average accuracy of 0.7775 –0.7350 for English and 0.8200 for Spanish.

6 Conclusions

In this paper, we described the submitted models for the Profiling Fake News Spreaders on Twitter task at PAN 2020. These consist of SVM as a classifier, and TF-IDF of char and word n-grams as features.

Similar to previous editions of the Author Profiling task in PAN's conferences, 2020 research shows that SVM classifiers with n-grams and TF-IDF features performed positively as our proposed models achieved one of the two best average accuracies.

Therefore, despite the similarity of the methods applied to the Author Profiling task in previous years, the best results for 2020 are for Spanish tweets.

The accuracy obtained by our models in the early bird dataset differs from the accuracy obtained during training. Consequently, to get a more accurate estimation of the model's performance, we shifted from 5-fold cross-validation to 10-fold cross-validation for the training.

Finally, the use of hyperparameter tuning tools was a crucial step to obtain positive results during the model construction process.

References

1. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. arXiv preprint arXiv:1707.03764 (2017)
2. Bergstra, J., Yamins, D., Cox, D.D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures (2013)
3. Bessi, A., Ferrara, E.: Social bots distort the 2016 us presidential election online discussion (2016)
4. Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Wiegmann, M., Zangerle, E.: Shared tasks on authorship analysis at pan 2020. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 508–516. Springer International Publishing, Cham (2020)
5. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
6. Daneshvar, S., Inkpen, D.: Gender Identification in Twitter using N-grams and LSA—Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France. CEUR Workshop Proceedings, CEUR-WS.org (Sep 2018), http://ceur-ws.org/Vol-2125/
7. Dickerson, J.P., Kagan, V., Subrahmanian, V.S.: Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). pp. 620–627 (Aug 2014). https://doi.org/10.1109/ASONAM.2014.6921650
8. Ghanem, B., Ponzetto, S.P., Rosso, P.: Factweet: Profiling fake news twitter accounts (2019)
9. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and linguistic computing **17**(4), 401–412 (2002)
10. Metaxas, P.T., Mustafaraj, E.: Social media and the elections. Science **338**(6106), 472–473 (2012). https://doi.org/10.1126/science.1230456, https://science.sciencemag.org/content/338/6106/472
11. Mustafaraj, E., Metaxas, P.T.: From obscurity to prominence in minutes: Political speech and real-time search (2010)
12. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. ACL '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004). https://doi.org/10.3115/1218955.1218990, https://doi.org/10.3115/1218955.1218990
13. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. pp. 79–86. EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1118693.1118704, https://doi.org/10.3115/1118693.1118704

14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
15. Pizarro, J.: Using N-grams to detect Bots on Twitter. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CEUR-WS.org (Sep 2019), http://ceur-ws.org/Vol-2380/
16. 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)
17. 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 Workshop Proceedings (Sep 2020), CEUR-WS.org
18. 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)
19. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs. vol. 6, pp. 199–205 (2006)
20. Stella, M., Ferrara, E., De Domenico, M.: Bots increase exposure to negative and inflammatory content in online social systems. Proceedings of the National Academy of Sciences **115**(49), 12435–12440 (2018). https://doi.org/10.1073/pnas.1803470115, https://www.pnas.org/content/115/49/12435
21. Tellez, E.S., Miranda-Jiménez, S., Moctezuma, D., Graff, M., Salgado, V., Ortiz-Bejar, J.: Gender identification through multi-modal tweet analysis using microtc and bag of visual words. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) (2018)
22. Twitter, Inc.: Q1 2019 Selected Company Metrics and Financials (2019), https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Selected-Company-Metrics-and-Financials.pdf
23. Yang, K.C., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies **1**(1), 48–61 (2019). https://doi.org/10.1002/hbe2.115, https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.115