Fake News Spreader Detection on Twitter using Character N-Grams
Notebook for PAN at CLEF 2020

Inna Vogel and Meghana Meghana

Fraunhofer Institute for Secure Information Technology SIT
Rheinstrasse 75, 64295 Darmstadt, Germany
{Inna.Vogel, Meghana.Meghana}@SIT.Fraunhofer.de

Abstract The authors of fake news often use facts from verified news sources and mix them with misinformation to create confusion and provoke unrest among the readers. The spread of fake news can thereby have serious implications on our society. They can sway political elections, push down the stock price or crush reputations of corporations or public figures. Several websites have taken on the mission of checking rumors and allegations, but are often not fast enough to check the content of all the news being disseminated. Especially social media websites have offered an easy platform for the fast propagation of information. Towards limiting fake news from being propagated among social media users, the task of this year's PAN 2020 challenge lays the focus on the fake news spreaders. The aim of the task is to determine whether it is possible to discriminate authors that have shared fake news in the past from those that have never done it. In this notebook, we describe our profiling system for the fake news detection task on Twitter. For this, we conduct different feature extraction techniques and learning experiments from a multilingual perspective, namely English and Spanish. Our final submitted systems use character n-grams as features in combination with a linear SVM for English and Logistic Regression for the Spanish language. Our submitted models achieve an overall accuracy of 73% and 79% on the English and Spanish official test set, respectively. Our experiments show that it is difficult to differentiate solidly fake news spreaders on Twitter from users who share credible information leaving room for further investigations. Our model ranked 3rd out of 72 competitors.

Keywords: Author Profiling, Fake News Spreader, Fake News Detection, Deception Detection, Social Media, Twitter

1 Introduction

Author profiling uses information of people's writing style to determine specific characteristics such as the author's gender, age, personality, or cultural and social context, like

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.

CEUR-WS.org/Vol-2696/paper_59.pdf

mother tongue and dialects [12]. Author profiling is not only used in criminal investigations and in the security sector [11] but also in marketing by specifying the target group. This year, the author profiling task of PAN 2020 was designed to investigate whether the author of a Twitter feed is a fake news spreader or not[1] [9]. The dataset provided by the organizers covers two languages: English and Spanish.

Fake news poses a serious threat to our society. They can destroy reputations of corporations and public figures, can push down the stock price and manipulate peoples opinions and therefore also their actions. Social media has become an ideal place for fake news propagation as user-generated content reaches very quickly a broad audience. Fraudsters use those networks to deceive users and shape specific opinions by making the reader believe a certain political or social agenda. The sheer mass of false information spread on the internet has reached new heights and cannot be handled by manual fact-checking alone. However, automatic recognition of fake news is a challenging task. Knowledge-based and context-based approaches to combat fake news can be applied, but only after the fake in the news has been verified by experts. This is often not fast enough as fake news spread very quickly and reach a broad audience, especially on social media websites.

Style and content-based approaches are a viable alternative [14,13,3,6,8] and have been proven to be effective in addressing the problem of author profiling in social networks [2,1]. Style-based approaches analyze how the author expresses himself while writing, whereas the content-based approaches consider the topic of the text. We propose a content-based approach by identifying possible fake news spreaders on Twitter as a first step towards preventing fake news from being propagated among online users. We investigate whether it is possible to discriminate authors that have shared fake news in the past from those who share credible information. We conduct different learning experiments for the English (EN) and Spanish (ES) language. The performance of our system is ranked by accuracy. The best-performed models achieve an overall accuracy of 73% and 79% on the English and Spanish corpus, respectively. The results show that it is not an easy task to differentiate solidly fake news spreaders from users spreading credible information. Our model ranked 3rd out of 72 competitors.

In the following, we describe our approach for the author profiling task at PAN 2020. After a review of related work in Section 2, Section 3 details the Twitter data that was provided by the PAN organizers and shows some key statistics observed in the corpus. The preprocessing steps and features used to train our models are detailed in Section 4. Our models and classification results are discussed in Section 5. We also provide some information about our alternatively tested methods (Section 6) and conclude our work in Section 7.

2 Related Work

Potthast et al. [8] used the manually fact-checked *BuzzFeed news corpus*[2] and extended it with linked articles, ratings and other metadata. The enriched *BuzzFeed-Webis Fake*

[1] PAN at CLEF 2020 "Profiling Fake News Spreaders on Twitter": https://pan.webis.de/clef20/pan20-web/author-profiling.html

[2] https://github.com/BuzzFeedNews/2016-10-facebook-fact-check

News Corpus[3] was then used to analyze the writing style of different news creators, namely mainstream, hyperpartisan and satire news. Hyperpartisan refers to extremely left-wing or right-wing standpoints. Using the unmasking method, which was originally proposed for authorship verification by Koppel et al. [4], Potthast et al. [8] showed that the writing style of extremely one-sided news and satire can be distinguished from the writing style of mainstream news (F_1 78%). Fake news, on the other hand, could not be detected by their style alone [8].

Liu and Wu [5] proposed a method to early detect fake news on social media. Therefore, a propagation path of each news was constructed as a multivariate time series. Each tuple in the path is a numerical vector which represents user characteristics who engaged in spreading the news story. The user features (e.g. length of the user name, age, followers, account verification) were extracted from the profile and transformed into a fixed-length sequence. A time series classifier was built incorporating RNN and CNN to capture the user's characteristics and to predict whether a given news story is fake or true. Experiments on two Twitter datasets and a SinaWeibo[4] corpus showed that the model can detect fake news within five minutes after it started to spread. The model achieved an accuracy of 85% on the Twitter data and 92% on the SinaWeibo corpus.

Zhou et al. [15] studied different features of fake news being spread on social networks, which refer to the news itself, the spreaders of the fake news and the relationship among the engaged users. Therefore, they analyzed features like the frequency and number of news that have been spread, the distance of the fake news spreaders in a network, or the number of user engagements. The existence of the selected patterns validated in empirical studies that fake news spread farther and attract more readers than true news. Additionally, fake news spreaders are more connected and engaged than other users. The accounts of the Twitter users derived from *PolitiFact*[5] and *BuzzFeed*[6]. The extracted features were additionally used to train classifiers such as SVM, KNN, Random Forests etc. Random Forests performed best among all the other classifiers achieving an F_1-Score of 93% on *PolitiFact* and 84% on the *BuzzFeed* corpus.

3 Dataset and Corpus Analysis

To train our system, we used the PAN 2020 author profiling corpus[7] proposed by Rangel et al. [10]. The corpus consists of 300 English (EN) and Spanish (ES) Twitter user accounts each. The tweets of every Twitter user are stored in an XML file containing 100 tweets per author. Every tweet is stored in a <*document*> XML tag. The tweets were manually collected and fact-checked. The dataset is balanced which means the data refers to an equal distribution of class instances. Half of the documents per language folder are authors that have been identified sharing fake news. The other half are texts from credible users. Table 1 shows excerpts from the data. Every author received an

[3] https://zenodo.org/record/1239675#.XrVvwWgzaUm
[4] https://www.weibo.com
[5] https://www.politifact.com
[6] https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data
[7] https://zenodo.org/record/3692319#.XrlnomgzZaQ

alphanumeric author-ID which is stored in a separate text file together with the corresponding class affiliation. For training and testing, we split the data in the ratio 70/30. The gold-standard can only be accessed through the TIRA [7] evaluation platform provided by the PAN organizers. The results are hidden for the participants.

Table 1. English (EN) and Spanish (ES) excerpts from the PAN 2020 Twitter "Fake News Spreader" data.

EN and ES True News Tweets	EN and ES Fake News Tweets
"RT #USER#: Best dunk of the contest no doubt about it. Aaron Gordon robbed again #URL#"	"Jay-Z Must Give Beyonce $5 Million Per Child They Have Together Due to Crazy Prenup... #URL#"
"RT #USER#: Sure would be an interesting day to read a book that examines Trump's obsession with the king-like powers of his offic..."	"RT #USER# #USER# When Obama was tapping my phones in October, just prior to Election!"
"A Data-Driven Approach Aims to Help Cities Recover After Earthquakes #URL#"	"Why Trump lies, and why you should care - The Boston Globe #URL#"
"Javier Cámara ya es el líder más valorado de los españoles por delante de Pedro Sánchez, según una encuesta #URL# #URL#"	"Dictadura pura y dura toma tasas y todos felicices #URL#"
"Me gusta la foto. Una foto con variedad, diversidad. Me da la impresion que con más sonrisas que otras. #URL#"	"GANAR DINERO AHORA ES FACIL – Google te paga 15 dólares por contestar encuestas #URL# #URL#"
"Navidad en RD: son 3 días gozando, luego 362 llorando y deseando mal a los demás. Dejen su hipocresía !!"	"Ortega Smith: 'VOX expulsará de España a todos los inmigrantes ilegales' #URL#"

As can be seen in Table 1, the Twitter specific tokens hashtags, URLs and user mentions were replaced by the providers with the following placeholders: *#HASHTAG#*, *#URL#* and *#USER#*. Prior to the feature engineering, we analyzed the distribution of different tokens. Additionally, we determined the sentiment of each tweet (positive, negative, or neutral) using *TextBlob*[8]. For recognizing the named entities (NER), we used the Python library *spaCy*. Table 2 shows some key insights for both languages.

The observations of the corpus content were the following:

- Fake news spreaders:
  - mention other Twitter users less often (*#USER#*[9]).
  - utilize fewer hashtags (#HASHTAG#).
  - re-post fewer tweets (RT).
  - share slightly more URLs (#URL#).
- Spanish speaking authors use more emojis than English speaking Twitter users.
- Half of the English tweets are based on factual information and most of the Spanish tweets (90%) are free of emotions.

[8] https://textblob.readthedocs.io/en/dev
[9] e.g. "@Username"

Table 2. Feature distribution of the fake news (Fake) and true news (True) spreaders

	English		Spanish	
Features	**True**	**Fake**	**True**	**Fake**
Unique Tokens	24,050	23,809	32,802	27,932
Emojis Total	1,614	522	3,867	1,629
Emojis Unique	325	145	603	301
Neutral Tweets	6,857	7,061	14,228	14,261
Positive Tweets	6,173	5,464	571	488
Negative Tweets	1,970	2,475	201	251
Uppercased Tokens Total	38,519	32,467	36,388	30,177
Uppercased Phrases Total	861	1,019	406	953
#URL# Token	16,565	17,018	10,887	13,900
#HASHTAG# Token	6,739	4,715	5,905	1,580
#USER# Token	5,628	2,279	10,668	5,949
Retweets (RT)	2,383	1,158	4,289	1,977
NER ORG	8,340	7,299	2,617	2,595
NER PERSON	7,742	9,801	4,845	5,573
NER LOC	188	222	5,337	5,214

– Fake news tend to be more often negative.
– Tweets of true news spreaders tend to be more often positive.
– By counting the named entities no significant difference between the classes could be established.
– Fake news spreaders tend to tweet slightly more often about other people.
– Uppercased tokens are shared equally by true news and fake news spreaders.
– Spanish fake news spreaders make more often use of capitalized phrases.

4 Preprocessing and Feature Extraction

The preprocessing pipeline was performed for both languages (EN and ES) basically. The steps for cleaning and structuring the data were performed as follows:

1. First, we extracted the text from the original XML document of each user and concatenated all 100 tweets to a single text.
2. White space between tokens were normalized to a single space.
3. URLs, hashtags and user mentions were left untouched as they are already replaced by placeholders by default.
4. Numbers and emojis were replaced by the placeholders #NUMBER# and #EMOJI#.
5. Irrelevant signs, e.g. *"+,*,/,"* were deleted.
6. Sequences of repeated characters with a length greater than three were normalized to a maximum of two letters (e.g. "LOOOOOOOOL" to "LOOL").
7. Words with less than three characters were ignored.
8. Stopwords were deleted by using the NLTK (Natural Language Toolkit) library[10] for each language separately.

[10] https://www.nltk.org/

9. From the NLTK library we additionally used the *TwitterTokenizer* to tokenize the words. The tokenizer is suitable for Twitter and other casual speech that is often used in social networks. Additionally, *TwitterTokenizer* contains different regularization and normalization features. We made use of the lowercaser.

After the Twitter texts were preprocessed, we tested different vectorization techniques with manual hyperparameter tuning, and by employing scikit-learn's grid search function. The hyperparameters were tuned separately for English and Spanish, but the features we used were mainly language-independent which means that the same set of features can be used in multi-language domains. The selected features were presented in Section 3 (e.g. counts of tokens or named entities). The only language dependant feature we experimented with was the sentiment polarity calculated separately for every tweet (whether it is positive, negative, or neutral). Besides the handcrafted features, we also experimented with automatically learned features i.e. term frequency distribution (tf) and character and word n-grams. Additionally, we made use of *Feature Union*[11] to experiment with feature concatenation. To convert the tokens to a numerical matrix in order to build a vector for each language, we made use of:

(1) Scikit-learn's term frequency-inverse document frequency (TF-IDF)
(2) *GloVe*[12] (Global Vectors for Word Representation) word vectors pre-trained on Twitter data as well as custom trained *word2vec*[13] word embeddings
(3) Scikit-learn's Count Vectorizer

All tested features and their representations are summarized in Table 3.

Table 3. Features, vectorization techniques and model hyperparameters used for training purposes

Features	**Vectorizer**	**Hyperparameters / ranges**
Tokens	Word Embeddings	n-gram_range: $[1;3]$,$[2;7]$,$[3;7]$
Token n-grams	TF-IDF	min_df: 1,2,3
Character n-grams	Count Vectorizer	max_features: $[1,000;50,000]$

5 Methodology

We defined the author profiling task as a binary problem predicting whether a tweet was composed by a fake news spreader or a reliable Twitter user. For each language (EN and ES) a separate classification model was trained. As mentioned before, for training and testing, we split the data in the ratio 70/30. We tested different features, vectorization techniques and dimensionality sizes in combination with a Support Vector Machine

[11] https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html
[12] https://nlp.stanford.edu/projects/glove
[13] https://radimrehurek.com/gensim/models/word2vec.html

(SVM) and Logistic Regression of which we report the best performed ones. For the final SVM, we used a linear kernel with default hyperparameter values[14]. Logistic Regression was also trained by utilizing default hyperparameters[15].

The performance of the fake news spreader author profiling task was ranked by accuracy. Table 4 shows the scores for our final system performed on the official PAN 2020 test set on the TIRA platform [7]. Accuracy scores were calculated individually for each language by discriminating between the two classes. Each model was trained on 70% of the training data. Hyperparameters were tuned on the remaining 30% split. As the data set is hidden, the four confusion matrix values (TP, TN, FP and FN) and other metrics like Precision and Recall cannot be provided. Therefore, we display these classification results and accuracy scores which we achieved on the 30% test dataset (see Table 5). The highest accuracy in English was obtained using SVM with TF-IDF weighted character n-grams with range $[1;3]$ and top 3,000 features. In Spanish, the best results were achieved using Logistic Regression employing a feature union of TF-IDF weighted character n-grams with range $[1;3]$ and top 5,000 features and a vector consisting of character n-gram counts with range $[3;7]$ and top 50,000 features. The submitted models achieve an overall accuracy of 73% and 79% on the English and Spanish corpus, respectively.

Table 4. Accuracy (Acc.) scores of the final submitted systems on the official PAN 2020 test dataset on Tira

Model	Features	Language	Acc.
SVM	TF-IDF char n-grams [1;3] 3,000 features	EN	**0.73**
Logistic Regression	Feature union TF-IDF char n-grams [1;3] 5,000 features and char n-gram counts [3;7] 50,000 features	ES	**0.79**

Table 5. Evaluation results on the test split of the submitted systems for every language (EN and ES) with the metrics Precision (P), Recall (R), Accuracy (Acc.) and F_1-Score

			Confusion Matrix							
Model	Features	Language	TP	TN	FP	FN	P	R	F_1	Acc.
SVM	TF-IDF char n-grams [1;3] 3,000 features	EN	35	35	10	10	0.78	0.78	0.78	**0.78**
Logistic Regression	Feature union TF-IDF char n-grams [1;3] 5,000 features and char n-gram counts [3;7] 50,000 features	ES	42	36	9	3	0.92	0.80	0.86	**0.87**

[14] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[15] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

6 Other Tested Methods and Features

In this Section, we report our experiments with alternatively tested feature selections and representation techniques which were not able to keep up with the systems described above in terms of performance (see Section 5). Besides character n-grams, we also experimented with word n-grams in the range of [1;7]. Other selected features comprised counts of emojis, uppercase tokens and phrases, hashtags, user mentions, URLs and retweets. Additionally, we incorporated sentiment analysis in our vector by using *TextBlob*. The selected features we presented in Section 4 and Table 3.

Besides TF-IDF, we tested term frequencies (tf) and word embeddings as feature representations. Therefore, we utilized *GloVe* word vectors pre-trained on Twitter data as well as custom trained *word2vec* word embeddings. To combine the different features in one vector, the inner product space of two vectors was required. First, all texts of the fake news spreaders were concatenated and vectorized. Then, the cosine similarity of this vector and every twitter user was determined. The resulting vector comprising a varying number of features was standardized (using *StandardScaler* [16]). The final vector was then forwarded to train the SVM and Logistic Regression models. Our aim was to test whether emotions and sentiments, emojis, or uppercase tokens in fake news could improve the classification performance. The training results showed that none of those features or feature combinations could improve the performance in both languages. The accuracy has even slightly decreased.

7 Discussion and Conclusion

In this paper, we described our participation in the author profiling task at PAN 2020. The goal was to develop a system for profiling fake news spreaders on Twitter as a first step towards preventing the propagation of fake news among online users. For our experiments, we used the PAN 2020 author profiling corpus provided by the organizers. We conducted different learning experiments from a multilingual perspective, namely English and Spanish. We evaluated different features, most of them language-independent. The features were extracted and had their importance evaluated in the detection task. We provided some corpus statistics that showed that there are differences between fake and true news spreaders. We experimented with different features, vectorization techniques and dimensionality sizes.

For the English language, our model performed best using SVM with TF-IDF weighted character n-grams with range $[1;3]$ and top 3,000 features. For the Spanish language, the best results were achieved using Logistic Regression employing a feature union of TF-IDF weighted character n-grams with range $[1;3]$ and top 5,000 features and a vector consisting of character n-gram counts with range $[3;7]$ and top 50,000 features. The submitted models achieve an overall accuracy of 73% and 79% on the English and Spanish corpus, respectively. Our model ranked 3rd out of 72 competitors.

The results showed that it is challenging to detect fake news spreaders in Twitter data. It was challenging in two ways. First, not every tweet of a fake news spreader is

[16] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

false but a mixture of true and false information. Second, Twitter data is short, noisy and incorporates platform-specific features (such as user mentions and retweets). The biggest challenge is the orthography. The tweets are strewn with spelling mistakes and grammatical errors. Word-level based approaches perform poorly compared to approaches based on character n-grams.

In the future, we first want to experiment with style-based approaches in order to determine whether fake news spreaders can be identified by the writing style alone. Finally, we plan to experiment with different standardization and pre-processing techniques as our submitted system does not consider misspelled words.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE and under grant agreement "Lernlabor Cybersicherheit" (LLCS) for cyber security research and training.

References

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Meza, I.: Evaluating topic-based representations for author profiling in social media. In: Montes y Gómez, M., Escalante, H.J., Segura, A., Murillo, J.d.D. (eds.) Advances in Artificial Intelligence - IBERAMIA 2016. pp. 151–162. Springer International Publishing, Cham (2016)
2. Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J.W.: Lexical predictors of personality type. In: Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America (01 2005)
3. Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 877–880 (2019)
4. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Brodley, C.E. (ed.) Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4–8, 2004. ACM International Conference Proceeding Series, vol. 69. ACM (2004), http://doi.acm.org/10.1145/1015330.1015448
5. Liu, Y., Wu, Y.B.: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 354–361. AAAI Press (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16826
6. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3391–3401. Association for Computational Linguistics (2018), http://aclweb.org/anthology/C18-1287

7. 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 - Lessons Learned from 20 Years of CLEF. Springer (2019)
8. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. In: The 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). Association for Computational Linguistics (2018), http://arxiv.org/abs/1702.05638
9. 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
10. Rangel, F., Rosso, P., Ghanem, B., Giachanou, A.: Profiling fake news spreaders on twitter. In: PAN at CLEF 2020 Fake News Spreader Twitter Dataset. Zenodo (Feb 2020), https://doi.org/10.5281/zenodo.3692319
11. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)
12. Russell, C.A., Miller, B.H.: Profile of a terrorist. Studies in conflict & terrorism 1(1), 17–34 (1977), https://doi.org/10.1080/10576107708435394
13. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: Fever: a large-scale dataset for fact extraction and verification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 809–819. Association for Computational Linguistics (2018), http://aclweb.org/anthology/N18-1074
14. Wang, W.Y.: Liar, liar pants on fire: A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 422–426. Association for Computational Linguistics (2017)
15. Zhou, X., Zafarani, R.: Network-based fake news detection: A pattern-driven approach. SIGKDD Explor. Newsl. 21(2), 48–60 (Nov 2019), https://doi.org/10.1145/3373464.3373473