Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter

Francisco Rangel[1], Anastasia Giachanou[2], Bilal Ghanem[2,1] and Paolo Rosso[2]

[1]Symanto Research, Germany
[2]Universitat Politècnica de València, Spain

Abstract. This overview presents the Author Profiling shared task at PAN 2020. The focus of this year's task is on determining whether or not the author of a Twitter feed is keen to spread fake news. Two have been the main aims: (i) to show the feasibility of automatically identifying potential fake news spreaders in Twitter; and (ii) to show the difficulty of identifying them when they do not limit themselves to just retweet domain-specific news. For this purpose a corpus with Twitter data has been provided, covering the English and Spanish languages. Altogether, the approaches of 66 participants have been evaluated.

1 Introduction

The rise of social media has given the opportunity to users to publish and share content online in a very fast way. The easiness of publishing content in social media has led to an increase in the amount of misinformation that is published and shared. The propagation of fake news that is shown to be faster than the one of real news [64] is causing several negative consequences in the society. One of the most recent cases is the large amount of misinformation that was propagated related to the origin, prevention, diagnosis, and treatment of COVID-19 pandemic and that affected the society in different ways. For example, fake news about the effectiveness of the chloroquine led to an increase of cases of chloroquine drug overdose [11]. The influence of fake news is also evident in other domains as for example in the political domain, and researchers have drawn attention to their influence regarding elections and referendums outcomes [6].

Understanding whether a piece of news is fake or not is a very challenging task for users who, in their majority are not experts. In addition, fake news usually contain a mixture of real and fake claims in an attempt to further confuse users. In an effort to raise awareness and inform users about which pieces of news contain fake information, several platforms (e.g., Snopes[1], PolitiFact[2],

[1] https://www.snopes.com
[2] https://www.politifact.com/

Leadstories[3]) have been developed. These platforms employ journalists or other domain experts who thoroughly examine the information presented in various articles before they label them based on their factuality.

Our hypothesis is that users who do not spread fake news may have a set of different characteristics compared to users who tend to share fake news. For example, they may use different linguistic patterns when they share posts compared to fake news spreaders. This is what we aim at investigating in this year's author profiling shared task where we address the problem of fake news detection from the author profiling perspective. The final goal is profiling those authors who have shared some fake news in the past. This will allow for identifying possible fake news spreaders on Twitter as a first step towards preventing fake news from being propagated among social media users. This should help for their early detection and, therefore, for preventing their further dissemination.

The remainder of this paper is organized as follows. Section 2 covers the state of the art, Section 3 describes the corpus and the evaluation measures, and Section 4 presents the approaches submitted by the participants. Sections 5 and 6 discuss results and draw conclusions respectively.

2 Related Work

Fake news detection has recently received significant research attention. Among others, researchers have focused on fake news [52,59,29], bots [48,14] and clickbaits [2,46] detection. Some of the previously proposed approaches have explored the effectiveness of linguistic patterns such as the number of pronouns and punctuation marks on the detection of fake news. For example, Rashkin *et al.* [51] analysed various linguistic features such as personal pronouns and swear words that were incorporated into a Long Short Term Memory (LSTM) network to address credibility detection. Other researchers, proposed to use the emotions expressed in the piece of news. In this direction, Giachanou *et al.* [23] proposed emoCred, an LSTM-based neural network that utilised emotions from text, whereas Ghanem *et al.* [21] proposed to incorporate emotions extracted from text into an LSTM network and showed that emotions are useful for the classification of the different types of fake news. Guo *et al.* [29] proposed a dual emotion-based fake news detection framework to learn content and comment emotion representations for publishers and users respectively, whereas Wang [65] proposed a hybrid convolutional neural network to combine user metadata with text for fake news detection.

Although the detection of fake news, and credibility in general, has received a lot of research attention [23,65,29,59], there are only few studies that have addressed the problem from a user or author profiling perspective. One of the studies that focused on users was presented by Shu *et al.* [60] who analyzed different features, such as registration time, and found that users that share fake news have more recent accounts than users who share real news. Vo and Lee [62]

[3] https://leadstories.com/

analyzed the linguistic characteristics (e.g., use of tenses, number of pronouns) of fact-checking tweets and proposed a deep learning framework to generate responses with fact-checking intention. Recently, Giachanou *et al.* [22] explored the impact of the personality traits of users in discriminating between users who tend to share fake news and fact-checkers. In their study, they proposed a model based on a Convolutional Neural Network (CNN) that combines word embeddings from the text with features that represent users' personality traits and linguistic patterns and showed that those features are useful for the task. Ghanem *et al.* [20] proposed a model that utilizes chunked timelines of tweets and a recurrent neural model in order to infer the factuality of a Twitter news account.

With regards to author profiling, early attempts focused on profiling the authors of blogs and formal text [3,35]. However, with the rise of social media researchers proposed methodologies to profile the authors of social media posts where the language is more informal [10,56]. Previous author profiling tasks at PAN have tried to profile different characteristics of users. In 2019 the PAN author profiling task aimed to classify an author of a tweet feed as a bot or human [48], whereas in 2018 the task focused on multimodal gender identification in Twitter for which images were also provided together with the text for the classification [50]. Since 2013 a wide range of approaches have been developed and tested in the author profiling tasks. Maharjan *et al.* [38] proposed a MapReduce architecture to address the gender identification task with 3 million features on the PAN-AP-2013 corpus, whereas Bayot and Gonçalves [7] showed that word embeddings work better than TF-IDF for gender detection on the PAN-AP-2016 corpus. At PAN 2019 the best results in bots detection in English was obtained by Johansson [33] who used Random Forest with a variety of stylistic features such as term occurrences, tweets length or number of capital and lower letters, user mentions etc., whereas in gender identification in English the best result was obtained by Valencia et al. [61] with Logistic Regression and n-grams. Finally, in Spanish, Pizarro [44] achieved the best results in both bots and gender identification with combinations of n-grams and Support Vector Machines.

3 Evaluation Framework

The purpose of this section is to introduce the technical background. We outline the construction of the corpus, introduce the performance measures and baselines, and describe the idea of so-called software submissions.

3.1 Corpus

To build the PAN-AP-2020 corpus[4] we have proceeded as follows. Firstly, we have reviewed fact-checking websites such as PolitiFact or Snopes to find news

[4] We should highlight that we are aware of the legal and ethical issues related to collecting, analysing and profiling social media data [47] and that we are committed to legal and ethical compliance in our scientific research and its outcomes.

labelled as fake[5]. Secondly, we have searched for these news on Twitter. We downloaded all the possible tweets containing some information related to the identified fake news (e.g., the topics mentioned on the news) and manually inspected them to discard those not actually referring to the news. We also manually inspected the collected tweets to re-label them as supporting or not the fake news. With this step, we label as real news those tweets where the user warns about the fake news. Thirdly, for the identified users, we collected their timelines and checked all the tweets with the list of fake news identified in the first step together with an overall manual inspection. If the user has shared at least one fake news, we labelled it as keen to spread fake news. Otherwise, if the user, to the best of our knowledge, had not shared any fake news in her timeline, we labelled the user as real news spreader. Finally, we have ordered the users by the number of shared fake news and picked up the ones with the highest ranking. We balanced the corpus picking up the same number of real news spreaders. To ensure that the classifiers will not bias towards the identified fake news topics, we have removed the tweets containing them from the whole corpus.

Language	Training	Test	Total
English	300	200	500
Spanish	300	200	500

Table 1: Number of authors in the PAN-AP-20 corpus created for this task.

Table 1 presents the statistics of the corpus that consists of 500 authors for each of the two languages, English and Spanish. For each author, we retrieved via the Twitter API her last 100 Tweets. The corpus for each language is balanced, with 250 authors for each class (fake and real news spreaders). We have split the corpus into training and test sets, following the 60/40 proportion.

3.2 Performance Measure

The performance of the systems has been ranked by accuracy. For each language, we calculated individual accuracy in discriminating between the two classes. Finally, we averaged the accuracy values per language to obtain the final ranking.

3.3 Baselines

As baselines to compare the performance of the participants with, we have selected:

– *RANDOM.* A baseline that randomly generates the predictions among the different classes.

[5] We have manually reviewed these "fake news" to ensure that there was not political manipulation behind them and that the news is clearly fake.

- *LSTM.* An Long Short-Term Memory neural network that uses FastText[6] embeddings to represent texts.
- *NN + w nGrams.* Word n-grams with values for n from 1 to 3, and a Neural Network.
- *SVM + c nGrams.* Character n-grams with values for n from 2 to 6, and a Support Vector Machine.
- *SYMANTO (LDSE)* [49]. This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDSE is a weight, representing the probability of a term to belong to one of the different categories: fake news spreader / real news spreader. The distribution of weights for a given document should be closer to the weights of its corresponding category. LDSE takes advantage of the whole vocabulary.
- EIN. the Emotionally-Infused Neural (EIN) network [21] with word embedding and emotional features as the input of an LSTM.

3.4 Software Submissions

Similar to previous year, we asked for software submissions. Within software submissions, participants submit executables of their author profiling softwares instead of just the output of their softwares on a given test set. For the software submissions, the TIRA experimentation platform was employed [27,28], which renders the handling of software submissions at scale as simple as handling run submissions. Using TIRA, participants deploy their software on virtual machines at our site, which allows us to keep them in a running state [26].

4 Overview of the Submitted Approaches

This year, 66 teams participated in the Author Profiling shared task and 33 of them submitted the notebook paper. We analyse their approaches from three perspectives: preprocessing, features used to represent the authors' texts and classification approaches.

4.1 Preprocessing

With the aim at preventing bias towards some URLs, user mentions or hashtags, the corpus was provided with these elements already masked. In the same vein, some participants cleaned other Twitter-specific elements such as RT, VIA, and FAV[7] reserved words [24,30,43], as well as emojis and other non-alphanumeric characters [9,45,63,24,16,39,37,57], numbers [45,63,24,16,30,57] or

[6] https://fasttext.cc/docs/en/crawl-vectors.html

[7] RT is the acronym for *retweet*; VIA is a way to give the authorship to a user (e.g., "via @kicorangel"); and FAV stands for *favourite*.

punctuation signs [63,34,24,16,30,37,57]. Various participants lower-cased the texts [9,45,63,43], removed stop words [63,34,24,17,30,37,57] or treated character flooding [63,36]. Finally, some users got rid of short texts [63], stemmed or lemmatised [24,30,37,57] and tokenised [63,36,18,17,37,57,5]. Some users also removed infrequent terms [32].

4.2 Features

The participants have used a high variety of different features and their combinations, albeit we can group them into the following main groups: *(i)* n-grams; *(ii)* stylistics; *(iii)* personality and emotions; and *iv)* embeddings. As every year, one of the most used features has been the combination of n-grams. For example, Pizarro [45] and Espinosa *et al.* [16] combined character and word n-grams. TF-IDF n-grams have been used by Vogel *et al.* [63], Koloski *et al.* [34], López-Fernández *et al.* [18] and Vijayasaradhi *et al.* [43]. Regarding combinations of stylistic-based features, Manna *et al.* [40] combined the average number of emojis (classified by category such as affection, emotion, sceptical, concerned, etc.), the number of URLs, spaces, digits, punctuation marks, tags, quotes, etc., and lexical features such as groups of words expressing personal opinions in addition to personal pronouns or verbs and expressions related to clickbait headlines.

However, most participants combined n-grams with stylistic-, personality- and emotional-based features. For instance, Buda and Bolonyai [9] combined n-grams with some statistics from the tweets, such as their average length or their lexical diversity. Lichouri *et al.* [37] combined TF-IDF word and character n-grams with POS, stemmed and lemmatised tokens. Justin *et al.* [19] combined personality, emotions, style-based features with word embeddings. For personality extraction, they used a classifier to obtain the MBTI[8] indicator [8], while for the emotions they used the NRC emotion lexicon [41]. Finally, they also computed the frequencies of different grammatical constructs such as the frequency of auxiliaries, verbs, pronouns, adjectives, punctuation, etc. Niven *et al.* [42] obtained the frequencies of adverbs, impersonal and personal pronouns, and all the function words. They also used a constituency tree parser to measure the sentence complexity, taking the average branching factor, the average max noun phrase and verb phrase heights, etc. Finally, they combined all the previous features with a measure of emotional content by means of SentiWordNet. Russo [53] has combined stylistic features such as type/token ratio, number of mentions, URLs, hashtags, celebrities counting, punctuation signs or replies with emotional features. Hörtenhuemer [31] combined 8 different feature sets: TF-IDF, average word length, sentence embedding, POS tagging, recognition of named entities, sentiment analysis (positive/negative), emotional analysis (10 emotions) and readability scores. Espinosa *et al.* [17] extracted psychographic features with the Symanto API[9]. Concretely: *(i)* personality traits of the author

[8] Myers-Briggs Type Indicator

[9] `https://developers.symanto.net/`

of the tweets (emotional vs. rational); *(ii)* communication styles of the Twitter user (self-revealing, action-seeking, information-seeking, fact-oriented); and *(iii)* sentiment analysis (positive/negative). They combined the previous features with other linguistic features, Twitter action features and headline analysis data. Cardaioli *et al.* [12] used 10 stylometric features aiming to summarise the writing style of the author. In particular, the diversity score, readability score, hashtags average, user mentions average, URLs average, retweets, lower and upper cases, punctuation signs, etc. They combined the previous features with the Big Five personality traits[10] obtained with Watson Personality Insights by IBM[11].

Some participants also combined the previous types of features with different types of embeddings. For example, Spezzano *et al.* [58] combined: *(i)* style, such as the average number of words, characters, lower and upper case words and characters, stop words, punctuation symbols, hashtags, URLs, user mentions, emojis and smiles; *(ii)* n-grams, obtaining TF-IDF for words and characters; *(iii)* tweet embeddings, computed using BERT; and *(iv)* sentiment analysis, by means of the Valence Aware Dictionary and sEntiment Reasoner (VADER) [25]. Agirrezabal *et al.* [1] combined word embeddings, including also the standard deviation of each dimension of the vectors, with bag-of-pos, the average length of the tweets, or the ratio of upper-cased characters. Fahim *et al.* [54] used a combination of word embeddings (Glove Twitter 5D) with stylistic features obtained from the hashtags, elongated words, emphasis words, curse words or emoticons. Ogaltsov *et al.* [4] combined TF-IDF features with hand-crafted ones such as whether the tweet contained the name of a celebrity, Shashirekha *et al.* [57] ensembled TF-IDF n-grams of different size with Doc2vec embeddings, and Babaei [24] combined TF-IDF and word n-grams with ConceptNet word embeddings. Labadie *et al.* [36] combined the relationship between the use of grammatical structures like the number of nouns, adjectives, lengths of words and function words, with the encoding of a dense vector at the word- and character-level. Hashemi *et al.*. [30] combined word embeddings, with TF-IDF vectors and statistical features such as the ratio of retweets, the average number of mentions/URLs/hashtags per tweet, and the average length of the tweet.

Other participants approached the task only with embeddings. Cilet *et al.* [13] used a multilingual sentence encoder to feed their pre-trained CNN. Similarly, Majumder [39] used Google's universal sentence encoder for their LSTM approach. The popular BERT has been used by Kaushik *et al.* [15], Baruah *et al.* [5], and Chien *et al.* [66].

Finally, a couple of participants approached the task from different perspectives. Moreno-Sandoval *et al.* [55] obtained social network tokens such us hashtags, URLs and user mentions, and then analysed the statistics of central tendency metrics. They combined the previous features with sentiments and emotions. Ikae *et al.* [32] estimated the occurrence probability difference of terms in both classes and generated a couple of clusters with them, reducing the dimensionality with a chi-square method.

[10] `https://en.wikipedia.org/wiki/Big_Five_personality_traits`
[11] `https://personality-insights-demo.ng.bluemix.net/`

4.3 Approaches

Regarding the classification approaches, most participants used traditional approaches, mainly Support Vector Machines (SVM) [45,63,34,16,18,30,37,1,19], Logistic Regression [9,63,34,31,43,1,40], or a combination of both depending on the language. Random Forest [12,17,30,1,55,40] is the third most used classification algorithm. Ensembles of classifiers have been used by various authors. For example, Decision Tree, Random Forest and XGB [32]; SVM, Logistic Regression, Random Forest and Extra Tree [58]; Linear SVM and Logistic Regression [57]; or SVM, Random Forest and Naive Bayes with XGBoost [42].

The author of [1] has used Multilayer Perceptron and the authors in [5] a Neural Network with Dense layer. However, only a few participants went beyond to experiment with more deep approaches. For example, Fully-Connected Neural Networks [24], CNN [13], LSTM [39,36], or Bi-LSTM with self-attention [54]. Finally, the authors of [4] ensembled a GRU-based aggregation model with CNN.

5 Evaluation and Discussion of the Results

In this section, we present the results of the shared task, as well as we analyse the most common errors made by the best performing teams. Although we recommended to participate in both languages (English and Spanish), some participants only participated in English. We present the results for the two languages, and obtain the ranking by averaging them.

5.1 Global Ranking

In Table 3, the overall performance of the participants is presented. The results are shown in terms of accuracy for both languages, and the ranking is its average.

Table 2: Statistics on the accuracy per language.

STAT	EN	ES	AVG
Min	0.5250	0.5050	0.5150
Q1	0.6512	0.7250	0.6950
Median	0.6850	0.7450	0.7125
Mean	0.6733	0.7318	0.7039
SDev	0.0511	0.0650	0.0489
Q3	0.7100	0.7650	0.7325
Max	0.7500	0.8200	0.7775
Skewness	3.3252	7.6268	7.2518
Kurtosis	-1.0222	-2.1130	-1.7969
Normality (p-value)	1.41e-05	6.247e-10	5.074e-06

The best results have been obtained in Spanish (82% vs. 75%). The overall best result (77.75%) has been obtained in a tie by Pizarro [45] and Buda and Bolonyai [9]. Pizarro obtained the best result in Spanish (82% vs. 80.05%) while

Buda and Bolonyai did in English (75% vs. 73.5%). Pizarro approached the task with combinations of character and word n-grams and Support Vector Machines whereas Buda and Bolonyai approached the task with a Logistic Regression ensemble of five sub-models: n-grams with Logistic Regression, n-grams with SVM, n-grams with Random Forest, n-grams with XGBoost and XGBoost with features based on textual descriptive statistics such as the average length of the tweets or their lexical diversity.

	PARTICIPANT	EN	ES	AVG
1	bolonyai20	**0.750**	0.805	0.7775
1	pizarro20	0.735	**0.820**	0.7775
	SYMANTO (LDSE)	*0.745*	*0.790*	*0.7675*
3	koloski20	0.715	0.795	0.7550
3	deborjavalero20	0.730	0.780	0.7550
3	vogel20	0.725	0.785	0.7550
6	higueraporras20	0.725	0.775	0.7500
6	tarela20	0.725	0.775	0.7500
8	babaei20	0.725	0.765	0.7450
9	staykovski20	0.705	0.775	0.7400
9	hashemi20	0.695	0.785	0.7400
11	estevecasademunt20	0.710	0.765	0.7375
12	castellanospellecer20	0.710	0.760	0.7350
	SVM + c nGrams	*0.680*	*0.790*	*0.7350*
13	shrestha20	0.710	0.755	0.7325
13	tommasel20	0.690	0.775	0.7325
15	johansson20	0.720	0.735	0.7275
15	murauer20	0.685	0.770	0.7275
17	espinosagonzales20	0.690	0.760	0.7250
17	ikae20	0.725	0.725	0.7250
19	morenosandoval20	0.715	0.730	0.7225
20	majumder20	0.640	0.800	0.7200
20	sanchezromero20	0.685	0.755	0.7200
22	lopezchilet20	0.680	0.755	0.7175
22	nadalalmela20	0.680	0.755	0.7175
22	carrodve20	0.710	0.725	0.7175
25	gil20	0.695	0.735	0.7150
26	elexpuruortiz20	0.680	0.745	0.7125
26	labadietamayo20	0.705	0.720	0.7125
28	grafiaperez20	0.675	0.745	0.7100
28	jilka20	0.665	0.755	0.7100
28	lopezfernandez20	0.685	0.735	0.7100
31	pinnaparaju20	0.715	0.700	0.7075
31	aguirrezabal20	0.690	0.725	0.7075
33	kengyi20	0.655	0.755	0.7050
33	gowda20	0.675	0.735	0.7050
33	jakers20	0.675	0.735	0.7050
33	cosin20	0.705	0.705	0.7050

	Participant	En	Es	Avg
37	navarromartinez20	0.660	0.745	0.7025
38	heilmann20	0.655	0.745	0.7000
39	cardaioli20	0.675	0.715	0.6950
39	females20	0.605	0.785	0.6950
39	kaushikamardas20	0.700	0.690	0.6950
	NN + w nGrams	*0.690*	*0.700*	*0.6950*
42	monteroceballos20	0.630	0.745	0.6875
43	ogaltsov20	0.695	0.665	0.6800
44	botticebria20	0.625	0.720	0.6725
45	lichouri20	0.585	0.760	0.6725
46	manna20	0.595	0.725	0.6600
47	fersini20	0.600	0.715	0.6575
48	jardon20	0.545	0.750	0.6475
	EIN	*0.640*	*0.640*	*0.6400*
49	shashirekha20	0.620	0.645	0.6325
50	datatontos20	0.725	0.530	0.6275
51	soleramo20	0.610	0.615	0.6125
	LSTM	*0.560*	*0.600*	*0.5800*
52	russo20	0.580	0.515	0.5475
53	igualadamoraga20	0.525	0.505	0.5150
	RANDOM	*0.510*	*0.500*	*0.5050*

	Participant	En
54	hoertenhuemer20	0.725
55	duan20	0.720
55	andmangenix20	0.720
57	saeed20	0.700
58	baruah20	0.690
59	anthonio20	0.685
60	zhang20	0.670
61	espinosaruiz20	0.665
62	shen20	0.650
63	suareztrashorras20	0.640
64	niven20	0.610
65	margoes20	0.570
66	wu20	0.560

Table 3: Overall accuracy of the submission to the task on profiling fake news spreaders on Twitter: Teams that participated in both languages (English and Spanish) are ranked by the average accuracy between both languages, teams that participated only in English (bottom right) are ranked by the accuracy on English. The best results for each language are printed in bold.

We should highlight the high performance of the n-grams-based approaches on this task. The participants in the following positions also used this kind of features. Koloski *et al.* used Logistic Regression and Support Vector Machines,

depending on the language, with combinations of character and word n-grams, and Vogel *et al.* [63] also used TF-IDF and character n-grams to train a Support Vector Machine classifier. De-Borja and Higueras-Porras[12] approached the problem with TF-IDF word and character n-grams and Support Vector Machines and Naïve Bayes respectively. Babaei *et al.* [24] is the top-ranked participant who used some kind of deep learning approach. Concretely, they used a Fully Connected Neural Network combining a word embedding representation based on CoceptNet with TF-IDF word n-grams. Only Pizarro and Buda and Bolonyai outperformed the Symanto (LDSE) [49] baseline.

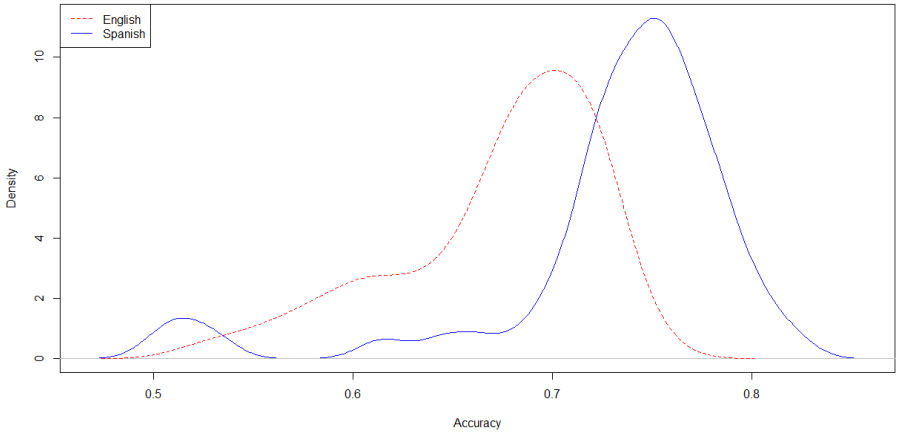

Fig. 1: Density of the results in the different languages.

As can be seen in Figure 1 and Table 2, the results for Spanish are higher than for English both in terms of average (73.18% vs. 67.33%) and maximum (82% vs. 75%) accuracies. Although the standard deviation is larger for Spanish (6.5% vs. 5.11%), the inter-quartile range is larger for English (5.88% vs. 4%), showing a slightly more sparse distribution in this last language. This might be due to the highest number of outliers in the Spanish distribution, as shown in Figure 2.

[12] Although the authors did not submit their working notes, they sent us a brief description of their system.

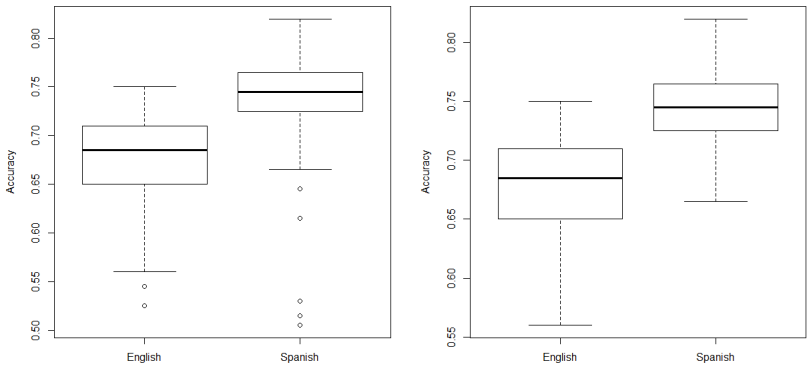

Fig. 2: Distribution of results in the different languages. The figure on the left represents all the systems. The figure on the right removes the outliers.

5.2 Error Analysis

We have aggregated all the participants' predictions for the fake news spreaders vs. real news spreaders discrimination task, except baselines, and plotted the respective confusion matrices for English and Spanish in Figures 3 and 4, respectively.

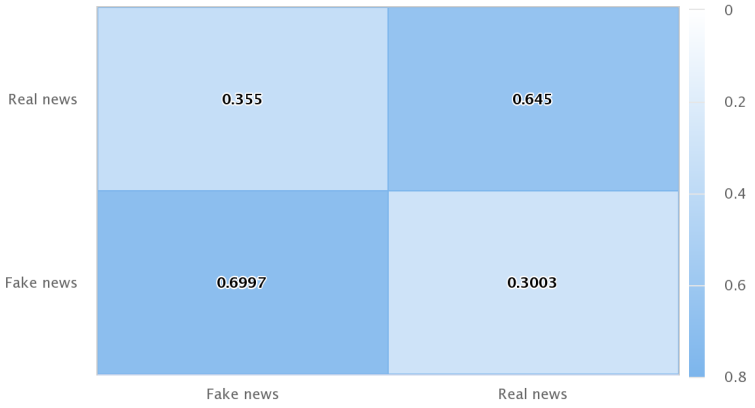

Fig. 3: Aggregated confusion matrix for fake news spreaders vs. real news spreaders discrimination in English.

In the case of English (Figure 3), the highest confusion is from real news spreaders to fake news spreaders (35.50% vs. 30.03%). This means that, for this language in this corpus, the number of false positives is higher than one-third. Regarding Spanish (Figure 4), the highest confusion is from fake news spreaders to real news spreaders (35.09% vs. 20.23%). In this case, the number of false negatives is higher, but the number of false positives is still high (one-fifth).

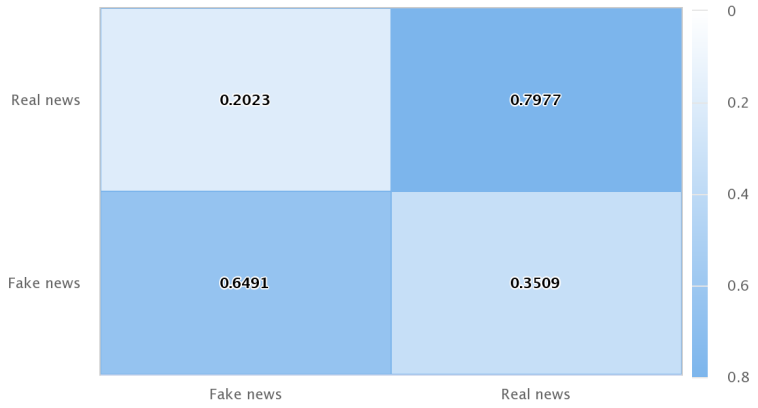

Fig. 4: Aggregated confusion matrix for fake news spreaders vs. real news spreaders discrimination in Spanish.

In both languages, there is a large number of false positives which should be taken into account in further research, since a misclassification might have some consequences for the profiled user and lead to ethical or legal implications [47].

5.3 Best Results

In Table 4 we summarise the best results per language. The best result in English (0.750) has been obtained with a combination of n-grams models and stylistic features in a Logistic Regression ensemble. The best result in Spanish (0.820) has been obtained with combinations of character and word n-grams with Support Vector Machines.

Table 4: Best results per language.

English	Spanish
Buda and Bolonyai [9] (0.750)	Pizarro [45] (0.820)

6 Conclusion

In this paper, we have presented the results of the 8th International Author Profiling Shared Task at PAN 2020, hosted at CLEF 2020. The participants had to discriminate from Twitter authors whether they are keen to spread fake news or not. The provided data cover the English and Spanish languages.

The participants used different features to address the task, mainly: *(i)* n-grams; *(ii)* stylistics; *(iii)* personality and emotions; and *(iv)* embeddings. Concerning machine learning algorithms, the most used ones were Support Vector Machines and Logistic Regression, or combinations of both. Nevertheless, few participants approached the task with deep learning techniques. In such cases, they used Fully-Connected Neural Networks, CNN, LSTM and Bi-LSTM with self-attention. According to the results, traditional approaches obtained higher accuracies than deep learning ones. The six teams with the highest performance [9,45,34,63][13] used combinations of n-grams with traditional machine learning algorithms such as SVM or Logistic Regression. The first time a deep learning approach appears in the ranking is with the seventh-best performing team [24]. They used a Fully-Connected Neural Network combining word embeddings based on ConceptNet with TF-IDF word n-grams.

The best results have been obtained in Spanish (0.820) by Pizarro [45] with combinations of character and word n-grams and Support Vector Machines. The best result in English (0.750) has been obtained by Buda and Bolonyai [9] with a Logistic Regression ensemble of combinations of n-grams and some textual descriptive statistics. The overall best result (0.775) has been obtained in a tie by them.

The error analysis shows that the highest confusion in English is from Real News spreaders to Fake News Spreaders (false positives) (35.50% vs. 30.03%), whereas in Spanish is the other way around, from Fake News Spreaders to Real News Spreaders (false negatives) (35.09% vs. 20.23%). In this second case, the difference is much higher (14.86% vs. 5.47%).

Looking at the results and the error analysis, we can conclude that: *(i)* it is feasible to automatically identify potential Fake News Spreaders in Twitter with high precision, even when only textual features are used; but *(ii)* we have to bear in mind *false positives* since especially in English, they sum up to one-third of the total predictions, and misclassification might lead to ethical or legal implications [47].

7 Acknowledgments

First of all we thank the participants: 66 this year, record in terms of participants at PAN Lab since 2009! We have to thank also Martin Potthast, Matti Wiegmann, and Nikolay Kolyada to help with the 66 Virtual Machines in the TIRA platform. We thank Symanto for sponsoring the ex aequo award for the

[13] Together with De Borja and Higueras-Porras, who did not submit their working notes

two best performing systems at the author profiling shared task of this year. The work of Paolo Rosso was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). The work of Anastasia Giachanou is supported by the SNSF Early Postdoc Mobility grant under the project Early Fake News Detection on Social Media, Switzerland (P2TIP2_181441).

References

1. Agirrezabal, M.: KU-CST at the Profiling Fake News spreaders Shared Task. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
2. Anand, A., Chakraborty, T., Park, N.: We used Neural Networks to Detect Clickbaits: You won't Believe what Happened Next! In: Proceedings of the 2017 European Conference on Information Retrieval. pp. 541–547. ECIR '17 (2017)
3. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. Text & Talk **23**(3), 321–346 (2003)
4. Bakhteev, O., Ogaltsov, A., Ostroukhov, P.: Fake News Spreader Detection using Neural Tweet Aggregation. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
5. Baruah, A., Das, K., Barbhuiya, F., Dey, K.: Automatic Detection of Fake News Spreaders Using BERT. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
6. Bastos, M.T., Mercea, D.: The Brexit Botnet and User-Generated Hyperpartisan News. Social Science Computer Review **37**(1), 38–54 (2019)
7. Bayot, R., Gonçalves, T.: Multilingual author profiling using word embedding averages and svms. In: 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA). pp. 382–386. IEEE (2016)
8. Briggs-Myers, I., Myers, P.B.: Gifts differing: Understanding personality type (1995)
9. Buda, J., Bolonyai, F.: An Ensemble Model Using N-grams and Statistical Features to Identify 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)
10. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309 (2011)
11. Busari, S., Adebayo, B.: Nigeria Records Chloroquine Poisoning after Trump Endorses it for Coronavirus Treatment. CNN (2020)
12. Cardaioli, M., Cecconello, S., Conti, M., Pajola, L., Turrin, F.: Fake News Spreaders Profiling Through Behavioural Analysis. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
13. Chilet, L., Martí, P.: 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)

14. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing **9**(6), 811–824 (2012)
15. Das, K.A., Baruah, A., Barbhuiya, F.A., Dey, K.: Ensemble of ELECTRA for Profiling Fake News Spreaders. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
16. Espinosa, D., Gómez-Adorno, H., Sidorov, G.: Profiling Fake News Spreaders using Character and Words N-grams. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
17. Espinosa, M.S., Centeno, R., Rodrigo, : Analyzing User Profiles for Detection of 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)
18. Fernández, J.L., Ramírez, J.A.L.: Approaches to the Profiling Fake News Spreaders on Twitter Task in English and Spanish. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
19. Fersini, E., Armanini, J., D'Intorni, M.: Profiling Fake News Spreaders: Stylometry, Personality, Emotions and Embeddings. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
20. Ghanem, B., Ponzetto, S.P., Rosso, P.: FacTweet: Profiling Fake News Twitter Accounts. In: Statistical Language and Speech Processing (SLSP). Lecture Notes in Computer Science. Springer, Cham (2020)
21. 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)
22. Giachanou, A., Ríssola, E.A., Ghanem, B., Crestani, F., Rosso, P.: The Role of Personality and Linguistic Patterns in Discriminating Between Fake News Spreaders and Fact Checkers. In: International Conference on Applications of Natural Language to Information Systems. pp. 181–192. Springer (2020)
23. 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)
24. Giglou, H.B., Razmara, J., Rahgouy, M., Sanaei, M.: LSACoNet: A Combination of Lexical and Conceptual Features for Analysis of 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)
25. Gilbert, C., Hutto, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf. vol. 81, p. 82 (2014)
26. Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Recent trends in digital text forensics and its evaluation. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 282–302. Springer (2013)
27. Gollub, T., Stein, B., Burrows, S.: Ousting ivory tower research: Towards a web framework for providing experiments as a service. In: Proceedings of the 35th

international ACM SIGIR conference on Research and development in information retrieval. pp. 1125–1126 (2012)
28. Gollub, T., Stein, B., Burrows, S., Hoppe, D.: Tira: Configuring, executing, and disseminating information retrieval experiments. In: 2012 23rd International Workshop on Database and Expert Systems Applications. pp. 151–155. IEEE (2012)
29. Guo, C., Cao, J., Zhang, X., Shu, K., Liu, H.: Dean: Learning dual emotion for fake news detection on social media. arXiv preprint arXiv:1903.01728 (2019)
30. Hashemi, A., Zarei, M.R., Moosavi, M.R., Taheri, M.: Fake News Spreader Identification in Twitter using Ensemble Modeling Notebook for PAN at CLEF 2020. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
31. Hörtennhuemer, C., Zangerle, E.: A Multi-Aspect Classification Ensemble Approach for 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)
32. Ikade, C., Savoy, J.: UniNE at PAN-CLEF 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)
33. Johansson, F.: Supervised Classification of Twitter Accounts Based on Textual Content of Tweets. 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/
34. Koloski, B., Pollak, S., Škrlj, B.: Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
35. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and linguistic computing **17**(4), 401–412 (2002)
36. Labadie, R., Castro, D.C., Bueno, R.O.: Fusing Stylistic Features with Deeplearning methods for Profiling Fake News Spreader. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
37. Lichouri, M., Abbas, M., Benaziz, B.: Profiling Fake News Spreaders on Twitter based on TFIDF Features and Morphological Process Notebook for PAN at CLEF 2020. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
38. Maharjan, S., Shrestha, P., Solorio, T., Hasan, R.: A straightforward author profiling approach in mapreduce. In: Ibero-American Conference on Artificial Intelligence. pp. 95–107. Springer (2014)
39. Majumder, S.B., Das, D.: Detecting Fake News Spreaders on Twitter Using Universal Sentence Encoder. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
40. Manna, R., Pascucci, A., Monti, J.: Profiling Fake News Spreaders through Stylometry and Lexical Features. UniOR NLP @PAN2020. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
41. Mohammad, S.M., Turney, P.D.: Nrc emotion lexicon. National Research Council, Canada **2** (2013)

42. Niven, T., Kao, H.Y., Wang, H.Y.: Profiling Spreaders of Disinformation on Twitter: IKMLab and Softbank Submission. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
43. Pinnaparaju, N., Indurthi, V., Varma, V.: Identifying Fake News Spreaders in Social Media. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
44. 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-WS.org (Sep 2019), http://ceur-ws.org/Vol-2380/
45. Pizarro, J.: Using N-grams to detect 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)
46. Potthast, M., Köpsel, S., Stein, B., Hagen, M.: Clickbait detection. In: European Conference on Information Retrieval. pp. 810–817. Springer (2016)
47. Rangel, F., Rosso, P.: On the implications of the general data protection regulation on the organisation of evaluation tasks. Language and Law= Linguagem e Direito **5**(2), 95–117 (2019)
48. Rangel, F., Rosso, P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)
49. Rangel, F., Rosso, P., Franco-Salvador, M.: A Low Dimensionality Representation for Language Variety Identification. In: In 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing'16. Springer-Verlag, LNCS(9624). pp. 156–169 (2018)
50. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. In: Working Notes Papers of the CLEF (2018)
51. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2931–2937. EMNLP '17 (2017)
52. Ruchansky, N., Seo, S., Liu, Y.: CSI: A Hybrid Deep Model for Fake News Detection. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 797–806. CIKM '17 (2017)
53. Russo, I.: Sadness and Fear: Classification of Fake News Spreaders Content on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
54. Saeed, U., Fahim, H., Shirazi, F.: Profiling Fake News Spreaders on Twitter using Deep Learning LSTM and BI-LSTM Approach. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
55. Sandoval, L.G.M., Puertas, E., Quimbaya, A.P., Valencia, J.A.A.: Assembly of Polarity, Emotion and User Statistics for Detection of Fake Profiles. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
56. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), e73791 (2013)

57. Shashirekha, H.L., Balouchzahi, F.: ULMFiT for Twitter Fake News Profiling. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
58. Shrestha, A., Spezzano, F., Joy, A.: Detecting Fake News Spreaders in Social Networks via Linguistic and Personality Features. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
59. Shu, K., Liu, H.: Detecting fake news on social media. Synthesis Lectures on Data Mining and Knowledge Discovery **11**(3), 1–129 (2019)
60. Shu, K., Wang, S., Liu, H.: Understanding User Profiles on Social Media for Fake News Detection. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 430–435. IEEE (2018)
61. Valencia, A.V., Gomez, H.A., Rhodes, C.S., Fuentes, G.P.: Bots and Gender Identification Based on Stylometry of Tweet Minimal Structure and n-Grams Model. 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/`
62. Vo, N., Lee, K.: Learning from Fact-checkers: Analysis and Generation of Fact-checking Language. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 335–344 (2019)
63. Vogel, I., Meghana, M.: Fake News Spreader Detection on Twitter using Character N-Grams. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
64. Vosoughi, S., Roy, D., Aral, S.: The Spread of True and False News Online. Science **359**(6380), 1146–1151 (2018)
65. 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. ACL '17 (2017)
66. Wu, S.H., Chien, S.L.: A BERT based Two-stage Fake News Spreader Profiling System. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)