UniNE at PAN-CLEF 2020

Profiling Fake News Spreaders on Twitter

Notebook for PAN at CLEF 2020

Catherine Ikae, Jacques Savoy

Computer Science Department, University of Neuchatel, Switzerland
{Catherine.Ikae, Jacques.Savoy}@unine.ch

Abstract. In our participation of the "Profiling Fake News Spreaders on Twitter" task (both in English and Spanish), our main objective is to be able to detect Twitter user accounts used to spread disinformation, fake news, as well as conspiracy theories. To automatically solve these questions based only on the tweets' contents, we suggest to reduce the number of features (isolated words) to a few hundred. This suggested approach is based on a two-stage method ignoring infrequent terms and ranking the others according to their occurrence differences between the two categories. Finally, a classifier is implemented combining decision tree, random forest, and boosting. Our first evaluation experiments indicate an overall accuracy around 70%.

1 Introduction[1]

Since 2013, CLEF-PAN has been generating test collections on author profiling with datasets extracted from social networks (e.g., blogs, tweets) [1]. During the last years, UniNE has participated in these text categorization tasks to identify some of the author's demographics (e.g., gender, age range, psychological traits, geographical origin) or to know if a set of tweets was created by bots or humans.

For this year, the participants need to implement a system identifying whether or not a set of 100 tweets was sent by a user spreading fake news (or junk news, pseudo-news, hoaxes, or in general disinformation). More precisely, the target task could be rephrased as knowing whether a set of tweets contains fake news (or misleading content). In fact, the available information is just the tweet contents, and the tweet context (e.g., number of likes, retweets, etc.) with the author source details (e.g., information about the Twitter account) not provided. Moreover, the multimedia elements are not included.

[1] Copyright (c) 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-15 September 2020, Thessaloniki, Greece.

The first step to solve this question is to define precisely what we mean by fake news. This is not an easy task, mainly because different variants of fake news can be encountered [2]. For example, satire or parody with its irony and sarcasm could represent the less harmful form of fake news (e.g., *Ig* (ignoble) *Nobel Prize*). For others, humor cannot be viewed as fake news because it is evident that the underlying information is not true. At a higher level, one can see a sentence extracted from its context (or embraced in the wrong context) while in its most sophisticated form the news is entirely fabricated (with additional multimedia elements).

Usually fake news [3] is rendered as normal customer reviews, political or financial news as well as advertising but with the objective to favor or undermine the image of a product or the reputation of a candidate. Their presence could be limited to a few seconds (e.g., flash ads), to the period of an electoral or advertising campaign, or can even stay visible longer (to support a conspiracy theory, even an extreme one such as "Hitler is alive on a Nazi moon base" [4]).

The identification of fake news is still a complex problem. One can take account of four main sources of evidence, namely a) the news content, b) the news creator or spreader, c) the post or social context, d) the target audience (e.g., users or news platforms).

The content of fake information tends to present more emotional words, usually to evoke anger and fear in the readers [5]. They employ more negative forms (e.g., not, never), usually with more uppercase letters or words, more spelling errors, more specific punctuation symbols (e.g., !, ?, as well as !!!), hashtags, mentions or hyperlinks. According to Pennebaker's studies [6], lying writers tend to use less *Self* words (I, me, my, mine), but more nouns, and some discrepancy verbs (would, should, could, ought). When telling the truth, the sentences are longer and more complex, containing more numbers, more details and more longer words.

The author's name and the URL of the source could also be pertinent during the identification. The user credibility could be estimated by his geolocation, the fact that the account was verified or not, or by the presence of weird tokens in the URL (as well as uncommon domains). Usually, creators of fake news (humans, bots or cyborgs) will send many posts during a short time interval. They also tend to have more friends and followers, and reply more often.

The social or post context can also provide some information indicating a fake news spreader such as a larger number of likes, a high intensity of retweets and more shares and comments than one would expect from a normal user. When monitoring the temporal activity, some patterns can identify a bot or cyborg activity.

The spreading of fake news presents several advantages for the sender. When receiving the same fake news many times (echo chamber effect) and particularly when received by friends, the misinformation is finally accepted as true. For example, analyzing Trump's tweets, the probability to see the word *fake* (of *faker*) just before (or after) CNN is high (more precisely, one can count 266 occurrences of CNN in which 88 times the term *fake news* appears in the short context). The same observation is valid for the *New York Times* or the *Washington Post*. After repeating this misinformation, only 9% of Republicans consider the *New York Times* as trustworthy [7]. Therefore, it is not surprising to observe that Conservatives tend to share fake news more often than Democrats and older persons [8]. And this trend continues to

undermine US politics. Now accepting (or spreading) conspiracy theories could be considered mandatory for a Republican candidate to win a primary election for the Congress or the Senate [9].

The rest of this paper is organized as follows. Section 2 describes the text datasets while Section 3 describes our feature selection procedure. Section 4 exposes our classifier and shows some of our evaluation results. A conclusion draws the main findings of our experiments.

2 Corpus

When limited to spreading action, one can focus only on the tweet contents. In our point of view, the problem is therefore to identify a set of tweets containing disinformation, leading to consider that the user generates and/or spreads fake news [10],[11]. This task will be performed both in English and Spanish.

When faced with a new dataset, a first analysis extracts an overall picture of the data, the relationship, and detects and explores some simple patterns related to the different categories. In the current study, two categories are provided (Category = 0 or 1), without further information about the precise meaning of the two values. When observing some examples of tweets reported in Tables 1, one can assume that Category = 1 means *fake news*.

England ease to World Cup win over France #HASHTAG# #HASHTAG# #HASHTAG# #URL# #URL#
Spain rescues 276 migrants crossing perilous Mediterranean #HASHTAG# #HASHTAG# #HASHTAG# #HASHTAG# #HASHTAG#…
Italy's Uffizi demands return of Nazi-looted painting #URL#
Trump invites congressional leaders to border security briefing #URL#

Table 1a: Examples of four tweets in the Category #0

#USER# Merkel is using her IMMIGRATION INVASION as a demographic weapon to destroy Germany. #HASHTAG# #HASHTAG# #HASHTAG# #HASHTAG# #HASHTAG# #USER# #USER# Trump is 1/2 Scottish and 1/2 German. Trump will smash the shyster rats. #HASHTAG# #HASHTAG# #HASHTAG#
With Obama's Approval, Russia Selling 130 Tons of Uranium to Iran #URL#
FBI admits illegal wiretapping of President Trump, issues apology #URL#

Table 1b: Examples of four tweets in the Category #1

As one can see in Tables 1, tweets in Category #0 describe facts without expressing many emotions. In tweets appearing under the second label, the terms belong to swear expressions (e.g., shyster rats) or tend to cause fear (or anger) (e.g., invasion, destroy).

The available tweets are included in a training corpus available in English and Spanish. As depicted in Table 2, the training data contains the same number of documents in the two categories and in the two languages.

As each document corresponds to 100 tweets, the mean number of tokens (composed only by letters) per document is around 1,260 for the English language. For the Spanish language, the mean length is around 1,508, with a significant difference between the two categories (Category #0: 1,655; Category #1: 1,361).

	English			Spanish		
	Cat. #0	Cat. #1	Test	Cat. #0	Cat. #1	Test
Nb. doc.	150	150	100	150	150	100
Nb tweets	15,000	15,000	10,000	15,000	15,000	10,000
Mean length	1,252	1,276		1,655	1,361	
|Voc|	20,509	19,851		29,137	24,825	
Hashtag	54,366	46,722		48,216	40,065	
URL	16,577	17,119		10,887	13,900	

Table 2: Overall statistics about the training data in both languages

As shown in Table 2, one can observe that the number of hashtags is larger in Category #0 than in the second one with a difference between them close to 20%. In addition, the number of URLs (hyperlinks) is higher in Category #1 than in Category #0. For the English language, the difference is small, but clearly larger for the Spanish corpus.

As text categorization problems are known for having a large and sparse feature set [12], Table 2 also indicates the number of distinct terms per category (or the vocabulary size denoted under the label |Voc|) which is 20,509 for the English Category #0. Fusing the two categories, the English corpus counts 29,521 distinct words (or 40,867 word-types for the Spanish collection).

For both languages, the vocabulary size is larger for Category #0 than for Category #1 (English: 20,509 vs. 19,851; Spanish: 29,137 vs. 24,825). The texts sent when spreading fake news are composed with a smaller *lexis* implying that the same or similar expressions are often repeated.

3 Feature Selection

To achieve a good understanding of the distinction between normal tweets and tweets containing fake news, a feature selection function must be applied. As a simple strategy, the term frequency (*tf*) or the document frequency (*df*) have been suggested, under the assumption that a higher frequency could indicate a more useful feature. Both functions return similar results and have been shown to be effective approaches for solving the authorship attribution problem [13]. For example, the Delta method [14] is based on the 50 to 200 most frequent words to determine the true author of a text. Identifying disinformation (lies) and authorship identification are however not the same question.

Moreover, when considering the most frequent words employed, very similar sets of terms appear in both categories (e.g., “URL”, “HASHTAG”, “the”, “to”, and some punctuations symbols (' , : ...)). Thus, a simple feature selection based on the *tf* information is not fully effective and the distinction between features associated with each category is not guaranteed.

However, it is always useful to ignore features having a low occurrence frequency. According to the Zipf's law, a large number of word-types appear just once or twice. According to statistics reported in Table 3, when removing words appearing only once, the vocabulary size of the English corpus (Category #0) decreases from 20,509 to 10,474 (a reduction of 48.9%). For the Spanish language (Category #0), the reduction is larger, from 29,137 to 12,882 (a decrease of 55.8%).

On the other hand, one can encounter terms having a relatively high occurrence frequency but appearing only in a few documents (one document = a set of 100 tweets). Thus, we also suggest to remove terms having a low document frequency (*df*), for example, with a $df > 3$. The effect on the vocabulary size is shown in the next to last row of Table 3. For example, for the English corpus (Category #0), the vocabulary decreases from 20,509 to 4,636, showing a reduction of 77.4%. Similar decreases can be observed for the other sub-collections. In this study, we have considered both frequency counts by ignoring terms having a $tf < 6$ and a $df < 4$ as indicated in the last row of Table 3.

	English		Spanish	
	Cat. #0	Cat. #1	Cat. #0	Cat. #1
|Voc|	20,509	19,851	29,137	24,825
$tf > 1$	10,474	10,672	12,882	11,514
$tf > 3$	5,797	6,184	6,573	6,001
$tf > 5$	4,136	4,431	4,463	4,150
$df > 3$	4,636	5,247	5,838	5,386
$tf > 5$ and $df > 3$	2,433	3,720	3,800	3,590

Table 3: Vocabulary size with different feature selection strategies

After removing infrequent terms, we propose a feature selection method that works in two stages. In the first, the term frequency (*tf*) information is taken into account. For each term, the discriminative power is computed by estimating the occurrence probability difference in both categories as indicated in Equation 1. In this case, tf_{i0} indicates the absolute frequency of the *i*th term in class c_0 (or Category #0), and n_0 the text length (in tokens) of all tweets belonging in class c_0 (and similarly with class c_1).

$$probD(t_i, c_0) = prob(t_i, c_0) - prob(t_i, c_1) = \frac{tf_{i0}}{n_0} - \frac{tf_{i1}}{n_1} \quad (1)$$

To determine terms able to describe the Category #0, only terms having a positive probD value are extracted. Of course, one can impose a stricter constraint by selecting terms having a probD larger than a threshold. Similarly, only words with a negative probD value are chosen to represent Category #1. This step generates two term clusters, one per class.

After this procedure, one can identify some terms more strongly associated to each category. For example, Table 4 reports the top fifteen words having the largest value for each language and category (the negative scores for Category #1 have been multiplied by -1). For both languages, tweets containing true information have more hashtags, retweets (rt) and user mentions. Tweets spreading fake news have more URL, "video" meaning that they tend to refer more often to other websites containing supporting information (in the form of text, video, etc.).

For the English language, the names of political leaders (“trump”, “obama”, “clinton”), or the adjective “new” are more recurrent in tweets spreading disinformation. This is also an indication that political news is more frequently spread than other domains. It is interesting to see the verb “says” as a feature indicating fake news (e.g., reporting a sentence spoken by a well-known person).

Some punctuation symbols (, ... ! : ? or ¿) appear more recurrently in normal tweets than in fake news (e.g., in the sequence RT #USER#:). The comma is more associated with longer sentences, usually indicating a real story [6] as well as the pronoun I.

English				Spanish			
Category #0		Category #1		Category #0		Category #1	
0.0144	USER	0.0054	'	0.0140	USER	0.0232	URL
0.0080	HASHTAG	0.0041	URL	0.0112	HASHTAG	0.0069	unete
0.0057	:	0.0037	trump	0.0097	.	0.0066	–
0.0051	rt	0.0024	-	0.0063	rt	0.0041	video
0.0043	.	0.0021	s	0.0061	:	0.0028	(
0.0026	the	0.0016	after	0.0059	,	0.0028	-
0.0025	,	0.0010	her	0.0036	...	0.0023	)
0.0017	...	0.0009	to	0.0034	que	0.0013	"
0.0016	i	0.0009	video	0.0027	no	0.0013	su
0.0016	this	0.0008	new	0.0015	es	0.0010	el
0.0016	!	0.0008	donald	0.0015	?	0.0010	vida
0.0013	and	0.0008	post	0.0014	¿	0.0009	fuerte
0.0012	your	0.0007	obama	0.0014	las	0.0009	para
0.0012	a	0.0007	says	0.0011	ha	0.0008	a
0.0011	read	0.0007	clinton	0.0009	qué	0.0008	tu

Table 4: The top fifteen terms having the largest probD values

After this step, one can stop the feature selection by considering the *k* terms (with *k* = 100 to 250) having the highest and smallest probD scores. To go further in this space reduction, the second step applies an additional feature selection procedure. In this perspective, previous studies have shown that the chi-square, odds ratio, or mutual information tend to produce effective reduced term sets for different text categorization tasks [12], [13]. In this study, the chi-square method was selected to reduce the feature space to a few hundred terms.

4 Evaluation

To define the different machine learning models, the `scikit-learn` library (Python) was applied [15]. The default setting defined in the library was chosen. The decision tree approach was applied to define our first model (with Gini function to measure the node impurity) [16]. As more complex classifiers, the random forest with 200 trees was applied and the final decision was acquired by majority voting. As another approach belonging to the ensemble learning, the bagging model forms one of our selected approaches. With boosting, represented by our last model, a set of weak learners is combined to produce a more effective assignment. More precisely, we chose

the extreme gradient boosting (XGB) [17], [18] based on a set of 100 decision trees (maximum depth was set to 2).

When computing the decision for a new set of tweets, the three classifiers determine a proposed attribution based on the same set of chosen features. To combine the three resulting decisions, a simple majority vote could be applied, giving the same importance to each of the three individual classifiers. This solution corresponds to a democratic vote.

However, each classifier returns not only the proposed decision but an estimated probability that the input set of tweets belongs to that category. Thus, our second approach, called *soft vote*, adds these three probabilities to determine the final assignment (this merging strategy was used in our early bird submission).

To compute the accuracy rates shown in Tables 5, only the training subset is used to select the feature sets and to generate the document surrogates (the same number of documents appears in both categories and languages). To achieve a fair evaluation, we randomly extracted 50 documents from each category to generate the test set.

From the results depicted in Tables 5, one can see that after reducing the feature set to a few hundred words one can still achieve a good overall effectiveness. Moreover, having more features does not imply obtaining a higher effectiveness level. For example, using in total only 150 features, our model achieves an accuracy rate of 0.81 (English corpus, majority vote) or 0.78 for the Spanish language (majority vote). Doubling the number of features does not always improve the overall effectiveness (English corpus: 0.81 vs. 0.75, Spanish: 0.788 vs. 0.79). It is interesting to know that even if, in mean, combining different classifiers provides a higher effectiveness, the best solutions for the Spanish corpus are often a single boosting model.

	Chi = 100	Chi = 150	Chi = 200	Chi = 250	Chi = 300
Random forest	**0.77**	0.76	0.77	0.71	**0.79**
Boosting	0.71	0.74	0.72	0.66	0.70
Decision Tree	0.68	0.72	0.63	0.64	0.63
Soft Vote	0.70	0.72	**0.78**	0.66	0.71
Majority Vote	0.72	**0.81**	**0.78**	**0.72**	0.75

Table 5a: Evaluation based on different feature sizes (English corpus)

	Chi = 100	Chi = 150	Chi = 200	Chi = 250	Chi = 300
Random forest	0.73	0.77	0.77	0.77	0.76
Boosting	0.77	**0.79**	0.78	**0.80**	**0.79**
Decision Tree	0.75	0.75	0.69	0.69	0.69
Soft Vote	**0.79**	0.73	0.74	0.76	0.71
Majority Vote	0.75	0.78	**0.79**	0.79	**0.79**

Table 5a: Evaluation based on different feature sizes (Spanish corpus)

Table 6 reports our official results achieved with the TIRA system [19] and using the official test subset of the data. Our first results called early bird results have been obtained under the soft vote scheme. They appear in the second row in Table 6. Our official performance was achieved with the majority scheme depicted in the last row in Table 6.

	TIRA test set	
Fusion	English	Spanish
Soft vote	0.675	0.700
Majority vote	**0.725**	**0.725**

Table 6: Official Evaluation of under different voting strategies (English and Spanish)

In both cases, the infrequent terms have been ignored ($tf > 5$ and $df > 3$). Then the top 150 terms having the highest chi-square values have been selected to define the feature set.

6 Conclusion

In our participation to the "Profiling Fake News Spreaders on Twitter" (CLEF PAN 2020) we have worked with tweets written in English and Spanish. Overall, we achieve the following main findings. First, we suggested a feature selection approach able to extract a reduced set of features (precisely 150). Based on such a reduced set, it is possible to identify those features more associated to normal tweets (e.g., *I*, *this*, *film*, *review*, *episode*, etc.). In addition, the conjunction *and* and the comma appears more often in normal posts, indicating the presence of longer sentences. In tweets spreading fake news, one can count more names of political leaders, as well as the terms *says*, *post*, *president*, *she*, *he*, *democrat*, etc. This is an indication of the presence of posts reporting opinions and words uttered by other persons.

Second, our analysis indicates that tweets containing fake news tend to include more references (URL) (see Table 2) to other webpages than normal tweets, references used to support the misinformation or to justify some conspiracy theory. On the other hand, normal tweets present more retweets and hashtags as shown in Table 2.

Third, our attribution approach is based on a model combining three individual attributions computed by a decision tree, a boosting, and a random forest classifier. It was a surprise to see that a simple majority scheme achieved a higher accuracy rate than a merging approach based on the probability estimates computed by each individual classifier.

References

[1] Potthast, M., Rosso, P., Stamatatos, E., Stein, B. (2019a). A Decade of Shared Tasks in Digital Text Forensics at PAN. *Proceedings ECIR 2019*, Springer LNCS # 11437, 291–303.

[2] Wardle, C. (2017). Fake News. It's Complicated. *First Draft.* February, 16. (URL: https://firstdraftnews.org/latest/fake-news-complicated/).

[3] Zhang, X., & Ghorbani, A.A. (2020). An Overview of Online Fake News: Characterization, Detection, and Discussion. *Information Processing & Management*, 57(2), 102025.

[4] Selk, A. (2018). No, Hitler isn't Alive on a Nazi Moon Base. *Washington Post,* May, 20.

[5] Hart, R.P. (2020). *Trump and Us. What he Says and Why People Listen*. Cambridge University Press.

[6] Pennebaker, J.W. (2011). *The Secret Life of Pronouns.* Bloomsbury Press, New York.

[7] Francia, P.L. (2017). Going Public in the Age of Twitter and Mistrust of the Media. In J. C. Baumgartnet, T.L. Towned (eds), *The Internet and the 2016 Presidential Campaign*, Lexington Books, Lanham.

[8] Guess, A., Nagler, J., & Tucker, J. (2020). Less than you Think: Prevalence and Predictors of Fake News Dissemination on Facebook. *Sciences Avdances,* 5, eaau4586.

[9] Philips, A. (2020). Why QAnon Supporters are Winning Congressional Primaries. *Washington Post,* June, 13th.

[10] Rangel, F., Giachanou, A. Ghanem, B., & Rosso, P. (2020). Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. *CLEF 2020 Labs and Workshops, Notebook Papers*.

[11] Ghanem, B., Rosso, P., & Rangel, F. (2020). An Emotional Analysis of False Information in Social Media and News Articles. *ACM Transactions on Internet Technology*, 20(2), 1-18.

[12] Sebastiani, F. (2002). Machine Learning in Automatic Text Categorization. *ACM Computing Survey*, 34(1), 1–27.

[13] Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text classification. *Journal of Machine Learning*, 3, 1289-1305.

[13] Savoy, J. (2015). Comparative Evaluation of Term Selection Functions for Authorship Attribution. *Digital Scholarship in the Humanities*, 30(2), 246–261.

[14] Burrows, J.F. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary & Linguistic Computing*, 17(3), 267–287.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Venderplas, J., Passos, A., Courapeau, D., Brucher, M., Perrot, M., & Duchernay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830 (2011).

[16] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.

[17] Shapire, R.E., & Freund, Y. (2012). *Boosting. Foundations and Algorithms*. MIT Press, Cambridge.

[18] Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Cambridge.

[19] Potthast, M., Gollub, T., Wiegmann, M., Stein, B. (2019b) TIRA Integrated Research Architecture. In N. Ferro, C. Peters (eds), *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*. Springer, Berlin.