Hong Kong Protests: Using Natural Language Processing for Fake News Detection on Twitter

Alexandros Zervopoulos[1(✉)], Aikaterini Georgia Alvanou[1], Konstantinos Bezas[1], Asterios Papamichail[1], Manolis Maragoudakis[2], and Katia Kermanidis[1]

[1] Department of Informatics, Ionian University, Corfu, Greece
{c19zerv,c19alva,c19beza,c19papa,kerman}@ionio.gr
[2] Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece
mmarag@aegean.gr

Abstract. The automation of fake news detection is the focus of a great deal of scientific research. With the rise of social media over the years, there has been a strong preference for users to be informed using their social media account, leading to a proliferation of fake news through them. This paper evaluates the veracity of politically-oriented news and in particular the tweets about the recent event of Hong Kong protests, with the aid of a dataset recently published by Twitter. From this dataset, Chinese tweets are translated into English, which are kept along with originally English tweets. By utilizing a language-independent filtering process, relevant tweets are identified. To complete the dataset, tweets originating from valid sources are used as the real portion, with journalists rather than news agencies being considered, which constitutes a novel aspect of the methodology. Well-known Machine Learning algorithms are used to classify tweets, which are represented by a feature value vector that is extracted, selected and preprocessed from the datasets and mainly revolves around language use, with word entropy being a novel feature. The results derived from these algorithms highlight morphological, lexical and vocabulary differences between tweets spreading fake and real news, which are for the most part in accordance with past related work.

Keywords: Fake news detection · Natural Language Processing · Machine Learning · Twitter · Hong Kong protests

1 Introduction

Social media, popular or not, allow both borderless communication and a plethora of information to be spread at a dizzying speed around the world,

Published by Springer Nature Switzerland AG 2020
I. Maglogiannis et al. (Eds.): AIAI 2020, IFIP AICT 584, pp. 408–419, 2020.
https://doi.org/10.1007/978-3-030-49186-4_34

justifying the choice of the largest percentage of them to keep up to date with domestic and global news events, via Facebook, Twitter and so on. However, the validity of the news is not guaranteed, as it may be hampered by conspiracy, political expediencies and interests. By extension, the spread of fake news contributes to a common and everyday phenomenon, which can undermine values and ideals and, thus, needs addressing.

The spread of fake news is particularly prevalent in politically oriented content, especially so on Twitter, where it has been found that the rate of dissemination of fake news is higher than that of real news [12,20]. In this context, computer science can also be used as the primary asset and tool for false detection in news releases from Twitter user accounts, helping to counteract and eliminate this phenomenon. Since the process of falsehood detection by conventional methods, such as the involvement of certified journalists, is a costly process, due to the financial costs and the lengthy periods of time required to complete it, technological approaches, starring Artificial Intelligence, have gained ground [11], instead. The recent (June 2019) events of the Hong Kong protests related to political controversy have been of great concern to the public because of the violent turn and the high turnout of citizens inside and outside China's borders [14]. As a result, a plethora of tweets were triggered, raising the question of the validity of their content. It is therefore important to study the extent of the fake news spread on Twitter about this event.

In this paper, the problem of automatically distinguishing between tweets spreading fake and real news is tackled through the use of Machine Learning (ML) algorithms. In order to accomplish this, an initial dataset published by Twitter[1] regarding the Hong Kong protests is used to represent the fake portion of the data used for classification. This dataset contains tweets in a multitude of languages, though only English and Chinese tweets are utilized, with the aid of machine translation. Relevant tweets are pinpointed through a filtering process that is language-independent, making selective use of machine translation. A collection of tweets is gathered to represent the real portion of the dataset, which are considered trustworthy based on the account posting the tweet. News agency and journalist accounts are considered trustworthy sources for the purposes of this study. The assembled dataset is publicly available for research purposes in Humanistic and Social Informatics Laboratory's website[2]. From the assembled dataset, a plethora of linguistic features are extracted, preprocessed and selected to be used as inputs for a variety of well-established ML algorithms, namely Naive Bayes, Support Vector Machines (SVMs), C4.5 and Random Forest. Twitter text has idiosyncrasies that render its linguistic processing quite interesting and that have been tackled in various contexts, the TraMOOC system being one of them [18]. The derived models indicate significant differences in morphological, lexical and vocabulary features between tweets spreading fake and real news. In contrast to previous studies, journalists are investigated

[1] https://transparency.twitter.com/en/information-operations.html.
[2] https://hilab.di.ionio.gr/index.php/en/datasets/.

regarding trustworthiness, rather than just news agencies, and word entropy is used as a novel feature, which plays an important role in classification.

The rest of this paper is organized as follows. An overview of related literature is presented in Sect. 2, while the applied methodology is described in Sect. 3. Section 4 specifies the produced results and finally, conclusions are drawn in Sect. 5.

2 Related Work

The detection of fake news, and, in particular, those that are spread through social media, has been extensively researched by the scientific community. Specifically, in [11,17], the technical challenges in automating fake news detection using Natural Language Processing (NLP) tactics are presented, while a comparison between the used datasets, features, models and their respective performances is provided, with the aim of facilitating future studies.

On the other hand, Ahmed et al. [1] approach the issue with the help of text analysis with N-gram attributes (up to 4-gram size) and by using 6 different machine learning techniques for classification. The model that reached a standing out performance is the linear SVM with the use of unigram attributes. Conroy et al. [5] focused on the detection of fake news, with the aim of presenting a hybrid approach based on a combination of linguistic and network-analysis techniques. For both categories, machine learning tools that ultimately contribute to successful detection are described.

In addition, Buntain and Golbeck [3] are occupied with automating the detection of fake news on Twitter via 2 existing datasets to analyze the structure and behavior of potentially fake Twitter threads, assessing their proximity to the thread with the help of the Buzzfeed dataset[3]. The aim is to determine the appropriate characteristics of training capable models for predicting falsehood.

While multiple well-established datasets exist, experimentation focusing on specific events regularly takes place. This poses an array of challenges, primarily due to the fact that expertly-annotated data is hard to come by. As such, attempts have been made to circumvent the need for experts' opinions by utilizing data-driven techniques. One such example is the work by Helmstetter [7], who consider the credibility of a tweet's source as a proxy for the trustworthiness of the tweet itself, achieving high prediction scores.

In [19], authors deliberate the classification of fake and verified news and the promotion of 4 categories of fake news: propaganda, satire, hoaxes and clickbait. The analysis and experimentation is based on Twitter and, in fact, the data collection takes place over a period of 2 weeks, during the terrorist attacks in Brussels in 2016. For classification, linguistically-infused neural network models are created, based on the content of tweets and social network interactions. Finally, they conclude that morphological and grammatical features are not efficient.

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

3 Methodology

3.1 Fake News Dataset

The initial fake news dataset is retrieved from Twitter's Election Integrity Hub[4], where three sets were disclosed in August and September 2019. In greater detail, this dataset consists of 13,856,454 tweets in total and includes 31 fields, which represent tweet-related features about both the tweet's text and the user. In the present study, Twitter is regarded as a reliable source and they have deemed these accounts to be "deliberately and specifically attempting to sow political discord in Hong Kong,"[5]; thus, this dataset is considered as ground truth with respect to the fake news portion of the assembled dataset.

However, as per Twitter's description of the dataset, the accounts involved tend to be fake, post spam and act in a coordinated manner, which has also been investigated in the literature [17]. Hence, not all tweets are relevant to the spread of fake news related to the Hong Kong protests, deeming mandatory an initial preprocessing step. Furthermore, due to the specificities of the events, which take place in China, it is assumed that most of the relevant tweets would be worded either in Chinese or English, as the latter is more prevalent in the Twitter platform.

To better visualize and understand the contents of the dataset, the tweets' text is preprocessed, from which word clouds are constructed. Namely, hashtags, mentions and URLs are removed from the texts in English and in Chinese. Those in Chinese are also translated into English, using Google's Translation API available through the Google Cloud[6] platform. Afterwards, a frequency word cloud is created for each language, with the aid of Python's word cloud[7] module, containing at most 400 words. It should be noted that these word clouds are derived only from the first two out of the three of Twitter's datasets, as described at the beginning of this subsection. The resulting word clouds are depicted in Fig. 1a, Fig. 1b. Chinese tweets are evidently more relevant to the events than their English counterparts.

To precisely identify the tweets spreading false information regarding Hong Kong protests, the ensuing filtering methodology is followed, which is largely based on the assumption that a tweet's hashtags also indicate the content of a tweet's text. It is worth pointing out that this process is language-independent, which is particularly advantageous in this case, as it is impractical to translate millions of Chinese tweets into English. Moreover, the presented filtering process is overall fairly efficient, requiring a short amount of time, typically a few minutes using commodity hardware.

The methodology can be broken down in the subsequent manner. First of all, a list of hashtags related to Hong Kong protests is manually constructed, comprising both English and Chinese hashtags. Afterwards, hashtags appearing

[4] https://transparency.twitter.com/en/information-operations.html.
[5] https://tinyurl.com/y3ffrblt.
[6] https://cloud.google.com/translate/.
[7] https://github.com/amueller/word_cloud.

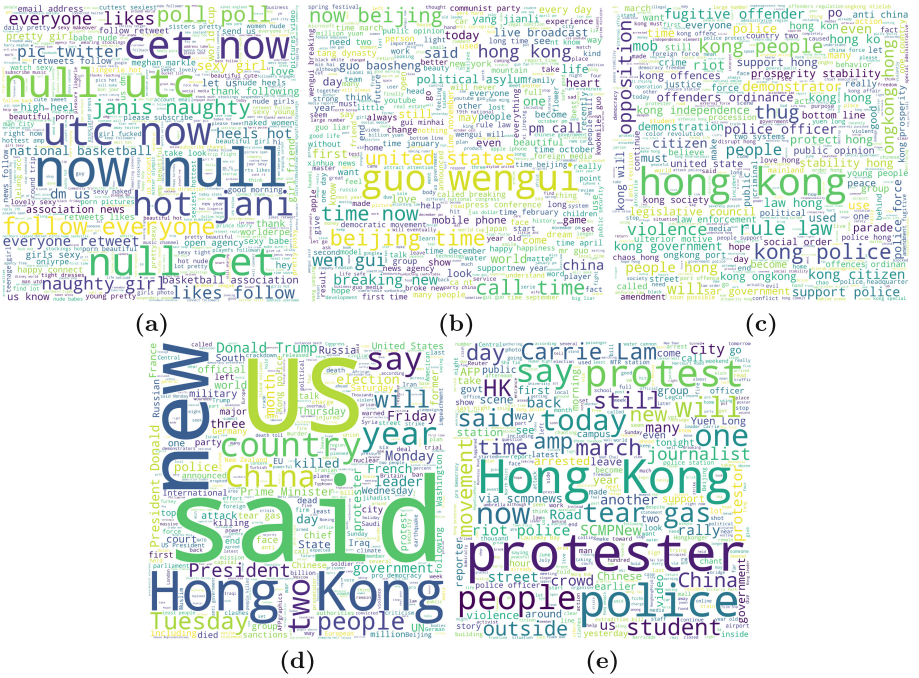

Fig. 1. Frequency word clouds formed from tweets: (a) In the fake news dataset worded in English. (b) In the fake news dataset worded in Chinese. (c) Resulting from the filtering process. (d) In the news agency dataset. (e) In the journalist dataset.

in tweets along with at least one of the previously mentioned hashtags are kept track of, and their co-occurrence counts are calculated across the entire dataset. Finally, tweets are deemed relevant if they contain a hashtag with a co-occurrence count higher than an arbitrary threshold (in this case, 100). It is evident that, due to this approach, tweets without hashtags cannot be considered relevant.

Upon the completion of the filtering steps, only 3,908 tweets, worded in English or Chinese, are considered to be relevant to the spread of fake news related to Hong Kong protests. The word cloud constructed by the collection of these relevant tweets is depicted in Fig. 1c.

3.2 Real News Dataset

As previously mentioned, the dataset corresponding to real news has not been retrieved in the same manner as the fake news, i.e. no real news dataset has been made publicly available by Twitter. Additionally, well-established datasets are not appropriate for this study, due to the fact that they do not specifically deal with events surrounding the Hong Kong protests. To address these challenges, the real news dataset has to be constructed from the ground up here. In more detail, since expert labeling is a costly process, news posts about Hong Kong

protests appearing in generally well-regarded news agencies are often assumed to be reliable sources[8]. Therefore, tweets from the accounts of news agencies are retrieved, rather than articles.

Moreover, one would expect news agencies' tweets to differ in style from those of personal accounts, e.g. more formal speech, fewer replies to other users, more references to the agency's articles, etc. As such, in this study, tweets from journalists of well-regarded news agencies are also considered as real news. The search and retrieval of relevant twitter accounts of news agencies and journalists employed by them takes place manually and, in this case, the following news agencies are considered: BBC News, Reuters, Bloomberg, BuzzFeed, Channel News Asia, CNN, Agence France-Presse, South China Morning Post, Wall Street Journal, The New York Times, The Associated Press, The Washington Post and Quartz. Having completed the search and retrieval of the aforementioned twitter accounts, 13 accounts of news agencies and 107 accounts of journalists are gathered. Using Twitter's user timeline API, 41,996 tweets from news agencies' accounts and 103,359 tweets from journalists' accounts are collected.

The retrieved data seem to be supporting the assumption that the tweets contained in the fake news dataset are more similar to those of journalists than those of news agencies. A few notable statistics derived from the unfiltered data and the first two fake news datasets are listed to showcase some differences and the aforementioned notion of 'similarity': On average, a tweet contains approximately 0.22 hashtags in the fake news dataset, 0.23 hashtags when posted by a journalist and 0.1 hashtags when posted by a news agency. Additionally, the mean number of urls in a tweet is 0.3 in the fake news dataset, 0.35 when posted by a journalist, and 0.82 when posted by a news agency. Lastly, on average, each of the accounts posting tweets have 4.1 followers in the fake news dataset, 15.14 in the journalist dataset, and 11,509.08 in the news agency dataset.

Frequency word clouds are also derived from relevant tweets found in these two datasets and are shown in Fig. 1d, Fig. 1e. While both are evidently relevant to Hong Kong events, the more objective, news-based narrative of the news agency dataset differs from the journalist and fake news dataset. Thus, it becomes clear that the tweets collected from journalist accounts are more similar to those in the fake news dataset, when considering both tweet content and account characteristics.

Using the filtering process described previously, 5,388 and 666 of the tweets posted by journalists and news agencies, respectively, are considered relevant, with the latter being fewer due to the fact that the filtering process relies purely on hashtags, which news agencies don't use as often, as was already showcased. Due to both the low number of tweets and dissimilarity to the fake news dataset, the news agency dataset is entirely dropped and not further studied. All in all, the assembled dataset consists of 3,908 and 5,388 tweets spreading fake and real news, respectively.

[8] https://www.4imn.com/news-agencies/.

3.3 Features

The selected features are purely linguistic in nature and they represent a single tweet. While the literature indicates that network-related features are worth investigating, most of the information about the fake news dataset has been made unavailable by Twitter and is no longer accessible on the platform, as the accounts involved in the disclosed datasets have been banned. Regarding the derivation of features, various preprocessing stages are necessary in some cases, which are mentioned when appropriate. The features add up to 38 in total, including the class label, and their Pearson correlation heatmap is depicted in Fig. 2. Even at this early stage, the features of tweet entropy, tweet length and type to token ratio (TTR) are highly correlated with the class label, so they are likely to be important in classification.

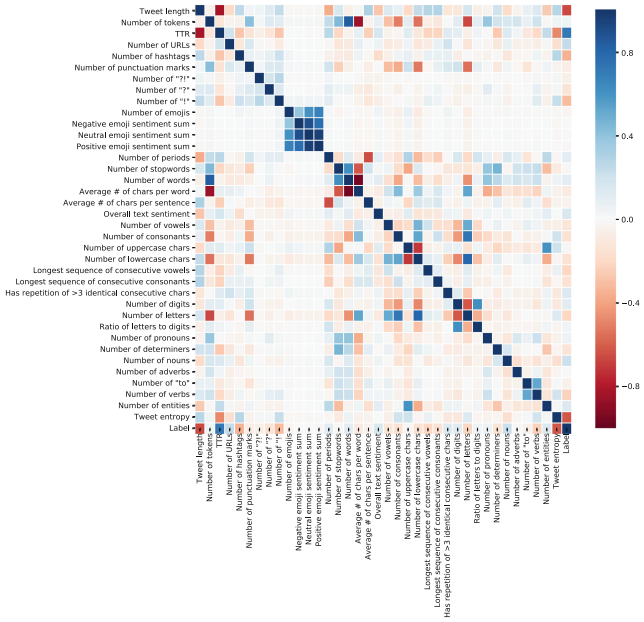

Fig. 2. Pearson correlation heatmap of features.

The TTR is calculated after the text has been tokenized as an indication of the tweet's richness in vocabulary. A plethora of features are used in the morphological level. These include a large assortment of counts regarding the tweet's: length in characters, tokens, included URLs, hashtags, certain punctuation marks ('?', '!', '?!') and total punctuation marks, emojis, periods, stopwords, words, vowels, consonants, upper and lower case characters, digits and letters. Since these counts are expected to be correlated with the tweet's length, they are normalized by dividing them with the latter. Furthermore, the average number

of characters per word and average number of words per period are calculated, along with the length of the longest sequence of vowels and consonants in a word. Finally, if a character's consecutive appearance takes place more than thrice, it is marked in the form of a boolean value. Last but not least, the entropy of a tweet is calculated through the equation: $S = -\sum_i P_i \log P_i$, where P_i is the probability of word i, which has been stemmed and converted to lowercase, appearing in the dataset. Entropy has been used in this case as an indicator of word importance for a tweet, but other similar metrics, such as word weighted frequency could be considered instead.

A number of Part of Speech (PoS) features are included, which keep track of the corresponding occurrences in the tweet: verbs, entities, pronouns, determiners, adverbs, as well as the proposition "to". These features are calculated with the aid of the NLTK module's [8] recommended PoS tagger and, much like the morphological features representing counts, they are normalized by the tweet's length.

A wide array of semantic features are used to provide higher-level information about the tweet. These include: positive, neutral and negative sentiment derived from text and emojis. The sentiment of emojis is calculated based on the list provided by [10] and summed up for each emoji found in the tweet text. Similarly, the text sentiment is calculated according to the AFINN word list [9], which is also available as a Python package[9].

3.4 Algorithms

In the sequel, the ML algorithms, feature preprocessing and selection methods are considered. Literature has deemed effective the use of Naive Bayes, SVMs and Decision Trees for predicting the veracity of news. As such, four different algorithms are used for the training and evaluation of classification models: Naive Bayes, SVM, C4.5 and Random Forests [2] of C4.5. The rather popular Scikit-Learn Python module [13] implements these algorithms and is being used for the purposes of this study. Regarding SVMs, the Radial Basis Function kernel is made use of and the tweaking of parameters gamma and C is optimized through the use of the grid search hyper parameter tuning technique.

Furthermore, certain algorithms require feature preprocessing or selection to become more effective. In all cases, feature selection can significantly reduce training time, and aids in the prevention of overfitting. As such, for all algorithms except Random Forest, all features are ranked according to their mutual information and the top 10 of those are selected. Due to the fact that decision trees are harder to understand if preprocessing is applied to the data, it is avoided in this study. However, in the case of Naive Bayes and SVMs, all feature values v have been normalized to values v_{norm} lying in the $[0, 1]$ range through the transformation $v_{norm} = (v - v_{min})/(v_{max} - v_{min})$, where v_{min}, v_{max} are the minimum and maximum values across all values of a given feature, respectively.

[9] https://pypi.org/project/afinn/.

4 Results

4.1 Feature Selection

Having completed the feature selection process for the collected datasets, the top 10 most significant features according to the mutual information metric are: tweet length, TTR, number of punctuation marks, number of periods, average number of characters per sentence, number of adverbs, number of "to", number of verbs, number of entities and tweet entropy.

Since the features are mostly linguistic, the goal is to identify patterns in language use between tweets spreading fake and real news. Thus, one would perhaps expect language to be more formal and structurally sound in tweets disseminating real news than those disseminating fake ones. The most significant features derived from the feature selection process seem to be in accordance with that theory. For instance, low TTR indicates repetition of words, which may be inversely associated with conceptual variance, while adverbs constitute a hard-to-interpret source of information [4].

4.2 Machine Learning Models

The ML algorithms utilized for the classification of tweets spreading fake and real news are trained using the collected datasets and the corresponding evaluation results are presented. The dataset consists of 3,910 and 5,388 tweets spreading fake and real news, respectively, with the majority class baseline being 57.9%. In all cases, 5-fold cross validation is used to increase reliability of results.

Table 1. Evaluation results of the ML algorithms used in the classification process.

Algorithm	Class	Precision	Recall	F1 Score
Naive Bayes	Fake	90.1%	85.4%	87.6%
	Real	89.7%	92.8%	91.2%
	Average	89.9%	89.1%	89.4%
SVM	Fake	96.0%	84.0%	89.6%
	Real	89.4%	97.5%	93.3%
	Average	92.7%	90.8%	91.4%
C4.5	Fake	94.7%	84.7%	89.3%
	Real	89.8%	96.6%	93.0%
	Average	92.3%	90.6%	91.2%
Random Forest	Fake	97.5%	84.3%	90.3%
	Real	89.7%	98.4%	93.8%
	Average	93.6%	91.3%	92.1%

As can be observed from Table 1, although all algorithms perform fairly well, the most highly performing algorithm is Random Forest, achieving a macro

average F1 Score of 92.4%. A noteworthy observation from these results is that all algorithms tend to score higher in precision in the Fake rather than the Real class, whereas the opposite is true for recall; recall scores are noticeably lower for the Fake class when compared to Real. Since the impact of a tweet spreading fake news could be considered significant, one could argue that models achieving higher recall scores would be preferred, even if precision scores suffered somewhat. Nevertheless, both scores are adequately high, such that no additional performance concerns are raised.

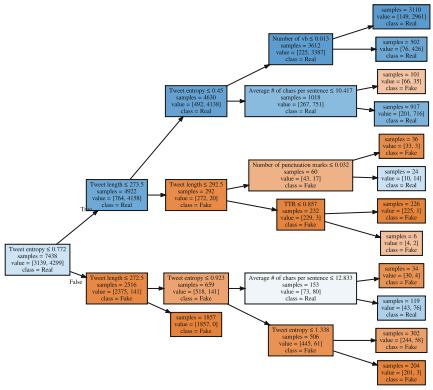

Fig. 3. A decision tree resulting from the C4.5 algorithm with maximum depth set to four.

In order to gain a better understanding of the factors that affect classification, an indicative decision tree resulting from C4.5 is depicted in Fig. 3. It is evident that tweet entropy contributes to an exceedingly high degree in the classification, as 75.7% and 96.7% of the tweets belonging to the Fake and Real class, respectively, are correctly identified within the first split. This would indicate that tweets in the Fake class contain more infrequent words than those in the Real class, which could be an aftereffect of the tweets being translated. Other notable features include tweet length, number of punctuation marks and a few of the morphological features, such as number of verbs. All things considered, according to such a model, a tweet spreading fake news would exhibit the following traits: unconventional vocabulary, longer length, fewer punctuation marks and shorter sentences.

The obtained results are compared to those found in the literature. Granik and Mesyura [6] scored an accuracy of 74% employing Naive Bayes. Ahmed et al. [1] scored an accuracy of 92% utilizing Linear SVM. Rubin et al. [16] achieved 87% F-score using SVM model. Sentiment features do not seem to be very important, as is also highlighted by literature [15]. Unlike the results presented here, Volkova et al. [19] conclude that morphological and grammar features are not important, although they do find that their results contradict previous work.

The overall high classification scores can be attributed to a number of factors resulting from the methodology. For one, even though journalists' tweets may be

more similar to the ones in the fake dataset than those of news agencies, they may still differ somewhat noticeably, making classification easier. Furthermore, the employed methodology relies purely on linguistic traits, with the ones in the fake news dataset being mostly translated to English from Chinese. Therefore, there is significant risk that the model may distinguish traits originating from the translation, which could explain the exceedingly high importance of the tweet's length and entropy. This is further accentuated if one considers the intricacies of the Chinese language, such as the fact that a single symbol corresponds to an English word. The consequences of such intricacies are evident in the translated dataset: a Chinese tweet that is originally 125 characters long is translated to 585 characters in English, which would normally be too long for a tweet.

5 Conclusions and Future Work

This study focused on detecting fake news in tweets related to Hong Kong protests, using ML algorithms. It used an initial dataset with fake news that has been publicized by Twitter, from which English and Chinese tweets are taken into account, with the latter being translated into English. Relevant tweets were identified through a language-independent process, utilizing hashtag information, enabling the selective use of machine translation. A whole new dataset with real news was built as well, which comes from Twitter accounts of worldwide news agencies. Interesting (in comparison to other studies) is the fact that real news were also retrieved from the personal Twitter accounts of the journalistic team of these agencies. The assembled datasets were used to train and evaluate ML algorithms, once the necessary feature extraction, preprocessing and selection was accomplished, to represent each tweet as a feature value vector. The obtained results achieved high classification scores and indicate that tweets spreading fake news and real news differ noticeably in linguistic features and most notably in tweet length, vocabulary. Word entropy was also deemed very important in classification, a feature not commonly used in similar studies.

Even though the obtained results are promising, there is still much room for improvement and experimentation in future work: (i) study the impact translation has on the results; (ii) include and compare different kinds of features, especially ones related to user and network characteristics; and (iii) utilize more modern ML algorithms, such as deep neural networks.

Acknowledgments. This project has received funding from the GSRT for the European Union's Horizon 2020 research and innovation programme under grant agreement No 644333.

References

1. Ahmed, H., Traore, I., Saad, S.: Detection of online fake news using N-gram analysis and machine learning techniques. In: Traore, I., Woungang, I., Awad, A. (eds.) ISDDC 2017. LNCS, vol. 10618, pp. 127–138. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69155-8_9

2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
3. Buntain, C., Golbeck, J.: Automatically identifying fake news in popular Twitter threads. In: 2017 IEEE International Conference on Smart Cloud (SmartCloud), pp. 208–215. IEEE (2017)
4. Conlon, S.P.N., Evens, M.: Can computers handle adverbs? In: The 15th International Conference on Computational Linguistics, COLING 1992, vol. 4 (1992)
5. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–4 (2015)
6. Granik, M., Mesyura, V.: Fake news detection using Naive Bayes classifier. In: 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 900–903, May 2017
7. Helmstetter, S., Paulheim, H.: Weakly supervised learning for fake news detection on Twitter. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 274–277, August 2018
8. Loper, E., Bird, S.: NLTK: the natural language toolkit. arXiv preprint cs/0205028 (2002)
9. Nielsen, F.Å.: A new ANEW: evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
10. Novak, P.K., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. PloS One **10**(12), e0144296 (2015)
11. Oshikawa, R., Qian, J., Wang, W.Y.: A survey on natural language processing for fake news detection. arXiv preprint arXiv:1811.00770 (2018)
12. Parmelee, J.H., Bichard, S.L.: Politics and the Twitter revolution: how tweets influence the relationship between political leaders and the public. Lexington Books (2011)
13. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**(Oct), 2825–2830 (2011)
14. Purbrick, M.: A report of the 2019 Hong Kong protests. Asian Aff. **50**(4), 465–487 (2019)
15. 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 (2017)
16. Rubin, V.L., Conroy, N., Chen, Y., Cornwell, S.: Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of the Second Workshop on Computational Approaches to Deception Detection, pp. 7–17 (2016)
17. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
18. Sosoni, V., et al.: Translation crowdsourcing: creating a multilingual corpus of online educational content. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
19. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.: Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on Twitter. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 647–653 (2017)
20. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)