Fake News Detection Regarding the Hong Kong Events from Tweets

Maria Nefeli Nikiforos, Spiridon Vergis, Andreana Stylidou, Nikolaos Augoustis, Katia Lida Kermanidis, Manolis Maragoudakis

Fake News Detection Regarding the Hong Kong Events from Tweets

Maria Nefeli Nikiforos[1][0000−0002−0118−8821], Spiridon Vergis[1][0000−0002−7642−1392], Andreana Stylidou[1][0000−0001−7816−9859], Nikolaos Augoustis[1][0000−0003−0934−6043], Katia Lida Kermanidis[1][0000−0002−3270−5078], and Manolis Maragoudakis[2][0000−0001−7701−0141]

[1] Ionian University, Department of Informatics, Corfu, Greece
{c19niki,c19verg,c19styl,c19avgo,kerman}@ionio.gr
[2] University of the Aegean, Department of Information and Communication Systems Engineering, Samos, Greece
mmarag@aegean.gr

Abstract. The rapid development of network services has led to the exponential growth of online information and the increasing number of social media users. These services are exploited by malicious accounts that spread fake news and propaganda in vast user networks. Consequently, an automated solution for fake news and deception detection is required. This paper introduces a new data set consisting of 2,366 tweets written in English, regarding the Hong Kong events (August, 2019), and a well-defined method for fake news detection that uses both linguistic and network features. Our approach is tested with experiments using 2 machine learning models, achieving high performance compared to previous research.

Keywords: Deception detection · Fake news detection · Linguistic Analysis · Twitter · Text Analysis.

1 Introduction

Due to the rapid development of network services and with the number of social media users constantly increasing, the size of information uploaded daily on the internet grows aggressively. This, combined with the ever growing number of malicious accounts created on social media, facilitates the spreading of fake news and propaganda in vast user networks [25]. Therefore, it is practically impossible for people to fact-check manually every piece of online information, and thus an automated solution based on modern technology is required.

Unlike previous work (to the authors' best knowledge), the approach proposed in this paper: a. uses both linguistic and network features to detect deceptive language in tweets for fake news detection, b. creates a new original data set which contains 2,366 tweets in English, regarding the Hong Kong events from August 2019, and c. identifies several attributes as determinant for fake news detection. Experiments with 2 machine learning models are conducted. The performance of each model is presented extensively in Section 6. The data set is available for research purposes at this address: `https://hilab.di.ionio.gr/wp-content/uploads/2020/02/HILab-Fake_News_Detection_For_Hong_Kong_Tweets.xlsx`

This paper is organized as follows: in Section 2 the past related work is discussed. In Section 3 the data set used for the fake news detection is analyzed. In Section 4 the linguistic and network features of the data set are presented. In Section 5 our methodology regarding SMOTE over-sampling and feature selection is outlined. In Section 6 the machine learning experiments and their results regarding the proposed approach are discussed. Finally, in Section 7 the paper is concluded and guidelines for future work are discussed.

2 Related Work

Due to the exponential growth of information created and uploaded on the internet daily, many researchers have focused on finding alternative ways of fact-checking that information. One of the first tools based on text analysis is the one presented by Houvardas and Stamatatos [12] for automatic authorship identification. Their tool can be considered a preliminary deception detector.

Along with the development of such tools, many researchers and journalists also started creating data sets with false and true news, in order to check the accuracy of these tools. Politifact [5], for example, is a website which offers short phrases and sentences, fact-checked by journalists. Another popular data set is FEVER [1] which consists of 185,000 short claims, created from Wikipedia sentences. Similar to FEVER, the LIAR [29] data set contains 13,000 short statements drawn from Politifact and categorised into 6 different classes (pants-fire, false, barely-true, half-true, mostly-true, true).

Recent papers regarding the automated fact-checking have created data sets using information from the social media. BUZZFEEDNEWS [11] contains 2,282 posts from 9 Facebook accounts of news agencies. Another data set containing posts from Facebook is Some-Like-It-Hoax [28], which consists of 15,500 posts from 32 Facebook pages. These posts are labeled based on the author, and not on the information contained within the post. Similar data sets based on Twitter are PHEME [9], which contains a collection of rumours and non-rumours, and CREDBank [17], which consists of 60 million Tweets labeled by human annotators.

Other data sets contain whole articles instead of short sentences and posts from social media. FakeNewsNet [24] contains whole false articles connected with Twitter posts. Another similar data set is ”Bluff The Listener and The Onion” [21], which consists of whole false articles, which are intended to be false by the authors (e.g. to express humor). BS Detector is also a data set that contains a list of unreliable news websites, and it labels automatically an article based on its information sources.

Another crucial factor for the accuracy of a fact-checking tool is the model used for the machine learning procedure. In Shojaee's et al. [23] work, Support Vector Machines and Naive Bayes classification algorithms are used for deception detection in reviews, using lexical and syntactic features, with accuracy 84% and 74%, respectively. Alowibdi et al. [6] conclude that a Naive Bayes-Decision Tree hybrid is the most effective classifier when it comes to deception detection using network features (accuracy 85%). Songram et al. [26] use k-nearest neighbour and Support Vector Machine, to detect messages leading to deception. The most accurate classifier was the Support Vector Machine with 99.21% accuracy.

More recent works suggest the use of neural networks to detect fake news, such as the model presented in Ruchansky's et al. [22] paper. Their model was tested with 2 data sets, and manages to surpass every other conventional classifier. Linguistic approaches have also been proposed in recent literature. Other research is that of Péerez-Rosas and Mihalcea [19], who present results from LIWC word class analysis used for deceptive text, and Hai et al. [10], who use logistic regression (semi-supervised learning) to detect fake news.

3 Data set

The data collected for the fake news detection consist of a set of 2,366 tweets written in English, regarding the Hong Kong events, and posted in August 2019 [2]. Our approach uses both linguistic and network features (Section 4). Certain network features, which are available for these tweets, are also collected. Additionally, certain linguistic features are extracted from the text of each tweet. A label feature is also added for each tweet, with value either "fake", or "real".

3.1 Fake News

In August 2019, Twitter disclosed 936 accounts originated from People's Republic of China, which were deliberately attempting to sow political discord in Hong Kong, including undermining the legitimacy and political positions of the protest movement on the ground. These accounts were suspended on violations of Twitter's manipulation policies; spam, coordinated activity, fake accounts, featured activity, and ban evasion. [2]

For the purposes of our research, tweets posted from these accounts were collected, in order to create a data set which contains false information, and are from now on referred to as "fake news". This data set originally consisted of 1,703,470 tweets. It is preprocessed, in order to extract tweets that actually: a. contain text, b. are written in English, and c. are about the Hong Kong political events (August, 2019). After preprocessing, the final fake news data set consists of 272 tweets in total.

3.2 Real News

For the purposes of our research, tweets posted from 9 Twitter accounts of renowned news agencies from August, 2019 to December, 2019 were collected, in order to create a data set which contains true and valid information, and are from now on referred to as “real news”. The news agencies are BBC Asia, BBC News (World), CCTV, China Daily, China Xinhua News, China.org.cn, Global Times, People’s Daily (China) and SHINE.

This data set originally consisted of 2,133 tweets. It is preprocessed, in order to extract tweets that actually: a. contain text, b. are written in English, and c. are about the Hong Kong political events (August, 2019). After preprocessing, the final real news data set consists of 2,094 tweets in total.

4 Features

The data set created in this research contains a plethora of features, both linguistic- and network-oriented. The final data set contains 23 features in total. The linguistic features are shown in Table 1 and the network features in Table 2.

Table 1. Linguistic features.

Feature	Tweet Example
Tweet text	Hong Kong Open: Wade Ormsby wins first European Tour title
Num syllables	33
Avg syllables	3
Avg Words in Sentence	10
Flesh-Kincaid	21.01
Num big Words	2
Num long sentences	0
Num short sentences	2
Num sentences	1
Num words	11
Rate adverbs adjectives	0.2727

4.1 Linguistic Features

A plethora of linguistic features is used in our research. (These features are selected and extracted according to [15] and [7]). The features selected in this work are based on Burgoon et. al. [15], and Li et. al. [7]. Both of these works use linguistic analysis for deception detection and attitude identification. From these works, the features best suited deception detection are used. The final set of features for each tweet of the data set includes: a. the tweet text, b. the number of syllables, c. the number of words, d. the number of sentences, e. the number of big words, f. number of syllables per word, g. number of short sentences, h. the number of long sentences, i. the Flesh-Kincaid level, j. the average number of words per sentence, and k. the rate of adjectives and adverbs (Table 1).

Table 2. Network features.

Feature	Tweet Example
User id	363345298
User display name	dalotbaba
User screen name	D3ZwcCm1Q1WNkKWaakHKQw=
Follower count	12033
Following count	10186
Account creation date	12/3/10 12:00 AM
Tweet time	12/11/17 12:00 AM
In reply to user id	2147483647
Like count	42
Retweet count	2
Num URLs	3

A feature that is widely used for fake news detection is the average number of syllables per word in the tweet text. The extraction of this feature is performed with the Loughran Master Dictionary [3], [16], which is used to determine which tokens are classified as proper English words, and their syllable count. In order to match the tweet text tokens with the words of the dictionary, tweet text is appropriately prepared; all punctuation marks, hashtags, at-mentions, and links are ignored. Due to the nature of tweets, tweet text may contain tokens that are either misspelled, or represent an assemblage of words, e.g. #behindthechair. To cover such cases, their respective syllable count value was replaced with the average syllable count of English words.

A python script is used to extract the rest of linguistic features. Twitter text has idioyncrasies that render its linguistic processing quite interesting and that have been tackled in various contexts, the TraMOOC system being one of them [27]. The Natural Language Toolkit (NLTK) library [4] is used to tag the text (PartOfSpeech-tagging) in order to identify the adjectives and adverbs. As for the long and short sentences, their threshold is set according to the research paper of Pennebaker et al. [18]. More specifically, sentences with less than 8 words are considered "short", while sentences with more than 25 words are "long". Similarly, words with more than 6 letters are considered "big".

A particularly unique and useful feature is the Flesh-Kincaid level [30]. It indicates how difficult it is to comprehend a particular piece of text; a high FK level means highly understandable text, and vice versa.

4.2 Network Features

The network features which are used in our research are collected directly from Twitter, and they are selected according to [13] and [14]. The first work presents a fake news data repository which consists of data collected from Twitter. The second work is a survey which reviews fake news detection approaches on social media. The final set of features for each tweet of the data set, which contains the features best suited for the proposed approach from both of the aforementioned works, includes: a. user id, b. user display name, c. user screen name, d. follower count, e. following count, f. account creation date, g. tweet time, h. in reply to user id (reply to specific user), i. like count, j. retweet count, and k. number of URLs (Table 2).

These features show how users form networks and interact within them on social media, based on their interests, topics and relations, which serve as the fundamental paths for information diffusion. Fake news dissemination processes tend to form an echo chamber cycle, highlighting the value of using network-based features to represent these types of network patterns for fake news detection.

5 Methodology

Following the preprocessing of the data set, as well as the collection and extraction of its network and linguistic features, comes the preparation of the data set for the machine learning experiments. Due to the imbalance between the number of fake news and real news learning examples (272 to 2,094, respectively), SMOTE over-sampling is performed, as defined in [8], and thereby "synthetic" fake news training examples are created. This imbalance has a bad effect on the performance of the minority class prediction, and therefore, over-sampling of the minority examples has been considered necessary. In real conditions, fake news is expected to be proportionally much less than real news. Consequently, imbalance of data will also exist in future research. Therefore, it is very likely that smoothing techniques, such as SMOTE, will be necessary. A feature selection is also performed, in order to identify the features which are more suitable and useful for fake news detection. The tool used for these tasks is the RapidMiner Studio (`https://bit.ly/2OaBX1N`).

5.1 SMOTE Over-sampling

The classification categories (labels), fake and real, of the data set are not equally represented (272 to 2,094 learning examples, respectively). Consequently, "synthetic" fake news training examples are created using SMOTE over-sampling.

According to Chawla et al. [8], the test set for the machine learning experiments must not include any "synthetic" examples. Therefore, prior to SMOTE over-sampling, the data set is split in 80% training set and 20% test set, with stratified sampling; it ensures that the label distribution in the subsets is the same proportionally as in the whole data set. The training set consisted of 1,893 examples: 1,675 with value "real" and 218 with value "fake" for the label feature. The test set to be used for the machine learning experiments consists of 473 examples: 419 with value "real" and 54 with value "fake" for the label feature.

The final step is "SMOTING" the minority label ("fake") of the training set. As a result, the final training set to be used for the machine learning experiments consists of 3,350 examples: 1,675 with value "real" and 1,675 (including 1,457 "synthetic" examples) with value "fake" for the label feature.

5.2 Feature Selection

In order to identify the features which are more suitable and useful for fake news detection, feature selection is performed. It is observed that certain network features are determinant for the fake news detection. These features, sorted from the most to the least determinant, are: a. user id, b. account creation date, c. following count, d. user display name, and e. user screen name. More specifically, it is observed that the machine learning algorithms take into consideration only one of these features (from most to least determinant), while completely ignoring the other features (linguistic and network). Therefore, to ensure that the models actually "learn", and thus obtain reliable results from the machine learning experiments, these features are removed from the training and test sets.

Consequently, the machine learning experiments (Section 6) are conducted with the training and test sets that were designed as described above, and without the aforementioned features, finally resulting in 18 remaining features.

6 Experiments and Results

The tool used to conduct the experiments with machine learning algorithms, the collection of results, the comparison of the models, and the identification of wrong predictions is the RapidMiner Studio. All models were trained with the SMOTE over-sampled training set and tested with the original test set, as described in Section 5.

An initial set of experiments were run using the Naive Bayes classifier, since it is commonly used in previous works [23,6] and can be used to compare the trained model's accuracy. Laplace correction is used in order to smooth the conditional probabilities. The produced model achieves 99.79% accuracy, which is high, compared to related work (Section 2). Precision and recall for each label are also high (Table 3). More specifically, the following are observed: a. all examples labelled as "real" are classified correctly, with a class recall of 100% and a class precision of 99.76%, and b. all examples labelled as "fake" are classified correctly except 1, with a class recall of 98.15% and a class precision of 100%. The wrong prediction concerning the only misclassified example occurs because, unlike the majority of the examples originally labelled as "fake", it does not contain long sentences or big words (linguistic features), while containing 2 URLs (network feature).

Table 3. Naive Bayes Confusion Matrix.

	true real	true fake	class precision
predicted real	419	1	99.76%
predicted fake	0	53	100.00%
class recall	100.00%	98.15%	

A second set of experiments using the Random Forest algorithm were implemented and applied. The Random Forest algorithm tries to minimize the overall error rate, so, in an unbalance data set, the larger class will get a low error rate while the smaller class will have a larger error rate [20]. Certain parameters need to be defined: a. the gain ratio is defined as the split criterion (no pruning), with maximal depth defined to 50, b. the number of trees to generate is set to 100, c. the subset ratio of randomly chosen features to test is set to 2, and d. the majority vote is defined as the voting strategy; it selects the label that is predicted by the majority of tree models. The produced model achieves 99.37% accuracy, which is high, compared to related work (Section 2). Precision and recall for each label are also high (Table 4). More specifically, the following are observed: a. all examples labelled as "real" are classified correctly, with a class recall of 100% and a class precision of 99.29%, b. all examples labelled as "fake" are classified correctly except 3, with a class recall of 94.44% and a class precision of 100%. The wrong predictions concerning the 3 misclassified examples occur because, unlike the majority of the examples originally labelled as "fake", they contain more than 7 words (linguistic feature) and 2 URLs (network feature).

Table 4. Random Forest Confusion Matrix.

	true real	true fake	class precision
predicted real	419	3	99.29%
predicted fake	0	51	100.00%
class recall	100.00%	94.44%	

7 Conclusions and Future Work

Unlike previous work (to the authors' best knowledge) the approach proposed in this paper: a. uses both linguistic and network features to detect deceptive language in tweets for fake news detection, b. creates a new original data set which contains 2,366 tweets in English, regarding the Hong Kong events from August 2019, and c. identifies several attributes as determinant for fake news detection.

To conclude, this paper described an innovative and well-defined method for detecting fake news in social media, by using both linguistic and network features to detect deceptive language, and identifying several attributes as determinant for fake news detection. The proposed method is tested on a new original data set consisting of 2,366 tweets in English, regarding the Hong Kong events (August 2019). During the experiment phase, 3 machine learning models are used. Both models managed to predict the examples labelled "real" and "fake" with high accuracy when compared with previous works. More specifically, the Naive Bayes model achieves an accuracy of 99.79% and the Random Forest model achieves an accuracy of 99.37%. This work draws guidelines for future work where more non-“synthetic” fake news will be utilized to train and test the machine learning models. Additionally, more models could be tested in order to improve the overall performance in such tasks. Finally, the features that turned out to be determinant could be further examined, to identify what makes them so.

Acknowledgements. 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. Fact extraction and verification, `http://fever.ai/`
2. Information operations directed at hong kong, `https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.html`
3. Loughran-mcdonald dictionary, `https://sraf.nd.edu/textual-analysis/resources/`
4. Natural language toolkit, `https://www.nltk.org/`
5. Fact-checking u.s. politics (2007), `https://www.politifact.com/truth-o-meter/`
6. Alowibdi, J.S., Buy, U.A., Philip, S.Y., Stenneth, L.: Detecting deception in online social networks. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). pp. 383–390. IEEE (2014)
7. Burgoon, J., Blair, J.P., Qin, T., Nunamaker, J.: Detecting deception through linguistic analysis. pp. 91–101 (06 2003). https://doi.org/10.1007/3-540-44853-5$_7$
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
9. Derczynski, L., Bontcheva, K.: Pheme: Veracity in digital social networks. In: UMAP workshops (2014)
10. Hai, Z., Zhao, P., Cheng, P., Yang, P., Li, X.L., Li, G.: Deceptive review spam detection via exploiting task relatedness and unlabeled data. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 1817–1826 (2016)
11. Hamza, S., Craig, S., Lauren, S., Ellie, H., Jeremy, S.V.: Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. Buzzfeed News (2016)
12. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. vol. 4183, pp. 77–86 (09 2006). https://doi.org/10.1007/11861461$_1$0
13. Kai Shu, Deepak Mahudeswaran, S.W.D.L., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media (2018)
14. Kai Shu, Amy Sliva, S.W.J.T., Liu, H.: Fake news detection on social media: A data mining perspective (2017)
15. Li, C., Guo, X., Mei, Q.: Deep memory networks for attitude identification. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 671–680. ACM (2017)
16. Loughran, T., McDonald, B.: When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance **66**(1), 35–65 (2011)

17. Mitra, T., Gilbert, E.: Credbank: A large-scale social media corpus with associated credibility annotations. In: Ninth International AAAI Conference on Web and Social Media (2015)
18. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of liwc2015. Tech. rep. (2015)
19. Pérez-Rosas, V., Mihalcea, R.: Experiments in open domain deception detection. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1120–1125 (2015)
20. Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P.: An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing **67**, 93–104 (2012)
21. Rubin, V.L., Conroy, N.J., Chen, Y.: Towards news verification: Deception detection methods for news discourse. In: Hawaii International Conference on System Sciences (2015)
22. 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 (2017)
23. Shojaee, S., Murad, M.A.A., Azman, A.B., Sharef, N.M., Nadali, S.: Detecting deceptive reviews using lexical and syntactic features. In: 2013 13th International Conference on Intellient Systems Design and Applications. pp. 53–58. IEEE (2013)
24. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 (2018)
25. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter **19**(1), 22–36 (2017)
26. Songram, P., Choompol, A., Thipsanthia, P., Boonjing, V.: Detecting thai messages leading to deception on facebook. In: International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making. pp. 293–304. Springer (2016)
27. Sosoni, V., Kermanidis, K.L., Stasimioti, M., Naskos, T., Takoulidou, E., Van Zaanen, M., Castilho, S., Georgakopoulou, P., Kordoni, V., Egg, M.: 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)
28. Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S., de Alfaro, L.: Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506 (2017)
29. Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017)
30. Wikipedia contributors: Flesch–kincaid readability tests Wikipedia, the free encyclopedia (2019), `https://en.wikipedia.org/w/index.php?title=Flesch%E2%80%93Kincaid_readability_tests&oldid=931233970`, [Online; accessed 21-January-2020]