computation

MDPI

Article

Deep Learning for Fake News Detection in a Pairwise Textual Input Schema

Despoina Mouratidis *,†, Maria Nefeli Nikiforos †and Katia Lida Kermanidis

Department of Informatics, Ionian University, 49100 Corfu, Greece; c19niki@ionio.gr (M.N.N.); kerman@ionio.gr (K.L.K.)

* Correspondence: c12mour@ionio.gr; Tel.: +30-266-1087756

† These authors contributed equally to this work.

Abstract: In the past decade, the rapid spread of large volumes of online information among an increasing number of social network users is observed. It is a phenomenon that has often been exploited by malicious users and entities, which forge, distribute, and reproduce fake news and propaganda. In this paper, we present a novel approach to the automatic detection of fake news on Twitter that involves (a) pairwise text input, (b) a novel deep neural network learning architecture that allows for flexible input fusion at various network layers, and (c) various input modes, like word embeddings and both linguistic and network account features. Furthermore, tweets are innovatively separated into news headers and news text, and an extensive experimental setup performs classification tests using both. Our main results show high overall accuracy performance in fake news detection. The proposed deep learning architecture outperforms the state-of-the-art classifiers, while using fewer features and embeddings from the tweet text.

Keywords: fake news detection; deception detection; machine learning; natural language processing; deep learning; social media; pairwise input

Citation: Mouratidis, D.; Nikiforos, M.N.; Kermanidis, K.L. Deep Learning for Fake News Detection in a Pairwise Textual Input Schema. *Computation* **2021**, *9*, 20. https://dx.doi.org/10.3390/computation9020020

Academic Editor: Yudong Zhang

Received: 31 December 2020
Accepted: 12 February 2021
Published: 17 February 2021

1. Introduction

In the past decade, the rapid spread of large volumes of online information among an increasing number of social network users is observed. It is a phenomenon that has often been exploited by malicious users and entities, which forge, distribute, and reproduce fake news and propaganda [1–19]. Fake news is intentionally forged information, which is distributed either to deceive and make false information believable, or to make verifiable facts doubtful [2,5,7–12,15,19–21]. Propaganda is another relative term for information which promotes specific political motives and other agendas [1,8–12,16,18,21,22].

The language used in forging fake news is deceptive, in the sense that it is intended to provoke and aggravate the users emotionally and lead them to spread the fake news [5,11,12,15–17,19,20,23], (e.g., "You thought this is on behalf of the people in Hong Kong. On the contrary, it is a rascality of putting the "false freedom" label on the will of most of Hong Kong people."). Another common indicator of deceptive language is the promotion of only one viewpoint, and thus being highly subjective [12,16,20,22], e.g., ("@feituji1994 I think we should supporting the Hong Kong Government."). Additionally, grammatical and spelling mistakes, as well as the use of the same limited set of words are characteristic properties of deceptive language [7,11,12,16]. The recent development of natural language processing (NLP), data mining, and machine learning tools has led to a more qualitative understanding of the features of deceptive language (linguistic features), as well as of the features of malicious users and entities (network account features) [1,2,4,5,7,8,11,12,14–19,22].

Fake news detection is the ability to define the truthfulness of information by analyzing its contents and related features [7,11]. Due to the unstructured and noisy data, the dynamic nature of news, and the increasing number of users, automated solutions

for fake news and deception detection in social networks are required [1,2,6,8,10,12,14–19,21,22]. Consequently, fake news and deception detection on social networks present unique challenges and has become an emerging research field, with future directions in data-oriented, feature-oriented, model-oriented, and application-oriented issues [1–5,8,11,12,15,16,18,22,24,25].

Unlike previous works, our work presents the following novelty and contribution:

- While the problem of fake news detection has been tackled in the past in a number of ways, most reported approaches rely on a limited set of existing, widely accepted and validated real/fake news data. The present work builds the pathway towards developing a new Twitter data set with real/fake news regarding a particular incident, namely the Hong Kong protests of the summer of 2019. The process of exploiting the provided fake tweets by Twitter itself, as well as the process of collecting and validating real tweet news pertaining to the particular event, are described in detail and generate a best practice setting for developing fake/real news data sets with significant derived findings.
- Another novelty of the proposed work is the form of the input to the learning schema. More specifically, tweet vectors are used, in a pairwise setting. One of the vectors in every pair is real and the other may be real or fake. The correct classification of the latter relies on the similarity/diversity it presents when compared to the former.
- The high performance of fake news detection in the literature relies to a large extent on the exploitation of exclusively account-based features, or to the exploitation of exclusively linguistic features. Unlike related work, the present work places high emphasis on the use of multimodal input that varies from word embeddings derived automatically from unstructured text to string-based and morphological features (number of syllables, number of long sentences, etc.), and from higher-level linguistic features (like the Flesh-Kincaid level, the adverbs-adjectives rate, etc.) to network account-related features.
- The proposed deep learning architecture is designed in an innovative way, that is used for the first time for fake news detection. The deep learning network exploits all aforementioned input types in various combinations. Input is fused into the network at various layers, with high flexibility, in order to achieve optimal classification accuracy.
- The input tweet may constitute the news text or the news header (defined in detail in Section 4). Previous works have used news articles headers and text as the two inputs for pairwise settings. However, this is the first time that tweets are categorized to headers and text based on their linguistic structure. This distinction in twitter data for fake news detection is made for the first time herein, accompanied by an extensive experimental setup that aims to compare the classification performance depending on the input type.
- Our work provides a detailed comparison of the proposed model with commonly used classification models according to related work. Additionally, experiments with these models are conducted, in order to assess and compare directly their performance with that of the proposed pairwise schema, by using the same input.
- Finally, an extensive review of the recent literature in fake news detection with machine learning is provided in the proposed work. Previous works with various types of data (news articles, tweets, etc.), different categories of features (network account, linguistic, etc.), and the most efficient network architectures and classification models are described thoroughly.

The rest of this paper is structured as follows: Section 2 discusses the recent related work regarding fake news detection from social networks, including the most common types of data and efficient machine learning techniques. Section 3 describes the creation and preprocessing of the data sets used in our experiments. Section 4 outlines the methodology regarding the feature set (Section 4.1), the embedding (Section 4.2), and the network architecture (Section 4.3). Section 5 presents the experiments' implementation, both for real header and real text input. Section 6 discusses the experiments' results, and compares

them to recent related work. Section 7 discusses the findings, concludes the paper, and presents some guidelines for future work.

2. Related Work

The spread of fake news has caused severe issues, having a great impact on major social events. Consequently, the recent related work regarding fake news detection from social networks is vast and several researchers have attempted to organize it and identify the most common types of data and machine learning techniques. Vishwakarma and Jain [8] listed the recent methods and data sets for fake news detection based on the content type of news they are applied to—the input data being either text or images. The review of Perera [22] offered an overview of the deep learning techniques for both manual and automatic fake news detection, identified 7 different levels of fake news based on the context, as well as on the motive for their creation and diffusion, and analyzed their processing by algorithms implemented for social media. Alam and Ravshanbekov [12] provided a definition for fake news and discussed the positive impact of combining NLP and deep learning techniques in automatic fake news detection. In a survey by Merryton and Augasta [4], baseline classifiers and deep learning techniques for fake and spam messages detection were overviewed, and the most common NLP preprocessing methods and tools, as well as the mostly used linguistic feature sets and data sets, were discussed. Han and Mehta [13] identified several fake news types and linguistic features, evaluated the performance of baseline classifiers and the performance of deep learning techniques regarding fake news detection, and compared them in the basis of balancing accuracy and lightweightness. Shu et al. [2] collected the existing definitions of fake news in the recent related work, identified the differences among the features, and the impact of fake news on social and on traditional media, and discussed the recent fake news detection approaches.

Regarding ensemble learning and reinforcement learning, there are certain works achieving high performance. Agarwal and Dixit [5] used the LIAR data set for the fake news class, and a data set from Kaggle, consisting of 20,801 news reports from the USA, for the real news class, resulting in a binary classification framework. They extracted credibility scores and other linguistic features from the text, and both data sets were normalized and tokenized. Python-based tools and libraries (Scikit-Learn, pandas, numpy, Keras, NLTK) were used for data preprocessing and the experiments. They created an ensemble, consisting of a Support Vector Machine (SVM), a Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), a k-Nearest Neighbor (KNN), and a Naive Bayes classifier, that used Bag of Words, Term Frequency–Inverse Document Frequency (TF-IDF), and n-grams. Their model achieved up to 97% accuracy with the LSTM. Wang et al. [6] developed the WeFEND framework for automatic annotation of news articles, which used user reports from WeChat as a form of weak supervision for fake news detection. They extracted textual and linguistic features from the data and conducted experiments with reinforcement learning, using the Linguistic Inquiry and Word Count (LIWC) and LSTM, reaching an accuracy value of up to 82%.

There are several approaches that explore the significance of textual and linguistic features for fake news detection. Nikiforos et al. [1] created a novel data set, consisting of 2366 tweets in English, regarding the Hong Kong protests of August 2019. Both network account and linguistic features were extracted from the tweets, while several features were identified as determinant for fake news detection. Their approach considered binary classification, and SMOTE over-sampling was applied to address class imbalance. The feature extraction, the SMOTE over-sampling and the experiments were conducted in the RapidMiner Studio. The performance of baseline classifiers, i.e., Naive Bayes and Random Forest, was evaluated, the final model achieving up to 99% accuracy. Zervopoulos et al. [18] also created a data set regarding the same events. It consisted of 3908 tweets in English, and Chinese translated into English (fake news class), and 5388 tweets in English from news agencies and journalists (real news class). They used exclusively linguistic features, translated Chinese tweets into English with Google's Translation API,

and identified linguistically relevant tweets. Python, Scikit-Learn, and NLTK were used for the preprocessing and the experiments. They evaluated the performance of Naive Bayes, SVM, C4.5, and Random Forest classifiers, achieving an average of 92.1% F1 score, and the best results were obtained with Random Forest. Jeronimo et al. [20] used a data set consisting of 207,914 news articles of 2 major mainstream media platforms in Brazil, collected from 2014 to 2017 (domains: Politics, Sports, Economy, and Culture), (real news class), and 95 news of 2 fact-checking services in Brazil (fake news class), collected from 2010 to 2017. The features were extracted by calculating the semantic distance between the data and 5 subjectivity lexicons (argumentation, presupposition, sentiment, valuation, and modalization) with Scikit-Learn. They conducted experiments with XGBoost, Random Forest (using Bag of Words and TF-IDF modeling), obtaining higher performance for inter domain scenarios. Mahyoob et al. [11] used 20 posts from PolitiFact as real news and 20 posts from Facebook as fake news, deriving 6 classes in total. They performed a qualitative and a quantitative data analysis with the QDA tool, comparing the posts on the basis of their linguistic features. Wang et al. [26] created LIAR, a new, publicly available data set for fake news detection. It consisted of approximately 12,800 manually labeled short statements of various topics from Politifact. Surface-level linguistic patterns were used for the experiments with hybrid CNNs, setting a benchmark for fake news detection on the novel data set. Shu et al. [27] presented a novel fake news data repository, FakeNewsNet. It contained 2 data sets with various features, including news content, social context, and spatiotemporal information. They also discussed the potential use of the FakeNewsNet on fake news and deception detection in social media. Ruchansky et al. [28] proposed a hybrid deep learning model for fake news and deception detection, by using features that included information regarding text and user behavior. They achieved up to 82.9% accuracy with experiments with a data set consisting of 992 tweets, 233,719 users, and 592,391 interactions.

Regarding deep learning, there are certain works achieving high performance. Sansonetti et al. [19] created a novel data set, consisting of 568,315 tweets that reference news indexed on PolitiFact, 62,367 news (34,429 fake news, 29,938 real news) referenced by tweets, and 4022 user profiles (2013 who publish mostly fake news, 2008 who publish mostly real news). They used both network account and linguistic features, and conducted experiments for offline and online analysis with CNN, LSTM, dense layer, and baseline classifiers (SVM, kNN), achieving up to 92% accuracy. Kumar et al. [16] compared different ensembles for binary classification on 1356 news from Twitter and 1056 real and fake news from PolitiFact. They created a data set per topic, and then they tokenized and encoded them. They used BeautifulSoup, Python, GloVe, and GPy. They conducted experiments with embeddings, CNN, and LSTM (ensemble and bidirectional). The CNN and bidirectional LSTM ensembled network with attention mechanism achieved the highest accuracy (88.78%). Alves et al. [21] created a novel, binary class data set, consisting of 2996 articles written in Brazilian Portuguese, collected from May to September 2018. The data set was normalized and tokenized, and Keras and TensorFlow were used. The experiments were conducted with a bidirectional and a regular LSTM and a dense layer. The 3-layer deep bidirectional LSTM with trainable word embeddings achieved accuracy up to 80%. Victor [3] used the PHEME data set and the LIAR data set, and conducted experiments with a deep two-path CNN and a bi-directional Recurrent Neural Network (RNN) for supervised and unsupervised learning, achieving up to 83% accuracy. Koirala [10] created a novel data set of 4072 news articles from Webhose.io, regarding fake news about COVID-19. They used linguistic features and conducted experiments with baseline classifiers, LSTM and dense layer, achieving an accuracy value between 70% and 80%.

Pairwise learning schemata are very popular in machine learning. The training data consist of lists of items that are specifically ordered within each list. Koppel et al. [29] presented a simple pairwise learning model for ranking. Experiments with the LETOR MSLR-WEB10K, MQ2007, and MQ2008 data sets were performed by using the Tensorflow library and its implementation of the Adam-Optimizer. Dong et al. [7] used the PHEME data set for semi-supervised, binary classification with baseline classifiers, LSTM, and a

deep two-path learning model containing 3 CNNs; both labeled and unlabeled data were used to train the model. Their performance was better than supervised learning models in the case where the distribution between the training and test data sets differed, and it proved to be more resistant to overfitting. Agrawal et al. [14] used tweets containing multimedia content; the training set consisted of approximately 5000 real news and approximately 7000 fake news, and the test set consisted of approximately 1200 real news and approximately 2500 fake news. They fused a pairwise ranking approach and a classification system, using image-based features, Twitter user-based features, and tweet-based features. For the classification a deep neural network, logistic regression, and SVM were used, along with n-grams and doc2vec vectors. The ranking was derived from the calculation of the distance between the features (contextual comparison) of tweets of the same topic (by hashtag). The ranking system outputs were incorporated within the classification system. They achieved accuracy up to 89% for real news and 78% for fake news. Bahad et al. [17] used 2 unstructured news data sets from the open machine learning repository (Kaggle) for binary classification. The experiments were conducted with LSTM, RNN, and CNN, using Python and TensorFlow. The highest accuracy, up to 98%, was achieved by the bi-directional LSTM-RNN. Abdullah et al. [15] used tokenized news from 12 distinct categories, and the prediction of the category determines the fake from the real news (12 classes). The experiments were conducted on Kaggle's cloud, with CNN, LSTM, and dense layer, achieving up to 97.5% accuracy. In a machine learning setting, Mouratidis et al. [30] presented a general deep learning architecture for learning to classify parallel translations, using linguistic information, of 2 machine translation model outputs and 1 human (reference) translation. They showed that the learning schema achieves the best score when information from embeddings and simple features are used for small data sets. Augenstein et al. [31] used a framework that combines information from embeddings in a multi-task learning experiment. They evaluated their approach on a variety of parallel classification tasks for sentiment analysis, and showed that, when the learning framework utilizes the ranker scores, the classification system outperforms a simple classification system.

More specifically, in this work, the learning schema is inspired by the architecture proposed for machine translation evaluation by Mouratidis et al. [30], and transferred to the domain of fake news detection, as described in Section 4. We define the input for this architecture based on the data set of [1] and according to the work of Augenstein et al. [31]. Augenstein et al. [31] have used news articles' headers and text as the two inputs for pairwise settings. However, this is the first time that tweets are categorized to headers and text based on their linguistic structure, as described in Section 3. The aim of this work was to identify the best practice setting for fake news detection. The proposed model exploits different input types (e.g., word embeddings, morphological and higher-level linguistic features) in various combinations. Input is fused into the model at various layers, with high flexibility, in order to achieve optimal classification accuracy. A detailed comparison of the proposed model with commonly used classification models according to related work is also presented.

3. Data

The data set used in our work is that of Nikiforos et al. [1]. It consists of 2363 tweets in English, regarding the Hong Kong protests of August 2019, and 23 features (described in Sections 4 and 4.1). The fake news tweets (fake tweets) of the data set (272 in total) were collected from 936 Twitter accounts that originated from the People's Republic of China, which were suspended in August 2019, due to violations of Twitter's manipulation policies, aiming to thwart the political convictions and notions of the Hong Kong protest movement. The real news tweets (real tweets) of the data set (2092 in total) were collected from 9 Twitter accounts of renowned news agencies: BBC Asia, BBC News (World), CCTV, China Daily, China Xinhua News, China.org.cn, Global Times, People's Daily (China), and SHINE. The aim was to include and represent true and valid information in the data set. The real tweets were originally 2133, posted from August 2019 to December 2019.

Preprocessing was considered necessary, in order to ensure that the tweets: (a) contain text, (b) are written in English, and (c) are relevant with the Hong Kong political movement of August 2019. 2092 remained after preprocessing.

The tweet text is used as input to the proposed neural network (described in Section 4). To this end, the tweets were divided into 4 distinct categories, depending on the class (real/fake) and the type of the tweet text (header/text). Therefore, the resulting categories are (a) real header, (b) real text, (c) fake header, and (d) fake text. As headers (real or fake), we consider the tweets that make a single-sentence statement (e.g., "Black terror: The real threat to freedom in Hong Kong"), in a form similar to newspaper headlines. Tweets that are longer than one sentence (e.g., "People with ulterior motives attempt to make waves in Hong Kong through the "color revolution", inciting student groups and Hong Kong citizens who do not know the truth, besieging the police headquarters and intending to undermine Hong Kong's stability") are considered as text (real or fake). There are two tweet inputs for the pairwise setting per experiment, *T1* and *T2*. For the first experiment, *T1* is a real header and *T2* can be either a real text, or a fake text, or a fake header. For the second experiment, *T1* is a real text and *T2* can be either a real header, or a fake text, or a fake header. Table 1 presents more details about the corpora. Imbalance between the two classes was observed, the fake tweet class being the minority class. Consequently, we applied the SMOTE filter to the minority class. Using SMOTE over-sampling [32], the total number of tweets increased from 2363 to 3766.

Table 1. Corpora details.

	Number
Real Header	1.027
Fake Header	127
Real Text	1.065
Fake text	144

4. Methodology

4.1. Feature Set

Similarly to Nikiforos et al. [1], both the network account and the linguistic features are used in our experiments. Every feature is scaled by the MaxAbsScaler [33]. The network account features were collected at the same time with the corresponding tweets from Twitter [1] to provide information about the account that posted the tweet and its connections throughout Twitter, as shown in Table 2. The network account features "user display name", "user screen name" and "in reply to user id" were not included in the final feature set, due to the large number of missing values. Regarding the account feature "account creation date", the dates were converted from text to numerical. Regarding the account feature "tweet time", the dates were converted from text to numerical, and the times were converted from 12-h mode to 24-h mode to avoid the ambiguity between p.m. and a.m. The linguistic features were extracted from the tweet text [1] to depict the specific language traits and forms per tweet, as shown in Table 2. The final feature set contains 18 features in total.

Table 2. Feature set.

Linguistic Features	Network Account Features
Num words	User id
Num syllables	Follower count
Avg syllables	Following count
Avg Words in Sentence	Account creation date
Flesh-Kincaid	Tweet time
Num big Words	Like count
Num long sentences	Retweet count
Num short sentences	Num URLs
Num sentences	
Rate adverbs adjectives	

4.2. Embedding Layer

In order to model the textual input, an embedding layer (automatically calculated) is used for the two different tweets per input pair (*T1*, *T2*). The embedding layer used is the one provided by the Keras library [34]. The encoding function applied is the one-hot function. The embedding layer size, in number of nodes, is 18. The input dimensions of the embedding layer are in agreement with the vocabulary of each input tweet text, taking into account the most frequent words.

4.3. Network Architecture

The fake news detection task is viewed as a binary classification problem. We propose a pairwise ranking approach in detecting tweets with fake content. Two tweets (*T1*, *T2*) are provided as input. The annotation for this problem is calculated as follows:

$$y = \begin{cases} 0, & \text{if } T1 \text{ is a real tweet and } T2 \text{ is a real tweet} \\ 1, & \text{if } T1 \text{ is a real tweet and } T2 \text{ is a fake tweet} \end{cases} \quad (1)$$

where y is the classification class label. The vectors (*T1*, *T2*) are used as input to the model, in a pairwise setting. Based on these tweets, the embedding vectors *EmbT1*, *EmbT2* were created on the embedding layer (described in Section 4.2). The MaxAbsScaler is used, as a preprocessing method for *EmbT1*, *EmbT2*. *EmbT1* and *EmbT2* were integrated in a parallel setting, and the vector (*EmbT1*, *EmbT2*) is thus created, and becomes the input to the hidden layer. The output of the hidden layer is the input to the last layer of the model. In this layer further input fusion takes place, i.e., a matrix *F[i,j]* is added, which is a 2D matrix with linguistic and network account features, as described in Section 4.1. The output label is modeled as a random variable in order to minimize the discrepancy between the predicted and the true labels, using maximum likelihood estimation, while the classification problem is modeled as a Bernoulli distribution. The model of the architecture is shown in Figure 1.

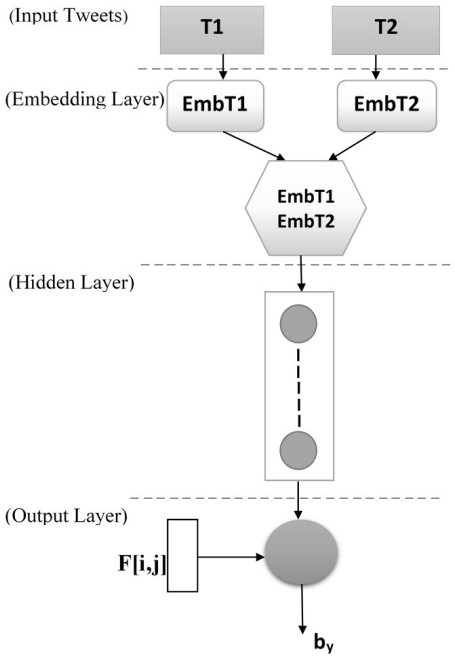

Figure 1. Model architecture.

5. Experiments

The present work investigates the modeling process that identifies real vs. fake tweets (text and headers) using the learning schema in Figure 1. For the first experiment (Experiment 1), *T1* is a real header and *T2* can be either a real text, a fake text, or a fake header. For the second experiment (Experiment 2), *T1* is a real text and *T2* can be either a real header, a fake text, or a fake header. The vector (*T1*, *T2*) is the input to the learning schema.

The model architecture for both experiments is defined as follows:

- Size of layers: Dense 1 and 2 with 128 hidden units, Dense 3 with 1 hidden unit (last layer).
- Output layer: Activation Sigmoid.
- Activation function of dense layers: 1 and 2 Relu, 3 Sigmoid.
- Dropout of dense layers: 0.4.

Table 3 presents additional parameters of the neural model.

Table 3. Parameters of the proposed model.

Parameter	Value
Optimizer	Adam [35]
Learning Rate	0.005
Loss function	Binary cross entropy

For all the experiments, we used 10-fold cross-validation, which is effective for small data sets. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

6. Results

In this section, the experiment results are presented. In order to quantify and evaluate the performance of the classifier, the Positive Predictive Value (Precision) and the Sensitivity (Recall) for both output labels were used as evaluation metrics. They are objective

measures, commonly used in classification tasks. The first metric shows which proportion of classifications is actually correct, whereas the second metric shows the proportion of actual positives that is classified correctly.

6.1. Accuracy Performance

Our main results are shown in Table 4. Regarding Experiment 1, prior to SMOTE over-sampling, Precision is 97% for real tweets and 100% for fake tweets and Recall is 95% for real tweets and 74% for fake tweets. After SMOTE over-sampling, Precision is 100% for real tweets and 100% for fake tweets and Recall is 100% for real tweets and 96% for fake tweets. Regarding Experiment 2, prior to SMOTE over-sampling, Precision is 99% for real tweets and 100% for fake tweets and Recall is 97% for real tweets and 93% for fake tweets. After SMOTE over-sampling, Precision is 100% for real tweets and 100% for fake tweets and Recall is 96% for real tweets and 96% for fake tweets.

It is observed that for both experiments the performance is increased after SMOTE over-sampling. Another observation is that for Experiment 2, in which the real text is *T1*, the performance is better than that of Experiment 1 prior to SMOTE. Consequently, the Experiment 2 setting is the most efficient for fake news detection, as it does not require SMOTE over-sampling to achieve better results. This also indicates that the correlation of the real text with text in general is greater than that of the real header. The correlation of the real header with the rest of the data is increased after SMOTE over-sampling, and thus for the framework of Experiment 1 the number of data affects the performance. Experiment 2 results also indicate that the real text (as *T1*) is highly correlated with data (*T2*, either real header, fake header, or text), compared to the respective correlation of the real header (as *T1*) with data (*T2*, either real text, fake header, or text) of Experiment 1. The latter correlation is slightly improved after SMOTE over-sampling, leading to the conclusion that the number of data affects the performance of the Experiment 1 framework. The proposed deep learning architecture achieves high overall accuracy performance, classifying mostly correctly both fake and real tweets, and thus shows great potential for successful fake news detection.

Table 4. Accuracy performance for Experiments 1 and 2.

	Experiment 1		**Experiment 2**	
Tweet	**Real**	**Fake**	**Real**	**Fake**
Prior_to_SMOTE_2.363 tweets segments				
Precision	97%	100%	99%	100%
Recall	95%	74%	97%	93%
Total Accuracy	95%		94%	
Average F1 score	99%		98%	
SMOTE_3.766 tweets segments				
Precision	100%	100%	100%	100%
Recall	100%	96%	96%	96%
Total Accuracy	98%		97%	
Average F1 score	100%		99%	

Figure 2 shows the accuracy performance according to training speed and batch size. By increasing the batch size and the epochs there is no significant accuracy increase. The best performance has been obtained for batch sizes 16 and 20 epochs.

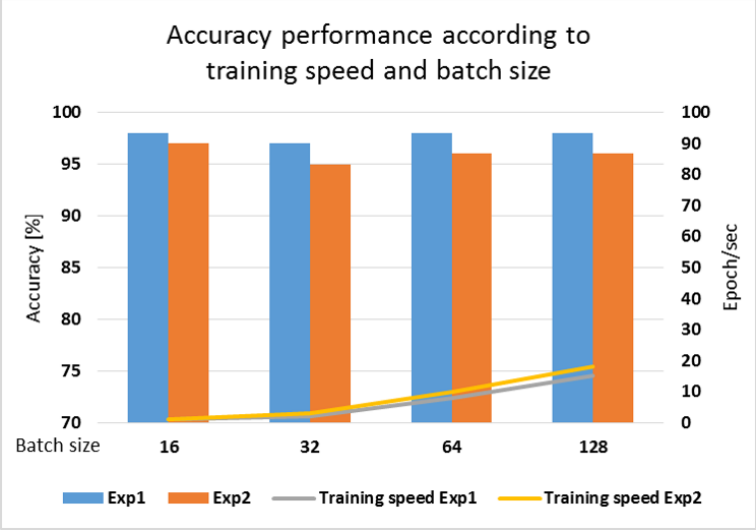

Figure 2. Accuracy performance according to training speed and batch size.

In Figure 3 the accuracy of the model is presented per experiment for both fake and real news, based on the network account feature user id. It is observed that the accuracy of prediction of real tweets (text and header) is not affected by this feature, while the accuracy of prediction of fake tweets (text and header) is reduced slightly when this feature is not used (1 to 2% decrease). Consequently, the network account feature user id does not affect the performance significantly.

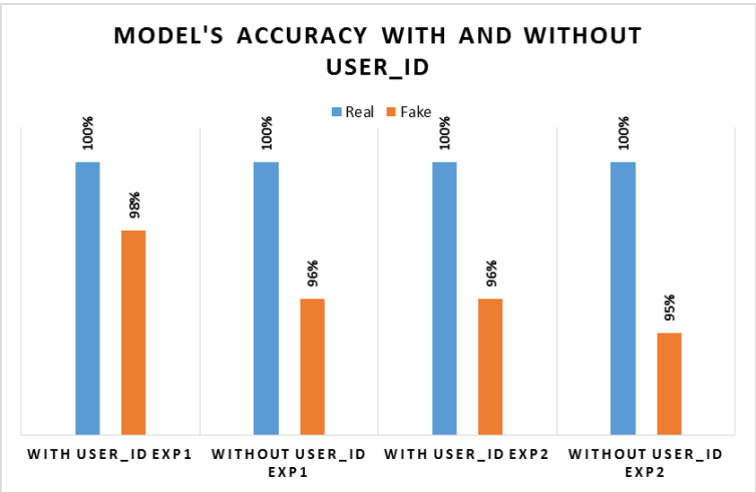

Figure 3. Model's accuracy with and without user_id feature.

6.2. Comparison to Related Work

In order to have a direct comparison of our experimental results with earlier work [1,18,36], additional experiments were run. Different configurations were experimented with, including Naive Bayes [1], Random Forest [18], and SVM, Logistic Regression [36] for Exp1 and Exp2. The WEKA framework was used as backend [37]. The evaluation metric used for the comparison is the harmonic mean of Precision and Recall, i.e., the F1 score.

It is observed that the proposed deep learning architecture outperforms the state-of-the-art classifiers for both experiments (up to 4% on average F1 score for Random Forest, up to 3% for Logistic Regression, up to 8% on average for SVM, and up to 15% for Naive Bayes) for both experiments. In addition, it is quite significant that the proposed deep learning architecture achieves a high F1 score for both fake and real tweets detection. The Random Forest classifier detected successfully all of the real tweets, and quite well the

fake tweets. The Naive Bayes and SVM classifiers faced problems in identifying real tweets from fake tweets (Figure 4).

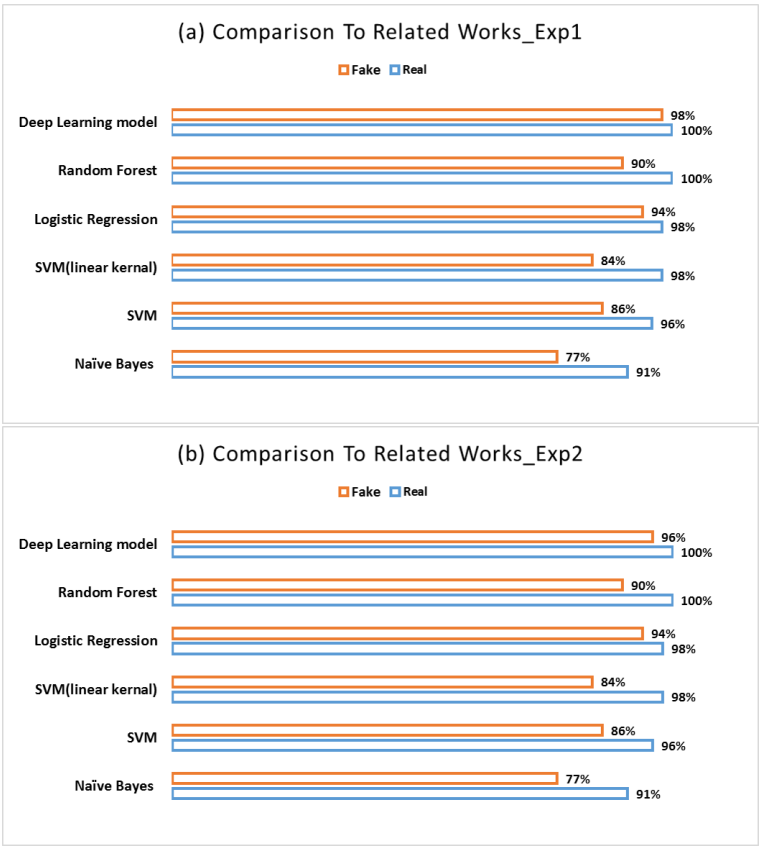

Figure 4. F1 score comparison for exp1 and exp2.

The accuracy metrics (Precision, Recall, F1 score) for each class (fake/real tweet) of our work is compared to those of recent related works, as shown in Table 5. More specifically, Zervopoulos et al. [18] used a larger data set (3908 fake and 5388 real tweets) concerning the same event (Hong Kong protest movement of summer, 2019). However, they used exclusively linguistic features. Their best results were obtained with Random Forest, achieving (on average) 93.6% Precision, 91.3% Recall, and 92.1% F1 score. Nikiforos et al. [1] used similar data set and feature sets with those used in our work, though a different feature selection methodology was applied. Their best results were obtained with Naive Bayes and SMOTE oversampling, achieving (on average) 99.8% Precision and 99% Recall. It is therefore observed that the model proposed in our work obtains better results and achieves higher performance, compared to these works.

Table 5. Accuracy comparison with related works.

	Tweet	Precision	Recall	F1 score
Deep Learning Model with SMOTE	Fake	100%	96%	98%
	Real	100%	100%	100%
	Average	100%	98%	99%
Random Forest [18]	Fake	97.5%	84.3%	90.3%
	Real	89.7%	98.4%	93.8%
	Average	93.6%	91.3%	92.1%
SVM [18]	Fake	96%	84%	89.6%
	Real	89.4%	97.5%	93.3%
	Average	92.7%	90.8%	91.4%
Naive Bayes [1]	Fake	100%	98.1%	-
	Real	99.7%	100%	-
	Average	99.8%	99%	-
Random Forest [1]	Fake	100%	94.4%	-
	Real	99.2%	100%	-
	Average	99.6%	97.2%	-

The above observations (Figure 4, Table 5) conclude that the proposed deep learning architecture, using 18 features and information (embeddings) from the tweet text, achieves the best accuracy results.

7. Conclusions

Unlike previous works, our work presents the following novelty and contribution. While the problem of fake news detection has been tackled in the past in a number of ways, most reported approaches rely on a limited set of existing, widely accepted, and validated real/fake news data. The present work builds the pathway towards developing a new Twitter data set with real/fake news regarding a particular incident, namely the Hong Kong protests of the summer of 2019. The process of exploiting the provided fake tweets by Twitter itself, as well as the process of collecting and validating real tweet news pertaining to the particular event, are described in detail and generate a best practice setting for developing fake/real news data sets with significant derived findings.

Another novelty of the proposed work is the form of the input to the learning schema. More specifically, tweet vectors are used, in a pairwise setting. One of the vectors in every pair is real and the other may be real or fake. The correct classification of the latter relies on the similarity/diversity it presents when compared to the former. The high performance of fake news detection in the literature relies to a large extent on the exploitation of exclusively account-based features, or to the exploitation of exclusively linguistic features. Unlike related work, the present work places high emphasis on the use of multimodal input that varies from word embeddings derived automatically from unstructured text to string-based and morphological features (number of syllables, number of long sentences, etc.), and from higher-level linguistic features (like the Flesh-Kincaid level, the adverbs-adjectives rate, etc.) to network account-related features.

The proposed deep learning architecture is designed in an innovative way, that is used for the first time for fake news detection. The deep learning network exploits all aforementioned input types in various combinations. Input is fused into the network at various layers, with high flexibility, in order to achieve optimal classification accuracy. The input tweet may constitute the news text or the news header (defined in detail in Section 4). Previous works have used news article headers and text as the two inputs for pairwise settings. However, this is the first time that tweets are categorized to headers and text based on their linguistic structure. This distinction in twitter data for fake news detection is made for the first time herein, accompanied by an extensive experimental setup that aims to compare the classification performance depending on the input type.

Our work provides a detailed comparison of the proposed model with commonly used classification models according to related work. Additionally, experiments with these models are conducted, in order to assess and compare directly their performance with that of the proposed pairwise schema, by using the same input. Finally, an extensive review of the recent literature in fake news detection with machine learning is provided in the proposed work. Previous works with various types of data (news articles, tweets, etc.), different categories of features (network account, linguistic, etc.), and the most efficient network architectures and classification models are described thoroughly.

More specifically, the deep learning architecture by Mouratidis et al. [30] is used as a basis to fake news detection, whereas the input for this architecture is based on the data set of [1], and defined according to the work of Augenstein et al. [31], who compared news headers and text through their pairwise framework to detect fake news text.

Our main results show high overall accuracy performance of the proposed deep learning architecture in fake news detection. For both experiments, the performance is increased after SMOTE over-sampling. For Experiment 2, where *T1* is real text, the performance is better than that of Experiment 1 prior to SMOTE. Consequently, the Experiment 2 setting is the most efficient for fake news detection, as it does not require SMOTE over-sampling to achieve better results. This also indicates that the correlation of the real text with text in general is greater than that of the real header. The correlation of the real header with the rest of the data is increased after SMOTE over-sampling, and thus for the framework of Experiment 1 the number of data affects the performance.

Experiment 2 results also indicate that the real text (as *T1*) is highly correlated with data (*T2*, either real header, fake header. or text), compared to the respective correlation of the real header (as *T1*) with data (*T2*, either real text, fake header, or text) of Experiment 1. The latter correlation is slightly improved after SMOTE over-sampling, leading to the conclusion that the number of data affects the performance of the Experiment 1 framework. Additional experiments with Naive Bayes, Random Forest, and SVM were also run, using the WEKA framework as backend [37], in order to compare directly our experimental results with earlier work. More specifically we achieved up to 99% accuracy with Naive Bayes [1], 92.1% average F1 score with Random Forest [18], and up to 92% accuracy with CNN [19]. The proposed deep learning architecture outperforms the state-of-the-art classifiers, while achieving high F1 score for both fake and real tweets detection. The Random Forest classifier detected successfully all of the real tweets and quite well the fake tweets. The Naive Bayes and SVM classifiers faced problems in identifying the real tweets from the fake ones. In conclusion, the proposed deep learning architecture, using 18 features and information (embeddings) from the tweet text, achieves the best accuracy results.

In future work, we will aim to test a different model configuration (e.g., different kinds of neural network layers). Apart from the pairwise classification schema that is used in this paper, we will test other classification schemata, for identifying fake content. In addition, the proposed model will be tested to a wider field of problems for fake content detection, e.g., spams. Finally, it is worth exploring further data sets and other content formats (e.g., multimedia content, photos, videos) in the proposed model.

Author Contributions: D.M., M.N.N., and K.L.K.; methodology, D.M.; software, D.M. and M.N.N.; validation, formal analysis, investigation, M.N.N.; data curation, D.M. and M.N.N.; writing—original draft preparation, K.L.K.; writing—review and editing, K.L.K.; supervision. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in [1].

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
KNN	k-Nearest Neighbor
LIWC	Linguistic Inquiry and Word Count
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
RNN	Recurrent Neural Networks
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency

References

1. Nikiforos, M.N.; Vergis, S.; Stylidou, A.; Augoustis, N.; Kermanidis, K.L.; Maragoudakis, M. Fake News Detection Regarding the Hong Kong Events from Tweets. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 177–186.
2. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
3. Victor, U. Robust Semi-Supervised Learning for Fake News Detection. Ph.D Thesis, Prairie View A&M University, Prairie View, TX, USA, 2020.
4. Merryton, A.R.; Augasta, G. A Survey on Recent Advances in Machine Learning Techniques for Fake News Detection. *Test Eng. Manag.* **2020**, *83*, 11572–11582.
5. Agarwal, A.; Dixit, A. Fake News Detection: An Ensemble Learning Approach. In Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13–15 May 2020; pp. 1178–1183.
6. Wang, Y.; Yang, W.; Ma, F.; Xu, J.; Zhong, B.; Deng, Q.; Gao, J. Weak supervision for fake news detection via reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 516–523.
7. Dong, X.; Victor, U.; Chowdhury, S.; Qian, L. Deep Two-path Semi-supervised Learning for Fake News Detection. *arXiv* **2019**, arXiv:1906.05659.
8. Vishwakarma, D.K.; Jain, C. Recent State-of-the-art of Fake News Detection: A Review. In Proceedings of the 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 5–7 June 2020; pp. 1–6.
9. Gill, H.; Rojas, H. Chatting in a mobile chamber: Effects of instant messenger use on tolerance toward political misinformation among South Koreans. *Asian J. Commun.* **2020**, *30*, 470–493. [CrossRef]
10. Koirala, A. COVID-19 Fake News Classification with Deep Learning. *Preprint* **2020**. [CrossRef]
11. Mahyoob, M.; Al-Garaady, J.; Alrahaili, M. Linguistic-Based Detection of Fake News in Social Media. *Forthcom. Int. J. Engl. Linguist.* **2020**, *11*, 99–109. [CrossRef]
12. Alam, S.; Ravshanbekov, A. Sieving Fake News From Genuine: A Synopsis. *arXiv* **2019**, arXiv:1911.08516.
13. Han, W.; Mehta, V. Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation. In Proceedings of the 2019 IEEE International Conference on Industrial Internet (ICII), Orlando, FL, USA, 11–12 November 2019; pp. 375–380.
14. Agrawal, T.; Gupta, R.; Narayanan, S. Multimodal detection of fake social media use through a fusion of classification and pairwise ranking systems. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 28 August–2 September 2017; pp. 1045–1049.
15. Abdullah, A.; Awan, M.; Shehzad, M.; Ashraf, M. Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory. *Int. J. Emerg. Technol. Learn.* **2020**, *11*, 209–212.
16. Kumar, S.; Asthana, R.; Upadhyay, S.; Upreti, N.; Akbar, M. Fake news detection using deep learning models: A novel approach. *Trans. Emerg. Telecommun. Technol.* **2020**, *31*, e3767. [CrossRef]
17. Bahad, P.; Saxena, P.; Kamal, R. Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. *Procedia Comput. Sci.* **2019**, *165*, 74–82. [CrossRef]
18. Zervopoulos, A.; Alvanou, A.G.; Bezas, K.; Papamichail, A.; Maragoudakis, M.; Kermanidis, K. Hong Kong Protests: Using Natural Language Processing for Fake News Detection on Twitter. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 408–419.
19. Sansonetti, G.; Gasparetti, F.; D'Aniello, G.; Micarelli, A. Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection. *IEEE Access* **2020**, *8*, 213154–213167. [CrossRef]
20. Jeronimo, C.L.M.; Marinho, L.B.; Campelo, C.E.; Veloso, A.; da Costa Melo, A.S. Fake News Classification Based on Subjective Language. In Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, Munich, Germany, 2–4 December 2019; pp. 15–24.

21. Alves, J.L.; Weitzel, L.; Quaresma, P.; Cardoso, C.E.; Cunha, L. Brazilian Presidential Elections in the Era of Misinformation: A Machine Learning Approach to Analyse Fake News. In *Iberoamerican Congress on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 72–84.
22. Perera, K. The Misinformation Era: Review on Deep Learning Approach to Fake News Detection. *Preprint* **2020**. [CrossRef]
23. Anshika, C.; Anuja, A. Linguistic feature based learning model for fake news detection and classification. *Expert Syst. Appl.* **2021**, *169*, 114–171.
24. De Oliveira, N.R.; Pisa, P.S.; Lopez, M.A.; de Medeiros, D.S.V.; Mattos, D.M.F. Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. *Information* **2021**, *12*, 38. [CrossRef]
25. Ranjan, S.S.; Gupta, B.B. Multiple features based approach for automatic fake news detection on social networks using deep learning. *Appl. Soft Comput.* **2021**, *100*, 106–983.
26. Wang, W.Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv* **2017**, arXiv:1705.00648.
27. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* **2020**, *8*, 171–188. [CrossRef]
28. 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, Singapore, 6–10 November 2017; pp. 797–806.
29. Köppel, M.; Segner, A.; Wagener, M.; Pensel, L.; Karwath, A.; Kramer, S. Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 237–252.
30. Mouratidis, D.; Kermanidis, K.L.; Sosoni, V. Innovative Deep Neural Network Fusion for Pairwise Translation Evaluation. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 76–87.
31. Augenstein, I.; Ruder, S.; Søgaard, A. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. *arXiv* **2018**, arXiv:1802.09913.
32. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
34. Keras, K. Deep Learning Library for Theano and Tensorflow. Available online: https://keras.io/ (accessed on 31 January 2021).
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Izonin, I.; Trostianchyn, A.; Duriagina, Z.; Tkachenko, R.; Tepla, T.; Lotoshynska, N. The combined use of the wiener polynomial and SVM for material classification task in medical implants production. *Int. J. Intell. Syst. Appl.* **2018**, *10*, 40–47. [CrossRef]
37. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]