Neural Computing and Applications
https://doi.org/10.1007/s00521-021-06230-0

SPECIAL ISSUE ON ADVANCES OF NEURAL COMPUTING PHASING CHALLENGES IN THE ERA OF 4TH INDUSTRIAL REVOLUTION

Deep learning for fake news detection on Twitter regarding the 2019 Hong Kong protests

Alexandros Zervopoulos[1] · Aikaterini Georgia Alvanou[1] · Konstantinos Bezas[1] · Asterios Papamichail[1] · Manolis Maragoudakis[1] · Katia Kermanidis[1]

Received: 13 November 2020 / Accepted: 13 June 2021

Abstract
The dissemination of fake news on social media platforms is an issue of considerable interest, as it can be used to misinform people or lead them astray, which is particularly concerning when it comes to political events. The recent event of Hong Kong protests triggered an outburst of fake news posts that were identified on Twitter, which were then promptly removed and compiled into datasets to promote research. These datasets focusing on linguistic content were used in previous work to classify between tweets spreading fake and real news using traditional machine learning algorithms (Zervopoulos et al., in: IFIP international conference on artificial intelligence applications and innovations, Springer, Berlin, 2020). In this paper, the experimentation process on the previously constructed dataset is extended using deep learning algorithms along with a diverse set of input features, ranging from raw text to handcrafted features. Experiments showed that the deep learning algorithms outperformed the traditional approaches, reaching scores as high as 99.3% F1 Score, with the multilingual state-of-the-art model XLM-RoBERTa outperforming other algorithms using raw untranslated text. The combination of both traditional and deep learning algorithms allows for increased performance through the latter, while also gaining insight regarding tweet structure from the interpretability of the former.

Keywords Fake news detection · Natural language processing · Deep learning · Machine learning · Convolutional neural networks · Long short-term memory · XLM-RoBERTa · Twitter · Hong Kong protests

1 Introduction

Social media is now an integral part of people's daily lives, providing their users with direct and borderless communication. At the same time, they are a source of information for current events that take place both domestically and globally. Nevertheless, many times a news item is not cross-referenced before it is disseminated to the public and, as a consequence, its validity is not guaranteed, which may be influenced by conspiracies, political interests and expediencies. Consequently, the phenomenon of spreading false news can be observed intensely and on a daily basis, making it necessary to address it, in order to protect values and ideals.

Fake news is often linked to political events and situations and is disseminated through a variety of media, including Twitter, where it appears that real news is being

✉ Alexandros Zervopoulos
c19zerv@ionio.gr

Aikaterini Georgia Alvanou
c19alva@ionio.gr

Konstantinos Bezas
c19beza@ionio.gr

Asterios Papamichail
c19papa@ionio.gr

Manolis Maragoudakis
mmarag@ionio.gr

Katia Kermanidis
kerman@ionio.gr

[1] Department of Informatics, Ionian University, Corfu, Greece

Published online: 03 July 2021

disseminated at a lower rate than fake news [22, 28]. In this context, computer science can also be used as a key asset and tool for crawling fake news feeds from Twitter user accounts, helping to tackle and eliminate this phenomenon. Detecting fake news with traditional methods, for example with the participation of certified journalists, is a costly process, both in terms of time and money. Therefore, more modern methods with the involvement of artificial intelligence are preferred [21]. The recent events of the Hong Kong protests in June 2019 related to political controversy have been of great concern to the public, due to the violent turn and the high turnout of citizens inside and outside China's borders [25]. As a result, a plethora of tweets was triggered, raising the question of the validity of their content. So, it is important to study the extent of the fake news spread on Twitter about this event.

In this paper, previous work [33] regarding the classification of fake news concerning the Hong Kong protests with the use of traditional machine learning (ML) algorithms is extended. More specifically, three diverse feature-sets, focusing on purely linguistic content (unlike most previous approaches that rely at least partly on user account information), are derived from the previously constructed dataset, which contains tweets in both English and Chinese, that are attempting to spread fake and real news originating from malicious users and trusted journalists, respectively. These feature-sets represent the content at different levels of abstraction, ranging from raw text to handcrafted features, and are fed as input to a multitude of modern deep learning algorithms. An evaluation is performed, comparing the results of the deep learning models to the previous ones, showing that the deep learning architectures outperform the traditional ones, achieving higher results across the board. In particular, a state-of-the-art multilingual model, XLM-RoBERTa [9], achieves the highest scores, which is trained over raw text in both English and Chinese, whereas the rest of the models utilize translated text. This fact provides confidence in the acquired results, showing that translation has not significantly impacted the performance of the other algorithms, while also confirming their capabilities at cross-lingual tasks. Regarding feature-sets, a feature that has not been investigated so far in this setting, that stands out from the handcrafted feature-set, is tweet entropy, which serves as an indicator of word importance.

The rest of this paper is structured in the following manner. An overview of previous related work is presented in Sect. 2, focusing on deep learning algorithms and relevant text representation techniques. The applied methodology is described in Sect. 3, including dataset construction and the algorithms used. In Sect. 4, the produced results are reported and discussion is made, comparing the different algorithms and feature-sets. Finally, conclusions are drawn in Sect. 5.

2 Related work

Fake news detection, using natural language processing (NLP), has been extensively researched, especially in the current era, where there is a dissemination of distorted information through social media. More precisely, Oshikawa et al. [21] explain the difference between detecting fake news and other similar concepts, such as rumor detection and present existing datasets, features and models, while Khan et al. [18] provide a detailed description of some advanced deep learning models. Zhou et al. [34] categorize fake news detection in four different groups, depending on the aspects being focused on, including a text's writing style and source, analyzing the various techniques used in each group.

Regarding the techniques and methodologies used to identify falsehood, Bajaj [4] uses fake news articles from the Kaggle dataset and authentic articles from the Signal Media News dataset as part of his study, with the ultimate goal of creating a classifier that can predict the validity of news based on their content. In addition, the performance of a variety of models is examined, with a recurrent neural network with gated recurrent units standing out.

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 [15], who consider the credibility of a tweet's source as a proxy for the trustworthiness of the tweet itself, achieving high prediction scores.

Long et al. [19], in their study, propose a model that extends the lexical-based analysis method for fake news detection. This is achieved by integrating speaker profile information into an attention-based long short-term memory (LSTM) model. The evaluation of the proposed model uses the dataset provided by Wang [29]. The results show an increased accuracy of 14.5% higher compared to the most advanced hybrid convolution-based models.

Wang et al. [30] propose a framework named event adversarial neural network (EANN) which overcomes a barrier introduced by existing approaches, the event-specific feature extraction. EANN utilizes two different sources for feature extraction purposes, image and text, both of which are extracted from posts. Next, the analysis continues with the event discriminator which removes the event-specific features. From the experimental procedure, it is proven that this method performs better than the state-of-the-art methods. Furthermore, the final features can be utilized for other events.

Kaliyar et al. [17] propose a deep convolutional neural network, called FNDNet, which uses GloVe's word-embedding vectors and multiple hidden layers, in order to automatically learn the discriminatory features. Their architecture is based on the concept of neural networks with multiple parallel channels and variable size, while their results are high and promising for fake news detection. Moreover, it is noteworthy that Conneau et al. [10] use a Transformer-based masked language model on one hundred languages, including Chinese and English, with the ultimate goal of improving cross-lingual language understanding.

3 Methodology

This section lays out the employed methodology, starting from the retrieval of the initial dataset, containing tweets spreading fake news, and the process of collecting tweets from sources considered trustworthy enough to be spreading real news. Afterward, the feature-sets extracted from the collected tweets are listed, along with the algorithms used for classification and evaluation.

3.1 Fake news dataset

The initial dataset is available from Twitter and specifically from Twitter's Election Integrity Hub[1] and dates back to August and September 2019. This dataset consists of three smaller datasets, which contain data from user accounts, which are considered to be "deliberately and specifically trying to sow political discord in Hong Kong."[2] In more detail, the aforementioned dataset consists of 13,856,454 tweets and includes 31 fields that represent tweet-related features. These features contain information about the tweet, the account that posted it and network-related interactions, which are listed in Table 1 along with the number of fields in each category. Therefore, and considering Twitter as a trustworthy source, this dataset is considered as ground truth with respect to the fake news portion of the assembled dataset.

Based on the description of the dataset from Twitter, the accounts involved tend to be fake, post spam and act in a coordinated manner, which has also been investigated in the literature [26]. So there is a reasonable possibility that not all tweets are related to the falsehood regarding Hong Kong protests, reinforcing the need for an initial preprocessing step. Moreover, given that these events are taking place in China, it is assumed that most of the related tweets will be written in either Chinese or English, with the latter being considered due to it being globally used and being the most popular language on the Twitter platform.

Due to the large size of these datasets, word clouds are constructed to provide some intuition about the tweets' contents and the topics they're concerned with. To construct these word clouds, preprocessing is performed on the tweet's text, including the removal of hashtags, mentions and URLS from the Chinese and English tweets found in the first two out of the three datasets published by Twitter. Furthermore, the Chinese tweets are translated into English through the Google Translation API available on the Google Cloud[3] platform. A word cloud is presented for tweets written in each of the two languages: Fig. 1a for English, Fig. 1b for Chinese. It is apparent that the phrases appearing in the former mostly consist of advertisements and spam, while those in the latter are more politically oriented, overall.

To precisely identify the tweets spreading false information regarding Hong Kong protests, the filtering methodology presented in Algorithm 1 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. In fact, using the appropriate data structures, the average time complexity of this algorithm is $O(n)$, where n is the number of tweets, due to the fact that there is a maximum number of hashtags in a tweet as a result of a tweet's character limit. Thus, the nested for loop has a small upper bound of iterations that does not scale with n.

The methodology can be broken down as follows. First of all, a list of curated hashtags related to Hong Kong protests is manually constructed, comprising both English and Chinese hashtags. Afterward, hashtags appearing 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 (lines 2–15). Finally, tweets are deemed relevant if they contain a hashtag with a co-occurrence count higher than an arbitrary threshold (lines 16–23). In this case, the value 100 is used as the co-occurrence threshold across all experiments. It is evident that, due to this approach, tweets without hashtags cannot be considered relevant.

[1] https://transparency.twitter.com/en/information-operations.html.

[2] https://tinyurl.com/y3ffrblt.

[3] https://cloud.google.com/translate/.

Table 1 Categorization of the 31 fields present in the fake news dataset retrieved from Twitter's Election Integrity Hub

Category	Indicative fields	Number of fields
Account-related	Display name, # of followers/following users	11
Tweet-related	Tweet text, hashtags, URLs	14
Network-related	Quote, like, reply and retweet count	6

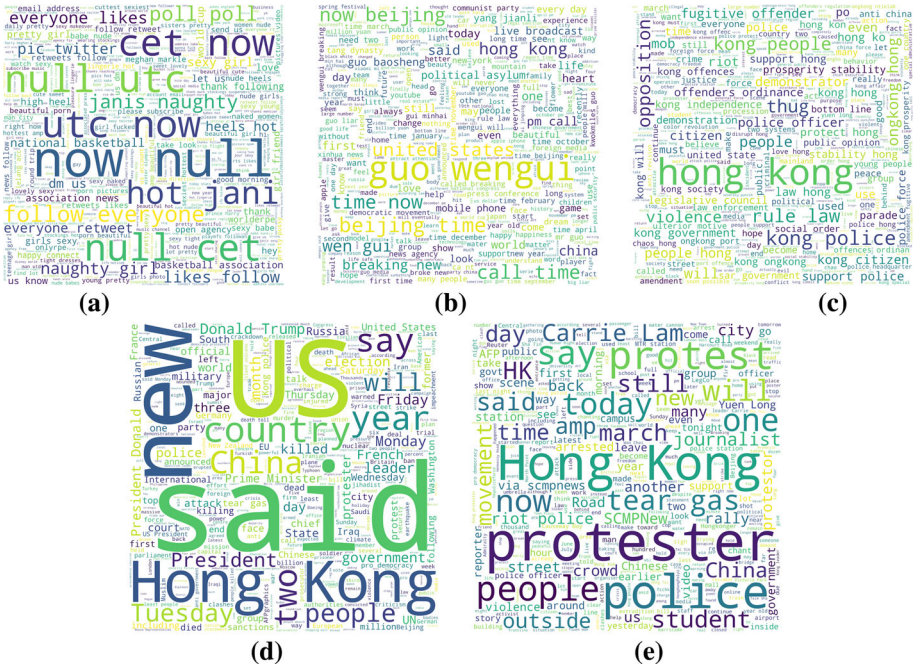

Fig. 1 Frequency word clouds formed from tweets: **a** In the fake news dataset worded in English. **b** In the fake news dataset worded in Chinese. **c** Resulting from the filtering process. **d** In the news agency dataset. **e** In the journalist dataset

Algorithm 1: The algorithm used to determine which tweets are relevant using a list of curated hashtags and a co-occurrence threshold.

Input : List of input tweets: *tweets*, curated hashtags: *curated*, co-occurrence threshold: *thr*
Output: Relevant tweets: *relevant*

begin
  `// Initialize hashtag co-occurrence count.`
  $counter \leftarrow dictionary()$;
  foreach t *in tweets* **do**
    if $t.hashtags \cap curated == \emptyset$ **then**
      continue;
    end
    foreach h *in* $t.hashtags$ **do**
      if h *in counter* **then**
        $counter[h] + +$;
      else
        $counter[h] \leftarrow 1$;
      end
    end
  end
  `// Keep only the hashtags that co-occur frequently enough`
  $frequent \leftarrow \{h \text{ for } h \text{ in } counter \text{ if } counter[h] > thr\}$;
  $relevant = list()$;
  foreach t *in tweets* **do**
    if $t.hashtags \cap frequent \neq \emptyset$ **then**
      $relevant.append(t)$;
    end
  end
end

Upon the completion of the filtering steps, only 3908 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. In comparison to the unfiltered set of tweets presented in the previous word clouds, it is undeniable that the tweets remaining after the filtering process are more relevant to the Hong Kong protests, with phrases such as "Hong Kong police" and "Hong Kong independence" appearing.

Alternative methods that could be employed instead of the presented algorithm belong to the field of topic modeling. The most commonly used approach is latent Dirichlet allocation (LDA), but it has been documented that this struggles with short sequences of text [16] and it generally uses text, not hashtags, making multilingual analysis even more difficult and time-consuming. Experimentation with LDA did not yield results that were as satisfactory as those of Algorithm 1. Multilingual variants of LDA exist (e.g., [3]), although these are not as prevalent.

3.2 Real news dataset

In order to perform binary classification, it is necessary to include a portion of the dataset labeled as real news. Since the datasets published by Twitter's Election Integrity Hub only included malicious accounts, different sources need to be identified for the real news. Other datasets, commonly used in the fake news classification domain, are not applicable in this case, as this study focuses on a very specific event, which is not addressed by any existing datasets. Thus, the approach employed to identify the real news portion of the dataset is described in the sequel.

Ideally, expert labeling is the most appropriate method to guarantee the correctness of labels. However, for a platform as large and diverse as Twitter, it is unlikely that the manual annotation of tweets by experts would be feasible for any reasonably popular topic. Therefore, since this study focuses on automatic classification of fake news, using a tweet's author to determine its validity is investigated as an alternative. It is not unreasonable to assume that validating the author can be considered a simpler task than validating each individual tweet. After all, an author can be identified as malicious by considering a wider variety of factors, such as suspicious behavior (rapidly posting, spam, etc.) or history (user reports, previously posting invalid content, etc.)

Additionally, existing systems, such as Twitter's "verified accounts"[4] can potentially be taken advantage of as well, acting as a form of user curation. This approach is also more in line with the fake news portion of the dataset, as it consists of tweets posted by accounts that were identified as malicious and were then deemed to be relevant to Hong Kong protests, specifically. Last but not least, combining datasets from different sources for each label is not uncommon in the fake news classification literature [2, 6, 8].

Considering the above and the nature of the event being investigated, news agencies seem like the first option to consider as reliable sources of information regarding the Hong Kong protests. Thus, the Twitter accounts of news agencies that are generally considered as trustworthy[5] are identified and their tweets are retrieved. However, the content that news agencies post is likely to be vastly different from the tweets of the malicious accounts present in the fake news part of the dataset. For instance, one would expect most of the news agencies' posts to contain references to news articles based on their websites, which would not be the case for the tweets of malicious accounts. This poses a risk for the classification of the tweets, as it may be based on the style rather than the content of the tweets.

In order to address this issue, journalists are also investigated as a trustworthy source for tweets. In particular, Twitter accounts of journalists employed by the previously identified trustworthy news agencies are also identified manually. The selected journalists have been selected for having written at least one article related to the Hong Kong protests and their official accounts have been identified through the agency they are affiliated with.

Overall, the accounts of 13 news agencies with a global outreach are identified, as opposed to the 107 accounts of journalists that are gathered. Using Twitter's user timeline API endpoint, 41,996 tweets from news agencies' accounts and 103,359 tweets from journalists' accounts are collected during the period of December 2019–February 2020. The selected agencies and the number of identified journalists employed by each agency are listed in Table 2, as well as the number of tweets collected by each agency and the corresponding set of journalists.

Having collected tweets from both news agencies and journalists, it is possible to compare a few of their characteristics to assess how similar they are. Hence, 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

[4] https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.

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

Table 2 The identified list of trustworthy news agencies and the number of journalists employed by each agency in the real news dataset, along with the number of tweets initially collected by each

News agency	Agency tweets	# Journalists	Journalist tweets
BBC News	3247	2	400
Reuters	3220	10	4412
Bloomberg	3250	4	3620
BuzzFeedNews	3247	9	17,942
China News Asia	3234	2	1596
CNN	3214	6	5384
Agence France-Presse	3214	16	14,685
South China Morning Post	3249	8	5209
Wall Street Journal	3217	24	16,032
New York Times	3240	9	11,048
The Associated Press	3214	3	3812
The Washington Post	3246	12	12,382
Quartz	3204	3	6837

0.82 when posted by a news agency. Lastly, on average, each of the accounts posting tweets has 4.1 followers in the fake news dataset, 15.14 in the journalist dataset, and 11,509.08 in the news agency dataset. 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.

Frequency word clouds are also derived from relevant tweets found in these two datasets and are shown in Fig. 1d, e. 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, 5388 and 666 of the tweets posted by journalists and news agencies, respectively, are considered relevant. The lower number of news agency tweets that remain after the filtering step is attributed to the lower number of hashtags present in their content, as previously mentioned in the comparison to journalist tweets. 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 3908 and 5388 tweets spreading fake and real news, respectively.

3.3 Features

From the selected datasets, three feature-sets are extracted so they can be fed into the classification models. It is worth noting that all feature-sets represent purely linguistic information, despite the fact that literature indicates that network-related features are worth investigating, as most of the information about the fake news dataset has been made unavailable by Twitter and is no longer accessible on the platform, with the accounts involved in the disclosed datasets having been banned. The feature extraction process for each feature-set is described in the sequel.

In the first feature-set, Feature-set 1, a feature engineering approach is followed, with the selected features being handcrafted. These features, which are thoroughly described in previous work [33], are purely linguistic in nature and they represent a single tweet. Since this feature-set does not contain any tweet text or account-specific information, it has been made publicly available on the website of the Ionian University's Humanistic and Social Informatics Laboratory[6] under the label "Tweets for Fake News Detection." The features add up to 38 in total, including the class label, are listed in Table 3 and their Pearson correlation heatmap is depicted in Fig. 2.

The features in Feature-set 1 span various categories, including morphological (e.g., part of speech), vocabulary (e.g., type-to-token ratio), semantic (e.g., text and emoji sentiment) and lexical features (e.g., number of pronouns). One feature, which is not commonly encountered in the literature, is tweet entropy, which is derived 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. Even at this early stage, the features of tweet entropy, tweet length and type to token ratio stand out, as they are highly correlated with the class label; thus, they are likely to be important for classification.

In Feature-set 2, on the other hand, the text of each tweet is tokenized and converted into word embeddings. Word embeddings are a prominent method of representing words as fixed-size numerical vectors, which are then fed as input to neural networks. In this case, each word is

[6] https://hilab.di.ionio.gr/index.php/en/datasets/.

Table 3 The features present in Feature-set 1, along with their IDs

Features	
1—Tweet length	20—Number of vowels
2—Number of tokens	21—Number of consonants
3—TTR	22—Number of uppercase chars
4—Number of URLs	23—Number of lowercase chars
5—Number of hashtags	24—Longest sequence of consecutive vowels
6—Number of punctuation marks	25—Longest sequence of consecutive consonants
7—Number of "?!"	26—Has repetition of >3 identical consecutive chars
8—Number of "?"	27—Number of digits
9—Number of "!"	28—Number of letters
10—Number of emojis	29—Ratio of letters to digits
11—Negative emoji sentiment sum	30—Number of pronouns
12—Neutral emoji sentiment sum	31—Number of determiners
13—Positive emoji sentiment sum	32—Number of nouns
14—Number of periods	33—Number of adverbs
15—Number of stopwords	34—Number of "to"
16—Number of words	35—Number of verbs
17—Average # of chars per word	36—Number of entities
18—Average # of chars per sentence	37—Tweet entropy
19—Overall text sentiment	38—Label

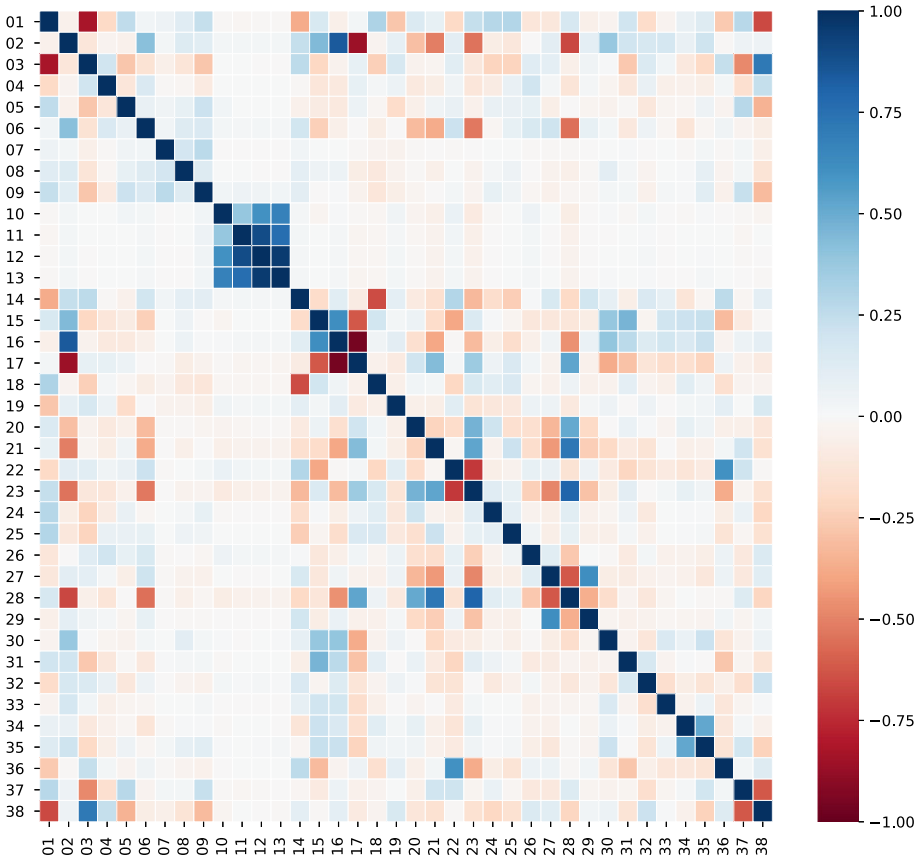

Fig. 2 Pearson correlation heatmap of the features present in Feature-set 1. The features can be identified through Table 3

mapped to an 100-dimensional vector, according to the pre-trained GloVe embeddings [24], which have been derived from a Twitter-based corpus, containing 27 billion tokens. Even though each word is mapped to a fixed-size vector, tweets still vary in word count; to counteract this, tweets are post-padded (i.e., padding occurs at the end of a tweet) to match the longest tweet, which in this case is 118 tokens long. As such, a tweet in Feature-set 2 is represented by an 118×100 matrix.

Last but not least, Feature-set 3, rather than constructing an explicit representation of tweets, the current work relies on the representational capabilities of state-of-the-art transformer-based models, such as BERT [11]. These types of models attempt to build a pre-trained language representation model from unlabeled text, that can be customized with additional neural network layers to achieve exceptional results in a wide array of tasks. Interestingly enough, XLM-RoBERTa [9], a recent model that builds on the success of BERT, has been pre-trained on over 100 languages, making it adept at cross-lingual tasks. Utilizing this model, it is possible to perform fake news classification utilizing the original text, before it was even translated. As such, Feature-set 3 contains the tweets remaining from the filtering and preprocessing previously described in Sect. 3.1.

3.4 Algorithms

In the sequel, the ML algorithms, feature preprocessing and selection methods are considered. Regarding more traditional algorithms, the 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 [5] of C4.5. All of these algorithms operate over Feature-set 1, containing the set of handcrafted features. The rather popular Scikit-Learn Python module [23] 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 hyperparameter tuning technique.

In this work, further experimentation takes place with deep neural network architectures. In particular, three different architectures have been selected as they are quite prominent in the literature when it comes to text classification [18]. These architectures operate over Feature-set 2, containing word embeddings. They have been implemented using Tensorflow [1] and are graphically depicted in Fig. 3. They all share the same loss function, that being binary cross-entropy, as well as the output layer, which consists of a single neuron with a sigmoid *activation function*. Different optimizers and associated parameters, including learning rate and the number of epochs, are experimented with for each architecture. Furthermore, the *Early Stopping* method is used to terminate training if evaluation accuracy does not increase significantly for more than 10 epochs, which helps avoid overfitting.

Convolutional neural network (CNN): the first layer is the convolution layer which is initialized as a 1-dimension convolution layer with a *filter size* equal to 128 and a *kernel size* equal to 3. Next, a max pooling layer with a *pooling size* of 2, followed by a flattening layer are included. A dropout layer is added right before the final dense output layer with a dropout rate value of 0.8.

Bidirectional Long Short-Term Memory (BiLSTM): This architecture is a sequential model comprising two BiLSTM layers with 32 and 16 units. After the BiLSTM layers, a dense layer of 64 units with the rectified linear unit (RELU) as the activation function is used. Next, the model consists of a dropout layer with dropout rate equal to 0.2 and then fed to the output dense layer with one unit using the sigmoid activation function.

Convolutional LSTM (C-LSTM): The first layer included in the C-LSTM architecture is the convolution layer initialized with a filter size of 128, a kernel size equal to 3 and a RELU activation function. The subsequent layer is max pooling with a pool size of 2. An LSTM layer is added next with 100 units and a dropout rate equal to 0.2.

Finally, the XLM-RoBERTa model is utilized, which is then fine-tuned with the help of the Transformers [31] and Simple Transformers[7] modules for Python. Through these modules, the pre-trained transformer model is adjusted so it is suitable for binary classification and then fine-tuned for a single epoch over Feature-set 3, containing raw English and Chinese text.

4 Results

In this section, the predictive performance of the ML algorithms is evaluated, detailing the hyperparameter selection process to avoid overfitting and evaluating the results acquired for the various feature-sets, while comparing them with results found in related work.

4.1 Algorithm evaluation

The ML algorithms utilized for the classification of tweets spreading fake and real news are trained using the extracted feature-sets of the collected datasets, and the corresponding evaluation results are presented. The dataset consists of 3,910 and 5388 tweets spreading fake and real news,

[7] https://simpletransformers.ai/.

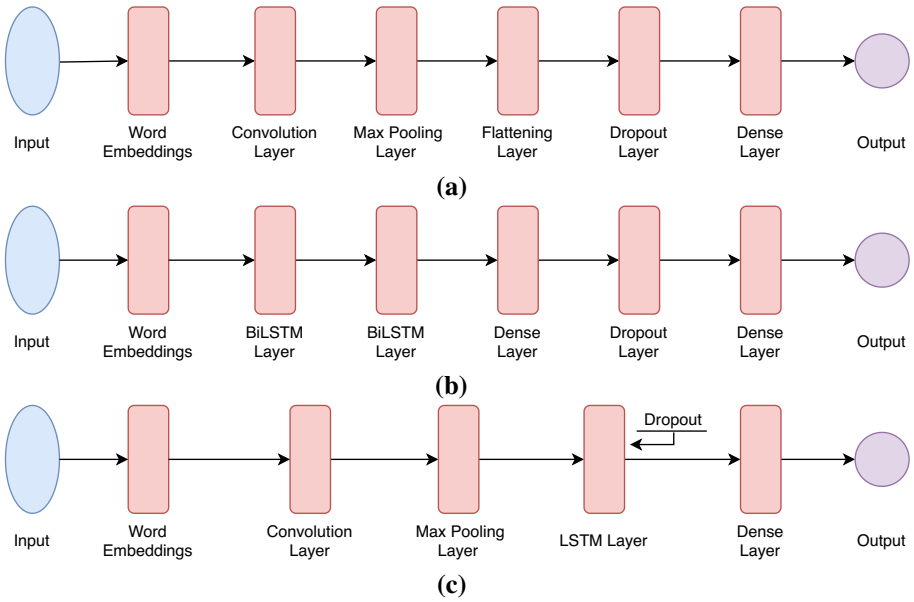

Fig. 3 Deep neural network architectures trainer over Feature-set 2: **a** CNN. **b** BiLSTM. **c** C-LSTM

respectively, with the majority class baseline being 57.9%. The evaluation metrics presented include macro-average and per-class precision, recall and F1 Score. For all ML algorithms, fivefold cross-validation is used to increase reliability of results. The evaluation results are summarized in Table 4, with the highest performing algorithm per feature-set listed in bold.

For Feature-set 1, the highest performing algorithm is random forest, achieving an average F1 Score of 92.1%. The rest of the traditional ML algorithms perform similarly, with the lowest F1 score being 89.4% achieved by Naive Bayes. These algorithms are the least computationally intensive, while also processing the simplest of the three feature-sets, yet they achieve adequate performance. They also have the benefit of providing interpreting results, which are further analyzed in a later subsection.

In contrast, all of the deep neural networks using Feature-set 2 manage to achieve higher scores than the traditional ML algorithms across the board, with BiLSTM achieving 96.9% average F1 Score. All three networks yield very similar performance, so no one architecture particularly stands out. However, they do differ in their training speeds, with C-LSTM taking 112.8 seconds on average per fold, CNNs taking 138.4 seconds and BiLSTMs taking 203.6 seconds. These training times are still vastly larger than the traditional algorithms, although the neural networks do provide a notable increase in predictive performance.

Last but not least, the transformer-based model, XLM-RoBERTa, outperforms all other models, achieving 99.3% average F1 Score. This is a very significant increase compared to the rest of the deep learning models, further increasing the difference from the traditional ML algorithms. The training times are also comparable to the deep learning architectures, even though XLM-RoBERTa is a much larger model. This is attributed to the fact that the model comes pre-trained and merely needs to be adjusted to the present task. As such, the integration of this state-of-the-art transformer-based model seems to significantly increase performance in discriminating between the tweets spreading fake and real news that make up the collected datasets.

One issue worth paying attention to is that the recall scores of the fake class are generally lower than those of the real class. This is arguably suboptimal, since the impact of a tweet spreading fake news could be considered significant. Thus, one could argue that models achieving higher recall scores would be preferred, even if precision scores suffered somewhat. However, the classification scores are still very high, so this should not be a significant concern.

4.2 Feature-set evaluation

Comparing across the different feature-sets, it seems clear from the evaluation results that feature learning tends to supersede feature engineering regardless of the architecture on the collected datasets, as ML algorithms using Feature-sets 2 and 3 outperform those using Feature-set 1. This represents an interesting trade-off between computation costs and engineering costs, as handcrafted features generally require domain-specific knowledge, whereas learned features are computationally derived from the data.

Another important observation that can be made concerns Feature-set 3, which contains raw untranslated text. A potential issue with fake news classification on

Table 4 Evaluation results of the ML algorithms used in the classification process

Feature-set	Algorithm	Class	Precision (%)	Recall (%)	F1 Score (%)
Feature-set 1	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**
Feature-set 2	CNN	Fake	97.5	93.8	95.6
		Real	95.6	98.3	96.9
		Average	96.5	96.0	96.3
	BiLSTM	Fake	97.2	95.5	96.3
		Real	96.8	98.0	97.4
		Average	**97.0**	**96.8**	**96.9**
	C-LSTM	Fake	96.8	95.6	96.1
		Real	96.9	97.6	97.2
		Average	96.8	96.6	96.7
Feature-set 3	**XLM-RoBERTa**	Fake	99.9	98.4	99.2
		Real	98.9	99.9	99.4
		Average	**99.4**	**99.2**	**99.3**

platforms, such as Twitter, is that extracting the required features is generally not addressed in a multilingual setting, whether it be handcrafted features or word embeddings. This makes it difficult to tackle classification using a single model, unless translation is made use of, which is generally an expensive process and can introduce other issues in the data that may complicate matters further. This can be observed in the collected dataset: in one instance, 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. This is particularly hard to address when using handcrafted features, such as those included in Feature-set 1. The fact that XLM-RoBERTa performed the best provides further confidence that such issues have not significantly skewed the results of the other algorithms.

On one hand, feature engineering allows for the interpretation of the acquired results. In this case, the top ten most significant features selected according to the mutual information metric are listed in Table 5. According to this ranking and utilizing some of the rule-generating models trained over Feature-set 1, a profile can be formed for tweets spreading fake news. These tweets are longer in length containing shorter sentences, they contain fewer punctuation marks and they use uncommon vocabulary, resulting in higher entropy. It is worth noting, though, that the difference in length may be caused by the translation from Chinese to English.

Table 5 The top ten most significant features according to the mutual information metric

Top ten features	
Tweet length	Type-to-token ratio
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	Tweet entropy

Nonetheless, it is undeniable that learned features can significantly increase performance. Currently, multilingual models, such as XLM-RoBERTa boast exceptional representational capabilities. Moreover, they can be made available pre-trained and then fine-tuned to accommodate different tasks, without requiring retraining from scratch. These benefits do come at a cost though, as such large models also require care when training to not overfit and their large size leads to increased inference times, making

online deployment harder. Overall, there is an interesting trade-off between predictive performance, computational complexity and interpretability between the various feature-sets and the corresponding algorithms.

4.3 Hyperparameter selection

Particular attention needs to be paid to overfitting in the case of deep neural networks trained over Feature-set 2. To address this, their training and validation performance is monitored over the course of their training, which is collectively presented across different optimizers in Figs. 4, 5 and 6, averaged over the 5 folds for the maximum number of epochs required. Do note when comparing across the different subfigures that the value range of the y-axis is different, so that the fluctuations are more clearly visible. Moreover, the results obtained from the experimentation with different optimizers and learning parameters are detailed.

Various experiments were conducted using the Adam, AdaGrad and Adadelta optimizers with learning rates

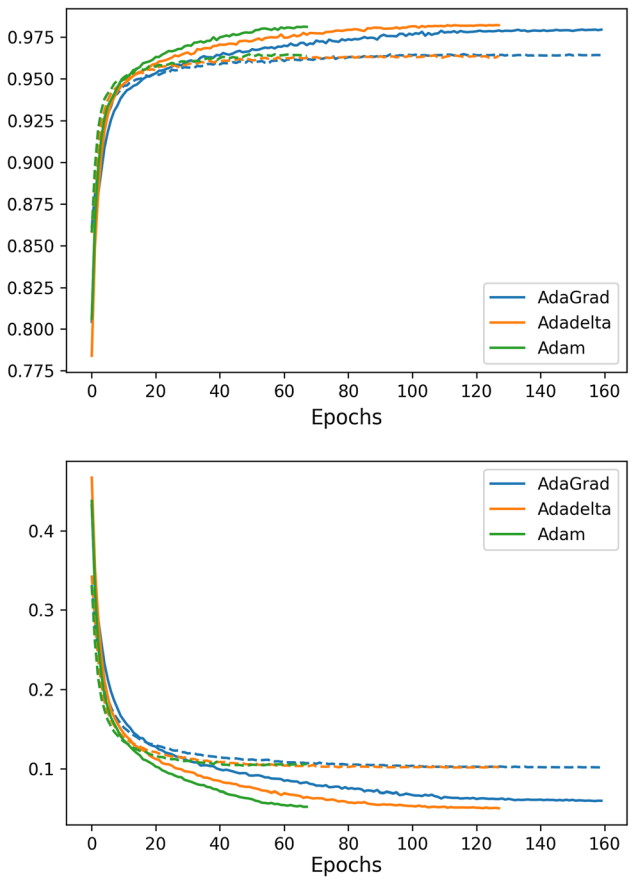

Fig. 4 The accuracy and loss during the training and validation of the CNN architecture for various optimizers. The dashed lines represent the validation scores, whereas the solid lines represent the training scores

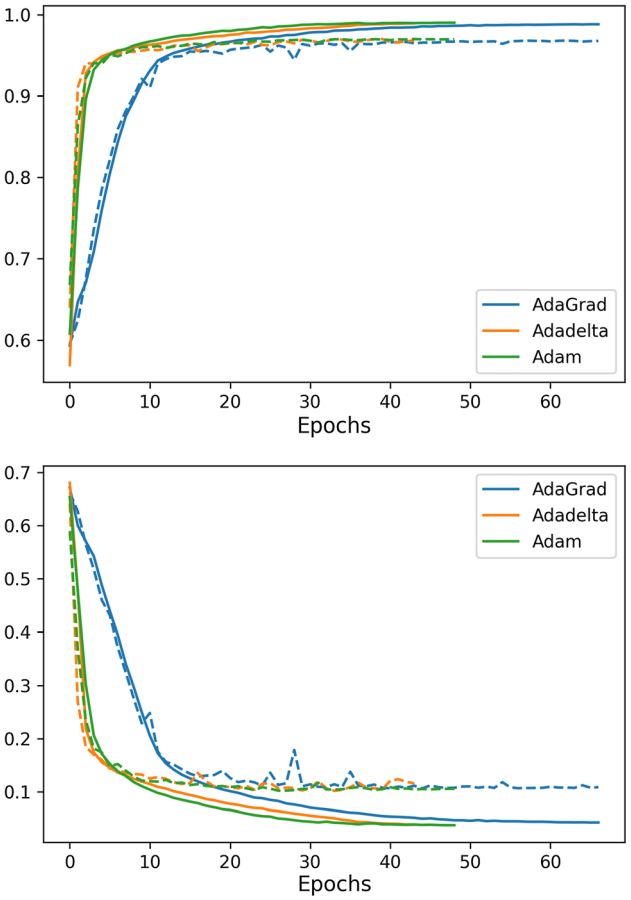

Fig. 5 The accuracy and loss during the training and validation of the BiLSTM architecture for various optimizers. The dashed lines represent the validation scores, whereas the solid lines represent the training scores

ranging from 0.1 to 0.0001. The learning rate presented in the previous Figures is set to 0.0001 for Adam, 0.1 for Adadelta and 0.01 for AdaGrad. Each architecture is trained for a different number of epochs, with a maximum of 200, until no significant performance improvement is observed and training is terminated by the Early Stopping method. Once all optimizers have converged, their performance is almost equivalent, so training time becomes an important metric. For all architectures, the Adam optimizer yielded the most consistent results while requiring fewer epochs, so it is the one being used for all of the presented results in the previous subsection. Considering that the Early Stopping method is utilized and the fact that the validation accuracy does not consistently drop during the latter training epochs, no obvious overfitting seems to be taking place. In this regard, do remember that all presented results have been averaged over 5 folds, so it is even less likely that overfitting has occurred.

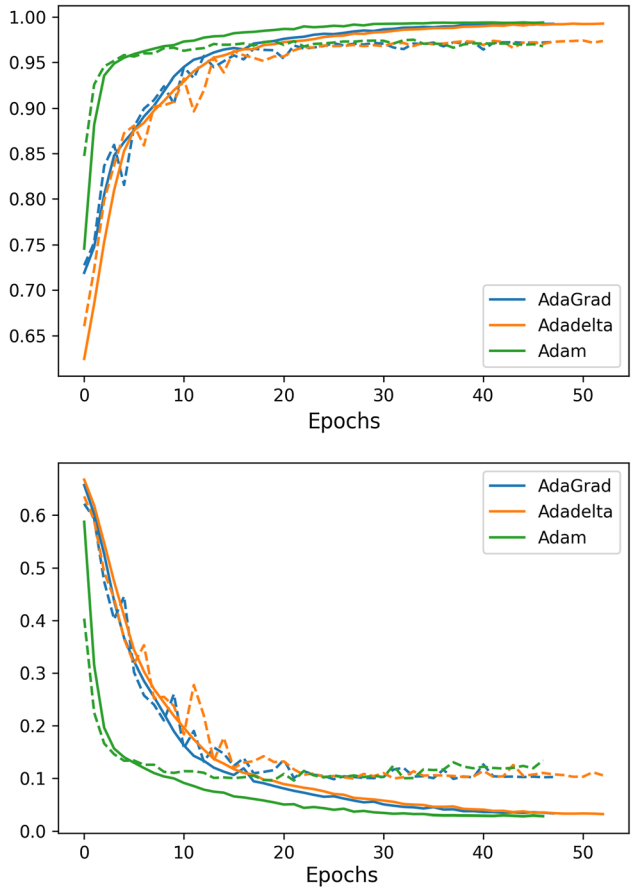

Fig. 6 The accuracy and loss during the training and validation of the C-LSTM architecture for various optimizers. The dashed lines represent the validation scores, whereas the solid lines represent the training scores

4.4 Comparison to related work

While the results of the presented approach are quite high, they are comparable to those of other approaches found for related tasks in the literature. Do note that direct comparisons cannot be made, as different datasets are being utilized. Furthermore, a review of the literature indicates that fake news classification tasks are diverse in the form of text they classify, with many focusing on articles, rather than tweets. Even among the approaches utilizing tweets, very few are those that focus on purely linguistic content, as they usually make use of other features, often extracted from images or network-related interactions, to perform classification. After all, tweets can only convey a very limited amount of linguistic information, due to the enforced character constraints, making it a more challenging issue for approaches that rely on purely linguistic information. A brief comparison to related works is made in the sequel, focusing on both quantitative and qualitative differences.

Nikiforos et al. [20] also focus on fake news detection on tweets related to the Hong Kong protests, using similar data collection methods, but different tweet filtering criteria, discarding any non-English ones, and feature selection methodologies, as they also include network and account-related features. As a consequence, their dataset is smaller and more unevenly distributed, which is balanced out using SMOTE oversampling [7], with their results yielding 99.8% accuracy. Tarek et al. [14] study fake news classification exclusively through user-related features and network interactions, with their experiments achieving as high as 98% F1 score using graph embeddings. In [13], the authors focus on extracting language-independent features and compare the performance of traditional ML algorithms over various feature representations and multilingual datasets of both articles and tweets, achieving F1 Scores ranging from 76 to 95% depending on the dataset. Singhal et al. [27] develop a multi-modal framework that utilizes BERT-like language models to learn text features in conjunction with a pre-trained model for image feature extraction to classify fake news, achieving accuracy scores ranging from 77.77 to 89.23% depending on the dataset.

Much work has also been conducted in fake news classification on articles, which shares some similarities as the primary focus is on text, even though it is of a different form. Kaliyar et al. [17] develop an intricate CNN architecture trained over the Kaggle dataset focusing on fake news detection for articles, achieving 98.12% F1 Score. Yang et al. [32] also focus on fake news classification related to articles, showcasing the importance of image information in classification, utilizing a unified model capable of analyzing both text and image information, reaching as high as 92.1% F1 Score. Fang et al. [12] scored a 95.5% precision, utilizing a CNN and a self multi-head attention mechanism and dealing with whole articles again, rather than tweets. Moreover, Khan et al. [18] scored as high as 95% F1 Score in various datasets, one of which was assembled by them, using some news agencies that have been used in the present work, but focusing on agencies' articles and not on tweets.

5 Conclusions

The tendency of people to use Internet technologies and, in particular, social media is undeniable. Their increasing popularity and the ever-evolving services and capabilities they offer have reinforced users' preference for information from posts on social media. This new reality leads to a weakening of traditional measures and practices for the validity of news, which contributes to the spread of fake news. The phenomenon of fake news is heavily observed

and has even concerned social media providers themselves, who advise users on ways to evaluate the news veracity.[8]

This paper focused on comparing different kinds of features and ML algorithms for detecting fake news in tweets, aiming at identifying patterns in both the linguistic content and structure of tweets. In order to accomplish this, a previously constructed dataset was utilized, which was assembled using a custom filtering process that is based on hashtag co-occurrence. The methodology followed for acquiring, preprocessing and filtering the dataset can serve as a generic but scalable process that could aid in future research to assemble similar datasets. This dataset is used to evaluate the performance of various modern deep learning architectures that are commonly encountered in text classification tasks, including CNNs and LSTM networks. These algorithms are trained over diverse feature-sets, with vastly different forms of text representation, ranging from raw text to handcrafted features. The results seem to be very promising, achieving as high as 99.3% F1 Score, significantly outperforming the more traditional approaches used in previous work, utilizing the state-of-the-art XLM-RoBERTa model. The other deep learning architectures performed admirably well, too, with their F1 Scores ranging from 96.3 to 96.9%.

Perhaps, the most interesting aspect of the presented experiments concerns the different feature-sets and the trade-offs they each present. Handcrafted features and traditional ML algorithms perform reasonably well, but they offer the benefit of interpretability, providing insight into the patterns and rules that can help us distinguish between fake and real news, while also allowing for validation by domain experts to prevent mispredictions and overfitting. On the other extreme, transformer-based models are trained over raw text, achieving the best results, while requiring the least amount of engineering time and domain-specific knowledge, but the most computational resources. More conventional deep learning architectures, such as LSTMs, strike an interesting balance between computational resources, engineering time and performance.

Even though traditional approaches may be somewhat lacking in performance, it is important to mention a few additional concerns regarding automated systems operating in the social realm. When it comes to the automatic identification of fake news, it is not as easy to accurately pinpoint which approach to employ as picking the most highly performing algorithm. As the political and social landscape is increasingly being shaped by conversations taking place online and the spread of fake news presents a greater risk of undermining long-held values, a case can be made that understanding their structure is more relevant. In this regard, feature engineering can shed light by confirming suspicions about the tactics used to spread fake news, while also revealing patterns that are not intuitively obvious; in this case, the entropy of a tweet turned out to be one of the most important features.

Interpretability and explainability of automated decisions are an equally important factor in a world, where it can be said that privately held corporations have control over political dialogue and free speech. Thus, power over user suspensions in social media entails the abstinence from such topics, which, in principle, should be open to debate to the public. The consequences of such power over public affairs have likely not been made apparent yet, but the permanent suspension of USA's former president from Twitter in 2020[9] has sparked debates over free speech violations and the need for social-media regulations. Thus, having the option of understanding and explaining why an automated decision, such as suspending a user due to detecting the spread of fake news, is made could prove important in the future, especially if there is the possibility that it is to be questioned in the court of law.

In summary, it is undeniable that the dissemination of fake news has become an important social issue, arguably influencing the outcome of political events through shaping the public's opinion. Various studies, including this one, have shown that automated tools and ML possess the capability of aiding in the early identification and stop of the spread of fake news. Fortunately, these techniques have advanced to the point that this can be accomplished using multiple approaches, each with its own set of pros and cons. Understanding and showcasing these approaches in practice over real data are an essential step toward their informed application in a social setting.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. (2016) Tensorflow: a system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI}16), pp 265–283
2. Ahmed H, Traore I, Saad S (2018) Detecting opinion spams and fake news using text classification. Secur Priv 1(1):e9

[8] https://www.facebook.com/help/188118808357379.

[9] https://blog.twitter.com/en_us/topics/company/2020/suspension.html.

3. Amara A, Taieb MAH, Aouicha MB (2021) Multilingual topic modeling for tracking covid-19 trends based on facebook data analysis. Appl Intell 1–22
4. Bajaj S (2017) The pope has a new baby! fake news detection using deep learning. Tech. rep., Technical Report, Stanford Univ
5. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
6. Cao J, Sheng Q, Qi P, Zhong L, Wang Y, Zhang X (2019) False news detection on social media. arXiv preprint arXiv:190810818
7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
8. Chu SKW, Xie R, Wang Y (2020) Cross-language fake news detection. Data Inf Manag 5(1):100–109
9. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2019) Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:191102116
10. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116
11. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805
12. Fang Y, Gao J, Huang C, Peng H, Wu R (2019) Self multi-head attention-based convolutional neural networks for fake news detection. PLoS ONE 14(9):1–13. https://doi.org/10.1371/journal.pone.0222713
13. Faustini PHA, Covões TF (2020) Fake news detection in multiple platforms and languages. Expert Syst Appl 158:113503
14. Hamdi T, Slimi H, Bounhas I, Slimani Y (2020) A hybrid approach for fake news detection in twitter based on user features and graph embedding. In: International conference on distributed computing and internet technology. Springer, Berlin, pp 266–280
15. Helmstetter S, Paulheim H (2018) 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
16. Jónsson E, Stolee J (2015) An evaluation of topic modelling techniques for twitter. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (short papers), pp 489–494
17. Kaliyar RK, Goswami A, Narang P, Sinha S (2020) Fndnet—a deep convolutional neural network for fake news detection. Cogn Syst Res 61:32–44. https://doi.org/10.1016/j.cogsys.2019.12.005
18. Khan JY, Khondaker MTI, Iqbal A, Afroz S (2019) A benchmark study on machine learning methods for fake news detection. arXiv:1905.04749
19. Long Y, Lu Q, Xiang R, Li M, Huang CR (2017) Fake news detection through multi-perspective speaker profiles. In: Proceedings of the eighth international joint conference on natural language processing (volume 2: short papers), pp 252–256
20. Nikiforos MN, Vergis S, Stylidou A, Augoustis N, Kermanidis KL, Maragoudakis M (2020) Fake news detection regarding the Hong Kong events from tweets. In: IFIP international conference on artificial intelligence applications and innovations. Springer, Berlin, pp 177–186
21. Oshikawa R, Qian J, Wang WY (2018) A survey on natural language processing for fake news detection. arXiv preprint arXiv:181100770
22. Parmelee JH, Bichard SL (2011) Politics and the Twitter revolution: how tweets influence the relationship between political leaders and the public. Lexington Books
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 2:2825–2830
24. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Empirical methods in natural language processing (EMNLP), pp 1532–1543
25. Purbrick M (2019) A report of the 2019 Hong Kong protests. Asian Aff 50(4):465–487
26. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. SIGKDD Explor Newsl 19(1):22–36
27. Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Satoh S (2019) Spotfake: a multi-modal framework for fake news detection. In: 2019 IEEE Fifth international conference on multimedia big data (BigMM). IEEE, pp 39–47
28. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151. https://science.sciencemag.org/content/359/6380/1146.full.pdf
29. Wang WY (2017) "liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv preprint arXiv:170500648
30. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) Eann: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. ACM, pp 849–857
31. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M et al (2019) Huggingface's transformers: state-of-the-art natural language processing. ArXiv pp arXiv–1910
32. Yang Y, Zheng L, Zhang J, Cui Q, Li Z, Yu PS (2018) TI-CNN: Convolutional neural networks for fake news detection. arXiv:1806.00749
33. Zervopoulos A, Alvanou AG, Bezas K, Papamichail A, Maragoudakis M, Kermanidis K (2020) 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, pp 408–419
34. Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv (CSUR) 53(5):1–40