Multimedia Tools and Applications (2022) 81:33801–33821
https://doi.org/10.1007/s11042-022-12428-8

Fake news detection on social media using a natural language inference approach

Fariba Sadeghi[1] · **Amir Jalaly Bidgoly**[1] · **Hossein Amirkhani**[1]

Received: 19 February 2021 / Revised: 5 August 2021 / Accepted: 25 January 2022 /
Published online: 21 April 2022

Abstract

Fake news detection is a challenging problem in online social media, with considerable social and political impacts. Several methods have already been proposed for the automatic detection of fake news, which are often based on the statistical features of the content or context of news. In this paper, we propose a novel fake news detection method based on Natural Language Inference (NLI) approach. Instead of using only statistical features of the content or context of the news, the proposed method exploits a human-like approach, which is based on inferring veracity using a set of reliable news. In this method, the related and similar news published in reputable news sources are used as auxiliary knowledge to infer the veracity of a given news item. We also collect and publish the first inference-based fake news detection dataset, called FNID, in two formats: the two-class version (FNID-FakeNewsNet) and the six-class version (FNID-LIAR). We use the NLI approach to boost several classical and deep machine learning models, including Decision Tree, Naïve Bayes, Random Forest, Logistic Regression, k-Nearest Neighbors, Support Vector Machine, BiGRU, and BiLSTM along with different word embedding methods including Word2vec, GloVe, fastText, and BERT. The experiments show that the proposed method achieves 85.58% and 41.31% accuracies in the FNID-FakeNewsNet and FNID-LIAR datasets, respectively, which are 10.44% and 13.19% respective absolute improvements.

Keywords Fake news detection · Natural language inference · Social media · Content features

✉ Amir Jalaly Bidgoly
jalaly@qom.ac.ir

Fariba Sadeghi
f.sadeghi@stu.qom.ac.ir

Hossein Amirkhani
amirkhani@qom.ac.ir

[1] University of Qom, Qom, 3716146611, Islamic Republic of Iran

1 Introduction

News and information is the tool and the basis of society's awareness and actions. Traditionally, news agencies have been the source of news. However, the rapid growth and attractiveness of online social media such as online social networks, messengers, and blogs have led to a significant amount of news being broadcast and disseminated through these platforms today. These Internet platforms are currently the most popular media in the world, so that even ordinary people have the opportunity to monitor the latest information and observations of each other at any time and communicate with each other. Every day a considerable amount of political, social, economic, health, art, information technology, or other news is produced [63]. Social media allows the audience to follow the news in their favorite areas instantly and republish it in the media as soon as they see an interesting one. That is why the present decade has been called the *information age*. Every person in society is consciously or unconsciously involved in the production and dissemination of news and information, and the news is published more quickly than ever before.

Fast publishing is one side of the story. On the other side, the publication of unconfirmed and unprofessional news by individuals may intentionally or accidentally contain false information. The news in these media, unlike traditional media, is published without supervision and verification,so recognizing this news's correctness has become a challenge in online social media. This misinformation may have been inadvertently propagated. Some individuals and organizations may deliberately spread fake news in the media for purposes such as profiteering, unhealthy competition, or even entertainment. Fake news is usually more interesting than real ones; hence they will be shared and spread more quickly throughout society [62]. They may cause irreparable damage to individuals, organizations, and governments, which can have devastating effects, such as increased social anxiety, reduced productivity, and crippling of the economic cycle. News experts and volunteer individuals are trying to reduce the destructive effects of fake news by identifying and reporting them. Websites such as PolitiFact[1], Snopes[2], and FactCheck[3] are well-known examples in this field that identify and publish fake news daily in various fields. The identification mechanism in these websites is manually based on individual reports or approaches such as *crowdsensing* [40]. However, this mechanism is not suitable for the high volume of fake news published on online social media. Therefore, to detect fake news and deal with its excessive publishing, there is a desperate need to automate this process.

Various methods have already been proposed to identify fake news. The main approach in these methods is to use machine learning. In the mainstream of this work, having a labeled data set of correct and fake news, a classification model is trained on news features and then used to predict a news item's correctness. The features used in these methods may fall into two categories: 1) content-based features, and 2) context-based features. Content-based features refer to those features that are extracted from the text or the content of the news itself [1, 14, 46]. In contrast, context-based features are based on news context such as the publisher, the stance of other individuals in the network, and propagation structure to indicate whether the news is fake or not. These methods have been able to achieve good results [52, 65], but they often need information that is hard to gather in the moment of receiving a fake news item. They only work when fake news has affected the community.

[1] www.politifact.com
[2] www.snopes.com
[3] www.factcheck.org

For example, stance detection in news comments, which is one important method in fake news detection, is only applicable when the network users take a stance against news and write their idea about it [41]. In fact, these methods exploit the knowledge of the other users in the network, which means that they have to wait for at least a part of the network members to investigate the correctness of a news item.

In the the previous works, all has been looking for some patterns to detect fake news. The features used in these works were all based on the context or content of the news. In this paper, we propose a novel method for fake news detection based on *Natural Language Inference (NLI)* approach. The main idea is to imitate the way news experts follow to detect fake news. If the new news contradicts the confirmed news, it is labeled fake. However, if it corresponds to the confirmed news, it is labeled real. This method is innovative in two ways. The first is that a data source outside the content and context of the news has been used, and the second is that inference approach has been used for the first time for fake news detection. In the NLI task, which is one of the most important subfields of natural language processing (NLP), a received claim (hypothesis) is classified in one of the classes true (entailment), false (contradiction), or undetermined (neutral) based on initial knowledge (premise). The approach we used in this study is also similar. Considering the existing confirmed news as a *"premise"*, we infer the new news as a *"hypothesis"* and predict whether it is fake or real. The most important effect of this method in the process of detecting fake news is that we can check the a piece of received news through the previously confirmed news, even in the first moments of publishing, and determine whether it is true or false. This allows us to prevent the spread of fake news in the community and its destructive effects very quickly. Another advantage of this method is the automation of the news review and analysis process, which eliminates the manual process of this process, reduces costs and speeds up the news review process. We use this approach to boost a couple of classical and deep models including Decision Tree [47], Naïve Bayes [26], Random Forest [7], Logistic Regression [15], k-Nearest Neighbors [29], Support Vector Machine [45], BiGRU [12], and BiLSTM [21] along with different word embedding methods including Word2vec [38], GloVe [44], FastText [5], and BERT [13]. The results show considerable improvements in the accuracy of fake news detection using the auxiliary knowledge based on the NLI approach. We also introduce a new NLI-based dataset according to the *FakeNewsNet (Politifact)* [50] and *LIAR* [57] datasets, which has been made freely available [4]. In summary the contribution of the paper is as follows:

- Propose an Natural Language Inference approach for fake news detection for the first time.
- Utilize previously verified news as a source outside the social network to identify fake news.
- Outperform state of the are methods and reached a maximum accuracy of 90.19% and 39.65% in the two-classes and five-classes fake news classification respectively.
- Introduce a new dataset named *FNID* by extending LIAR and FakeNewsNet (Politifact) datasets.

The paper continues as follows. In the next section, related research and datasets on fake news detection are discussed. Section 3 reviews the NLI task and its methods. The proposed method and the collected dataset are described in Sections 4 and 5, respectively.

[4] https://ieee-dataport.org/open-access/fnid-fake-news-inference-dataset

The experimental results are presented and discussed in Sections 6; and finally, the paper concludes in Section 7.

2 Related work

Many research papers have been published in the field of fake news and rumors, which can be divided into four categories: *fake news detection* (e.g. [27, 35]), *fake news spreaders detection* (e.g. [3, 39]), *fake news propagation modelling* (e.g. [4, 64]), and *fake news mitigation* (e.g. [16, 19, 31]). Although these works are related and all of them are common in the field of fake news, but they are different in terms of methods and goals. For instance, in identifying a rumor spreader, the goal is to identify the person who spreads the rumor. As an example work, Bakhteev et al. [3] proposed a ensemble method which take the whole set of published tweets by a user to decide whether the user can spread fake news. Identifying the fake news spreader may later be used as a feature to identify the fake news, but the main purpose of this method is not to identify the fake news itself. In this study, we have just focused on fake news detecting. The state of the art methods in this category are mainly based on deep learning methods, in which fake news detection has been seen as a binary classification problem (e.g. *Real* & *Fake* classes) or multi-class problem (e.g. *False*, *Half-True* & *True* classes). In this section, considering the importance of datasets in the machine learning methods the most important works and available fake news datasets will review. The available machine learning-based methods in fake news detection use either the *content-based* or *context-based* features or both of them.

Content-based features these features are extracted from the textual or visual content of news items or social media messages. These features may include lexical, textual, syntactic, semantic, visual, emotional, or link ones. For example, one study introduced a method called *Event Adversarial Neural Network (EANN)* that extracts features from multi-modal data and used both textual and visual features to detect fake news [58]. In another work, used sentiment analysis in twitter posts for rumor and fake news detection [1], or in the other work used combined stylometric features with word vector representations to predict fake news [46]. In a study used a BERT-based deep learning approach by combining different parallel deep Convolutional Neural Networks for fake news detection [27].In another study, labeled and unlabeled data were used to detect fake news. In this study, a model based on self-learning semi-supervised deep learning network is proposed for fake news detection [35]. In another study, researchers first extracted important features from fake news datasets, then classified the news using the ensemble learning method. They achieved high accuracy in fake news detection [20].

Context-based features these features are mainly based on social communication and interaction in the network. They may include the users' profile, the news propagation network features, or spreading structure. For example, in a research used the propagation network between news publishers and subscribers based on the assumption that fake news have a different propagation pattern than other types of news [65]. User profiles are also used in fake news detection methods. In one study, the ability to detect fake news increased by separating fake news publishers from other publishers [52].

Several studies have used both types of these features. For example, one study has used a threshold on the number of user interactions in a post to decide which type of feature should be used. Content features are used if the number of interactions is less than the threshold,

while context features are used if the number of user interactions exceeds the threshold [11]. In another study, researchers proposed a new method using both publishing and friendship networks and combined them with content features to more accurately detect fake news [25]. Table 1 shows a summary of the reviewed research, based on the type of features used.

From another point of view, the lack of sufficient labeled data in supervised learning is an important challenge. To solve this problem, some researchers propose methods other than supervised learning. For instance, in one study presented a *semi-supervised* method with a two-path deep model, one path for supervised learning to learn from a limited labeled dataset and another for unsupervised learning to learn from an abundant amount of unlabeled [14]. Despite some efforts in this line, most of the proposed methods in this field are still classification-based.

To enable the supervised learning, there are several famous datasets in this field which are reviewed in the following. Vlachos and Riedel published a dataset in 2014 from *Politifact* and *Channel4* websites; this dataset is a collection of 221 samples that are labeled in five classes: *true*, *mostly true*, *half true*, *mostly false*, and *false* [59]. In 2016, *BuzzFeedNews* dataset collected and published by a group of journalists of the BuzzFeed website. The dataset includes 2,282 news items published on Facebook which are classified into four classes: *mostly true*, *mixture*, *mostly false*, and *no factual content* [53].

In 2017, Horne and Adali introduced three new datasets of *satire*, *fake*, and *real news* articles from different political and non-political news sources. The datasets include 120, 225, and 4233 labeled samples in two, three, and four classes, respectively [23]. In another study in the same year, published a dataset called *LIAR* which includes 12,800 statements and related metadata. Statements in this dataset are labeled in six classes: *pants-fire*, *false*, *barely true*, *half true*, *mostly true*, and *true*, collected from the Politifact website [57]. In 2018, *Fake News vs. Satire* dataset was introduced in which 486 political news items have been collected [17]. In the same year, *FakeNewsNet* dataset was introduced to conduct fake news detection research through the analysis of news texts and social networks. In this dataset, 1,056 and 22,856 samples are collected from *Politifact* and *Gossip Cop* websites, respectively. These samples are labeled in two classes fake and true [50].

The features used in most works on fake news detection are based on the content of fake news and the profile of the spreader. These features often indicate some statistical patterns that are more common in fake news. These patterns may change over time or in different datasets, so the output obtained usually only works well in the training dataset. On the other

Table 1 A summary of some previous research in the field of fake news detection

Refrence	Year	Content-based features	Context-based features
Wang et al. [58]	2018	✓	
Vedova et al. [11]	2018	✓	✓
Ajao et al. [1]	2019	✓	
Zhou & Zafarani [65]	2019		✓
Shu et al. [52]	2019		✓
Jiang et al. [25]	2019	✓	✓
Reddy et al. [46]	2020	✓	
Kaliyar et al. [27]	2021	✓	
Li et al. [35]	2021	✓	

Table 2 An example of natural language inference

premise	Permanent members of the UN Security Council are the five governments of China, France, Russia, Britain and the United States.	
hypothesis	The United States is a permanent member of the United Nations Security Council.	Entailment
	One of the five permanent members of the UN Security Council is the German government.	Contradiction
	The permanent members of the Security Council are all allies who won World War II.	Neutral

hand, in some works, contextual features such as propagation speed have been addressed. These features are very hard to find. Access to them requires access to the general state of the network, and even at the beginning of the fake news release they are not yet usable. For example, the propagation speed can be calculated and used after a period of the news release. This causes them to waste golden time to prevent the spread of fake news and rumors. All of the current works focused only on the content or context information of the social network and what has been neglected is the knowledge source outside the social network. Reliable news sources are one the most important knowledge sources. They publish a significant source of confirmed news which can be used to distinguish the fake news from the truth. In this study, for the first time, we presented a method based on natural language inference, which can be used to distinguish between true and false news using the set of published and confirmed news.

One of our challenges in this research is the lack of a suitable data set to detect fake news by inferring new news from previously confirmed news. For this purpose, we prepared this data set called FNID and used it in the current research, which led to improving the accuracy of detecting fake news. We have developed the FNID dataset based on two datasets, FakeNewsNet and LIAR. This dataset is available to researchers for free.

3 Natural language inference

Natural Language Inference (NLI) is one of the tasks in natural language processing which is also known as "*Recognizing Textual Entailment*" (RTE). It is believed to be close to the ultimate goal of natural language processing, namely "*Natural Language Understanding*" [37]. The task is to determine the inference relationship between two given phrases called *premise* (p) and *hypothesis* (h). A hypothesis may be inferable from a given premise (entailment), contradicts with premise (contradiction), or indeterminate (neutral). In Table 2, an example is presented for each of these classes.

The state-of-the-art methods in NLI are deep learning-based which learn to automatically extract features from vast amount of data. For this aim, large datasets in English language have been developed and introduced, including *"SNLI"* [6] *"MultiNLI"* [60], and *"SciTail"* [30], as well as datasets in non-English languages like *"FarsTail"* [2] and *"OCNLI"* [24].

Figure 1 shows the scheme of a typical NLI model [10]. The input premise and hypothesis are encoded to fixed-length numeric vectors using a neural encoder like a bidirectional LSTM. The obtained vectors u and v are then concatenated along with their element-wise product and absolute difference, resulting in a representation which captures information

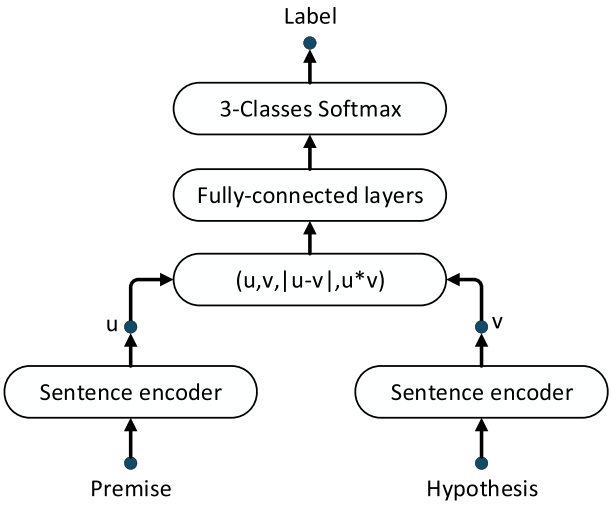

Fig. 1 The scheme of a typical NLI model

from both premise and hypothesis. This vector is then passed to a 3-class classifier consisting of multiple fully-connected layers. Along with this typical architecture, researchers have also come up with a variety of more sophisticated models to get better performance in this task [8, 33, 34, 36, 42, 54, 61].

The significant advances of NLI have led researchers in many fields to use this task to solve various problems and apply it to applications that require inference between two expressions. These include question answering [56], fact extraction [55], generating video captions [43], and judging textual quality [22] and etc.

In this work, we use NLI to detect fake news in a similar way to humans. The detection of fake news by humans is mainly based on inferring the veracity using a set of reliable news rather than by merely statistical features within the news content or context. In the proposed approach, the news item that we intend to verify is considered as a hypothesis, and the available set of reliable news plays the role of the premise. The inference relationship between this premise set and the intended hypothesis reveals the reliability of the news item.

4 Proposed method

Suppose that h is the news item whose veracity is under investigation, and p is the set of related confirmed news received from trusted sources. Based on the standard definition of NLI problem mentioned in Section 3, three situations can be considered. The news item h can be assumed *true* if $p \vdash h$, that is, p entails h. On the other hand, this news item is proved to be *fake* if $p \vdash \overline{h}$, i.e., h contradicts the previously verified news. In *neutral* case that neither *entailment* nor *contradiction* of h is distinguishable from p, we can not definitively accept or reject that news item.

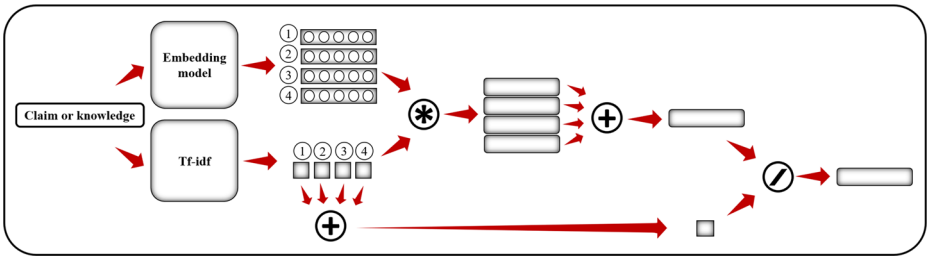

Fig. 2 Phrase representation by word embedding and tf-idf

We consider two versions of this problem. In the first version, we have a two-class problem with *fake* and *real* as labels which is compatible with the *FakeNewsNet* dataset [50]. In the second version, a six-class problem is considered with *pants-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true* as fine-grained labels. This is compatible with the *LIAR* dataset [57]. The details are presented in Section 5.

We use the proposed approach along with classical machine learning models as well as neural network models, which are described below.

4.1 Classical machine learning models

In these models, the feature extraction phase is performed before model training. These two steps are detailed below:

- **Feature extraction:** To represent the premise and hypothesis, we use the bag-of-words approach, which delivers an average of the constituting words' representations as the sentence representation. To reduce the effect of stop words in long premises, we weight each word based on its *tf-idf*. This increases the impact of more important words on the final representation. The weighted sum of the word vectors is then normalized by the sum of *tf-idf* values. The used word embedding methods in our experiments are Word2vec [38], GloVe [44], FastText [5], and BERT [13]. The normalized, weighted average of word vectors for the premise and hypothesis are then concatenated to deliver the final sample representation. Figure 2 shows an overview of the mentioned phrase representation process.
- **Model training:** In many past content-based studies, only the claims have been used to detect the fake news, ignoring the previous relevant news as the auxiliary knowledge. We bridge this gap by the NLI approach. To measure the effectiveness of using the NLI approach in detecting fake news, we first train the models only using the generated vectors for the claims (hypotheses). These models are called *simple* in our experiments. Then, by concatenating the premise and hypothesis vectors, we train a so-called *NLI* model, which is designed to infer the claim's correctness based on the previous knowledge (premises). Figure 3 illustrates the aforementioned process.

4.2 Neural network models

In recent years, deep neural network models have shown excellent performance in supervised learning tasks [32]. They benefit from feature learning for the input representation, reducing the needs of feature engineering.

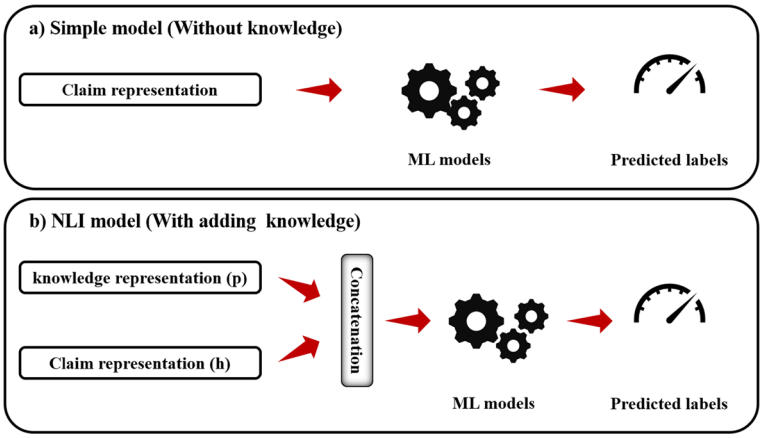

Fig. 3 Simple and NLI models based on classical machine learning models

In this section, a NLI-based model is designed using Bidirectional LSTM [21] and Bidirectional GRU [9] neural networks to investigate the correctness of a given claim based on the previously confirmed related news. Similar to the previous section, firstly, we use only the claims (hypotheses) to train a simple neural network model. Then, the NLI-based model is trained to infer the claim's correctness from previous knowledge (premises). By comparing the results of these two models, we evaluate the effectiveness of the proposed NLI-based approach in detecting fake news. Figure 4 shows a schematic view of this process.

5 Data acquisition and preprocessing

Since there is not a complete dataset available including premises to be used in the NLI setting, we have collected a new appropriate dataset. It has been gathered in a way that is compatible with FakeNewsNet and LIAR datasets as two well-known and frequently used datasets in this field. The required data for training an NLI system should include premise,

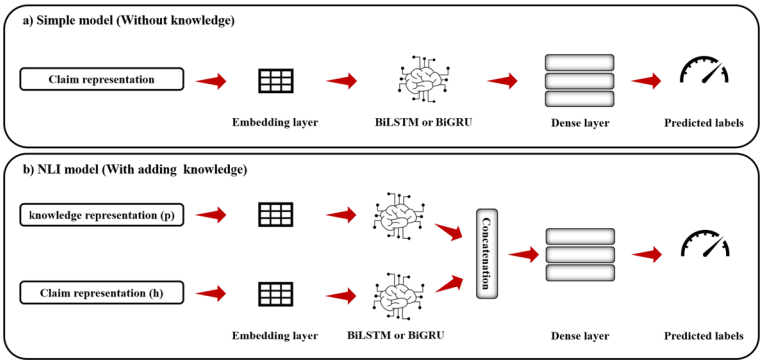

Fig. 4 Simple and NLI models based on neural network models

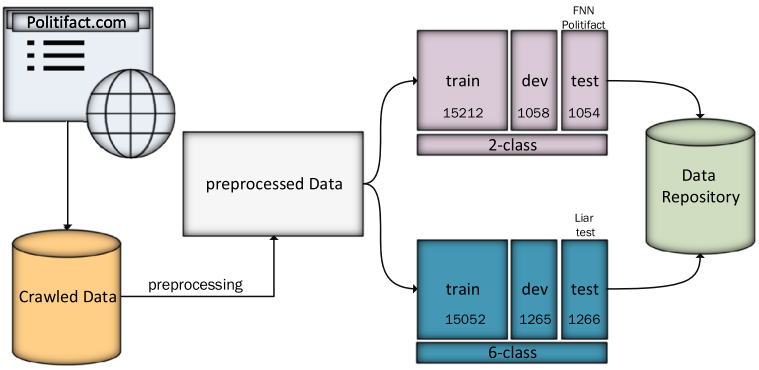

Fig. 5 The overview of our dataset construction

hypothesis, and label fields. We consider the news as hypothesis, the confirmed related news as premise, and the veracity of the news item as the label.

The overall steps of data acquisition and preprocessing are illustrated in Fig. 5.

Table 3 Fields of PolitiFact published articles

No.	Field	Description
1	*Statement*	A claim published in the media by a person or an organization which has been investigated in PolitiFact.
2	*Title*	The title of the article published by PolitiFact about the claim.
3	*Time*	The publication time of this article on the PolitiFact website.
4	*Speaker*	The person or organization to whom the *Statement* relates.
5	*Content*	The text of the Politifact article including parts of the past and present news related to the statement which is selected by Politifact's experts and can be used to investigate the accuracy of the statement. Also, at the end of this section, the experts' final opinions on the statement are given according to the sources mentioned as Our Ruling... and We Rate....
6	*Sources*	The news' URL related to the *Statement* as well as the sources' URL used in the *Content* section.
7	*Label*	The *Statement*'s tag suggested by the expert team among nine labels: Mostly-True, True, Half-True, False, Mostly-False, Pants on Fire, No Flip, Half Flip, and Full Flop.

5.1 Data collection

The dataset is collected using PolitiFact website API[5]. This website is a reputable source of fact-finding in which a team of experts evaluate political news articles published in various sources (including *CNN*, *BBC*, and *Facebook*). Each published article on this website consists of seven sections listed in Table 3. All the articles published until **April 26, 2020** are crawled and collected in our dataset.

Since LIAR and FakeNewsNet datasets use also the PolitiFact website to collect their data records, we establish a mapping between the items in our dataset and those datasets. This eases the comparison between the proposed approach and previous methods. To this aim, we use as the test set the part of our data that is also available in FakeNewsNet or LIAR datasets. As the development set, a random subset of the remaining samples is selected whose size is proportional to the size of the test set. The remaining samples are considered as the train set.

In the FakeNewsNet dataset, there are two different labels: *fake* and *real*, while in the LIAR dataset, the number of classes is six: *pants-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true*. On the other hand, the total number of unique labels in the PolitiFact articles is 9 (last row of Table 3). We publish our dataset as two different folders which are compatible with FakeNewsNet and LIAR datasets, respectively.

Based on the FakeNewsNet article, we consider the label *real* instead of *true*, *mostly-true*, and *half-true* labels. We also consider *fake* instead of *pants-fire*, *false*, and *barely-true* labels. We ignore *no-flip*, *half-flip*, and *full-flop* which do not have a corresponding label in FakeNewsNet dataset. For LIAR dataset, along with the six labels which are common between LIAR and PolitiFact, we replace the *no-flip*, *half-flip*, and *full-flop* labels with *true*, *half-true*, and *false* labels, respectively. This labeling is the same as presented in the LIAR article.

5.2 Preprocessing

To clean the collected articles from PolitiFact website, HTML and CSS tags as well as extra spaces and characters were removed from the text. The last sections of each article that were about the rules of the website (i.e. *Our ruling...*) and the final opinion of the experts about the veracity of news (i.e. *we rate ...*) were also removed. The remaining content is the text of the news collection that has been reviewed by experts to get the veracity of the intended news. This data is stored in two modes: sequences of paragraphs and a single text (joint paragraphs) in columns *Paragraph-based-content* and *FullText-based-content*, respectively. In this work, *FullText-based-content* is used, but *Paragraph-based-content* can be exploited in paragraph-based NLI in future research.

The NLI task requires dataset to include three distinct fields: *premise*, *hypothesis*, and *label*. Accordingly, we select following fields for this aim:

- **Premise**: We use *FullText-based-content* field as the premise which contains the text of news related to the news under investigation.
- **Hypothesis**: The *Statement* field is considered as hypothesis (see Table 3). It is a claim published in the news media, and now its integrity is under investigation.
- **Label**: *Label-FNN* and *Label-LIAR* are used as the label of data.

[5] https://www.politifact.com/api/factchecks

Table 4 FNID data statistics

Total number of news	17583	
Average number of statement characters	111.083	
Average number of statement words	22.564	
Average number of content characters	4670.107	
Average number of content words	903.791	
Average number of content paragraphs	21.602	
Number of labels based on FNN (PolitiFact)	fake	8557
	real	8767
Number of labels based on LIAR	pants-fire	2012
	false	3809
	barely-true	2897
	half-true	3339
	mostly-true	3096
	true	2430

The final dataset, called Fake News Inference Dataset (FNID) [48], is publicly available for future research[6]. Some statistics of this dataset are presented in Table 4.

6 Experiments and results

6.1 Setup

In this section, we evaluate our proposed method on the *FNID-FakeNewsNet* and *FNID-LIAR* datasets. As mentioned in Section 4, two models are compared to evaluate the effectiveness of the NLI-based approach in fake news detection. The first one, called *simple model*, uses only *statements* (*hypotheses*); while the other one, called *NLI model*, exploits *fullText-based-contents* (*premises*) along with *statements* (*hypotheses*). As classical machine learning models, we use Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR), k-Nearest Neighbors (KNN), and Support Vector Machine (SVM) algorithms; while as neural networks, we use BiLSTM and BiGRU models. For representing the words, Word2vec, GloVe, fastText, and BERT are used.

The used evaluation measures are *accuracy* and *F1-score*, along with the confusion matrices for more detailed investigations. In the following, we review the definition of the used evaluation measures.

Accuracy: It measures the percentage of correctly classified samples:

$$Accuracy = \frac{\sum_{i=1}^{n} TP_i}{N} \tag{1}$$

where n is the number of classes, TP_i indicates the number of true positives in class i, and N is the total number of samples.

[6] https://ieee-dataport.org/open-access/fnid-fake-news-inference-dataset

F1-score: To better evaluate the performance of a classifier in imbalanced problems, it is better to use the F1-score, since accuracy may be misleading. Particularly, in the fake news context, the number of fake news is often significantly less than real news. F1-score is defined as the harmonic mean of *Precision* and *Recall*:

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{3}$$

$$F1\text{-}score_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{4}$$

where TP_i, FP_i, and FN_i are, respectively, True Positive, False Positive, and False Negative samples in class i.

Macro-F1 : This metric gives an overview of the model performance in all classes, which is obtained by averaging the F1-scores of the classes:

$$Macro\text{-}F1 = \frac{\sum_{i=1}^{n} F1\text{-}score_i}{n} \tag{5}$$

6.2 Results

The results of simple and NLI models on the *FNID-FakeNewsNet* dataset are given in Table 5. The best obtained accuracies by different models are also depicted in Fig. 6. As can be seen, the best obtained results for all models, except Naïve Bayes, have been improved by the NLI model. The best results in both the simple and NLI models have been obtained by BiLSTM neural network using BERT embedding. By the way, comparing the best simple and NLI models shows that using the NLI approach has made 10.44 and 10.34 absolute improvements in terms of accuracy and Macro-F1 scores, respectively. Figure 7 shows the confusion matrices of the best simple and NLI models on *FNID-FakeNewsNet* dataset.

Figure 7 shows the prediction improvement in both fake and real classes. By considering "fake class" as Positive (P) class and "real class" as Negative (N) class, we find that the inference model was able to reduce 14 samples from False Negative (FN) part and add this number to the True Positive (TP) part, Which has led to an increase in the TP part from 394 to 408 samples. Also, this inference model has been able to increase the number of True Negative (TN) samples from 398 samples to 494 samples by subtracting 96 samples from the False Positive (FP) part and adding this value to the True Negative (TN) part. These changes in the confusion matrix, in addition to improving Accuracy, have led to improved F1-score in both classes, as a result, the Macro-F1 score has also increased.

Table 6 shows the evaluation results of simple and NLI models on the *FNID-LIAR* dataset. The best obtained accuracies are also depicted in Fig. 8. These results show that the best results for all classifiers are obtained by the NLI approach. Also, the best overall result in both simple and NLI models is obtained by the BiLSTM neural network using BERT embedding. Using the NLI approach has made 13.19 and 14.33 absolute improvements in terms of the best obtained accuracy and Macro F1 scores, respectively. Figure 9 shows the confusion matrices of the best simple and NLI models on *FNID-LIAR* dataset.

Figure 9 shows the prediction improvement in all classes. Considering each of the classes as a Positive (P) class and the other classes as a Negative (N) class, we find that the inference model was able to the inference model has been able to improve the Accuracy and Macro-F1 evaluation metrics by reducing the samples of False Negative (FN) parts and adding this

Table 5 The obtained results on FNID-FakeNewsNet dataset

Models		Simple model				NLI model			
		Word2vec	GloVe	fastText	BERT	Word2vec	GloVe	fastText	BERT
DT	Acc	0.5702	0.5655	0.5844	0.5750	0.5380	0.5787	0.5797	**0.6157**
	Macro-F1	0.5647	0.5631	0.5822	0.5739	0.5360	0.5785	0.5792	**0.6158**
NB	Acc	0.4004	**0.6926**	0.6509	0.6708	0.4099	0.6803	0.6641	0.6670
	Macro-F1	0.2923	**0.6923**	0.6497	0.6691	0.3129	0.6778	0.6629	0.6623
RF	Acc	0.6328	0.6537	0.6556	0.6584	0.6613	0.6689	0.6471	**0.7144**
	Macro-F1	0.6327	0.6520	0.6548	0.6567	0.6605	0.6666	0.6435	**0.7132**
LR	Acc	0.3966	0.6850	0.6879	0.6954	0.3966	0.7353	0.7068	**0.8007**
	Macro-F1	0.2840	0.6850	0.6873	0.6949	0.2840	0.7351	0.7056	**0.8007**
KNN	Acc	0.5892	0.6698	0.6157	0.6451	0.5465	0.6755	0.6499	**0.7457**
	Macro-F1	0.5835	0.6688	0.6149	0.6448	0.5459	0.6755	0.6493	**0.7458**
SVM	Acc	0.3975	0.7002	0.7135	0.6784	0.4127	0.7324	0.7258	**0.7609**
	Macro-F1	0.2870	0.7001	0.7135	0.6766	0.3201	0.7322	0.7254	**0.7607**
BiGRU	Acc	0.7144	0.7125	0.7021	0.7116	0.7960	0.8102	0.8140	**0.8178**
	Macro-F1	0.7143	0.7125	0.7015	0.7112	0.7956	0.8097	0.8138	**0.8175**
BiLSTM	Acc	0.7400	0.6243	0.7106	0.7514	0.8397	0.8520	0.8463	**0.8558**
	Macro-F1	0.7399	0.6170	0.7106	0.7514	0.8390	0.8512	0.8458	**0.8548**

The best results are marked with bold

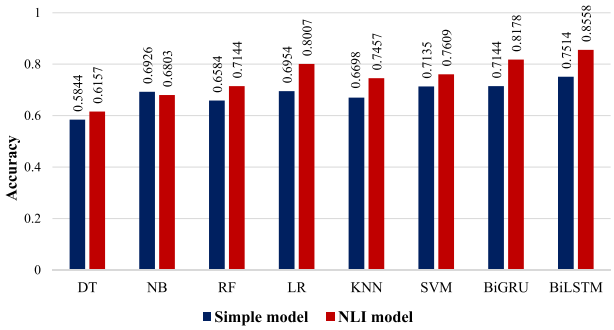

Fig. 6 The best obtained accuracies by different models on the FNID-FakeNewsNet dataset

samples to the True Positive (TP) part. This reduction of FN and addition to TP in classes *pants-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true* is equal to 24, 17, 10, 33, 36 and 47 samples, respectively.

To compare the proposed approach with the baseline methods reported by Shu et al. [50] and the SAF/S [51] method on FakeNewsNet (PolitiFact) data, we performed an experiment under a similar condition. Since the reported results by these works are based on 1,054 samples, we also trained our best model, which is BiLSTM (BERT) according to Table 5, on the same data. The samples were divided into 80%, 10%, and 10% for training, validating, and testing, respectively. The last row of Table 7 shows the average accuracy of our approach over five experiments. The other results were extracted from the references.

Similarly, we compared our approach with the baseline models reported by Wang et al. [57] and the method proposed by Karimi et al. [28] on LIAR dataset. Note that the work of Karimi et al. [28] combines information from multiple sources beyond the news content. The last row of Table 8 shows the accuracy of our best achieved model, i.e., BiLSTM (BERT), with the same number of data samples as the baseline models, which is 10,268 samples for training, 1,284 samples for validation, and 1,266 samples for testing. According to Tables 7 and 8, our proposed method, which exploits the verified news using a NLI approach, outperforms the baselines by a considerable margin. This improvement is specially noticeable for the FakeNewsNet (PolitiFact) dataset which has less training data, showing the effectiveness of the auxiliary knowledge specially in the low-resource situations.

Fig. 7 Confusion matrices of the best simple and NLI models on the FNID-FakeNewsNet dataset

Table 6 The obtained results on FNID-LIAR dataset

Models		Simple model				NLI model			
		Word2vec	GloVe	fastText	BERT	Word2vec	GloVe	fastText	BERT
DT	Acc	0.1667	0.2062	0.1793	0.1872	0.1801	0.1991	0.2085	**0.2172**
	Macro-F1	0.1596	0.1956	0.1739	0.1761	0.1712	0.1911	0.2018	**0.2107**
NB	Acc	0.2014	0.2204	0.2141	0.2393	0.0805	0.2235	0.2314	**0.2551**
	Macro-F1	0.0700	0.1927	0.1973	0.2331	0.0444	0.2168	0.2216	**0.2507**
RF	Acc	0.2227	0.2480	0.2346	0.2291	0.2512	0.2504	0.2275	**0.2812**
	Macro-F1	0.1834	0.2167	0.2068	0.1997	0.2196	0.2246	0.2047	**0.2630**
LR	Acc	0.0727	0.2330	0.2346	0.2646	0.0727	0.2583	0.2749	**0.3081**
	Macro-F1	0.0226	0.2154	0.1955	0.2538	0.0226	0.2493	0.2493	**0.3091**
KNN	Acc	0.1856	0.2085	0.2188	0.2338	0.1848	0.2299	0.2243	**0.2409**
	Macro-F1	0.1727	0.2011	0.2131	0.2305	0.1753	0.2217	0.2190	**0.2403**
SVM	Acc	0.1967	0.2567	0.2654	0.2575	0.2014	0.2678	0.2670	**0.3002**
	Macro-F1	0.0561	0.2081	0.2032	0.2089	0.0733	0.2298	0.2262	**0.2764**
BiGRU	Acc	0.2694	0.2765	0.2707	0.2812	0.3815	0.3594	0.3863	**0.4013**
	Macro-F1	0.2493	0.2651	0.2383	0.2686	0.3904	0.3570	0.3972	**0.4061**
BiLSTM	Acc	0.2551	0.2591	0.2417	0.2812	0.3799	0.3949	0.3878	**0.4131**
	Macro-F1	0.2302	0.2205	0.1594	0.2715	0.3830	0.4126	0.4002	**0.4148**

The best results are marked with bold

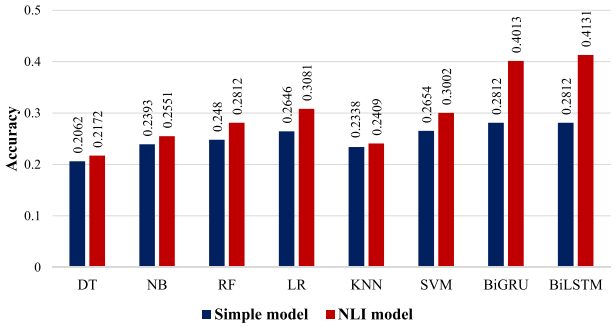

Fig. 8 The best obtained accuracies by different models on the FNID-LIAR dataset

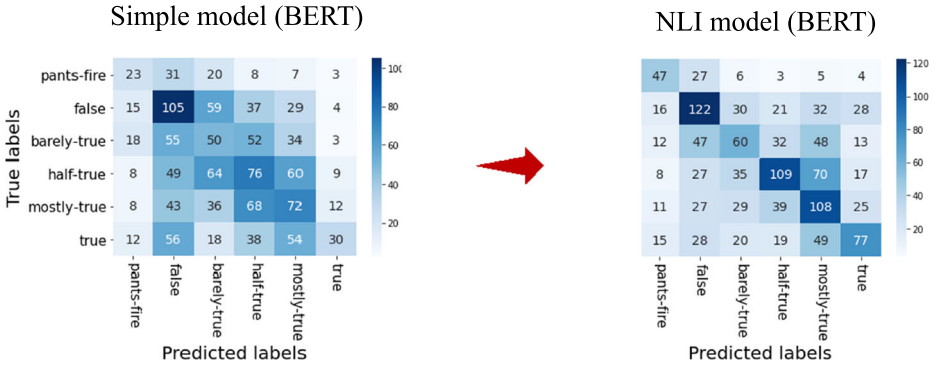

Fig. 9 Confusion matrices of the best simple and NLI models on the FNID-LIAR dataset

Table 7 The accuracy of baseline methods on FakeNewsNet (PolitiFact) dataset as well as the accuracy of the proposed method on FakeNewsNet-compatible version of FNID dataset

Method	Accuracy
SVM [50]	0.580
Logistic Regression [50]	0.642
Naïve Bayes [50]	0.617
CNN [50]	0.629
SAF/S [51]	0.633
Our method (BiLSTM (BERT))	**0.9019**

The best result is marked with bold

Table 8 The accuracy of baseline methods on LIAR dataset as well as the accuracy of the proposed method on LIAR-compatible version of FNID dataset

Method	Accuracy
Majority [57]	0.208
SVM [57]	0.255
Logistic Regression [57]	0.247
Bi-LSTMs [57]	0.233
CNN [57]	0.270
MMFD[28]	0.3881
Our method (BiLSTM (BERT))	**0.3965**

The best result is marked with bold

7 Conclusion and future works

Most methods for detecting fake news use post-publication effects on the community to determine whether the news is true or false. In other words, these methods cannot work in the early stages of the publication of news and can only be used when the news has spread in the community and has left its harmful effects. In this study, we present a method that uses previously verified news to detect fake news instead of using only the content or context of the news. We have designed this method based on the natural language inference task, in which to verify a new news item as a hypothesis, previous similar verified news is used as a premise. The proposed method enables us to detect fake news in the early moments of publication. One of the most critical challenges in this study was the need for verified news similar to the news item under review, but there was no suitable dataset for this purpose. Therefore, we created the first Fake News Inference Dataset (FNID) in a rigorous process and published it for free. The results of this study show an increase in the accuracy of detecting fake news using the proposed approach.

Although the proposed method has been able to overcome other previous methods, but it has its own weaknesses and limitations. These limitations fall into two general categories: the retrieving the set of confirmed and related news, and the limitations of the inference method. From the first limitation point of view, the proposed method requires a set of verified news related to fake news. In this work, it is assumed that this set is already available and no mechanism is provided to automate the extraction process of this set. From the second limitation point of view, NLI methods which have been employed here for inferring the correctness of the news, have weakness for understanding long texts. In the future, we intend to work on these limitation to further improve the proposed method. We also want to make an online tool that finds similar news items to the given news from reputable sources and uses them as the premise input to the NLI model trained to detect fake news. Investigating other more complex and specialized NLI models for use in the approach presented in this research is another of our future plans.

References

1. Ajao O, Bhowmik D, Zargari S (2019) Sentiment aware fake news detection on online social networks. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2507–2511. IEEE

2. Amirkhani H, AzariJafari M, Pourjafari Z, Faridan-Jahromi S, Kouhkan Z, Amirak A (2021) FarsTail: A Persian Natural Language Inference Dataset, arXiv:2009.08820
3. Bakhteev O, Ogaltsov A, Ostroukhov P (2020) Fake News Spreader Detection using Neural Tweet Aggregation. CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org
4. Behzad B, Bheem B, Elizondo D, Marsh D, Martonosi S (2021) Prevalence and Propagation of Fake News, arXiv:2106.09586
5. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Ass Comput Linguist ics, 5:135–146. MIT Press
6. Bowman SR, Angeli G, Potts C, Manning DD (2015) A large annotated corpus for learning natural language inference, arXiv:1508.05326
7. Breiman L (2001) Random forests: Machine learning, vol 45. Springer, pp 5–32
8. Chen Q, Zhu X, Ling Z, Wei S, Jiang H, Inkpen D (2016) Enhanced lstm for natural language inference. arXiv:1609.06038
9. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv:1406.1078
10. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017) Supervised learning of universal sentence representations from natural language inference data (EMNLP)
11. Della Vedova ML, Tacchini E, Moret S, Ballarin G, DiPierro M, de Alfaro L (2018) Automatic online fake news detection combining content and social signals. 2018 22nd Conference of Open Innovations Association (FRUCT), pp 272–279. IEEE
12. Dey R, Salemt FM (2017) Gate-variants of gated recurrent unit (GRU) neural networks. 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE, pp 1597–1600
13. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp 4171–4186
14. Dong X, Victor U, Qian L (2020) Two-path Deep Semi-supervised Learning for Timely Fake News Detection, arXiv:2002.00763
15. Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inf 35, 352–359. Elsevier
16. Farajtabar M, Yang J, Ye X, Xu H, Trivedi R, Khalil E, Li S, Song L, Zha H (2017) Fake News Mitigation via Point Process Based Intervention: International conference on machine learning, pp 1097–1106, PMLR
17. Golbeck J, Mauriello M, Auxier B, Bhanushali KevalH, Bonk C, Bouzaghrane MA, Buntain C, Chanduka R, Cheakalos P, Everett JennineB et al (2018) Fake news vs satire: A dataset and analysis. Proceedings of the 10th ACM Conference on Web Science, pp 17–21
18. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning word vectors for 157 languages. Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp 2018
19. Hakak S, Khan WZ, Bhattacharya S, Reddy GT, Choo K-R (2020) Propagation of fake news on social media: challenges and opportunities. International Conference on Computational Data and Social Networks, pp 345–353. Springer
20. Hakak S, Alazab M, Khan S, Gadekallu TR, Maddikunta PKR, Khan WZ (2021) An ensemble machine learning approach through effective feature extraction to classify fake news. Fut Gener Comput Syst 117:47–58. Elsevier
21. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780. MIT Press
22. Holtzman A, Buys J, Forbes M, Bosselut A, Golub D, Choi Y (2018) Learning to Write with Cooperative Discriminators. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Assoc Comput Linguist:1638–1649
23. Horne BD, Adali S (2017) This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Eleventh International AAAI Conference on Web and Social Media
24. Hu H, Richardson K, Xu L, Li L, Kuebler S, Moss LS (2020) OCNLI: Original Chinese Natural Language Inference, arXiv:2010.05444
25. Jiang S, Chen X, Zhang L, Chen S, Liu H (2019) User-Characteristic Enhanced Model for Fake News Detection in Social Media. CCF International conference on natural language processing and chinese computing, pp 634–646. Springer
26. Jiang L, Wang D, Cai Z, Yan X (2007) Survey of improving naive bayes for classification. International conference on advanced data mining and applications. Springer, pp 134–145

27. Kaliyar RK, Goswami A, Narang P (2021) FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed Tools Appl 80(8):11765–11788. Springer
28. Karimi H, Roy P, Saba-Sadiya S, Tang J (2018) Multi-source multi-class fake news detection. Proc 27th Int Conf Comput Linguisti:1546–1557
29. Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 4:580–585.IEEE
30. Khot T, Sabharwal A, Clark PS (2018) A textual entailment dataset from science question answering. Thirty-Second AAAI Conference on Artificial Intelligence
31. Kumar PJS, Devi PR, Sai NR, Kumar S, Benarji T (2021) Battling Fake News A Survey on Mitigation Techniques and Identification. 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, pp 829–835
32. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. Nature Publishing Group
33. Li P, Yu H, Zhang W, Xu G, Sun X (2020) SA-NLI: A supervised attention based framework for natural language inference, Elsevier, Neurocomputing
34. Liu X, He P, Chen W, Gao J (2019) Improving multi-task deep neural networks via knowledge distillation for natural language understanding, arXiv:1904.09482
35. Li X, Lu P, Hu, Wang X, Lu L (2021) A novel self-learning semi-supervised deep learning network to detect fake news on social media. Multimedia Tools and Applications. Springer, pp 1–9
36. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach, . arXiv:1907.11692
37. MacCartney B (2009) Natural language inference. Stanford University
38. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf PSyste:3111–3119
39. Moreno-Sandoval LG, Del Puertas EAP, Quimbaya AP, Alvarado-Valencia JA (2020) Assembly of Polarity: Emotion and user statistics for detection of fake profiles. CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org
40. Noureen J, Asif M (2017) Crowdsensing: socio-technical challenges and opportunities. IJACSA 8:363–369
41. Pamungkas EW, Basile V, Patti V (2019) Stance classification for rumour analysis in Twitter: Exploiting affective information and conversation structure, arXiv:1901.01911
42. Parikh AP, Täckström O, Das D, Uszkoreit J (2016) A decomposable attention model for natural language inference. arXiv:1606.01933
43. Pasunuru R, Bansal M (2017) Reinforced video captioning with entailment rewards. CoRR, arXiv:1708.02300
44. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
45. Pradhan A (2012) Support vector machine-a survey, vol 2
46. Reddy H, Raj N, Gala M, Basava A (2020) Text-mining-based Fake News Detection Using Ensemble Methods. International journal of automation and computing, pp 1–12 Springer
47. Ross QJ. (1986) Induction of decision trees. Mach Learn 1:81–106. Springer
48. Sadeghi F, Bidgoly AJ, Amirkhani H (2020) FNID: Fake News Inference Dataset. IEEE Dataport. https://doi.org/10.21227/fbzd-sw81
49. Shabani S, Sokhn M (2018) Hybrid machine-crowd approach for fake news detection. 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), pp 299–306. IEEE
50. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2018) FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media, arXiv:1809.01286
51. Shu K, Mahudeswaran D, Liu H (2019) Fakenewstracker: a tool for fake news collection, detection, and visualization. Comput Math Organ Theory 25:60–71. Springer
52. Shu K, Zhou X, Wang S, Zafarani R, Liu H (2019) The role of user profiles for fake news detection. Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 436–439
53. Silverman C, Strapagiel L, Shaban H, Hall E, Singer-Vine J (2016) Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. Buzzfeed News 20
54. Talman A, Yli-Jyrä A, Tiedemann J (2018) Natural language inference with hierarchical bilstm max pooling architecture, arXiv:1808.08762
55. Thorne J, Vlachos A, Cocarascu O, Christodoulopoulos C, Mittal A (2018) The fact extraction and VERification (FEVER) shared task proceedings of the first workshop on fact extraction and VERification (FEVER). Assoc Comput Linguist:1–9

56. Trivedi H, Kwon H, Khot T, Sabharwal A, Balasubramanian N (2019) Repurposing Entailment for Multi-Hop Question Answering Tasks, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Assoc Comput Linguist:2948–2958
57. Wang W, Yang L (2017) Liar pants on fire: A new benchmark dataset for fake news detection, arXiv:1705.00648
58. 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
59. Vlachos A, Riedel S (2014) Fact checking: Task definition and dataset construction. Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pp 18–22
60. Williams A, Nangia N, Bowman SR (2017) A broad-coverage challenge corpus for sentence understanding through inference, arXiv:1704.05426
61. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. Advances in Neural Inf Process Syst:5753–5763
62. Zhou X, Zafarani R (2018) A survey of fake news: Fundamental theories, Detection Methods, and Opportunities, arXiv:1812.00315
63. Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R (2018) Detection and resolution of rumours in social media: A survey, ACM Computing Surveys (CSUR), vol 51. ACM, New York, pp 1–36
64. Zhao Z, Zhao J, Sano Y, Levy O, Takayasu H, Takayasu M, Li D, Wu J, Havlin S (2020) Fake news propagates differently from real news even at early stages of spreading. EPJ Data Sci 9:11–14. SpringerOpen
65. Zhou X, Zafarani R (2019) Network-based Fake News Detection: A Pattern-driven Approach. ACM SIGKDD Explor Newslett 21, 2, 48–60. ACM, New York