Fake News Detection on Social Media Using A Natural Language Inference Approach

Fariba Sadeghi
University of Qom
Amir Bidgoly (jalaly@qom.ac.ir)
University of Qom
Hossein Amirkhani
University of Qom

Research Article

Keywords: FNID, online, Detection, correctness.

Posted Date: November 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-107893/v1

Fake News Detection on Social Media using a Natural Language Inference Approach

Fariba Sadeghi[1], **Amir Jalaly Bidgoly**[1,*], **and Hossein Amirkhani**[1]

[1]Department of Information Technology and Computer Engineering, University of Qom, Qom, Iran
[*]corresponding.jalaly@qom.ac.ir

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.

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[1]. Social media allows the audience to follow the news in their favorite areas instantly and republish the news 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 the 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 the individuals may intentionally or accidentally contain false information. Given that the news in these media, unlike traditional media, is published without supervision and verification, 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[2]. 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*[3]. 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 their excessive publication, they always seek 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

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

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[4–6]. 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[7,8], 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. 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[9]. 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 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. They usually try to find a contradiction or correspondence between a given news item and other existing, confirmed ones. If the given item contradicts confirmed news, it is considered to be fake; while if it corresponds to the confirmed news, it is labeled as true. The NLI approach does the same by deciding about the inference relationship between the given news item and available confirmed ones. The NLI task is a critical subfield in Natural Language Processing (NLP) with a significant progress in recent years using deep learning methods. Its goal is to determine whether a *"hypothesis"* is true (entailment), false (contradiction), or undetermined (neutral) based on a given *"premise"* as initial knowledge. We use this approach to boost a couple of classical and deep models including Decision Tree[10], Naïve Bayes[11], Random Forest[12], Logistic Regression[13], k-Nearest Neighbors[14], Support Vector Machine[15], BiGRU[16], and BiLSTM[17] along with different word embedding methods including Word2vec[18], GloVe[19], FastText[20], and BERT[21]. The results show considerable improvements in the fake news detection accuracy using the auxiliary knowledge based on the NLI approach. We also introduce a new NLI-based dataset according to the *FakeNewsNet (Politifact)*[22] and *LIAR*[23] datasets, which has been made freely available [4].

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

2 Related Work

Many research articles have been proposed on fake news detection. In these papers, which are mainly based on deep learning methods, 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[24]. In another work, used sentiment analysis in twitter posts for rumor and fake news detection[4], or in the other work used combined stylometric features with word vector representations to predict fake news[6].

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[7]. 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[8].

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[25]. 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[26].

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

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

and another for unsupervised learning to learn from an abundant amount of unlabeled[5]. 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*[27]. 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*[28].

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[29]. In another study in this 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[23]. In 2018, *Fake News vs. Satire* dataset was introduced in which 486 political news items were collected[30]. 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 classified into two classes of fake and true labels[22].

One of the main problems in detecting fake news is the lack of a dataset that can cover all methods. For example, there is no valid dataset that can use prior knowledge in inference approaches. Therefore, in this work, we also introduce a new dataset that can be used to detect fake news using the inference approach and prior knowledge. This dataset is based on two datasets, FakeNewsNet and LIAR.

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*"[31]. 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 1, an example is presented for each of these classes.

Table 1. 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

The state-of-the-art methods in NLI are deep learning-based which learn to automatically extract features from vast amount of data. Large datasets have been developed and introduced for this aim including *"SNLI"*[32] *"MultiNLI"*[33], and *"SciTail"*[34], as well as datasets in non-English languages like *"FarsTail"*[35] and *"OCLNI"*[36].

Figure 1 shows the scheme of a typical NLI model[37]. 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 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[38–44].

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[45], fact extraction[46], generating video captions[47], and judging textual quality[48] 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

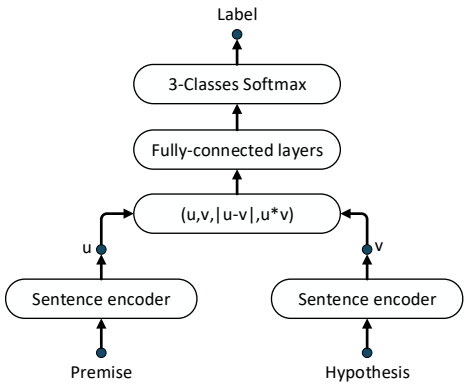

Figure 1. The scheme of a typical NLI model.

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.

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[22]. 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[23]. 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[18], GloVe[19], FastText[20], and BERT[21]. 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.

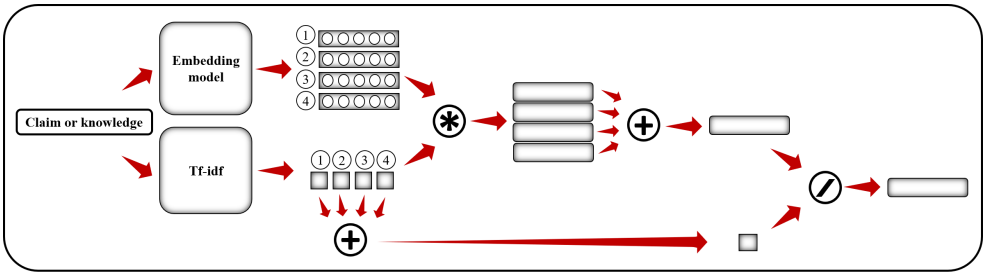

Figure 2. Phrase representation by word embedding and tf-idf.

- **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.

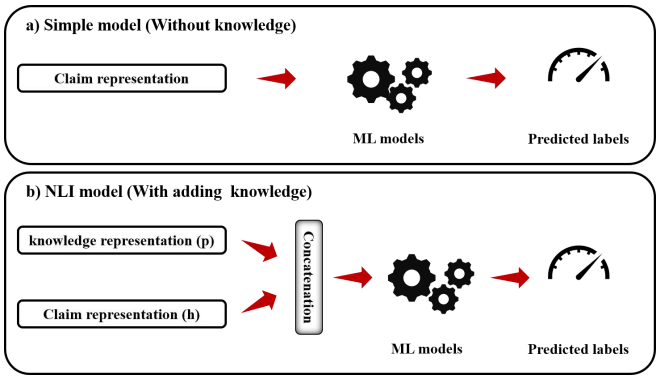

Figure 3. Simple and NLI models based on classical machine learning models.

4.2 Neural network models

In recent years, deep neural network models have shown excellent performance in supervised learning tasks[49]. They benefit from feature learning for the input representation, reducing the needs of feature engineering.

In this section, a NLI-based model is designed using Bidirectional LSTM[17] and Bidirectional GRU[50] 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.

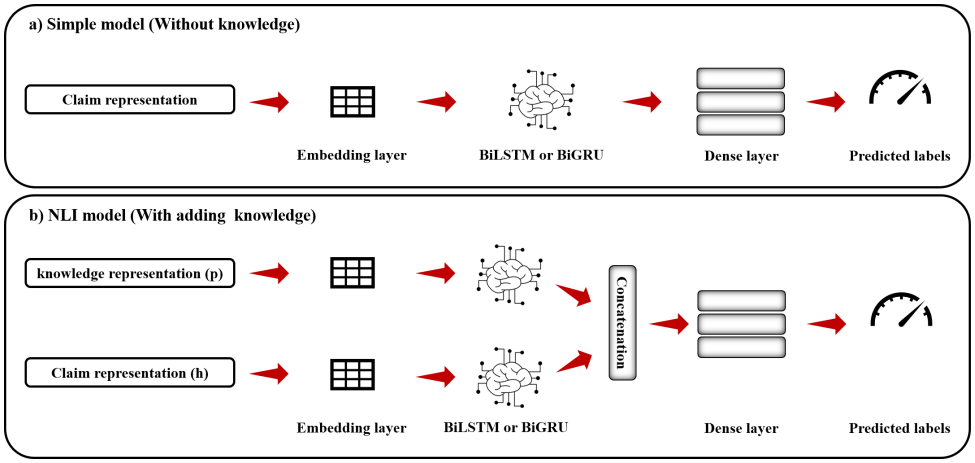

Figure 4. Simple and NLI models based on neural network models.

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, 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 Figure 5.

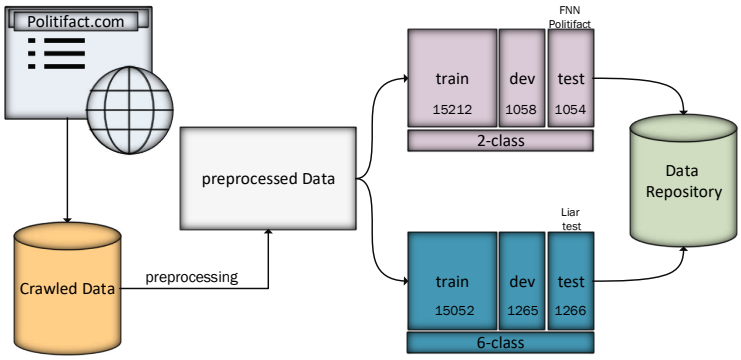

Figure 5. The overview of our dataset construction.

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 2. All the articles published until **April 26, 2020** are crawled and collected in our dataset.

Table 2. 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.

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.

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

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 2). 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 2). 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.

The final dataset, called Fake News Inference Dataset (FNID)[51], is publicly available for future research[6]. Some statistics of this dataset are presented in Table 3.

Table 3. 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

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

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.

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 4. The best obtained accuracies by different models are also depicted in Figure 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.

Table 5 shows the evaluation results of simple and NLI models on the *FNID-LIAR* dataset. The best obtained accuracies are also depicted in Figure 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.

To compare the proposed approach with the baseline methods reported by Shu et al.[22] and the SAF/S[52] 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 4, on the same

Table 4. The obtained results on FNID-FakeNewsNet dataset.

ML 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**

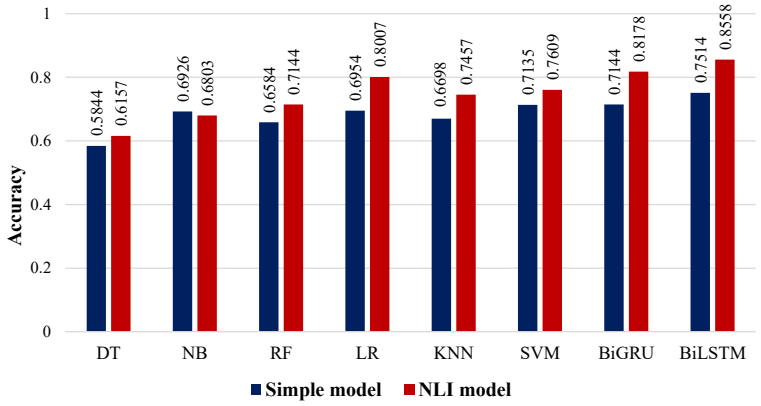

Figure 6. The best obtained accuracies by different models on the FNID-FakeNewsNet dataset.

data. The samples were divided into 80%, 10%, and 10% for training, validating, and testing, respectively. The last row of Table 6 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.[23] and the method proposed by Karimi et al.[53] on LIAR dataset. Note that the work of Karimi et al.[53] combines information from multiple sources beyond the news content. The last row of Table 7 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 6 and 7, 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.

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

Table 5. 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**

7 Conclusion and Future Work

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 introduced a method based on Natural Language Inference task that can be used to verify a news item using the previously verified news items. This method enables us to detect fake news in the first stages of the news spread. In addition, the proposed method is more similar to the way the humans verify a news item by comparing its relation to the previously verified items rather than investigating just the news content. We made a freely available dataset called FNID which can be used to train such systems.

In the future, we intend to use the content paragraphs in the paragraph-based-content column of the dataset as the premise to compensate for the weakness of NLP models for understanding long texts. 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 sophisticated and specialized NLI systems is another future direction of this research.

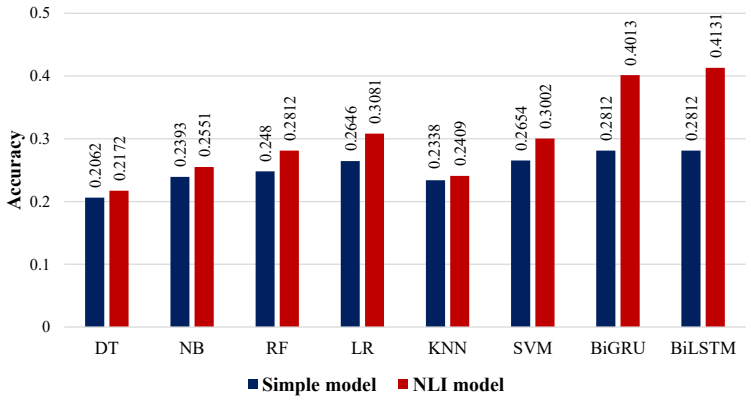

Figure 8. The best obtained accuracies by different models on the FNID-LIAR dataset.

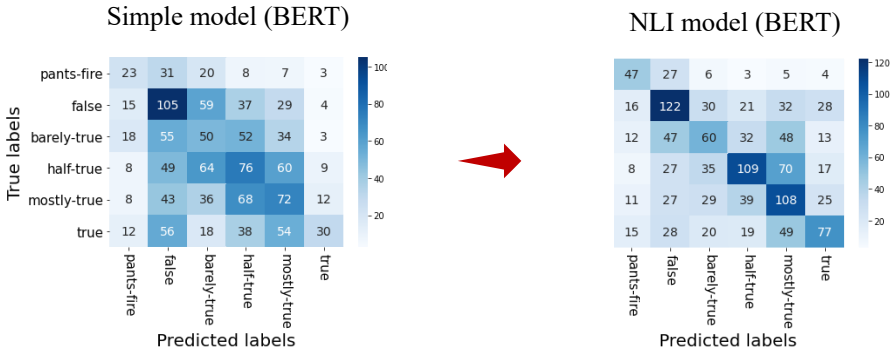

Figure 9. Confusion matrices of the best simple and NLI models on the FNID-LIAR dataset.

Table 6. 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[22]	0.580
Logistic Regression[22]	0.642
Naïve Bayes[22]	0.617
CNN[22]	0.629
SAF/S[52]	0.633
Our method (BiLSTM (BERT))	**0.9019**

References

1. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv. (CSUR)* **51**, 1–36 (2018).
2. Zhou, X. & Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315* (2018).
3. Noureen, J. & Asif, M. Crowdsensing: socio-technical challenges and opportunities. *IJACSA* **8**, 363–369 (2017).
4. Ajao, O., Bhowmik, D. & Zargari, S. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2507–2511 (IEEE, 2019).

Table 7. 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[23]	0.208
SVM[23]	0.255
Logistic Regression[23]	0.247
Bi-LSTMs[23]	0.233
CNN[23]	0.270
MMFD[53]	0.3881
Our method (BiLSTM (BERT))	**0.3965**

5. Dong, X., Victor, U. & Qian, L. Two-path deep semi-supervised learning for timely fake news detection. *arXiv preprint arXiv:2002.00763* (2020).
6. Reddy, H., Raj, N., Gala, M. & Basava, A. Text-mining-based fake news detection using ensemble methods. *Int. J. Autom. Comput.* 1–12 (2020).
7. Zhou, X. & Zafarani, R. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explor. Newsl.* **21**, 48–60 (2019).
8. Shu, K., Zhou, X., Wang, S., Zafarani, R. & Liu, H. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 436–439 (2019).
9. Pamungkas, E. W., Basile, V. & Patti, V. Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure. *arXiv preprint arXiv:1901.01911* (2019).
10. Quinlan, J. R. Induction of decision trees. *Mach. learning* **1**, 81–106 (1986).
11. Jiang, L., Wang, D., Cai, Z. & Yan, X. Survey of improving naive bayes for classification. In *International Conference on Advanced Data Mining and Applications*, 134–145 (Springer, 2007).
12. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).
13. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. biomedical informatics* **35**, 352–359 (2002).
14. Keller, J. M., Gray, M. R. & Givens, J. A. A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, cybernetics* 580–585 (1985).
15. Pradhan, A. Support vector machine-a survey. *Int. J. Emerg. Technol. Adv. Eng.* **2**, 82–85 (2012).
16. Dey, R. & Salemt, F. M. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, 1597–1600 (IEEE, 2017).
17. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119 (2013).
19. Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543 (2014).
20. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. *Transactions Assoc. for Comput. Linguist.* **5**, 135–146 (2017).
21. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, 4171–4186, DOI: 10.18653/v1/N19-1423 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
22. Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286* (2018).

23. Wang, W. Y. Liar, liar pants on fire: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

24. Wang, Y. *et al.* 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*, 849–857 (2018).

25. Della Vedova, M. L. *et al.* Automatic online fake news detection combining content and social signals. In *2018 22nd Conference of Open Innovations Association (FRUCT)*, 272–279 (IEEE, 2018).

26. Jiang, S., Chen, X., Zhang, L., Chen, S. & Liu, H. User-characteristic enhanced model for fake news detection in social media. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 634–646 (Springer, 2019).

27. Vlachos, A. & Riedel, S. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22 (2014).

28. Silverman, C., Strapagiel, L., Shaban, H., Hall, E. & Singer-Vine, J. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. *Buzzfeed News* **20** (2016).

29. Horne, B. D. & Adali, S. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh International AAAI Conference on Web and Social Media* (2017).

30. Golbeck, J. *et al.* Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, 17–21 (2018).

31. MacCartney, B. *Natural language inference*. Ph.D. thesis, Stanford University (2009).

32. Bowman, S. R., Angeli, G., Potts, C. & Manning, C. D. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

33. Williams, A., Nangia, N. & Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).

34. Khot, T., Sabharwal, A. & Clark, P. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

35. Amirkhani, H. *et al.* Farstail: A persian natural language inference dataset. *arXiv preprint arXiv:2009.08820* (2020).

36. Hu, H. *et al.* Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444* (2020).

37. Conneau, A., Kiela, D., Schwenk, H., Barrault, L. & Bordes, A. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP* (2017).

38. Chen, Q. *et al.* Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038* (2016).

39. Parikh, A. P., Täckström, O., Das, D. & Uszkoreit, J. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933* (2016).

40. Talman, A., Yli-Jyrä, A. & Tiedemann, J. Natural language inference with hierarchical bilstm max pooling architecture. *arXiv preprint arXiv:1808.08762* (2018).

41. Li, P., Yu, H., Zhang, W., Xu, G. & Sun, X. Sa-nli: A supervised attention based framework for natural language inference. *Neurocomputing* (2020).

42. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

43. Yang, Z. *et al.* Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763 (2019).

44. Liu, X., He, P., Chen, W. & Gao, J. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482* (2019).

45. Trivedi, H., Kwon, H., Khot, T., Sabharwal, A. & Balasubramanian, N. Repurposing entailment for multi-hop question answering tasks. In *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)*, 2948–2958, DOI: 10.18653/v1/N19-1302 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).

46. Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C. & Mittal, A. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 1–9, DOI: 10.18653/v1/W18-5501 (Association for Computational Linguistics, Brussels, Belgium, 2018).

47. Pasunuru, R. & Bansal, M. Reinforced video captioning with entailment rewards. *CoRR* **abs/1708.02300** (2017). 1708.02300.

48. Holtzman, A. *et al.* Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1638–1649, DOI: 10.18653/v1/P18-1152 (Association for Computational Linguistics, Melbourne, Australia, 2018).

49. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).

50. Cho, K. *et al.* Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

51. Sadeghi, F., Bidgoly, A. J. & Amirkhani, H. FNID: Fake news inference dataset, DOI: 10.21227/fbzd-sw81 (2020).

52. Shu, K., Mahudeswaran, D. & Liu, H. Fakenewstracker: a tool for fake news collection, detection, and visualization. *Comput. Math. Organ. Theory* **25**, 60–71 (2019).

53. Karimi, H., Roy, P., Saba-Sadiya, S. & Tang, J. Multi-source multi-class fake news detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1546–1557 (2018).

Figures

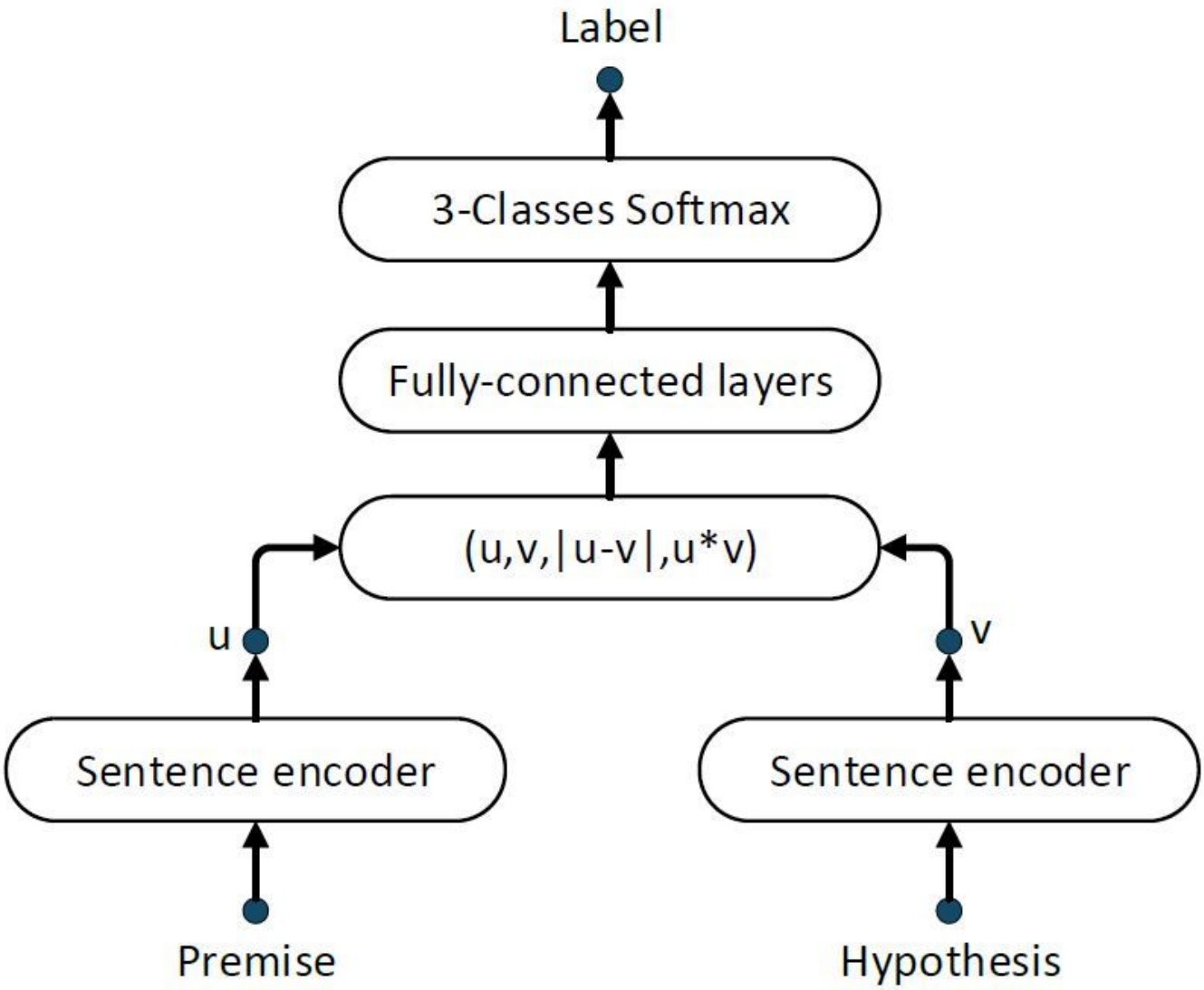

Figure 1

The scheme of a typical NLI model.

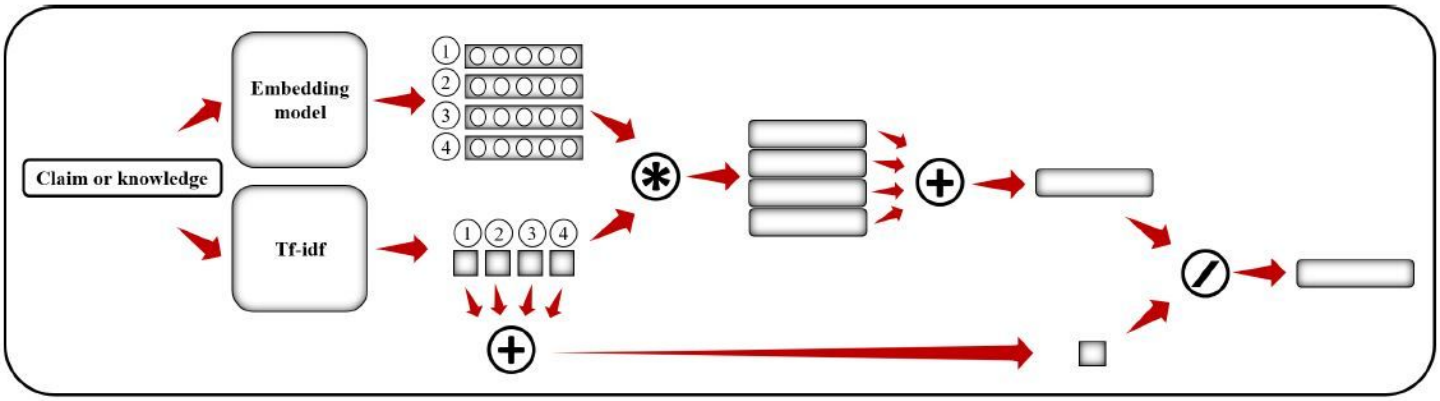

Figure 2

Phrase representation by word embedding and tf-idf.

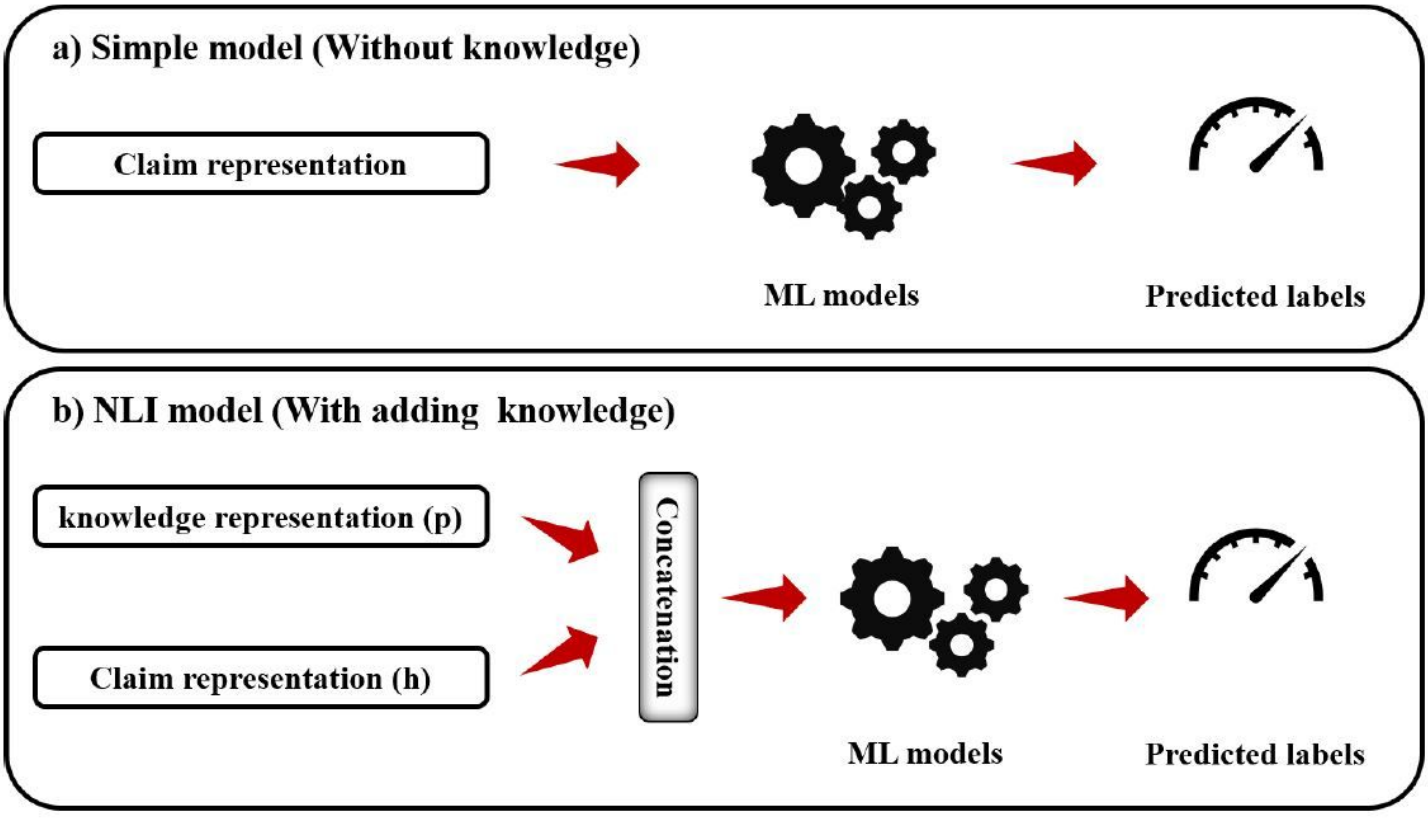

Figure 3

Simple and NLI models based on classical machine learning models.

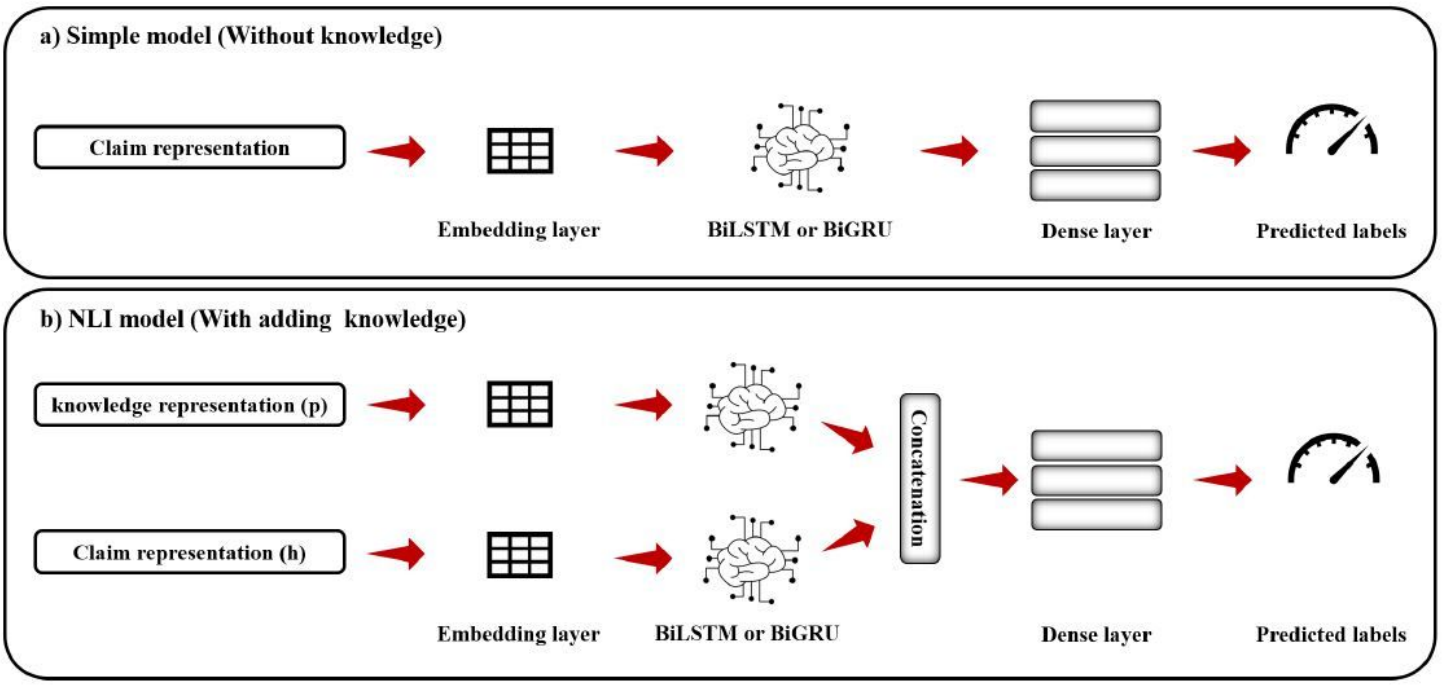

Figure 4

Simple and NLI models based on neural network models.

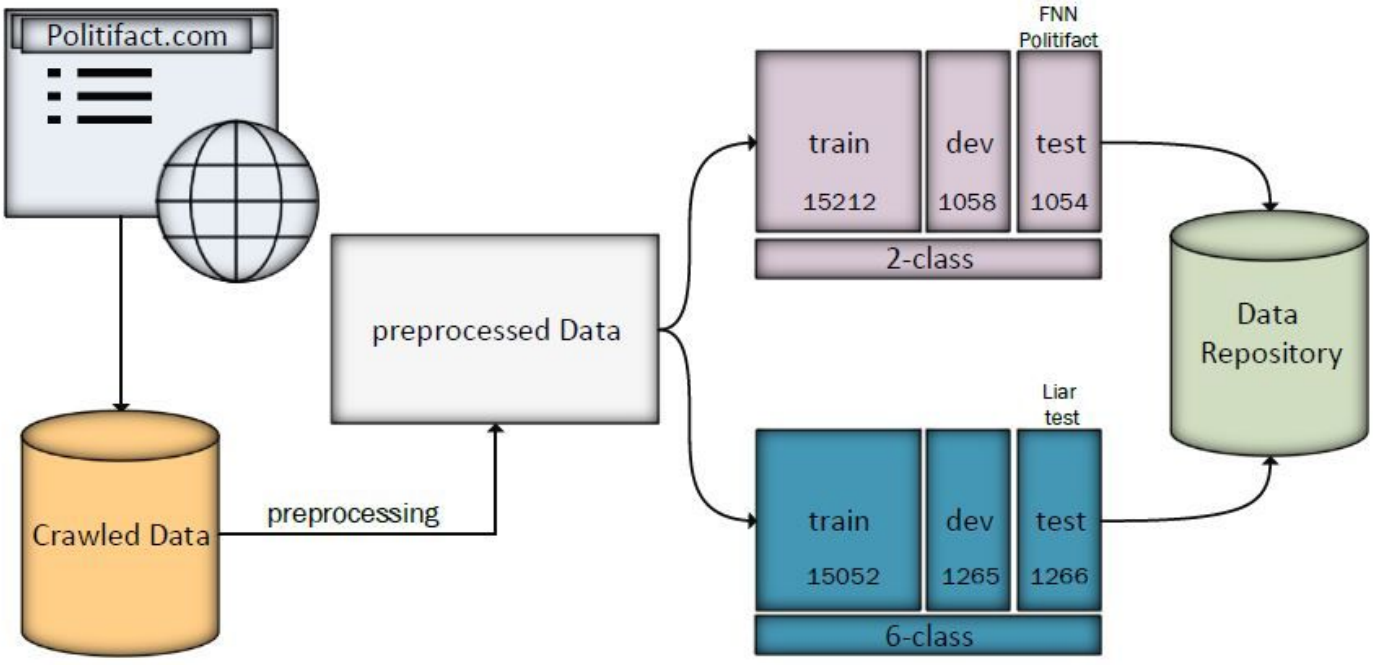

Figure 5

The overview of our dataset construction.

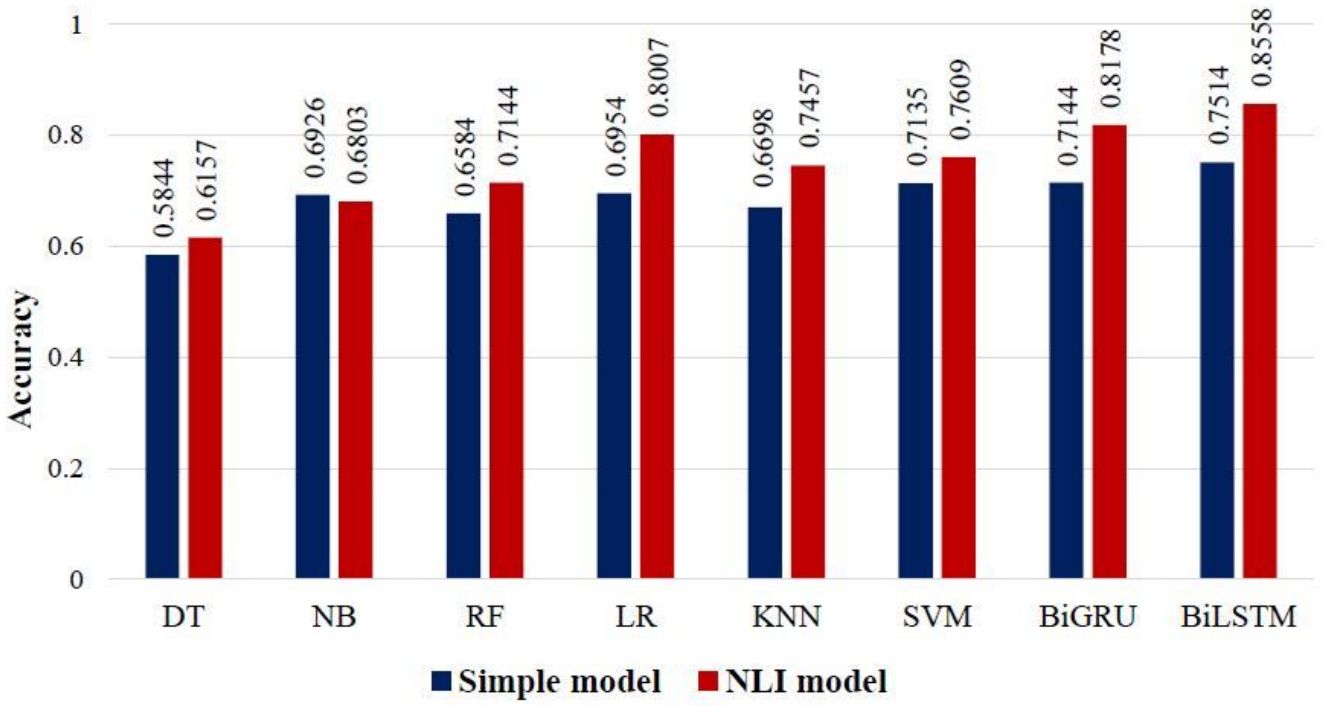

Figure 6

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

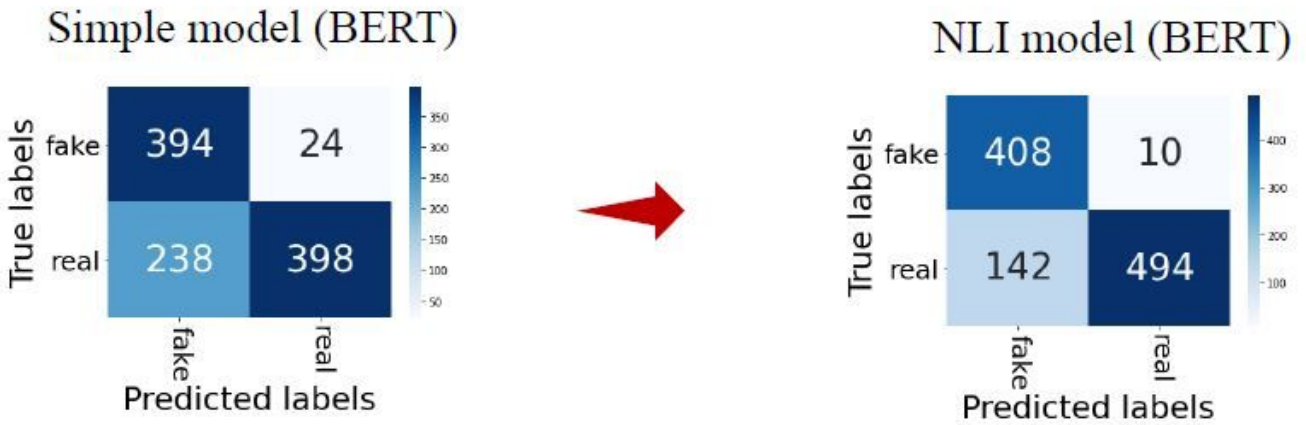

Figure 7

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

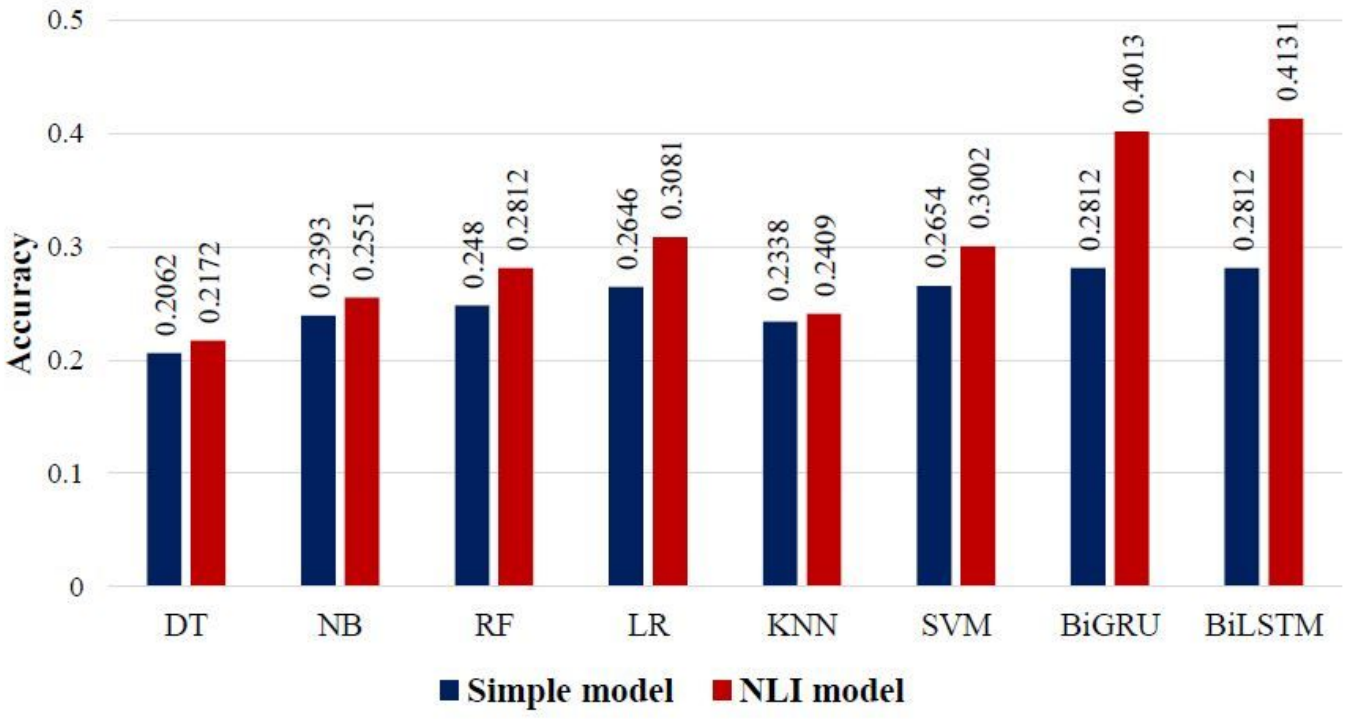

Figure 8

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

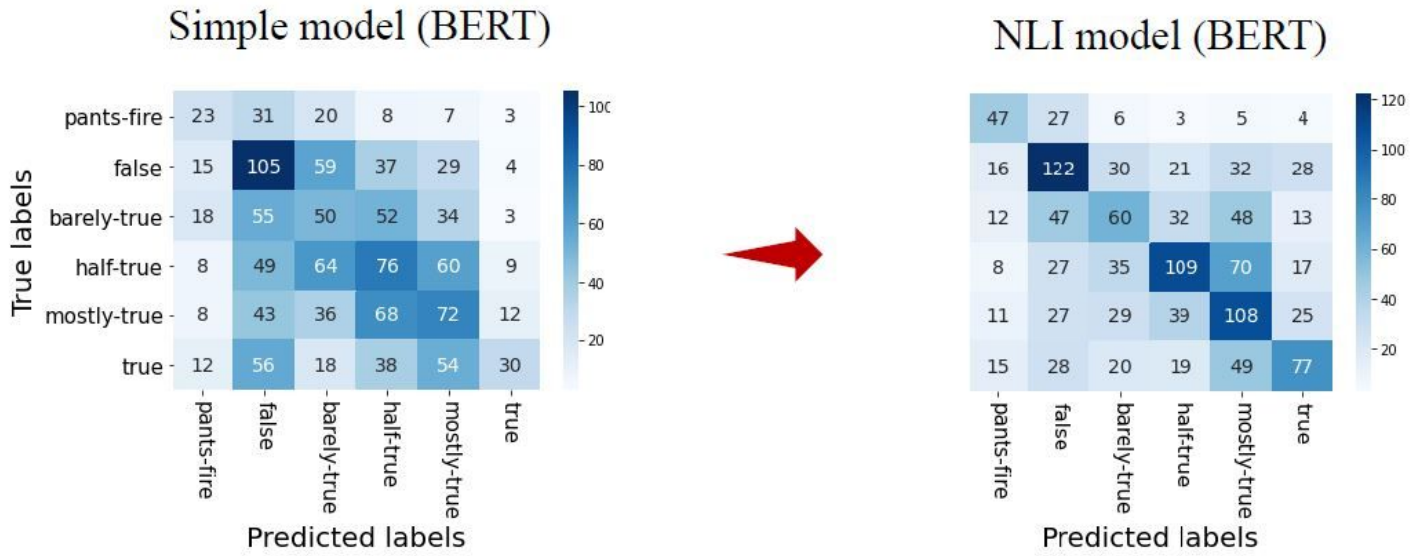

Figure 9

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