Array 19 (2023) 100309

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda

Liyakathunisa Syed [a,*], Abdullah Alsaeedi [a], Lina A. Alhuri [a], Hutaf R. Aljohani [b]

[a] Department of Computer Science, College of Computer Science & Engineering; Taibah University, Madinah, Saudi Arabia
[b] Department of Computer Science, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ARTICLE INFO

Keywords:
Fake news detection
Feature extraction
Deep learning
Weakly supervised learning

ABSTRACT

Due to the emergence of social networking sites and social media platforms, there is faster information dissemination to the public. Unverified information is widely disseminated across social media platforms without any apprehension about the accuracy of the information. The propagation of false news has imposed significant challenges on governments and society and has several adverse effects on many aspects of human life. Fake News is inaccurate information deliberately created and spread to the public. Accurate detection of fake news from cyber propagation is thus a significant and challenging issue that can be addressed through deep learning techniques. It is impossible to manually annotate large volumes of social media-generated data. In this research, a hybrid approach is proposed to detect fake news, novel weakly supervised learning is applied to provide labels to the unlabeled data, and detection of fake news is performed using Bi- GRU and Bi-LSTM deep learning techniques. Feature extraction was performed by utilizing TF-IDF and Count Vectorizers techniques. Bi-LSTM and Bi-GRU deep learning techniques with Weakly supervised SVM techniques provided an accuracy of 90% in detecting fake news. This approach of labeling large amounts of unlabeled data with weakly supervised learning and deep learning techniques for the detection of fake and real news is highly effective and efficient when there exist no labels to the data.

1. Introduction

Social media platforms such as Twitter, Facebook, and WhatsApp have revolutionized information dissemination and have become the source of all aspects of information transmission in an era when the Internet is frequently the primary source of information. According to a survey conducted in 2022, more than 50% of adults in the United Kingdom, France, Germany, the Netherlands, Poland, Bulgaria, Portugal, Romania, Hungary, and Croatia use social media as a source of news [1]. And 82% of the population uses social media for news, despite believing it is untrustworthy. In the United States, several young generations read daily news on social media [1]. However, social networks contribute to the spread of misinformation while also accelerating data dissemination. False information, conspiracies, and rumors can spread quickly on social media, especially in disaster-like situations where alerts and risk-related information are constantly disseminated. The transparency of social networking platforms, combined with automated technology, creates unprecedented challenges by allowing disinformation to spread quickly to many people.

Cyber propaganda is the intentional or unintentional dissemination of false or misleading information through social networking platforms using technology and tools [2]. Cyber propaganda encompasses many concepts, including fake news, disinformation, rumors, spam, hoax, and cyberbullying. All share the same concept of spreading incorrect messages through social media, which can cause distress and negative consequences, mainly when timely intervention is not provided. Fig. 1 depicts the various terminologies adapted for cyber propaganda.

This study focuses on the dissemination of false information through social media as cyber propaganda. Fake news is widely spread through social media networks, such as Twitter, Facebook, and WhatsApp [3]. The increasing popularity of social media raises the possibility of creating and disseminating false information. Users can share and express their views and opinions in a straightforward and unchecked way across social media networks. Some news reports posted or shared on Facebook and Twitter have more views than those reported directly on the news website. Massachusetts Institute of Technology (MIT) researched disseminating fake news on social media sites. Findings revealed that fake news stories propagated 70% faster than real news

* Corresponding author.
E-mail address: lansari@taibahu.edu.sa (L. Syed).

https://doi.org/10.1016/j.array.2023.100309
Received 1 April 2023; Received in revised form 5 July 2023; Accepted 17 July 2023
Available online 20 July 2023

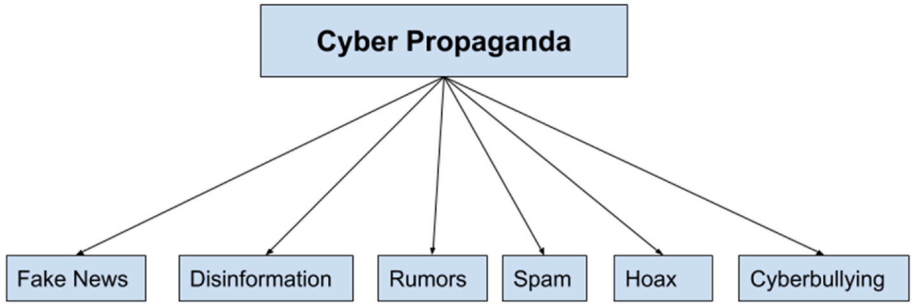

Fig. 1. Cyber propaganda.

stories [4]. It was also discovered that misleading tweets reached people six times faster than genuine tweets [4]. As a result, it has several negative consequences, particularly on the political, social, and economic front lines, and affects many people. Fake news seems to be more probable in inherently unpredictable situations, which include outbreaks of infectious diseases, pandemics, elections, disasters, also fatalities [5]. Even during the 2016 presidential election in the United States, fake and biased news was more prevalent than news from 19 credible sources, according to a recent report [6]. During the COVID-19 pandemic, social media platforms disseminated misinformation about the disease, its causes, potential treatments, and government responses [7]. Therefore, Artificial Intelligence tools and techniques for automated detection of cyber propaganda are significantly required to enforce security to reduce the severity of the problem.

The main contributions of this study are summarized below.

1. We introduce transductive learning, a novel weakly supervised learning technique to annotate unlabeled instances.
2. Feature extraction was performed using TF-IDF and Count Vectorization techniques to convert the text into a vector form.
3. Supervised Machine learning models such as SVM was applied to train and provide labels for unlabeled data through weak supervision techniques. SVM linearly separates fake and real news, providing higher accuracy than other models.
4. Glove word embeddings were employed for the labeled data to provide input to the deep learning models.
5. BiLSTM and Gi_GRU deep learning models were employed to detect fake and real news as they provide high accuracy compared to other models.

The main goal of this study is to detect fake news from cyberpropaganda using our novel weakly supervised learning approach that labels significant amounts of unlabeled data. Deep learning techniques are then utilized to categorize the news as fake or true.

This paper is structured as follows: The second section discusses the most recent related work for detecting fake news with weakly supervised and deep learning techniques. Section 3 describes the methods used to detect fake news in this research study, focusing on weakly supervised learning techniques. Section 4 discusses the suggested Hybrid approach for identifying fake news. The experimental findings are presented in Section 5, and the suggested investigation is wrapped up in Section 6.

2. Related work

Several methods for recognizing false information have been proposed in the literature, employing deep learning and machine learning approaches. Massive volumes of labeled data are required for deep learning approaches in order to train the models [8–12]. In the early stages, collecting such a large volume of labeled training data from social media platforms is difficult. The lack of labeled data has become a prevalent issue in the detection of false information. Hence there is a need for new approaches, such as weakly supervised learning for labeling unlabeled data.

A Hybrid CNN and RNN model was proposed in Ref. [13] to recognize fake news from Twitter data. A private data set of 58,000 tweets from five different rumor stories was utilized. A two-step approach was adopted for fake news detection: first, features were extracted from Twitter posts without prior information about the topic, and then fake news was predicted from Twitter posts comprising text and images using an LSTM and CNN hybrid deep learning model.

The authors in Ref. [14] presented weak supervision with a reinforcement learning technique for the detection of false news. Data was collected from WeChat's official accounts along with users' feedback as reports. The framework included a fake news detector, a reinforced selector, and an annotator. By assigning weak labels to unlabeled news articles in response to user input, the annotator was trained on a small collection of labeled fake news. High-quality samples were selected, and low-quality samples were filtered out through a reinforced selector. Additionally, the CNN model was fed with the extracted features to classify false information.

Another Weakly Supervised Learning technique on Twitter was presented in Ref. [15]. The training dataset was created by the first compilation of reliable and untrustworthy sources. The Twitter API was used to retrieve tweets from each source, every tweet from a reputable source was labeled as true news, and it was labeled as false news from an untrustworthy source. A tiny, hand-marked dataset from the PolitiFact website has been used for evaluation purposes. Decision Trees, Naive Bayes, Neural Networks, SVM, Random Forest, and XGBoost, were applied as generic classifiers for model training. With XGBoost, the highest accuracy was obtained.

In [16], a weak social supervision technique for identifying disinformation was proposed. FakeNewsNet benchmark dataset was utilized, which was collected from PolitiFact and GossipCop fact-checking websites. Weak labels were generated through three statistical measures such as sentiment, bias, and credibility. In sentiment based on whether the user's sentiment scores a standard deviation is higher than the threshold, the news is weakly branded as fake news. In Bias-based systems, if the average user's absolute bias score is higher than the threshold, it is weakly classified as fake news. In Credential-driven, users have been clustered using hierarchical clustering. The average credibility score below the threshold was weakly labeled as false news. For the classification of false information, CNN and robustly optimized BERT (RoBERT) model, along with multi-source Weak Social Supervision (MWSS), were used. CNN-MWSS provided an accuracy of 77% for GossipCop data and 82% for PolitiFact data, whereas RoBERTa-MWSS provided an accuracy of 80% for GossipCop data and 82% for PolitiFact data.

Early rumor detection using a neural language model and weakly supervised learning for data augmentation was presented in Ref. [17]. Publicly available rumor datasets-Pheme, CrisisLexT26, Twitter event dataset, SemEval-2015 task1 data, CREDBANK, and SNAP data were used on six events. The input corpus was divided into two collections: "References" and "Candidates." A small portion of the source tweets was utilized as reference data (Candidate) to provide weak supervision to

Table 1
Comparative analysis of related work.

Reference	Data	Data Annotation	Classification Models	Accuracy	Limitation
[13]	Twitter	–	CNN-LSTM	80%	Despite using labeled data, the neural network model's inadequate training resulted in poor performance.
[14]	WeChat feedbacks	Weak Supervision	Reinforcement Learning, CNN	82%	comprises a lot of nosy data, and including weak supervision may generate more false positives. As a result, a reinforcement selector was utilized to enhance performance.
[15]	Twitter	Weak Supervision	XGBoost	89%	Inaccurate labels
[16]	GossipCop data, PolitiFact data	Weak Supervision	RoBERTa-MWSS	80% 82%	Inaccurate unclean dataset
[17]	Rumor Dataset	Weak Supervision	bidirectional Language Model	75%	Thresholding was considered to enhance performance because normalizing the text decreased the neural language model's performance.
[18]	New Articles	Weak Supervision	XGBoost, Logistic Regression ALBERT, XLNet, and RoBERTa	79.3%	The performance of the classifier decreased whenever additional weak labels were included.
[19]	LAIR dataset	–	Ensemble Bi-LSTM – GRU model	89%	Publicly available open-source data set was used. Data annotation was performed through weak supervised learning

unlabeled tweets. Furthermore, they were domain-specific and classify non-rumors and rumors from candidate tweets. An accuracy of 63%–75% was obtained for PHEME5 for the six events.

The authors in Ref. [18] presented a weakly supervised learning technique based exclusively on content features for fake news detection. They prefer news articles to social media posts. A probabilistic weak labeling system was employed to label the content features, and a test set from the fact-examination groups - PolitiFact3 and Snopes2 was used. Five machine learning classifiers on content features were applied for the classification of fake news with and without weak labels to explore the efficacy of applying weakly supervised learning. For the test set, RoBERTa with weak labels achieved the highest accuracy of 79.3%.

Using the labeled LIAR dataset, authors in Ref. [19] demonstrated an ensemble-based deep learning model to recognize news as real or false. Two deep learning methods, Bi-LSTM and GRU, were utilized for the statement attribute, while dense, deep learning models were applied for the other attributes. With the suggested approach, an accuracy of 89% was attained.

It was observed from Table 1 and related studies that existing weakly supervised techniques adversely affects the classifier performance due to noisy data and erroneous labels [14–16,18], the limitations of each technique have also been provided in Table 1. Whereas in Refs. [13,19], data annotation and weakly supervised learning was not performed to label unlabeled data, furthermore the performance did not increase significantly even after applying deep learning techniques to the open-source data. The limitation of the existing techniques motivated us to propose novel weakly supervised learning techniques to label large amounts of unlabeled data. Additionally, it was also observed in Ref. [19] that ensemble BiLSTM-GRU deep learning techniques provided better performance compared to ALBERT and RoBERTa. The key benefits of Bi-LSTM and BI-GRU deep learning approaches are that they eliminate vanishing gradient issues and are designed for sequential data processing, as well as preserve information for long-term dependencies. As a result, in this research work, BiLSTM and BiGRU deep learning approaches were adopted for detecting false information.

3. Methodology

This section discusses the weakly supervised learning technique used for data annotation.

3.1. Weakly supervised learning

Most efficient deep learning and machine learning techniques in several tasks require providing ground-truth labels for a large set of training data. But, due to the high expenses of the data labeling procedure, accurate supervision information may be challenging to attain. As a result, machine learning and deep learning techniques must be able to work with limited supervision. Weakly supervised learning is a machine learning approach that uses partially annotated data to build models. There are three kinds of weak supervision: The first is incomplete supervision, which involves labeling only a subset of training data while leaving the rest unlabeled. The second is incorrect supervision that includes labeled training data not as reliable as expected, and the third category is inaccurate supervision which contains only some labels with errors in the training data [20]. In this research, incomplete supervision is used for the labeling of unlabeled Twitter data.

Incomplete Supervision: There is a partial amount of labeled data available, but there is an enormous amount of unlabeled data that would be insufficient to train a good learner. Formally, given a limited number of labeled instances, the task is to train unlabeled instances. The detection of fake news is viewed as a problem of binary classification

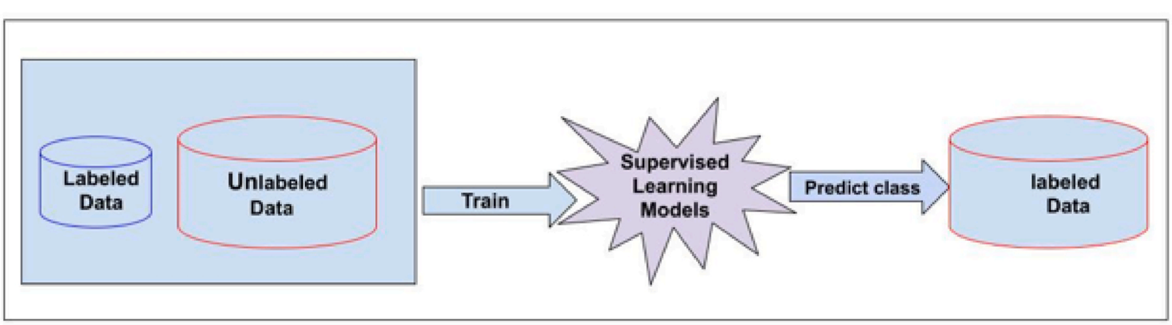

Fig. 2. Transductive learning.

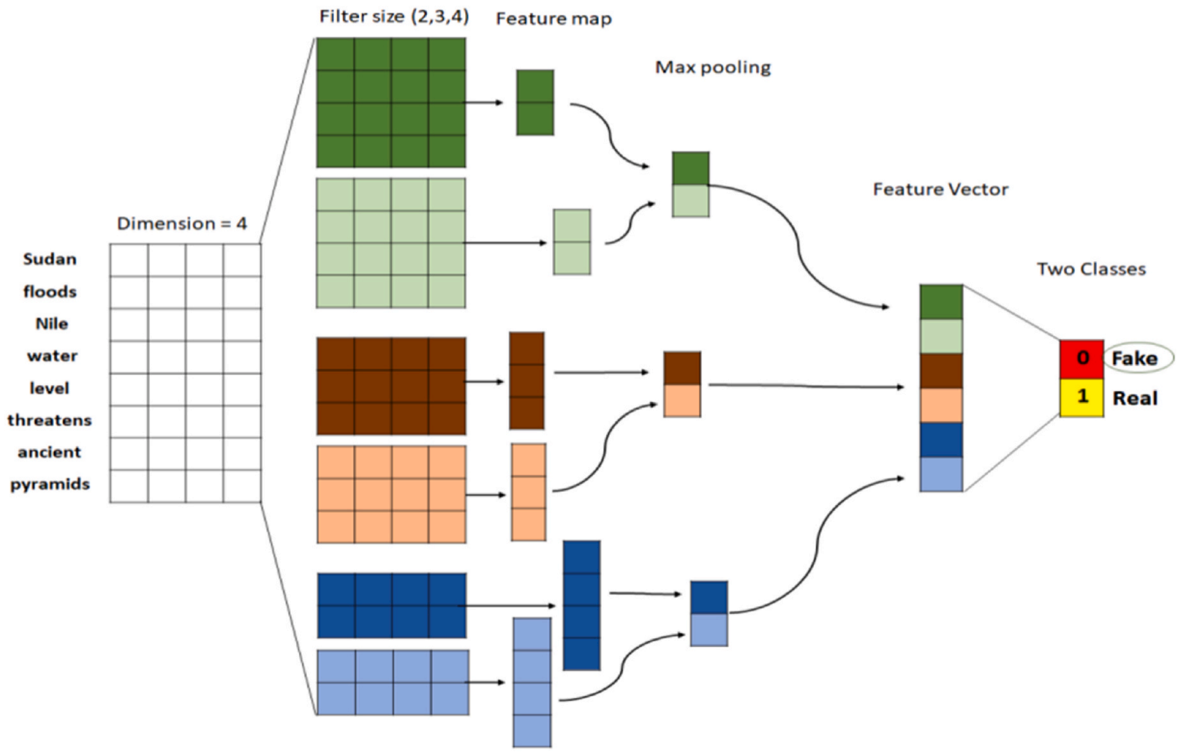

Fig. 3. Convolution Neural Network architecture for fake news classification.

with two interchangeable classes of 'fake' and real labels. The formal task is to learn $f: U \rightarrow V$ from dataset D presented in Eq. (1).

$$D = f\{(U_1, V_1)\ldots(U_n, V_n), U_{n+1}, V_{n+1}, \ldots, U_m\} \tag{1}$$

Where U_l represents the labeled training data and V = n-l represents the unlabeled data.

Two major techniques are adopted by incomplete supervision to label unlabeled data, namely active learning [21] or semi-supervised learning [22–24]. Active learning uses domain experts to provide labels for unlabeled samples. Semi-supervised learning, on the other hand, seeks to improve learning performance by leveraging both labeled and unlabeled data through techniques such as generative models, label propagation, and transductive learning.

In this study, we will use transductive learning to predict labels for unlabeled data. Transductive learning uses supervised learning algorithms to train the models [25] and predicts the categories for the data. Fig. 2 shows the transductive learning approach.

In this research, supervised machine learning models, such as Random Forest, SVM, Logistic Regression, and Naïve Bayes, are utilized to predict class and provide labels as Fake or Real for unlabeled Twitter data from social media.

3.2. Deep learning techniques

A detailed description of different deep neural network architectures utilized in this study are discussed in the following sections.

3.2.1. Convolutional Neural Networks (CNN)

Convolutional Neural Networks are made up of neurons with learnable parameters, namely biases and weights. CNN assumes that each input has a grid pattern, such as images, which enables certain features to be encoded into the architecture. CNN is very similar to dense neural networks, except the hidden layers consist of layers of convolution and pooling accompanied by one or more fully connected layers (i.e. dense layers) at the end. Feature extraction in CNN is performed through the convolution layer, which consists of a stack of mathematical operations [26]. In computer vision, the filters typically slide over patches of an image (i.e. pixels), but in NLP and text classification the filters slide over a matrix where the width is similar to the width of the tweet and the height may vary. Fig. 3 illustrates the CNN architecture for fake news classification. The input will be a word, sentence, or even a whole document. To inject the text into the matrix, the input must be represented as vectors, these vectors typically are word embedding. Word embedding is a method for mapping each word into a vector space and represented as real-valued vectors based on the use of this word. There are a variety of techniques that can be used to obtain word embedding from text such as Word2vec and Global Vectors (GloVe).

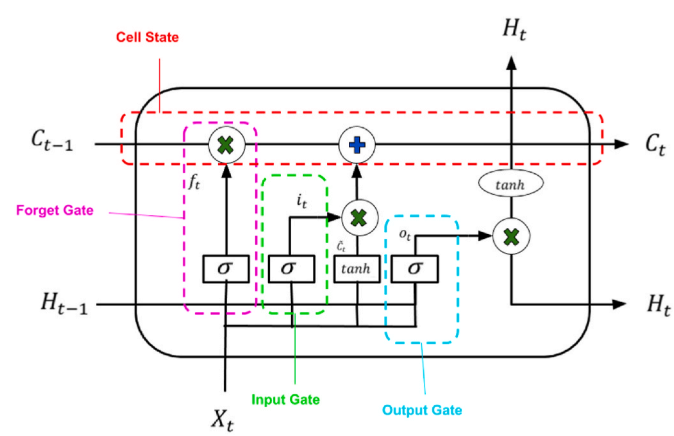

Fig. 4. Lstm neural network.

3.2.2. Long short-term memory (LSTM)

LSTM is a deep neural network with sequential processing that allows information to persist. LSTM shows superior ability when handling long-term dependencies of data. The LSTM units attempt to solve the

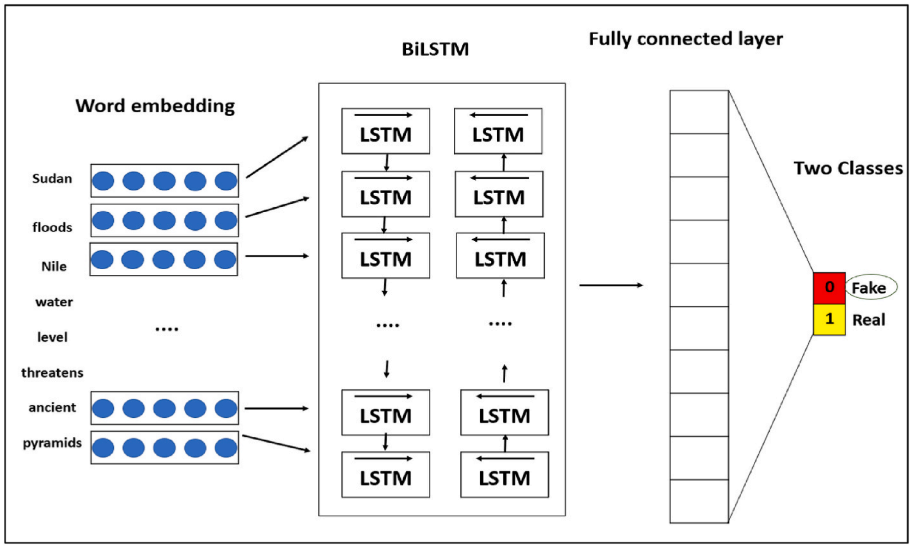

Fig. 5. Bi-LSTM architecture for fake news classification.

problem of standard RNN following gating mechanisms, therefore, the units are composed of several gates that give them the ability to remember past information. The LSTM cells are types of recurrent units that can control the flow of information by controlling input gates, forget gates, and output gates, and then use different functions to update the memory state. This architecture has been widely used in NLP tasks since the LSTM units can handle long sequential data. LSTM is one of the most popular RNN architectures and has attained excessive success with sequential data such as text data, and time-series data. The gating mechanism attempts to solve the gradient vanishing problem using four gates: input, output, forget, and memory. Fig. 4 presents the internal structure of the LSTM cell.

The main idea behind LSTM is to control how much information to preserve from the previous state. Given the previous memory (C_{t-1}), the new memory (i.e., cell state) is computed using Eq. (2).

$$C_t = f_t * C_{t-1} + i_t * \widehat{C_t} \tag{2}$$

Forget gate: decides which information from the prior memory is required in the current memory. Eq. (3) shows how to compute each cell state at the t time step.

$$f_t = \sigma_g\left(W_f \bullet [X_t, H_{t-1}] + b_f\right) \tag{3}$$

The next two gates determine which information is to be processed in a cell state. The input gate decides which information needs to be updated as new candidate values, while the memory gate generates a vector of the new candidate values.

Input gate:

$$i_t = \sigma_g\left(W_i \bullet [X_t, H_{t-1}] + b_i\right) \tag{4}$$

Memory gate:

$$\widehat{C_t} = tanh\left(W_c \bullet [X_t, H_{t-1}] + b_c\right) \tag{5}$$

Output gate: decides which information from the cell states are forwarded to output, it is calculated as illustrated in Eq. (6).

$$o_t = \sigma_g(W_o \bullet [X_t, H_{t-1}] + b_o) \tag{6}$$

Then, the new hidden state is updated by putting the cell state through *tanh* activation function, and multiply it by the output gate, computed using Eq. (7).

$$H_t = o_t * tanh(C_t) \tag{7}$$

The LSTM can be modified to address limitations in other areas by using Bidirectional LSTM (Bi-LSTM) layers. This technique not only tries to feed the sequences forward but also feeds them backward. Bi-LSTM can better interpret words because of the intuition that future words alter the meaning of existing words (e.g., polysemous words) in language. The Bi-LSTM also shows some improvement in computational complexity compared to the LSTM [27].

The tweets must be represented in a vector space like CNN, where each entry is an index of a word in a GloVe dictionary. A diagram of Bi-LSTM architecture for fake news classification is shown in Fig. 5.

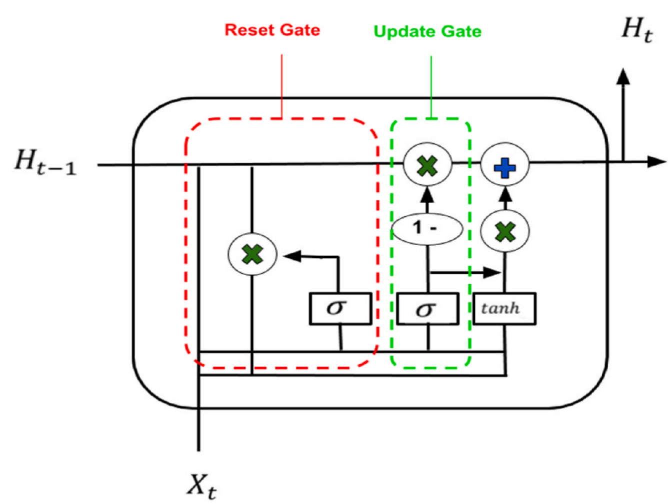

Fig. 6. Gru neural network.

3.2.3. Gated Recurrent Unit

Gated Recurrent Units (GRU) is very similar to LSTM but easier to implement. The same gating mechanism implemented in the LSTM is followed in this network. They have shown comparable performance, but GRU reduces the gating signals to two from the LSTM as shown in Fig. 6. Conversely, with a less external gating signal, it can train a bit faster than other Recurrent Neural Networks [28,29].

GRU computes only two gateways to monitor the flow of information. Given the cell state at previous time step t and the output of reset gate (H_{t-1}), the new memory is computed as illustrated in Eq. (8).

$$\widehat{H_t} = tanh\left(W \bullet [r_t * H_{t-1}, X_t]\right) \tag{8}$$

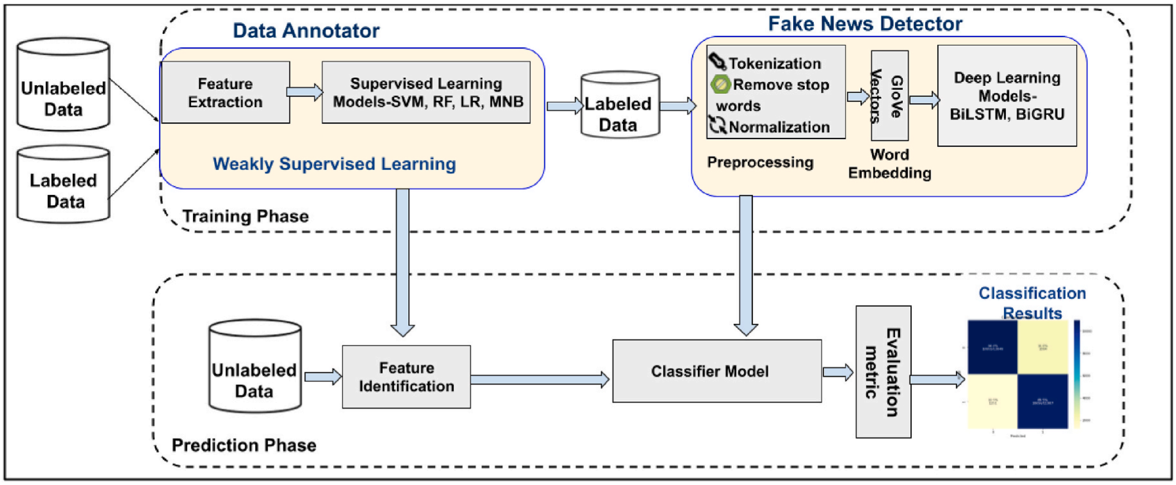

Fig. 7. Framework for detection of fake news.

Table 2
Dataset collection.

Fake	Real	Keyword	Collection date
Com	Gov	Biden	July 13, 2020–July 22, 2020
	Gov	Kanye West	
	Gov	Trump	August 9, 2020–July 22, 2020
	Edu	#DistanceLearning	August 1, 2020–March 1, 2021
	Edu	#OnlineClasses	
	Gov - WHO	#CoronaVaccine	
	Gov - WHO	Covid-19 Vaccine	
	Gov - WHO	Covid-19 (reference)	August 13, 2020–August 14, 2020
	Gov - Org	Israel	August 13, 2020–May 17, 2021
	Gov - Org	Palestine	
	Gov - Org	Lebanon	August 4, 2020–August 13, 2020
	Gov - Org	Sudan	September 7, 2020–September 9, 2020

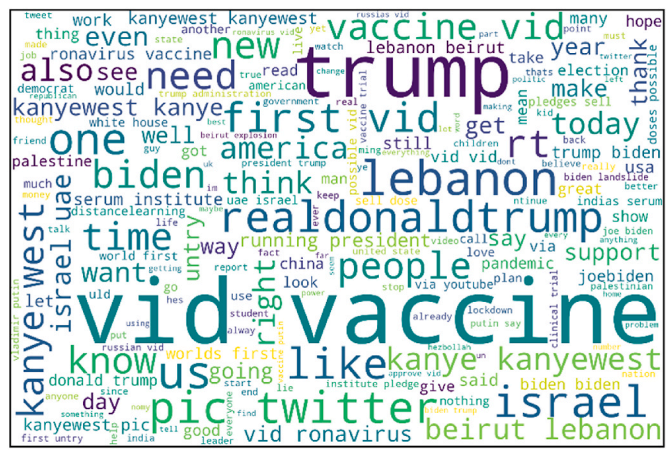

Fig. 8. WordCloud for the events selection.

Reset gate: computes the relevance of previous memory in the calculation of current memory.

$$r_t = tanh\ (W_r \bullet [H_{t-1}, X_t]) \tag{9}$$

The reset gate output is then used as a candidate for the updated candidate $\widehat{H_t}$ to calculate the current cell state.

$$H_t = (1 - z_t)\ * H_{t-1} + z_t * \widehat{H_t} \tag{10}$$

Update gate: decides which information from the updated candidate is required to the current memory.

$$z_t = \sigma\ (W_z \bullet [H_{t-1}, X_t]) \tag{11}$$

The GRU also can be modified to overcome the language limitations by using Bidirectional GRU layers (Bi-GRU). This approach is very similar to Bi-LSTM, where the model can understand the context better by preserving the sequence in two directions (from left to right, and vice versa). Moreover, there are other reduced gated RNN, e.g., the Minimal Gated Unit (MGU).

4. Fake news detection system

Due to the adaptive nature of social media data and the availability of large amounts of unlabeled data, analyzing Twitter data is a difficult task. This emerging situation has motivated us to propose weakly supervised and deep learning techniques for the detection of False News.

The proposed methodology compromises two phases: training and prediction. As shown in Fig. 7, the training phase includes two major components: (i) weakly supervised learning for data annotation and (ii) deep learning techniques for fake news detection. The following sections discuss the comprehensive description and implementation of each component of the training and prediction process.

4.1. Dataset collection

In this research study, we constructed our dataset to achieve our task. It contained tweets regarding some of the wildly spread news, which is the interest of the majority of the public. The dataset was collected using tweepy library that is used to access the twitter API. We collected data using keywords and hashtags related to the event chosen. The created dataset consists of 195,943 tweets in total. After eliminating noisy data, only 1200 of the tweets were classified into two categories Real and Fake News. The process of annotation was based on the URL and keywords included in each tweet. Following the set of URL criteria for fake and real news, shown in Table 2. After annotating the whole dataset, we cleaned and balanced the data to be utilized for model training and testing purposes. The events included in the data are mainly focused (USA election, COVID-19, Lebanon explosion, Israel and Palestine news, Distance Learning, Sudan floods) during the COVID era (2020–2021).

4.2. Weakly supervised learning for data annotation

Twitter data from social media platforms was collected for six

Table 3
Example of Count vector-matrix.

Trump	elections	promises	Biden	vaccine	asking	to
1	2	1	2	1	1	2

significant events. The events included in the data were selected based on WordCloud generated for the most widely spread news on social media platforms, illustrated in Fig. 8, which is the interest of the majority of the public. A small proportion of the data was labeled as "fake' and real' using URLs from trustworthy sites with com, gov or org domains. Labeled data was used to provide labels for unlabeled data, known as weak labels, without human intervention. The labels were generated automatically by supervised machine learning techniques using partially labeled data. This process is called weakly supervised learning. Transductive weakly supervised learning is adopted in the projected approach to provide labels for the unlabeled data, discussed comprehensively in section 3.1. The dataset was collected using the tweepy library that is used to access the Twitter API. Data was gathered using keywords and hashtags related to the event chosen. Some of the COVID-19 tweets were obtained from an available dataset provided by Ref. [30]. It has been retrieved using the Hydrator tool that extracts tweets from Twitter based on tweet I.D.s in the available coronavirus dataset. The created dataset consists of 195,943 tweets in total.

It is pretty challenging to train the model using textual data. To perform a text data analysis, it must be transformed into numerical features. TF-IDF and Count Vectorizer methods have been applied to convert the raw text to extract features.

TF-IDF Vectorizer: Term Frequency (T.F.) and Inverse Document Frequency (IDF) are two components of TF-IDF. Used to obtain a value that shows the significance of a word in a text. Therefore, its numerical value of it increases with the term that frequently occurs in the document. It consists of two values.

Term frequency: It records the number of times a word appears in a document donated by Eq. (12).

$$TF(x, Y) = f_{x,Y} \tag{12}$$

Inverse document frequency: It illustrates how common a term is and how much information the term contains, illustrated in Eq. (13).

$$IDF(x, Y)\frac{N}{|\{y \in Y : x \in y\}|} \tag{13}$$

Providing us with the final equation of TF-IDF provided in Eq. (14). The number of documents is represented by 'Y,' and the number of documents containing the corresponding term is represented by 'x.'

$$TFIDF(x, y, Y) = TF(x, y)\ .IDF(x, Y) \tag{14}$$

High TF-IDF has a high frequency in the document and is not a common word across other texts that are significant for the document in consideration.

Count Vectorizer (CV): It splits the text form of the document and returns the vector representation of the term along with the count of the occurrence of the tokenized terms in a matrix form, an example is shown in Table 3.

Various supervised machine learning (M.L) techniques such as Random Forest (R.F.), Logistic Regression (L.R.), Naive Bayes, and SVM were applied to the extracted features using TF-IDF and Count Vectorizer techniques to obtain weak labels through the use of transductive weakly supervised learning to label unlabeled data. Table 4 displays the results of data annotation through weakly supervised learning. The SVM model was the most accurate, with an accuracy of 85%, followed by Logistic Regression with the TF-IDF feature extractor, which had an accuracy of 84%. Multinomial Naive Bayes, on the other hand, achieved a lower accuracy score of 77%. The SVM weakly supervised transductive learning technique's generated labeled data is being considered for the detection of fake news through deep learning techniques.

Further statistical tests were conducted to verify the accuracy of the annotated labeled data. The statistical test determines whether the two samples or features are significantly different from the others. A *t*-test with null hypothesis was used. The aim of the *t*-test was to assess if the measured value of alpha was smaller compared to the assumed one (i.e., P-value = 0.05), if so, the null hypothesis was rejected, indicating that there was in fact a substantial variance between both of the features or samples. Alternatively, if the P-value was higher or equal to the assumed one, the null hypothesis was considered valid, indicating that no significant variation existed between the two features [31,32]. We obtained a P-value of 0.43 which was much higher than the assumed one of 0.05, hence the null hypothesis was accepted and is considered statistically significant.

4.3. Fake news detection using deep learning techniques

The fake news detector module consists of preprocessing, word embedding, and classification using deep learning techniques.

(i) Preprocessing

Text data needs proper preprocessing for the implementation of deep learning algorithms. Stopword removal is performed as a preprocessing stage. Stopwords are commonly used and make the majority of the tokens in the dataset, and do not contribute much to the overall meaning of the sentence like connectors, for example, "and," "or" and "but." These stopwords are removed from each tweet, along with URLs and numbers. Then, every tweet is split into a set of tokens, and each letter in the token is converted into a lower-case letter. These steps are necessary to increase the number of matching tokens between our dataset and GloVe pre-trained model tokens.

(ii) Word Embedding

Word embedding is an efficient method to generate vector representations of words and documents. Word2Vec [33] and Global Vectors (GloVe) [34] are the two popular deep-learning methods of word embedding. In this research, GloVe is utilized to obtain a vector representation of words. GloVe, which is a word taken from Global Vectors, is a pre-trained model for word representation. It is a count-based model that generates vectors by lowering the dimensionality in a co-occurrence matrix. The matrix holds the information on the frequency of each word. Then, this matrix is log-smoothed and normalized. One of the benefits of using GloVe as a pretraining model is that it trains fast. GloVe has five pre-trained models-Wikipedia 2014, Gigaword5, two common crawls, and Twitter. Each of these models is available in sizes ranging from 25 to 300 [35]. In our experiment, we will utilize the Twitter pre-trained

Table 4
Results of Data Annotation using Weakly Supervised Learning.

Feature extraction	ML-Models	Number of features	Accuracy	F1 score	Precision	Recall
CV	Multinomial Naïve Bayes	2000	77%	77%	71%	76%
CV	Random forest	2200	83%	82%	77%	80%
CV	Logistic Regression	2200	84%	84%	79%	84%
TF-IDF Vectorizer	SVM	2200	85%	85%	79%	85%

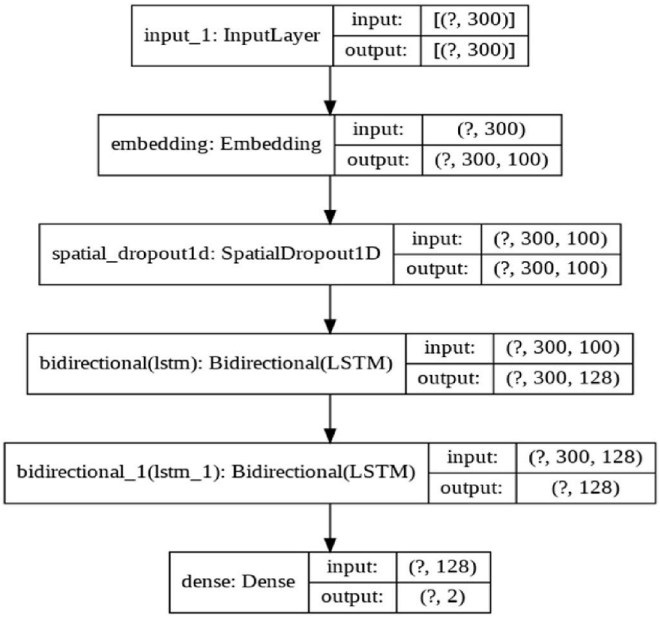

Fig. 9. Bi-LSTM model architecture.

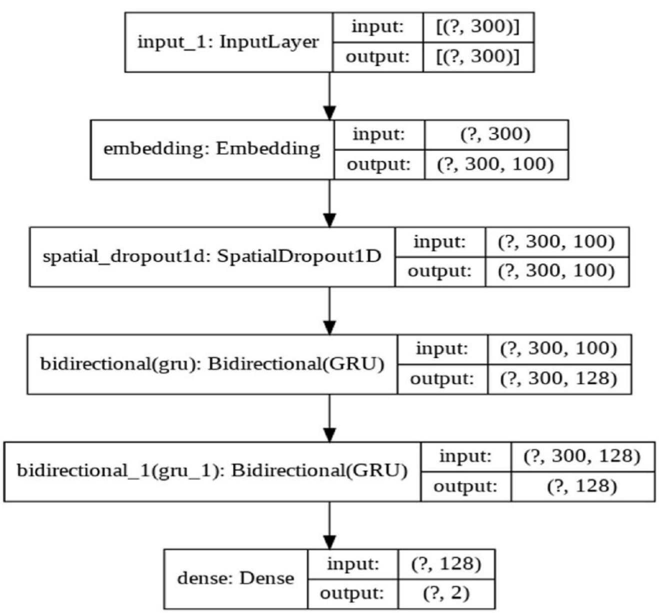

Fig. 10. Bi-GRU model architecture.

model with 100 dimensions. Utilizing the knowledge from an external word embedding (i.e., GloVe) can enhance the performance of deep learning algorithms.

(iii) Classification

After employing word embeddings (i.e., GloVe) to represent inputs as matrices, the results will be used to classify fake news through deep learning techniques such as BiLSTM and BiGRU discussed in section 3.2.2 and 3.2.3, furthermore it is compared with LSTM and CNN models. The embedding layer is fed as input to dense layers of the deep learning models to predict the target instances as ‘real’ or ‘fake’, illustrated in Figs. 9 and 10. Once the models are trained, we can then evaluate how well each model predicts accurate results with our dataset. This evaluation is performed by measuring a set of validation metrics such as Confusion Matrix, accuracy, recall, and precision [36].

Table 5
Performance of weakly supervised with deep learning models for fake news detection.

Weakly Supervised Learning Models for data annotation	Deep Learning Models for classification of fake news	Accuracy (%)	F1 score (%)	Precision (%)	Recall (%)
SVM	Bi-LSTM	89.95	89	89	89
	LSTM	86.19	86	86	86
	Bi-GRU	89.60	89	89	89
	CNN	82.06	82	82	82

5. Experimental results

Experiments were performed through our proposed hybrid approach, which includes transductive weakly supervised learning for data annotation and deep learning approaches for the classification of fake news. The data was split into a ratio of 80:20 for model training and testing. Using the TensorFlow framework. The deep learning techniques applied in the implementation are discussed below.

5.1. Bi-LSTM deep learning model

The BiLSTM model was trained for 40 epochs and the Adam optimizer was applied since it produced optimized results and was designed for 128 units. A SpatialDropout1D of (0.1) rate was added to the input layer to minimize the problem of overfitting. Followed by two LSTM layers to extract features from the input text. Both layers have bidirectional extensions. The model is able to recognize the context by retaining input information in the hidden state from the past to the future and from the future to the past. At every timestep, the network averages the outputs of the LTSM node to obtain hidden representation. Hidden representations are then sent to the fully connected layer to classify the input into one of the two classes, ‘Fake’ or ‘Real’. Fig. 9 illustrates the implemented Bi-LSTM model architecture.

5.2. Bidirectional-Gated Recurrent Unit (Bi-GRU) deep learning model

In this model, two bidirectional Gated Recurrent Unit (GRU) layers with 128 nodes were utilized to build the bidirectional extensions. To minimize the overfitting issue, a SpatialDropout1D rate (0.1) was added. The second GRU layers are further connected to dense and fully connected layers with sigmoid activation to generate predictions. The model was trained for 40 epochs and was optimized using Adam optimizer. A pictorial representation of the implemented Bi-GRU model architecture is shown in Fig. 10.

Table 5 presents the experimental results achieved through the application of the of our proposed novel weakly supervised learning for data annotation and deep learning techniques for the detection of fake news. Various machine learning techniques such as Naïve Bayes, Random Forest, SVM and logistic regression were applied to provide labels for unlabeled data using our proposed transductive weakly supervised learning technique, a detailed discussion is provided in section 4.2. SVM provided the highest accuracy compared to other machine learning models for data annotation hence was considered for data annotation. Furthermore, several deep learning algorithms such as BiLSTM, BiGRU, LSTM and CNN were applied for classification of fake news. It was observed that Bi-LSTM and BiGRU deep learning model provided a higher prediction accuracy of 90% than the other deep learning techniques, namely CNN and LSTM models. Bi-LSTM produces results slightly higher than Bi-GRU by almost 0.35%. CNN has low performance compared to other deep learning models. The decrease in

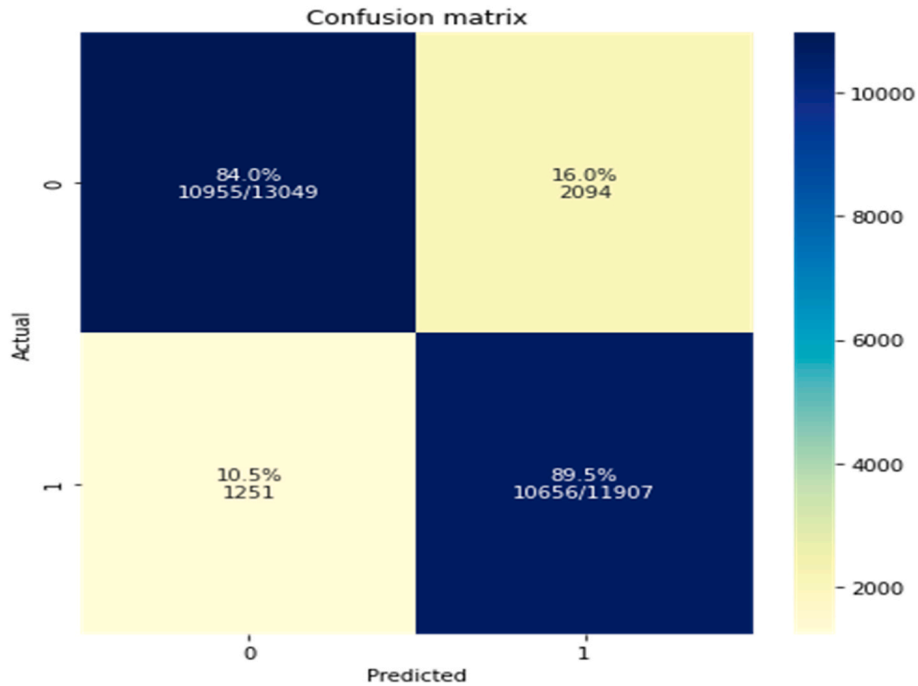

Fig. 11. Confusion matrix results with BiLSTM deep learning technique.

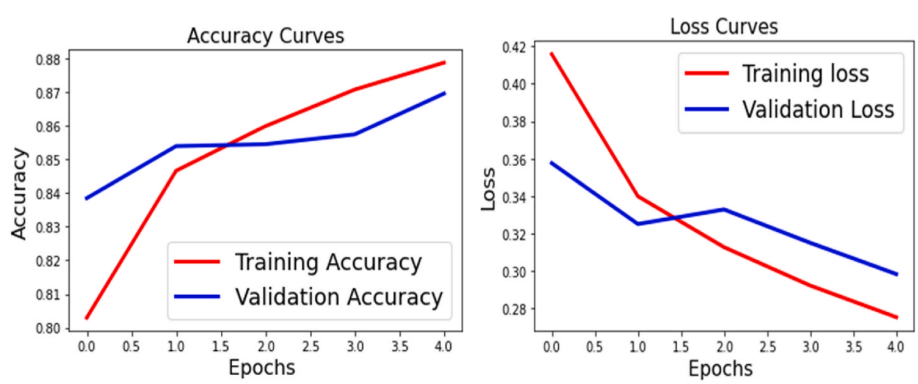

Fig. 12. Accuracy and loss curves of Bi-LSTM with Transductive SVM model.

performance with CNN is due to the complexity of the neural network. CNN extracts high-level features, LSTM captures long-term dependencies, and Bi-LSTM and Bi-GRU retain long-time information by propagating in both directions without duplicate representation. Hence Bi-LSTM and Bi-GRU performance is better. Bi-GRU is similar to Bi-LSTM with a lesser number of gates.

Fig. 11 presents the confusion matrix for detecting fake news from social media data through our proposed hybrid Bi-LSTM deep learning technique for classifying the news as 'Real' or 'Fake' and weakly supervised learning with an SVM model for data annotation. For the binary classification problem, our model accurately predicts 10,955 samples as Fake (0) and misclassifies 2094 twitter samples as false labels accounting for 16%. Whereas it accurately predicts 89.5% accounts for 10,656 news samples as Real (1) and misclassifies 1251 samples account for 10.5%. This shows approximately 84 of the correctly classified samples as 'Fake' as 0 and 89% of correctly classified samples as 'Real' as 1.

Fig. 12 presents the Accuracy and loss curves for Bi-LSTM with SVM transductive weakly supervised learning model. Training accuracy of 89% was achieved, the validation accuracy increased along with the training accuracy, and attained an accuracy of 87% for the test dataset Simultaneously the validation loss i.e. error rate started to reduce along with training loss. The error rate between the two plots is very minimal with 0.023, this shows that the model is neither overfitting nor underfitting, it is an optimized model. Similar performance was achieved for Bi-GRU with SVM model.

The limitation of the proposed weakly supervised learning is that, only machine learning techniques such as Naïve Bayes, Logistic Regression, Random Forest and SVM were considered for data annotation as it provided accurate results and are much simpler compared to the deep learning technique. In future we intend to investigate more sophisticated deep learning techniques in order to provide high-quality weak labels and enhance the performance of our models.

6. Conclusion

This paper proposes a hybrid approach utilizing deep learning and weakly supervised learning techniques for detecting fake news. SVM-based weakly supervised learning achieves higher accuracy and accurately labels large amounts unlabeled data through the proposed novel weakly supervised learning technique. Various deep learning approaches such as BiGRU, BiLSTM, LSTM, and CNN were investigated for detecting' Fake' and' Real' news. Experimental results demonstrate that

Bi-LSTM and BiGRU outperformed other deep learning models with an increase in accuracy of 89.5%. The performance of CNN was worse when compared to other deep learning models. Bi-LSTM and BiGRU with weakly supervised SVM are the best models for the classification of Fake news when compared to other state-of-the-art approaches using large amounts of weakly labeled data, hence the proposed approaches are significantly more effective and efficient at identifying false information.

Credit author statement

Liyakathunisa Syed: Conceptualization, Methodology, Software, Validation, Writing- Original draft preparation; Abdullah Alsaeedi.: Related Studies, Supervision, Visualization; Lina A. Alhuri-Data curation, Software; Hutaf R. Aljohani: Writing- Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Watson Amy. Social media as a news source worldwide. 2022. https://www.statista.com/statistics/718019/social-media-news-source/. December 2022.

[2] Goswami Manash, , Fake News, Propaganda Cyber. A study of manipulation and abuses on social media. Mediascpe in 21st Century 2018:535–44.

[3] Thota Aswini, Tilak Priyanka, Ahluwalia Simrat, Lohia Nibrat. Fake news detection: a deep learning approach. SMU Data Science Review 2018;1(No. 3). Article 10.

[4] Langin K. Fake news spreads faster than accurate news on Twitter—thanks to people, not bots. 2018. http://www.sciencemag.org.

[5] Al-Sarem M, Boulila W, Al-Harby M, Qadir J, Alsaeedi A. Deep learning-based rumor detection on microblogging platforms: a systematic review. IEEE Access 2019;7:152788–812. https://doi.org/10.1109/ACCESS.2019.2947855.

[6] Allcott H, Gentzkow M. Social media and fake news in the 2016 election. National Bureau of Economic Research; 2017. Technical report.

[7] Alsaeedi A, Al-Sarem M. Detecting rumors on social media based on a CNN deep learning technique. Arabian J Sci Eng 2020. https://doi.org/10.1007/s13369-020-04839-2.

[8] Hiramath CK, Deshpande GC. Fake news detection using deep learning techniques. In: 2019 1st international conference on advances in information technology (ICAIT). India: Chikmagalur; 2019. p. 411–5. https://doi.org/10.1109/ICAIT47043.2019.8987258.

[9] Kumar Sachin, Asthana Rohan, Upadhyay Shashwat, Upreti Nidhi, Akbar Mohammad. Fake news detection using deep learning models: a novel approach. Trans. Emerg. Telecommun. Technol 2020;31:2. https://doi.org/10.1002/ett.3767. February 2020.

[10] Ibrain Rodríguez Álvaro, Lloret Iglesias Lara. Fake news detection using Deep Learning". 2019.

[11] Krešňáková VM, Sarnovský M, Butka P. Deep learning methods for Fake News detection. In: 2019 IEEE 19th international symposium on computational intelligence and informatics and 7th IEEE international conference on recent achievements in mechatronics, automation, computer sciences and robotics (CINTI-MACRo); 2019. p. 143–8. https://doi.org/10.1109/CINTI-MACRo49179.2019.9105317. Szeged, Hungary.

[12] Pritika B, Preeti S, Raj K. Fake news detection using Bi-directional LSTM-recurrent neural network. Proc Comput Sci 2019;165:74–82.

[13] Ajao Oluwaseun, Bhowmik Deepayan, Zargari Shahrzad. Fake news identification on twitter with hybrid CNN and RNN models. In: Proceedings of the 9th International Conference on social media and society. New York, NY, USA: Association for Computing Machinery; 2018. p. 226–30. https://doi.org/10.1145/3217804.3217917.

[14] Wang Yaqing, Yang Weifeng, Ma Fenglong, Xu Jin, et al. Weak supervision for fake news detection via reinforcement learning. Association for the Advancement of Artificial Intelligence, 2020 Intelligence; 2020.

[15] Helmstetter Stefan, Paulheim Heiko. Weakly supervised learning for fake news detection on Twitter. In: Proceedings of the 2018 IEEE/ACM international conference on advances in social networks analysis and mining. IEEE Press; 2018. p. 274–7.

[16] Kai S, Guoqing Z, Yichuan L, Subhabrata M, et al. Early detection of fake news with multi-source weak social supervision". 2020, 01732. arXiv:2004.

[17] Han S, Gao J, Ciravegna F. Neural Language model-based training data augmentation for weakly supervised early rumor detection. In: 2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). Vancouver, BC: Canada; 2019. p. 105–12. https://doi.org/10.1145/3341161.3342892.

[18] Özgöbek Özlem, et al. Fake news detection by weakly supervised learning based on content features." nordic artificial intelligence research and development: 4th symposium of the Norwegian A.I. Society, NAIS 2022, oslo, Norway. May 31–June 1. p 52-64. 2022, Revised Selected Papers 2023. Cham: Springer International Publishing.

[19] Aslam Nida, et al. Fake detect: a deep learning ensemble model for fake news detection. Complexity 2021;2021:1–8.

[20] Zhou Zhi-Hua. A brief introduction to weakly supervised learning. Natl Sci Rev January 2018;5(Issue 1):44–53. https://doi.org/10.1093/nsr/nwx106.

[21] Settles B. Active learning literature survey. Wisconsin, WI: Department of Computer Sciences, University of Wisconsin at Madison; 2010. Technical Report 1648.

[22] Chapelle O, Schölkopf B, Zien A, editors. Semi-supervised learning. Cambridge, MA: MIT Press; 2006.

[23] Zhou Z-H, Li M. Semi-supervised learning by disagreement. Knowl Inf Syst 2010;24(3):415–39.

[24] Zhu X. Semi-supervised learning literature survey. Madison, WI: Department of Computer Sciences, University of Wisconsin at Madison; 2008. Technical Report 1530.

[25] https://medium.com/@datasciencemilan/weakly-supervised-learning-introduction-and-best-practices-c65f490d4a0a [Accessed July 2022].

[26] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 2020:1–62.

[27] Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning, vol. 1. MIT press Cambridge; 2016.

[28] S. Biswas, E. Chadda, F. Ahmad, Sentiment analysis with gated recurrent units, Department of Computer Engineering. Annual Report Jamia Millia Islamia New Delhi, India.

[29] Chollet F. Deep learning with Python. 2017. Manning.

[30] Lamsal R. Coronavirus (COVID-19) tweets dataset. 2021. https://doi.org/10.21227/781w-ef42 [Online]. Available: 2021.

[31] Mohammed M, Omar N. Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. PLoS One 2020 Mar 19;15(3):e0230442.

[32] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.

[33] Naili Marwa, Habacha Chaibi Anja, Hajjami Ben Ghezala Henda. Comparative study of word embedding methods in topic segmentation. Elsevier B.V. Proc Comput Sci 2017;112:340–9 *[Procedia Computer Science. Web]*

[34] Jeffrey Pennington RSCDM. GloVe: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing; 2014.

[35] Naili M, Chaibi AH, Ghezala HHB. Comparative study of word embedding methods in topic segmentation. Proc Comput Sci 2017;112:340–9.

[36] Syed L, Jabeen S, Manimala S, Elsayed HA. Data science algorithms and techniques for smart healthcare using IoT and big data analytics. In: Mishra M, Mishra B, Patel Y, Misra R, editors. Smart techniques for a smarter p lanet. Studies in fuzziness and soft computing, vol. 374. Cham: Springer; 2019. https://doi.org/10.1007/978-3-030-03131-2_11.