Light weight recommendation system for social networking analysis using a hybrid BERT-SVM classifier algorithm

Kiruthika Nallichery Subramanian[1]✉, Thailambal Ganapathy[2]

[1] Vels University, Chennai, Tamil Nadu, 600117, India
[2] Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, 600117, India

[1] sathishpoojaa5@gmail.com✉, https://orcid.org/0000-0001-6601-1341
[2] thaila.scs@velsuniv.ac.in, https://orcid.org/0000-0002-0043-2415

Abstract
Social media platforms, such as Twitter, Instagram, and Facebook, have facilitated mass communication and connection. Due to the development as well as the advancement of social platforms, the spreading of fake news has increased. Many studies have been performed for detecting fake news with machine learning algorithms; but these existing methods had several difficulties, such as rapid propagation, access method and insignificant selection of features, and low accuracy of the text classification. Therefore, to overcome these issues, this paper proposed a hybrid Bidirectional Encoder Representations from Transformers — Support Vector Machine (BERT-SVM) model with a recommendation system that used to predict whether the information is fake or real. The proposed model consists of three phases: preprocessing, feature selection and classification. The dataset is gathered from Twitter social media related to COVID-19 real-time data. Preprocessing stage comprises Splitting, Stop word removal, Lemmatization and Spell correction. Term Frequency Inverse Document Frequency (TF-IDF) converter is utilized to extract the features and convert text to binary vectors. A hybrid BERT-SVM classification model is used to predict the data. Finally, the predicted data is compared with the preprocessed data. The proposed model is implemented in MATLAB software with several performance metrics carried out, and these parameters attained better performance: accuracy is 98 %, the error is 2 %, precision is 99 %, specificity is 99 %, and sensitivity is 98 %. Therefore the better effectiveness of the proposed model than existing approaches is shown. The proposed social networking analysis model provides effective fake news prediction that can be used to identify the Twitter comments, either real or fake.

Keywords
social networking analysis, fake news detection, TF/IDF, BERT, SVM, hybrid BERT-SVM

Облегченная система рекомендаций для анализа социальных сетей с использованием гибридного алгоритма классификатора BERT-SVM

Налличери Субраманиан Кирутика [1]✉, Ганапати Тайламбал [2]

[1] Университет Велса, Ченнаи, шт. Тамил Наду, 600117, India
2 Институт науки, технологий и перспективных исследований Велса, Ченнаи, шт. Тамил Наду, 600117, Индия

[1] sathishpoojaa5@gmail.com✉, https://orcid.org/0000-0001-6601-1341
[2] thaila.scs@velsuniv.ac.in, https://orcid.org/0000-0002-0043-2415

Аннотация
Платформы социальных сетей, такие как Twitter, Instagram и Facebook, способствуют массовому общению и установлению связей. Развитие и продвижение социальных платформ приводит к увеличению распространения фейковых новостей. В настоящее время проведено большое количество исследований для обнаружения фейковых новостей с помощью алгоритмов машинного обучения. Существующие методы определения фейков имеют ряд трудностей: быстрое распространение фейков; различные методы доступа и незначительный выбор признаков, приводящие к невысокой точности классификации текста. Для преодоления данных трудностей предложена

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4

769

гибридная модель представления двунаправленного кодировщика трансформаторов – метод опорных векторов (BERT-SVM) с системой рекомендаций, которая используется для прогнозирования, является ли информация поддельной или реальной. Предложенная модель включает в себя три этапа: предварительная обработка, выбор признаков и классификация. Набор данных собран из социальных сетей Twitter, связанных с данными о COVID-19 в режиме реального времени. Этап предварительной обработки включает в себя разделение, удаление стоп-слов, лемматизацию и исправление орфографии. Преобразователь обратной частоты документа (TF-IDF) использован для извлечения признаков и преобразования текста в двоичные векторы. Гибридная модель классификации BERT-SVM применена для прогнозирования данных, которые сопоставлены с предварительно обработанными данными. Представленная модель реализована в программном пакете MATLAB. Рассчитанные показатели точности продемонстрировали следующие результаты: доля правильных ответов 98 %, ошибка 2 %, точность 99 %, специфичность 99 %, чувствительность 98 %. Полученные результаты показали эффективность предложенной модели по сравнению с существующими подходами. Возможность анализа социальных сетей обеспечивает эффективное предсказание фейковых новостей, которое можно использовать для идентификации комментариев в Twitter, как настоящих, так и поддельных.

Introduction

Social networking has developed dominant platforms to reveal their sentiments, opinions, reactions, and knowledge. Several social media, such as Twitter, Facebook, Instagram, etc., produce a massive amount of data each and every day. Emotion recognition or analysis of sentiment focuses on detecting polarization by using views or opinions from Twitter datasets. Sentiment polarity determines a user's reaction to a product, allowing businesses to take preventative and corrective actions to satisfy their expectations. Conversely, government criticism assists governments in analyzing public requirements and making crucial decisions. Fake news is becoming more common on social media sites like Twitter and Facebook. These platforms provide a venue for the general public to express themselves in an unfiltered and uncensored manner [1].

Misinformation is defined as information that is demonstrably false and is shared with the goal of misleading readers. For personal advantage, it is used to establish an economic, political and social bias in people's thoughts. Its goal is to manipulate and exploit individuals by creating bogus material that appears to be genuine. At its most severe cases, fake news has resulted in mob lynching and riots. As a result, it is critical to halt the proliferation of fraudulent content on social media sites. With the continuing Covid-19 scenario, preventing fake news is very vital [2]. The epidemic has done it very easy to deceive a psychologically stuck people excitedly anticipating the conclusion of this phase. Some people have reportedly committed suicide after being diagnosed with COVID-19 as a result of the misrepresentation of COVID-19 in society and even the mainstream media. The promotion of deceptive techniques will only exacerbate the COVID pandemic. Recently, researchers have begun to focus on the challenge of detecting fake news. Manual detection is the most reliable method, although it has speed constraints [3]. Manual verification is difficult due to the large amount of content published on the internet. Thus, automatic detection of bogus news has become increasingly important.

Different deep learning and machine learning algorithms were used to discover comments of social media information about its truthfulness. These false reports will not only lead individuals in the wrong route, but they will also take human lives. In these critical times of Covid-19, it is easy to deceive people and make them believe in fake information [4]. As a result, it's critical to spot fake news at the source and stop it from propagating to a wider audience. Many studies have been performed for detecting the fake news with machine learning approaches like Long Short-Term Memory (LSTM), Support Vector Machine (SVM), hybrid LSTM-SVM, Naive Bayesand Random Forest. From these existing methods, the accuracy of the text classification is low. So, to overcome the low accuracy, a hybrid Bidirectional Encoder Representations from Transformers — Support Vector Machine (BERT-SVM) model with a recommendation system is used to predict whether the information is fake or real. The main contribution of the paper is summarized as follows:

— An accurate Light Weight Recommendation System for Social Networking Analysis using a Hybrid BERT-SVM classifier.
— The preprocessing technique is utilized to improve accuracy and reduce the process complexity of the real time twitter dataset to provide better performance.
— Term Frequency Inverse Document Frequency (TF-IDF) is utilized for feature selection for word representations.
— The hybrid BERT-SVM classification model is designed for predicting the fake news in twitter to avoid spreading the fake news related on COVID-19 pandemic.
— The predicted news is given as a recommendation to the user for the awareness of the information about the fake news.

Literature review

Numerous studies have been performed using various techniques for recommendation system. Most of the existing techniques is designed based on LSTM, Random

770

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4

Forest, SVM, Bi-LSTM & Recurrent Neural Network (RNN), Convolution Neural Network (CNN), Neuro Fuzzy, CNN-LSTM, bidirectional Gated RNN (GRNN) and convolution RNN (CRNN); from that few of them are reviewed below.

Umer et al., [4] had presented the deep learning approach for detecting the fake news using CNN-LSTM. The dataset gathered from the website named "Fake News Challenges" and this website has four different types, such as unrelated, disagree, agree and discuss. Hakak et al., [5] had introduced a strategy for classify the fake news utilizing a feature extraction and ensemble machine learning technique. Abdullah et al., [6] had developed a multimodal method to achieve a fake news detection using a combination of CNN and LSTM. Huang et al., [7] had designed a fake news prediction strategy utilizing deep learning method. The developed model combined four separate models termed as embedding for fake news identification. The embedding models are LSTM, LIWC CNN, depth LSTM, and N-gram CNN. Paka et al., [8] had introduced a fake news prediction model for analyzing a large-scale COVID-19 Twitter dataset. Nasir et al., [9] had designed a novel hybrid strategy for fake news categorization using combined CRNN. Sabeeh et al., [10] had introduced a model for discovering fake news on social media platforms through Opinion mining and Trustworthiness of user and event that combines opinion mining on user comment and credibility investigation of twitter dataset. Bahad, P et al., [11] had designed a false news identification model utilizing Bi-LSTM and RNN. Misinformation is frequently created to entice and mislead readers for political and commercial purposes.

According to the above discussed literatures, the existing strategies still contain several difficulties, such as certain problems of poor accuracy, rapid propagation, access method and high cost [5, 6]. The poor accuracy is attained in many research according to many reasons, such as insignificant selection of features, imbalanced dataset, inefficient tuning of parameters and so on [7]. These difficulties arise during detecting fake news in social media. To deal with these issues, the proposed model utilizes a hybrid BERT-SVM classification strategy to predict the fake news very accurately compared to existing strategies.

Proposed methodology for fake news prediction

A light weight recommendation system for social media networking analysis using a hybrid BERT-SVM classification model is designed to predict whether the information is fake or real. In this proposed model, a hybrid BERT-SVM model is used to classify and recommend the information to user whether it is fake or real. It is illustrated in Fig. 1. The proposed architecture of fake news detection model comprises three phases, such as preprocessing, feature selection and classification for effective fake news prediction. Initially the user dataset is collected from twitter social media. The raw data is preprocessed in this phase which involved splitting, stop word removal, lemmatization and spell correction. The second phase uses the TF-IDF for converting the text data into meaningful representation of binary data which is utilized to fit the classifier for effective

prediction. At the third phase, these features are given to the input of the hybrid BERT-SVM classification model and it produced two classes which is either real or fake. This suggestion will be recommended on the particular tweet to the user helping aware the information about the fake news.

Data Gathering

A real time user data is collected from twitter social media to make a dataset. The dataset is entirely based on Covid-19 pandemic as well as the twitter dataset contains the corresponding label post ID and the posts which are either fake or real news on the COVID-19 pandemic[1]. It comprises 5000 manually gathered twitter commands equally on both fake and real news. These data are given into preprocessing for converting the raw data into a specified format, i.e., into the system readable format.

Data Pre-processing

The preprocessing phase is performed in four ways: stop word removal, splitting, spell correction and lemmatization. In this process, you need to eradicate the noise in the twitter dataset by normalizing or eliminate the unwanted data.

Splitting. Initially, in this stage, text sentences are split into individual words. String splitting is the technique of systematically dividing a text string into separate components that can be processed.

Stop word removal. In Natural Language Processing (NLP), stop word removal is a regularly used method. For instance, stop words such as 'a', 'the', 'an' are eliminating ones, the words that are often appearing in large numbers across all the documents, and it allows applications to focus on the important words instead.

Lemmatization. The algorithmic process of determining the lemma of a word based on its meaning is known as lemmatization. Lemmatization is a term that refers to accomplishing things correctly by analyzing words morphologically and using a vocabulary. Lemmatization is a term that relates to doing things correctly using a vocabulary and morphological analysis of words, with the goal of removing inflectional endings only and returning the base or dictionary form of a word, known as the lemma.

Spell correction. Spell checker is analyzing each and every word with thousands of proper spell words. The majority of techniques uses data from many sources of noisy as well as correct word mappings as training data for automatic spelling correction.

TF-IDF Feature Extraction

TF-IDF is a significant feature extraction and selection strategy used in text feature extraction. TF-IDF is used to find the significant features of the sentences that comprise the words and remove the bag incompetence of words strategies [12]. It is superior for text classification and is helpful for a machine reading words in numbers. One of the weight methods is Term Frequency (TF) that is used to determine the number of instances of term, i.e. word in a document. The Inverse Document Frequency (IDF) also presents the weight technique that used to find incidence terms in individual documents. The TF and IDF weight of the documents can be estimates using the given below expression.

---

[1] https://www.kaggle.com/c/sentiment-analysis-of-covid-19-related-tweets/ (accessed: 10.05.2022).

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4
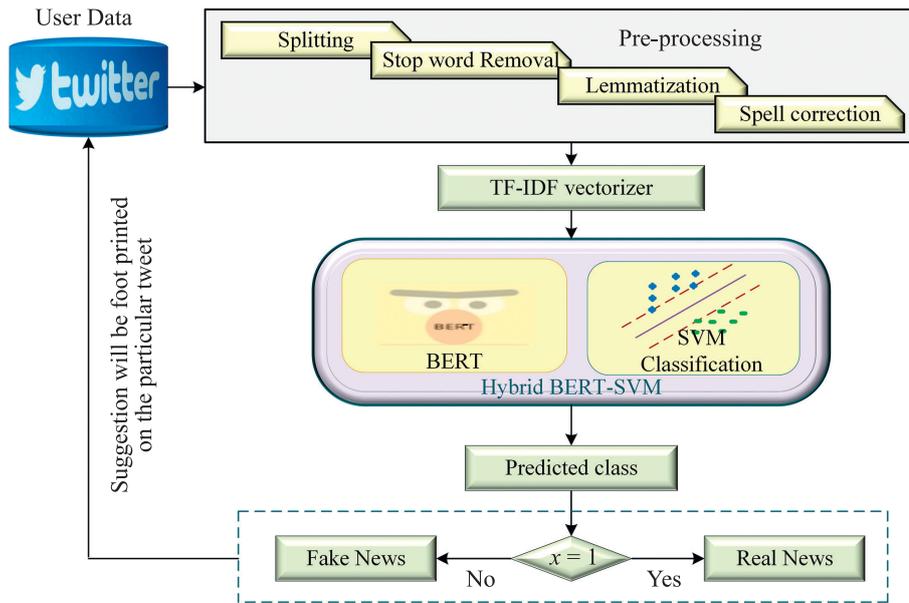
771

Fig. 1. Proposed architecture of social media analysis using a Hybrid BERT-SVM classifier, *x* is the class index

$$W_{dt} = TF(i, j) \times IDF(i).$$

The frequency of a feature appearing in a document in relation to the total number of features appearing in the text document is defined by TF. Simultaneously, IDF assesses a feature capacity to differentiate between categories. The categories in this case are the class label declared in the text documents. The following expressions are used to calculate TF and IDF:

$$TF(i, j) = \frac{Term\ i\ frequency\ in\ document\ j}{Total\ words\ in\ document\ j},$$

$$IDF(i) = \log_2 \frac{Total\ document}{document\ with\ term\ i},$$

where *t* is the term and *j* is the counting number of the document. The following expression computes the weight of each phrase using the TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right),$$

where $tf_{i,j}$ is the number of occurrence of *i* in *j*; df_i is the counting number of the document containing *i*; and *N* is the total number of documents.

Hybrid BERT-SVM for Fake News Detection

This proposed method is designed to predict the fake news in the twitter social media using hybrid BERT-SVM classification. Hence, the detail description of this hybrid classification approach is given below.

1) **BERT**

BERT is based on transformer encoded architecture which is one of the most significant words embedding model in sentimental analysis. An effective content illustration of a sentence is attained using a BERT sentence encoder. Mask Language Model (MLM) is used in the BERT encoder to eliminate the unidirectional restriction. Usually it masks several token in a random manner, and original vocabulary of the masked ID that is predicted based on the word. Moreover, compared to existing embedding strategies, the BERT can outperform because MLM can improve the BERT ability. It has the ability to handle unlabeled text by training together on both right context and left context in each layer [13]. Fig. 2 illustrates the basic BERT model for categorization.

BERT is regarded as the most advanced NLP technology. Pre-training and fine-tuning are the two phases in the BERT framework. At the initial stage, during pre-training in the model, the unlabeled large corpus is involved. At the second stage, labeled data is used to fine-tune every parameter for particular tasks. BERT's encoder is a multi-layer bidirectional Transformer encoder with multiple layers. This encoder consists of a stack of $N = 6$ identical layers, each with two sub-layers. The first layer is a basic position-wise completely connected feedforward network, and the second layer is a multi-head self-attention mechanism. Both sub-layers use a residual connection followed by layer normalization. Each sub-layer output is represented as Layer *Norm(x + Sublayer(x))*, where the sub-layer exaction function is denoted as *Sublayer(x)*. Before estimating multi-head self-attention, scaled dot product attention is required to define the following [14]:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V},$$

where **Q**, **K** and **V** are three matrices **Q** (Query), **K** (Key) and **V** (value). All come from the same input. To get the weight value, the Softmax operation is performed to normalize the output to a probability distribution for each row and then multiply it by the matrix **V**. d_k is the dimension of the **Q** and **K** matrices and queries matrix is represented as **Q**. The multi-head attention function can be expressed as

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, \ldots, head_h)W^0,$$

where $head_i = Attention(\mathbf{Q}W_i^Q, \mathbf{K}W_i^K, \mathbf{V}W_i^V)$. Multi-head attention contain estimation of queries, keys and

772

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4
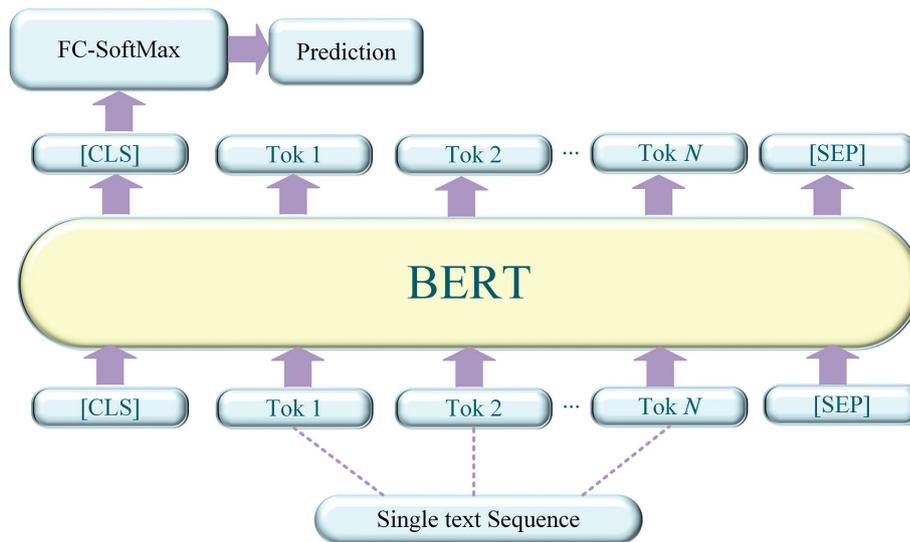
Fig. 2. BERT model for categorization, where [CLS] is BERT special classification token (Tok) and [SEP] is separation token

values; here these are denoted as **Q**, **K**, **V** and *h* is the repetition elements factor with different dimension of the value matrix such as d_q, d_k and d_v. Then, the attention function is performed in parallel with each of predicted versions of queries, keys and values as well as the findings of output values is d_v. At the end, these parameters are concatenated and estimated to produce the outcome of multihued function. BERT indicates a pair of sentences or a single sentence as a series of tokens based on BERT's use of Word Piece embedding. [CLS] is considered as a first classification token in the sequence. A pair of sentences is separated by [SEP] token in the sequence. Pre-training contains two phases: Masked LM and Next Sentence Prediction (NSP). The Masked LM involves randomly masking the input tokens like utilizing the [MASK] token. After performed masking, it starts predicting masked tokens. The second one is NSP that considers two sentences A and B. B is the true next statement which appears 50 % of the time after A (considered as IsNext). 50 % of the time, B is a random text from the corpus (considered

as NotNext). Because the Transformer's self-attention mechanism allows BERT to model multiple downstream processes, fine-tuning is simple. Then, just feed the specific inputs and outputs into BERT for each task and fine-tune all of the settings.

2) **SVM**

SVM is a machine learning technique for evaluating data and detecting patterns in order to classify what follows as a result of deciding which of two classes to assign a set of input data. A training set is necessary for learning SVMs, with each element of the set having an indication of a class it belongs to. The data from the training set are separated from each other by a border with the widest feasible margin, i.e. the distance from this hyperplane, in the SVM model. In binary classification, the SVM is a popular learning method. The primary goal is to identify the optimal hyperplane for separating data into its two classes. Multiclass classification has recently been achieved by combining multiple binary SVMs [15]. The basic architecture of SVM classifier is given in Fig. 3.
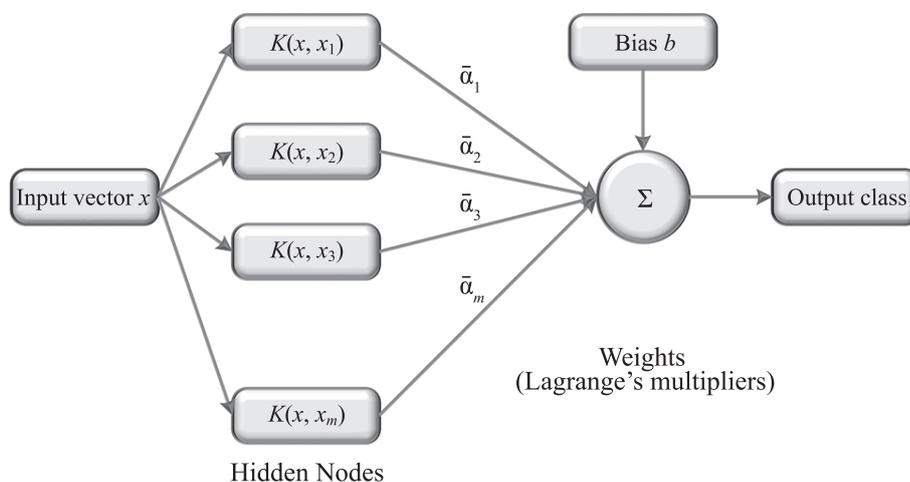

Fig. 3. Structure of SVM, where $K(u, v)$ is a kernel function satisfying Mercer's condition; *x* is a point of hyperplane; x_i are support vectors; $\bar{\alpha}_i$ are Lagrange's multiplies; *m* is SVM hyperplane dimension; and *b* is a bias

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4

773

Let x_i is considered as input; i is the training instances $\{x_i, y_i\}$, $i = 1, \ldots, l$; and each occurrence consists of x_i as well as label $y_i \in \{-1, 1\}$ [16]. A weight vector (w) and bias (b) both are used to parameterize every hyperplane which is expressed as given below:

$$w \times x + b = 0,$$

where x is a point lying on the hyperplane. To define the hyperplane function that classifies training and testing data, the following can be used

$$f(x) = sign(w \times x + b).$$

The analysis has so far been limited to the case when the training data can be separated linearly. Prior function can be given as equation when dealing with kernel

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i K(x_i, x) + b),$$

where x_i is the input of training instance; y_i is its corresponding class label; b is a bias; N is the number of training instances; and $K(x_i, x)$ is the used kernel function which maps the input vectors into an expanded features space. The coefficients α_i are attained subject to two constraints expressed as

$$0 \leq \alpha_i, \, i = 1, \ldots, N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0.$$

Except for a change in the bounds of the Lagrange multipliers, the solution to this minimization issue is similar to that of the separable case.

Hybrid BERT-SVM for Fake News Prediction

In this section, the proposed approach uses both BERT and SVM classifiers in a hybrid manner. BERT model is mainly selected for solving the difficulties of NLP models. BERT working is based on the transformers and is developed for text classification that is also termed as word embedding model in sentimental analysis. Moreover, this BERT approach performs NLP tasks as well as natural language understanding tasks. This strategy improves the ability of NLP models to implement data without having to maintain any order sequences. The BERT model input embedding contains tokens, segments and position components. SVM is a one of the significant machine learning approaches that able to handle high dimensional data as well as it attained better performance when using text classification. The proposed hybrid BERT-SVM architecture for expectations E_i is shown in Fig. 4. The extracted features are given to the proposed hybrid BERT-SVM classifier. Initially, the feature selected dataset with 80 % of twitter commands is given for training the classifier. The BERT model usually performs feature extraction as well as usual classification, but the proposed model uses the BERT model only for usual classification. To improve the BERT model, the SVM model is included in the FC-Softmax layer of the BERT.

Algorithm 1: Pseudo Code for Social Networking Analysis using a hybrid BERT-SVM

Input: A=Dataset; X1=Splitting; X2=Stop word removal; X3= Lemmatization; X4=Spell correction; Q=Feature extraction; P=Classification
Output: R=Fake news
Input dataset =A, pre-processed data =Z, User =M
 # Pre-processing
   X1 = Splitting (A) # splits combined words are separated for individual processing
   X2 = Stop word removal (X1) # stop words such as 'the', 'an', 'a' are removed
   X3 = Lemmatization (X2) # Grouping of words together with different derivatives
   X4 = Spell correction (X3) # Spell correction of wrong words
 # feature extraction
   Q = TD-IFD (X4) # convert text data into binary data
 # classification
   P = BERT-SVM (Q) # classification utilizing Hybrid BERT-SVM
 # compare predicted data with pre-processed data
   If (P = Z)–> real news; Else –> fake news
 # predicted fake news is given as a recommendation to the user
   R = fake news; M= R

Output= detect either real or fake news in the tweet information

The BERT model comprises two phases such as pre-training and fine-tuning. Initial stage of the classifier trained the information among various pre-trained problems using twitter dataset. This research uses the Base BERT model, which consists of 12 bidirectional self-attention heads and 12 encoder layer blocks. The model assumes a series of 512 tokens, with hidden vectors in the sequence being emitted. The final layer of BERT model is FC-Softmax layer. The proposed model uses the SVM classifier in FC-Softmax layer for accurate prediction outcome. After training the classifier, remaining 20 % of twitter commands are subjected into testing of the trained model. The classifier produces two classes such as 0 and 1. Here, class 0 is considered as fake news and class 1 is considered as real news. After predicted data, i.e. real or fake, this suggestion will be foot printed on the particular tweet to aware the information about COVID-19 in the twitter social media. In this way, the proposed model effectively predicts the news of its truth in twitter about COVID-19 pandemic. Algorithm 1 illustrates the pseudo code for network analysis using a hybrid BERT-SVM.

Result and discussion

The proposed light weight recommendation system for social media networking analysis using a hybrid BERT-SVM classification model is implemented in MATLAB software to validate its performance. The proposed system is tested on Intel(R) Core i5-10300H CPU, 4 GB Nvidia GeForce GTS 1650 GPU, with 16.0 GB RAM. The dataset is gathered from twitter social media[1]. It includes 5000 manually gathered twitter commands included both fake and real news. During preprocessing the sentences

---

[1] https://www.kaggle.com/c/sentiment-analysis-of-covid-19-related-tweets/ (accessed: 10.05.2022).

774

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4
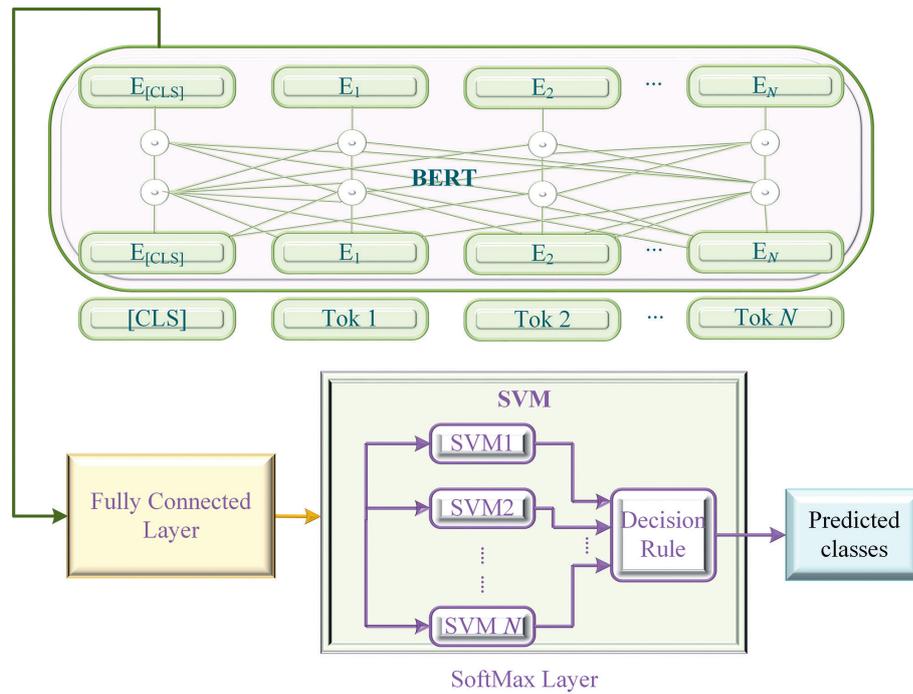
Fig. 4. Proposed Hybrid BERT-SVM Architecture

are extracted by using splitting, stop word removal, lemmatization, and spell correction. After that, the extracted features are given into TF-IDF that converts the text data into meaningful illustration of binary data. The extracted 3500 TF-IDF features are given to proposed Hybrid BERT-SVM classification for training. After training, the remaining 1500 features are utilized for testing.

Table 1 illustrates the extraction of input in the preprocessing phase; the data are initially fed into the splitting stage where you can split each and every word. After splitting process, the data are processed into stop

word removal, in this step the stop words are to be removed, such as 'the', 'an', 'a', 'if', 'are', etc. Then, the next stage is lemmatization; its aim is to take away inflectional suffixes and prefixes to bring out the word dictionary form. Finally, the data are spell corrected and the preprocessed findings produced. These performance metrics are estimated for the proposed BERT-SVM and existing fake news detection model in twitter social media. Several existing techniques based on fake news detection models are considered for comparison analysis, such as CNN-SVM, bidirectional GRNN, LSTM, CRNN, and SVM. Table 2 illustrates the

Table 1. Sample Input and Output attained in pre-processing

Input	Splitting	Stop word	Lemmatize	Spell correction
'BanMediaHouse whose is responsible for spreading Fake and communal	"BanMediaHouse" "whose" «is" "responsible" "for" "spreading» "Fake" "and" "communal"	"BanMediaHouse" "whose" "responsible" "spreading" "Fake" "communal"	"banmediahouse" "whose" "responsible" "spread" "fake" "communal"	"househusband whose responsible spread fake communal"

Table 2. Parameters estimated for proposed and existing techniques in fake news detection

Parameters, %	BERT-SVM (proposed)	CNN-SVM	GRNN	LSTM	CRNN	SVM
Accuracy	98	90	85	70	68	57
Error	2	10	15	30	31	45
Sensitivity	98	93	89	91	77	60
Specificity	99	92	86	78	61	60
Precision	99	95	82	87	78	55
False Positive Rate, FPR	2	1	15	27	32	35
F1_Score	99	81	72	75	60	58
Kappa	90	88	80	61	62	58

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4
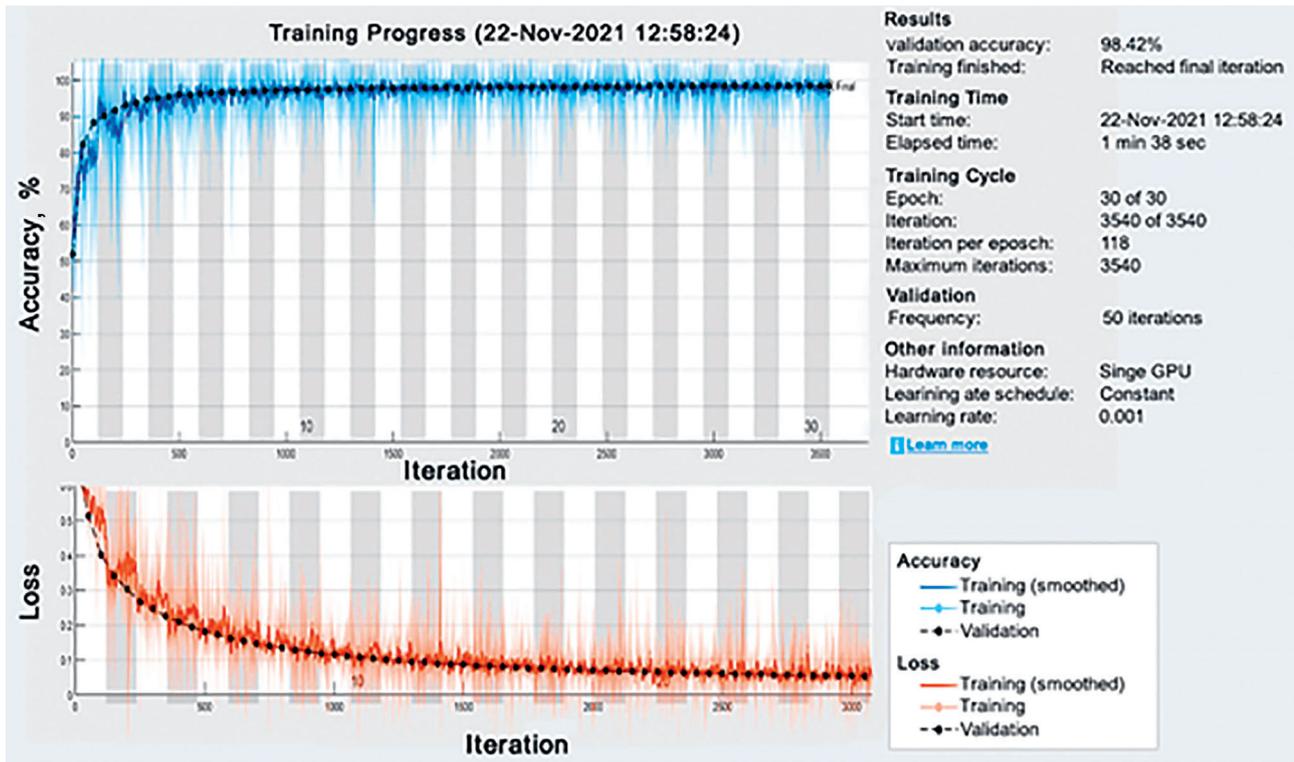
775

Fig. 5. Training process of accuracy and losses

parameter values estimated for the proposed and existing techniques for fake news detection.

Fig. 5 displays the accuracy and losses of the training process. The accuracy of the proposed method is 100 %, since it could more rapidly detect the error and eliminate it more effectively. The error of the proposed method is 0 % because the hybrid BERT-SVM classification model attained accurate prediction. Fig. 6 illustrates the confusion matrix obtained in this proposed model. Here, the X and Y labels represent the predicted class and the true class. The class 0 predicts 4298 data and class 1 predicts 3194 data. To provide an analytical assessment, a confusion matrix and performance measurement like accuracy, precision and so on is used.

The comparison analysis based on accuracy (%) among the proposed and existing strategies depended on fake news

detection on social media is shown in Fig. 6. The accuracy for the proposed method is 98 % which is greater compared to existing methods: CNN-SVM is 90 %, GRNN is 85 %, LSTM is 70 %, CRNN is 68 %, and SVM is 57 %. This can show that the performance of the proposed model works better compared to others. The comparison analysis based on error among the proposed and existing strategies depended on fake news detection on social media is shown in Table 2. The error found for the proposed method is 2 % which is less compared to existing methods: CNN-SVM is 10 %, GRNN is 15 %, LSTM is 30 %, CRNN is 31 % and SVM is 45 %.

This can show that the performance of the proposed model works better compared to others. Similarly, the comparison of performances, such as sensitivity, specificity, precision, FPR, F1_Score and kappa of the proposed and existing strategies, is given in Table 2. It clearly shows that the proposed technique has better performance than the other techniques. This can show that the performance of the proposed model works better compared to others.

Conclusion

A lightweight recommendation system is proposed for social networking analysis using a hybrid BERT-SVM classifier algorithm to improve the accuracy efficiently. Initially, a real-time dataset was collected from Twitter social media. These data are given into preprocessing, and splitting, stop word removal, lemmatization and spell correction have been performed. An effective feature extraction strategy is utilized for text feature selection and binary conversion. The converted features are classified with the hybrid BERT-SVM model. The predicted news
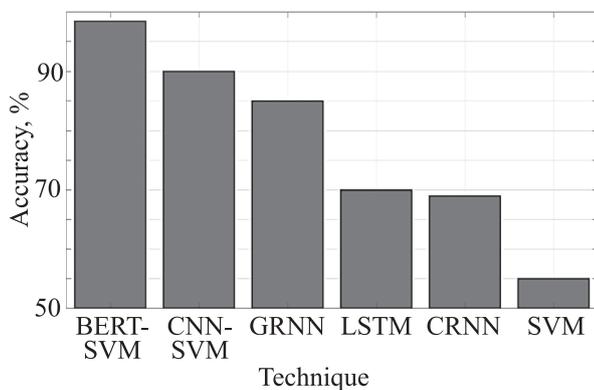

Fig. 6. Comparison of accuracy among proposed and existing approaches

776

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4

is given as a recommendation to the user to provide the awareness of the information about the fake news. The proposed model is executed on MATLAB for finding the performance metrics: accuracy is 98 %, the error is 2 %, sensitivity is 98 %, specificity is 99 %, precision is 99 %, FPR is 2 %, F1_Score is 99 %, and kappa is 90 %. The overall expected outcome of the recommendation system using hybrid BERT-SVM is turned out to be better compared to the existing techniques such as CNN-SVM, GRNN, LSTM, CRNN, and SVM. The proposed social networking analysis model delivers effective fake news detection that can be utilized to detect the Twitter comments related to the COVID-19 pandemic, either real or fake, and it is possible to suggest that the predicted news is given as a recommendation to the user to get acquainted with information about the fake news.

References

1. Kaur S., Kumar P., Kumaraguru P. Automating fake news detection system using multi-level voting model. *Soft Computing*, 2020, vol. 24, no. 12, pp. 9049–9069. https://doi.org/10.1007/s00500-019-04436-y
2. Kaliyar R.K., Goswami A., Narang P., Sinha S. FNDNet — a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 2020, vol. 61, pp. 32–44. https://doi.org/10.1016/j.cogsys.2019.12.005
3. Shim J.-S., Lee Y., Ahn H. A link2vec-based fake news detection model using web search results. *Expert Systems with Applications*, 2021, vol. 184, pp. 115491. https://doi.org/10.1016/j.eswa.2021.115491
4. Umer M., Imtiaz Z., Ullah S., Mehmood A., Choi G.S., On B.-W. Fake news stance detection using deep learning architecture (CNN-LSTM). *IEEE Access*, 2020, vol. 8, pp. 156695–156706. https://doi.org/10.1109/ACCESS.2020.3019735
5. Hakak S., Alazab M., Khan S., Gadekallu T.R., Maddikunta P.K.R., Khan W.Z. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 2021, vol. 117, pp. 47–58. https://doi.org/10.1016/j.future.2020.11.022
6. Abdullah, Yasin A., Avan M.J., Shehzad M.F., Ashraf M. Fake news classification bimodal using convolutional neural network and long short-term memory. *International Journal on Emerging Technologies*, 2020, vol. 11, no. 5, pp. 209–212.
7. Huang Y.-F., Chen P.-H. Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, 2020, vol. 159, pp. 113584. https://doi.org/10.1016/j.eswa.2020.113584
8. Paka W.S., Bansal R., Kaushik A., Sengupta S., Chakraborty T. Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. *Applied Soft Computing*, 2021, vol. 107, pp. 107393. https://doi.org/10.1016/j.asoc.2021.107393
9. Nasir J.A., Khan O.S., Varlamis I. Fake news detection: A hybrid CNN-RNN based deep learning approach. International *Journal of Information Management Data Insights*, 2021, vol. 1, no. 1, pp. 100007. https://doi.org/10.1016/j.jjimei.2020.100007
10. Sabeeh V., Zohdy M., Mollah A., Al Bashaireh R. Fake news detection on social media using deep learning and semantic knowledge sources. *International Journal of Computer Science and Information Security (IJCSIS)*, 2020, vol. 18, no. 2, pp. 45-68.
11. Bahad P., Saxena P., Kamal R. Fake news detection using bi-directional LSTM-recurrent neural network. *Procedia Computer Science*, 2019, vol. 165, pp. 74–82. https://doi.org/10.1016/j.procs.2020.01.072
12. Qaiser S., Ali R. Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 2018, vol. 181, no. 1, pp. 25–29. https://doi.org/10.5120/ijca2018917395
13. Pota M., Ventura M., Catelli R., Esposito M. An effective BERT-based pipeline for Twitter sentiment analysis: a case study in Italian. *Sensors*, 2021, vol. 21, no. 1, pp. 133. https://doi.org/10.3390/s21010133
14. Malla S., Alphonse P.J.A. COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets. *Applied Soft Computing*, 2021, vol. 107, pp. 107495. https://doi.org/10.1016/j.asoc.2021.107495
15. Goudjil M., Koudil M., Bedda M., Ghoggali N. A novel active learning method using SVM for text classification. *International Journal of Automation and Computing*, 2018, vol. 15, no. 3, pp. 290–298. https://doi.org/10.1007/s11633-015-0912-z

Литература

1. Kaur S., Kumar P., Kumaraguru P. Automating fake news detection system using multi-level voting model // Soft Computing. 2020. V. 24. N 12. P. 9049–9069. https://doi.org/10.1007/s00500-019-04436-y
2. Kaliyar R.K., Goswami A., Narang P., Sinha S. FNDNet — a deep convolutional neural network for fake news detection // Cognitive Systems Research. 2020. V. 61. P. 32–44. https://doi.org/10.1016/j.cogsys.2019.12.005
3. Shim J.-S., Lee Y., Ahn H. A link2vec-based fake news detection model using web search results // Expert Systems with Applications. 2021. V. 184. P. 115491. https://doi.org/10.1016/j.eswa.2021.115491
4. Umer M., Imtiaz Z., Ullah S., Mehmood A., Choi G.S., On B.-W. Fake news stance detection using deep learning architecture (CNN-LSTM) // IEEE Access. 2020. V. 8. P. 156695–156706. https://doi.org/10.1109/ACCESS.2020.3019735
5. Hakak S., Alazab M., Khan S., Gadekallu T.R., Maddikunta P.K.R., Khan W.Z. An ensemble machine learning approach through effective feature extraction to classify fake news // Future Generation Computer Systems. 2021. V. 117. P. 47–58. https://doi.org/10.1016/j.future.2020.11.022
6. Abdullah, Yasin A., Avan M.J., Shehzad M.F., Ashraf M. Fake news classification bimodal using convolutional neural network and long short-term memory // International Journal on Emerging Technologies. 2020. V. 11. N 5. P. 209–212.
7. Huang Y.-F., Chen P.-H. Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms // Expert Systems with Applications. 2020. V. 159. P. 113584. https://doi.org/10.1016/j.eswa.2020.113584
8. Paka W.S., Bansal R., Kaushik A., Sengupta S., Chakraborty T. Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection // Applied Soft Computing. 2021. V. 107. P. 107393. https://doi.org/10.1016/j.asoc.2021.107393
9. Nasir J.A., Khan O.S., Varlamis I. Fake news detection: A hybrid CNN-RNN based deep learning approach // International Journal of Information Management Data Insights. 2021. V. 1. N 1. P. 100007. https://doi.org/10.1016/j.jjimei.2020.100007
10. Sabeeh V., Zohdy M., Mollah A., Al Bashaireh R. Fake news detection on social media using deep learning and semantic knowledge sources // International Journal of Computer Science and Information Security (IJCSIS). 2020. V. 18. N 2. P. 45-68.
11. Bahad P., Saxena P., Kamal R. Fake news detection using bi-directional LSTM-recurrent neural network // Procedia Computer Science. 2019. V. 165. P. 74–82. https://doi.org/10.1016/j.procs.2020.01.072
12. Qaiser S., Ali R. Text mining: Use of TF-IDF to examine the relevance of words to documents // International Journal of Computer Applications. 2018. V. 181. N 1. P. 25–29. https://doi.org/10.5120/ijca2018917395
13. Pota M., Ventura M., Catelli R., Esposito M. An effective BERT-based pipeline for Twitter sentiment analysis: a case study in Italian // Sensors. 2021. V. 21. N 1. P. 133. https://doi.org/10.3390/s21010133
14. Malla S., Alphonse P.J.A. COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets // Applied Soft Computing. 2021. V. 107. P. 107495. https://doi.org/10.1016/j.asoc.2021.107495
15. Goudjil M., Koudil M., Bedda M., Ghoggali N. A novel active learning method using SVM for text classification // International Journal of Automation and Computing. 2018. V. 15. N 3. P. 290–298. https://doi.org/10.1007/s11633-015-0912-z
16. Zhu J., Tian Z., Kübler S. UM-IU@LING at SemEval-2019 task 6: Identifying offensive tweets using BERT and SVMs // Proceedings

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4

777

16. Zhu J., Tian Z., Kübler S. UM-IU@LING at SemEval-2019 task 6: Identifying offensive tweets using BERT and SVMs. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 788–795. https://doi.org/10.18653/v1/s19-2138

of the 13th International Workshop on Semantic Evaluation. 2019. P. 788–795. https://doi.org/10.18653/v1/s19-2138

Authors

Nallichery Subramanian Kiruthika — Research Scholar, Vels University, Chennai, Tamil Nadu, 600117, India, sc 55420781200, https://orcid.org/0000-0001-6601-1341, sathishpoojaa5@gmail.com

Ganapathy Thailambal — Associate Professor, Department of Computer science, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, 600117, India, sc 57189250428, https://orcid.org/0000-0002-0043-2415, thaila.scs@velsuniv.ac.in

Авторы

Кирутика Налличери Субраманиан — научный сотрудник, Университет Велса, Ченнаи, шт. Тамил Наду, 600117, Индия, sc 55420781200, https://orcid.org/0000-0001-6601-1341, sathishpoojaa5@gmail.com

Тайламбал Ганапати — доцент, Институт науки, технологий и перспективных исследований Велса, Ченнаи, шт. Тамил Наду, 600117, Индия, sc 57189250428, https://orcid.org/0000-0002-0043-2415, thaila.scs@velsuniv.ac.in

778

Научно-технический вестник информационных технологий, механики и оптики, 2022, том 22, № 4
Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no 4