Toxic Fake News Detection and Classification for Combating COVID-19 Misinformation

Mudasir Ahmad Wani, Mohammad ELAffendi, Kashish Ara Shakil, Ibrahem Mohammed Abuhaimed, Anand Nayyar, Amir Hussain, Ahmed A. Abd El-Latif,

***Abstract*—The emergence of COVID-19 has led to a surge in fake news on social media, with toxic fake news having adverse effects on individuals, society, and governments. Detecting toxic fake news is crucial, but little prior research has been done in this area. This study aims to address this gap and identify toxic fake news to save time spent on examining non-toxic fake news.**

To achieve this, multiple datasets were collected from different online social networking platforms such as Facebook and Twitter. The latest samples were obtained by collecting data based on the topmost keywords extracted from the existing datasets. The instances were then labelled as toxic/non-toxic using toxicity analysis, and traditional machine-learning techniques such as linear Support Vector Machine (SVM), conventional Random Forest (RF), and transformer-based machine-learning techniques such as Bidirectional Encoder Representations from Transformers (BERT) were employed to design a toxic-fake news detection and classification system.

As per the experiments, the linear SVM method outperformed BERT SVM, RF, and BERT RF with an accuracy of 92% and F_1-score, F_2-score, and $F_{0.5}$-score of 95%, 85%, and 87%, respectively. Upon comparison, the proposed approach has either suppressed or achieved results very close to the state-of-the-art techniques in the literature by recording the best values on performance metrics such as accuracy, F1-score, precision, and recall for linear SVM. Overall, the proposed methods have shown promising results and urge further research to restrain toxic fake news. In contrast to prior research, the presented methodology leverages toxicity-oriented attributes and BERT-based sequence representations to discern toxic counterfeit news articles from non-toxic ones across social media platforms.

***Index Terms*—Natural Language Processing, Machine learning, BERT, Toxicity Analysis, Emotion Extraction, Fake News**

I. Introduction

WITH the outbreak of COVID-19, the number of people being infected is around 4,31647782 and a total of 594671 deaths have been reported worldwide spanning 224 countries [1]. It has affected people's physiology as well as their psychology. During the entire period of COVID-19, real-time social media has played a quintessential role to obtain people's perception and communication of information [2]. Social media platforms provide users with versatile ways of expressing their opinions and ideas.

However, the positive impact of social media can be impeded by rumours and fake news thus can badly disturb the social and personal space of online users. Fake news is a piece of low-quality news that is intentionally wrong or false. Fake news can be of two types non-toxic fake news and toxic fake news. Non-toxic fake news is the news that has been spread for leisure or humour and has no negative impact on individuals, society, or government organizations [3]. In order to make it clear what constitutes toxic fake news, the following subsection presents and discusses some examples to distinguish these two fake news categories. This section also mentions the affected areas and likely targets for a particular toxic fake news.

A. Toxic Versus Non-toxic Fake News: Examples

A statement can be considered toxic if it expresses a bad attitude, behaviour, and narcissism towards someone with an intention to ruin a reputation, create harm, or destroy. For example, destroying political careers, defamation, identity threat, sexual abuse, etc. are some of the ill intentions behind toxic behaviour. For example, the viral news. "*Muslims are responsible for spreading COVID-19 infection in India*" is a type of fake news as it targets an individual, community, or organization. This fake news had caused social unrest and communal violence in the country as reported by [3]. Thus we categorize this as toxic-fake news rather than just fake news. And the news. "*The vibration generated by clapping together will destroy Coronavirus infection*" is also fake news about COVID-19 infection as reported by studies and medical experts. But on the other hand, this kind of news does not

Mudasir Ahmad Wani is with the EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia; (Email:mwani@psu.edu.sa)

Mohammed ELAffendi and Ibrahim Mohammed Abuhaimed are with EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia; (Email: affendi@psu.edu.sa)

Kashish Ara Shakil is with the Department of Computer Science, College of Computer and Information Sciences, -Princess Nourah bint, Abdulrahman University, Riyadh, 11671, Saudi Arabia; (Email:kashakil@pnu.edu.sa)

Anand Nayyar is with the School of Computer Science, Faculty of Information Technology, DuyTan University, DaNang, Vietnam (Email:anandnayyar@duytan.edu.vn)

Amir Hussain is with the Edinburgh Napier University, School of Computing, Merchiston Campus, Edinburgh EH10 5DT, UK. (email:A.Hussain@napier.ac.uk)

Ahmed A. Abd El-Latif is with the EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia, and with the Department of Mathematics and Computer Science, Faculty of Science, Menoufia University, 32511, Egypt (email: a.rahiem@gmail.com; aabdellatif@psu.edu.sa)

* Correspondence: mwani@psu.edu.sa; aabdellatif@psu.edu.sa; kashakil@pnu.edu.sa Tel.: +966 (0) 567 389 413

have any bad effect on an individual or a society in general, therefore categorized as non-toxic fake news. Similarly, "*Applying sanitizer on the whole body on two consecutive full moon nights can kill coronavirus*" has no negative impact on any individual or society. However, the sentence "*Drinking sanitizer can kill coronavirus*" is an example of toxic fake news and negatively impacts individuals and society. Table I shows examples of normal fake news and toxic fake news categories from social media. Please note that these examples are not specifically related to COVID-19.

In order to have more understanding of general fake and Toxic-fake news versus the COVID-19-related fake and toxic-fake news, we have framed some more examples for a few toxicity categories in Table II which are specifically related to COVID-19.

An identity attack is directed towards an individual or group of individuals. If this attack is fake news such as "*Dogs are responsible for the spread of coronavirus in India*", it will not have any harmful effect on individuals or groups. However, a toxic identity attack such as "*Tablighi jamaat event is responsible for the spread of coronavirus in India*", is targeted towards a religious group, and causes communal violence in India. The threat is another category of news if it's fake for example "*More than 50% of Asian sugar-items contain saffron!*", it is not going to have any impact on individuals' health or any other ill effect. Whereas if the news is toxic in nature such as "*Indians are smelly and unhygienic*", it is a racist comment directed to demean a particular ethnicity. Similarly, the sentence "*HIV/AIDS is divine punishment for homosexuality*" is toxic fake news to increase homophobia and thus inculcate an ill feeling towards homosexuals. Similarly, the comment "*The main ingredient for any cinema to entertain people is to telecast filthy, vulgar, and pornographic content*" is a general statement belonging to the obscene category and does not have any negative impact on an individual or place. However, the comment "*Reports reveal most of the adults are involved in pornography because of few illegally operated movie theatres in the country*" is a toxic comment directed towards movie theatres.

Thus, we can conclude that toxicity levels in speech may raise dangers and alarming situations for individuals, groups, or government organizations. In particular, during the middle of a global pandemic like COVID-19 where the users are physically and mentally disturbed such toxicity can act as a major roadblock to effective COVID-19 pandemic mitigation strategies and further aggravate the mental health of users. In 2020, the Reuters Institute for the Study of Journalism [4] carried out a study on Twitter tweets about COVID-19 for the months of January to April. As per their reports 21% of the conversation about the COVID-19 pandemic relating to WHO comprised of toxic messages. This percentage reached 25% after March. The toxicity in COVID-19 discussions may fuel the polarization of user opinions and threaten public health management measures. Therefore, it is highly desirable to control the spread of toxic fake news on social media.

B. Motivation

Toxic is a relational term for how someone affects another. Toxic posts and articles will leave you feeling bad: edgy, guilty, confused, frustrated, or overextended. Higher levels and continuous exposure to poisonous content may turn a peaceful online society into an agitated and noisy one and can induce feelings of aggression, violation, and exhaustion in individuals. Thus, galvanized by this topic's importance, this study's motive is to detect fake toxic content in social media data during COVID-19. Filtering out toxic fake content helps us in reducing the amount of time and effort spent studying all the categories of fake news. This work is one of the nascent steps toward analyzing toxic fake news as most of the prior literature has focused only on fake news.

C. Contributions

The following are the major contributions of this work:

- Designing of a Toxic Fake News Detection System (TFNDS) using conventional and Bidirectional Encoder Representations from Transformers (BERT) based Machine learning techniques.
- Creation of an Annotated dataset (will be publicly available for other researchers) which can be useful in several areas such as:
  - Fake News Detection and Classification in general.
  - COVID-19 related Fake News Detection and Classification.
  - Toxic Fake News Detection
  - Toxic and Fake comment classification.
  - Toxic and Fake review identification.
- Estimating toxicity score and performing toxicity analysis on false and real posts, tweets, and news articles circulating in social media and identifying if they are toxic or non-toxic.
- Classification techniques like Support Vector Machine (Linear SVM), BERT-based Support Vector Machine (BERT_SVM), Random Forest (RF), and BERT-based Random Forest (BERT_RF) have been trained and tested on our labeled dataset to predict toxic fake content
- Comparison of the proposed work based on toxicity-based features and BERT-based sequence representations, with existing literature based on performance metrics such as Accuracy (A), F-measure, Precision (P), Recall (R), F_2, F_1, and $F_{0.5}$.

Overall, this research has made a significant contribution to the field of fake news detection and classification, especially in the area of toxic fake news detection. Additionally, the publicly available annotated dataset can be a valuable resource for other researchers in related areas.

The rest of the paper is structured as follows: Section II summarizes recent work in fake news and toxicity analysis, while Section III describe general methodologies and datasets. Section IV explains the specific methods used in designing the Toxic-Fake News Detection System (TFNDS), along with its working architecture. Data collection, pre-processing, and feature engineering strategies are presented in Section V.

Categories	Fake News	Toxic-Fake news
Identity Attack	Life of an individual and business organizations in Kashmir is more comfortable after the scrapping of Article 370-Kashmir Vendors Organization.	Scrapping of Article 370 was the conspiracy of Islamist, pro-Pakistan leader SAS Geelani.
Threat	More than 50% of Asian sugar items contain saffron!	Most of the Indian confectionery products have pig-fat
Hate	LGBTs (lesbian, gay, bisexual, and transgender) are hated because they always wear something with a rainbow design	HIV/AIDS is divine punishment for homosexuality
Obscene	The main ingredient for any cinema to entertain people is to telecast filthy, vulgar, and pornographic content	Reports reveal most adults are involved in pornography because of few illegally operated movie theatres in the country

TABLE I
TOXIC AND NON-TOXIC FAKE NEWS EXAMPLES(GENERAL)

Categories	Fake News	Toxic-Fake news
Identity Attack	Pets such as Dogs and Cats are responsible for the spread of coronavirus in India	Tablighi jamaat event is responsible for the spreading of coronavirus in India
Threat	Clapping together will kill the Coronavirus and protect you from COVID-19 infection-India	Consuming cow dung and cow urine on a daily basis can protect you from COVID-19 infection
Hate	A few groups are spreading infection in the country	One should avoid welcoming china people to protect own people from COVID-19 infection
Obscene	Constant sex kills Coronavirus	Watching constant adult content makes your immune system strong to fight COVID-19 infection

TABLE II
COVID-19 RELATED TOXIC AND NON-TOXIC FAKE NEWS EXAMPLES

Section VI covers the experimental setup, model training, and validation, while Section VII presents the results of different experiments. Section VIII compares the proposed approach with four existing methods based on Accuracy, F-measure, Precision, and Recall. Limitations of the study are discussed in Section IX, and Section X concludes the work on designing the TFND.

II. BACKGROUND STUDY

In the digital age, the spreading of fake news and its consequences on social media has become a significant concern. Several studies have been conducted to address this problem. For example, the study in [5] investigated the causes of fake news spreading on digital media and developed a framework for managing fake news disasters on digital media to prevent the dangers of false information. Similarly, authors in [6] focused on detecting fake news surrounding COVID-19 in Arabic tweets. They collected more than seven million Arabic tweets related to the coronavirus pandemic and relied on two fact-checkers to extract a list of keywords related to misinformation and fake news topics. In another study, [7] the authors identified the relationship between big data analytics with context-based news detection on digital media. They found the trending approaches to detect fake news on digital media and explored the challenges of constructing quality big data to detect misinformation on social media.

As this study is being carried out to help in designing a toxic fake news detection and classification system. Therefore, we put our efforts into digging into the literature on fake news detection and toxicity classification. Both two sections have been very precisely discussed as follows.

A. Fake news in Social Media

Fake news is wrong or fictitious news that has been intentionally fabricated to make readers believe the false information that it provides. It can potentially have an extremely negative impact on individuals and organizations [8]. Falsehood and ambiguity are two characteristics of fake news, with falsehood leading to a loss of 2.11 million USD over a period of just 10 days [9]. Fake news can make people rely on false information, change people's attitudes toward true news, and loss of trust in the news system. A study about the spread of fake news related to COVID-19 [10] by using Technological Determinism theory amongst social media users of a state in Nigeria showed that 74% of respondents of a questionnaire agreed that social media aids in the spread of fake news. The trending topics influenced the spread of fake news and the consequence of this spread is non-adherence to precautions by individuals.

A plethora of research has been carried out to detect and reduce the spread of fake news. Literature has witnessed the employment of several techniques ranging from essential Machine Learning (ML) to advanced Deep Learning (DL), and Natural Language Processing (NLP) to efficiently differentiate between fake and real news about COVID-19. For example, a study [11] has used ten ML-based classification algorithms with seven feature extraction techniques to identify misinformation from the textual data collected through several authentic websites such as WHO, UNICEF, etc. Similarly, the study [12] proposed a two-step method for identifying fake news in social media using supervised artificial intelligence algorithms. Another similar study [13] has proposed a Deep Machine Learning (DML) approach for automatically detecting COVID-19 misleading information on Twitter. This approach apart from ML methods (Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest) also used DL-based algorithms such as LSTM (Long Short Term Memory), and GRU (Gated Recurrent Unit) for prediction and classification tasks. It has also been seen that the power of multiple ML-based algorithms has been used together to distinguish between fake and real content on social media. For example, the study in [14] has been carried out to verify the credibility of COVID-19-related tweets by employing an ensemble approach to tweet and user-based features. Literature has also employed Metaheuristic approaches to combat fake news on social media for example authors in [15]. In [16] authors have utilized metaheuristic algorithms such as the Grey Wolf Optimization (GWO) and Salp Swarm Optimization (SSO) along with other conventional machine learning techniques such as Decision Tree (DT), SVM, etc. to detect fake news on social media.

Talking about DL-based techniques for the analysis of fake news, there are a number of studies from different organizations from time to time. As an example, the authors in [17] proposed an AI and DL model for identifying depression on social media using hybrid features and CNN/LSTM models. Another study [18] proposed an LSTM neural network-based model to detect and differentiate between false and original news. For vector representation of words, a GloVe word embedding has been used. The authors achieved an accuracy of 99.8% and outperformed other algorithms such as BERT, multi-modal ConvNet, tensor decomposition-based deep neural network, Adaptive Salp swarm optimization algorithms [19] etc. A combination of ConvNet-RNN hybrid approach along with LSTM was performed to detect and differentiate fake news from real news [20]. In order to differentiate between untrustworthy news and real news a random forest algorithm along with NLP has been used in [21]. A BERT-based deep learning approach by combining it with CNN is used in [22]. The results show an accuracy of 98.9 % over other existing models. A Graph-aware Co-Attention network (GCAN), a neural network-based model was developed by authors to predict if a source tweet is fake or not and highlights the suspicious retweets [23]. Authors in [24] have used machine learning algorithms such as Random tree, Bayes network, logistic regression, and Naïve Bayes to identify and filter out sites posting fake news. Machine learning techniques along with deep learning have also been used to assess the user's activity on Facebook and thereby detect fake news. Thus, the majority of research in the literature focuses solely on detecting false news, with little preceding work in the field of toxic fake news.

B. Toxicity in Social Media

Social media leads to the unrestrained and unprecedented spread of abusive, rude, toxic, or hate speech. Hate speech is a subclass of toxicity [25], [26]. Toxicity is defined as the use of unwanted, rude, distasteful, and disrespectful language which can cause a user to leave a discussion [27]. Hate speech is targeted towards a particular group of people based on their race, ethnicity, etc. A study [28] carried towards developing a hate speech detection (HSD) system to prevent the spread of toxic posts particularly related to COVID-19 on the Twitter platform. Authors have employed ten machine and deep learning algorithms to classify hate/toxic speech and reported the best accuracy value out of all the employed methods as 90.30%, and 87.22% by Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN), respectively. Since this study is on hate speech which is a variant of toxic news, therefore, we believe that this study is more related to the proposed study in terms of toxic news detection. The results obtained by other methods in this study are given in the section. VIII where we made an attempt to compare our models with the existing approaches. In another study, [29], authors have proposed a metaheuristic-based hate speech detection system using the Ant Lion Optimization (ALO) and Moth Flame Optimization (MFO) algorithm. For feature representation, popular models such as Word2Vec, Bag of Words (BoW), and TF-IDF were employed.

Toxicity in cyberspace has an adverse effect on social network users. The toxic users with their toxic behavior can disrupt and disturbs the normal flow of discussions on any online platform and cause emotional anguish as well as lead people to resort to adverse measures such as suicide. Many prior studies have reported that the root cause of such behavior is boredom [30], to vent out feelings [31] or for the sake of enjoyment [32]. Authors in [33] have investigated Facebook data and shown that people interested in politics are more likely to express polarized opinions and toxic language. They also show that people who make comments on articles in the real-world use more toxic language than the average public. They also demonstrate that posting toxic language in a comment increases the toxicity in all successive comments on that post. Toxicity within YouTube video comments on pro- and anti-NATO channels has been identified by the authors in [27]. Their work focuses on scoring toxicity in user-generated content [27]. Their findings indicate that comments on anti-NATO channels are more toxic than the pro-NATO channels. They obtained toxic scores corresponding to attributes such as threats, insults, hate, sexually explicit, and identity-based attacks for each of the comments using Google's 'Project Jigsaw' and 'Counter Abuse Technology' teams - Perspective API [34]. It is based on Convolutional Neural Network (CNN) and is trained with a word vector. The comments with a toxicity score of less than 0.1 are considered to be non-toxic. In [35] a socio-computational approach has been proposed to examine toxicity propagation in a network. Topic modeling has been used to identify common themes and toxicity brokers and commenters on YouTube. They have shown how toxicity propagates and has an ill effect on other users in the community. The dataset used by them in this work was also based on COVID-19 discourse and comprised 544 channels and around 849,689 comments. Their work also showed that if toxic users are eliminated from the network the overall health of the network in terms of toxicity also improves.

1) Classes of Toxicity: Toxicity comprises different types and forms. Authors in [25] have described five different classes of toxicity. These classes include the following general categories:

- **Obscene Language/Profanity:** In this class a blacklist of disrespectful words such as swear or curse words is used for identification.
- **Insult:** This class comprises rude or offensive statements against an individual or a group. They comprise sentences that are usually directed towards a user or a group.
- **Threat:** This includes severe toxic comments which can be life-threatening for a user or their families. Statements describing pain, punishment, or damage to others are used in threats.
- **Hate Speech/Identity Hate/ Identity Attack:** This category presents an attack directed towards a group based upon their religion, ethnicity, gender, etc. Racist, homophobic, and misogynistic comments are part of this category.
- **Otherwise Toxic:** This class includes comments that are not part of the other four categories but are directed towards individuals in such a way that they make them leave a group or discussion fall in this class. Trolling and spamming are examples of this class.

Based on these base toxicity classes, several researchers and studies from time to time came up with a few more, or we can say fine-grained toxicity sub-classes. Detoxify [39] is a Python package to predict if a comment is toxic or not, it describes seven classes of toxicity. This package has also been used by us in this work. It includes a toxicity class comprising all the comments that are generally toxic in nature but not severe or life-threatening. Severe Toxicity class comprised of comments that are severely toxic and life-threatening. An obscene class comprising of comments that are morally offensive and talk about sexual matters. Identity attack class that aims to attack and ruin the reputation or goodwill of an individual or an organization. Insult class comprises rude or disrespectful comments on an individual or group or organization. Threat class comprises statements or comments intended to harm a person or organization and cause danger or harm. The sexually Explicit class comprises statements, descriptions, or pictures that are related to nudity and sexual acts such as sex and masturbation.

In order to make it easy for researchers we have presented the literature around fake news detection and hate speech detection in a tabular form in table III. This table shows the most popular and recent studies conducted under the umbrella of fake news on social media. It provides details such as methods employed, datasets used, and outcomes from each of these studies. Thus, provides assistance and makes it easy for researchers working in this field to identify research gaps and innovations.

III. Materials and Methods

Here in this section, we will talk about the dataset used for the proposed study and the methods employed to carry out the experiments. The experimental setup is presented separately in section VII.

A. The Data-sets

In general, we used four datasets *D1, D2, D3, and E1* in this study. *D1* and *D2* comprise old and latest fake news instances respectively, from several social media platforms such as Facebook and Twitter. The combination of *D1*, and *D2* gives dataset *D3* ($D1 = D2 + D3$) which contains all the instances (both old and latest) from news sources. Toxicity-based feature scores of each instance are calculated in the dataset *E1* and for BERT-based representation, the dataset *D3* has been used. The whole process of preprocessing these datasets is given in section V A quick glance at the four datasets is given as under:

- **Dataset 1 (D1-CoAID):** *post_id, post_category, article/news text, label (fake/non_fake)*
- **Dataset 2 (D2):** *post_id, post_category, article/news text, label (fake)- [based on top keywords from fake news articles]*
- **Dataset 3 (D3):** *post_id, post_category, All fake labelled instances from D1 + All instances from D2, label (fake)*
- **Dataset 4 (E1):** *post_id, post_category, toxicity_based features [calculated using Detoxify package [39].*

The whole scenario of creating the above datasets is shown in figure 3

B. Dataset Description

Dataset D1 comprises CoAID (COVID-19 healthcare misinformation Dataset), which is a COVID-19 healthcare misinformation dataset. It includes false news available on the internet and different social media platforms. It also contains information about the users' social involvement in such news. It comprises 5,216 news, 296,752 related user engagements, 958 social platform posts relating to COVID-19, and labels. D2 comprises data collected from different social media platforms such as Twitter, Facebook, and news portals. This data is collected based on top keywords from D1. This dataset comprises user ID, tweet/post/news_text, and a label representing a false or real post.

D3 dataset is obtained by combining together D1 and D2. It comprises the attributes: user ID or news URLs, tweet text or post text, and a label representing false or real posts. E1 dataset which is our experimental dataset comprises nine attributes user ID or news URLs, tweet text or post text and seven toxicity attributes. In addition to this, it comprises two labels: label one indicates whether the news is real or false, and label two indicates if it is toxic or non-toxic. The seven toxicity attributes include *Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat, and Sexually Explicit*. Toxicity attribute

Study/ Work	Data Source(s)	Approaches/ Techniques	Problem Domain	Performance Measures (Results)
Ozbay et al. [12]	Buzzfeed Political News Data Set, Random Political News Data Set	Decision Tree, Zeror, CVPS, WIHW-Algorithms	FND	Accuracy = 96.8%, Precision = 96.3%, Recall = 97.3% and F-measure = 96.8%

Shu et al. [36]	Fakenewsnet Buzzfeed And Politifact Datasets Based On News Article, Facebook Posts Or Tweet	Rhetorical Structure Theory (RST), Linguistic Inquiry Word Count (LIWC), Castillo	FND	Accuracy=86.4%,
				F-measure = 87%

Elhadad et al.[11]	Websites Of WHO, UNICEF, UN	DT, KNN, LR, LSVM, MNB, BNB, NN, ERF, Xgboost	FND	Accuracy = 99.68%, Error rate = 0.32% Area under curve = 99.47%, F1-Score = 99.89%
Abdelminaam et al.[13]	Twitter	RF, SVM, KNN, SVM, Bayesnet	FND	Accuracy = 97.8%, AUROC = 99.7%
Zhou et al. [37]	Real-World Datasets Politifact And Buzzfeed	Base Machine Learning models	FND	Accuracy = 89.2%, Precision = 87.7%, F-measure = 89.2%, Recall = 89.3%
Ozbay et al.[15]	Buzzfeed Political News, Random Political News, LIAR Dataset, ISOT Fake	Basic Machine Learning Algorithm, Salp Swarm Optimization	FND	Accuracy = 92.6%, Precision = 92%, Recall = 83.5%, and F-measure = 91%
Ozbay et al.[16]	Buzzfeed Political News, Random Political News, LIAR Dataset	Grey Wolf Optimization (GWO) and Salp Swarm Optimization (SSO)	FND	Accuracy = 92.6%, Precision = 100% Recall = 85.1%, and F-measure = 91%
T. Chauhan and H. Palivela [18]	Glove Twitter Data, Fake And Real News Dataset	LSTM	FND	Accuracy = 99.88%
F. Ma and G. Tan [20]	Snopes.com, PolitiFact, LIAR dataset,	ConvNet-RNN hybrid, Siamese LSTM	FND	Accuracy = 90%
Kaliyar et al.[22]	Real-world fake news dataset based on 2016 U.S. General Presidential Election https://www.kaggle.com	BERT-based deep Learning, CNN	FND	Accuracy = 98.9%
Y.-J. Lu and C.-T. Li [23].	Twitter 15 and Twitter 16	Graph-aware Co-Attention Networks	FND	Twitter 15: Accuracy = 87.67%, Precision = 82.57%, Recall = 82.95%, F1Score = 82.5% Twitter 16: Accuracy = 90.84%, Precision = 75.94%, Recall = 66.32%, F1-Score = 75.93%
Pérez-Rosa et al. [38]	Crowdsourcing-based dataset covering, Celebrity fake news datasets.	Linear SVM	FND	Accuracy = 81.1%, Precision = 74.2%, F-measure = 81.1%, Recall = 75.1%
He et al. [28]	Twitter Dataset	Ten machine and deep learning algorithms	HSD	Accuracy(RNN) = 90.30%, Accuracy(CNN) = 87.22%
S.-H. Lee and H.-W. Kim [31]	Three data sets based on Twitter and internet forums	Ant Lion Optimization (ALO) and Moth Flame Optimization (MFO) algorithm	HSD	Accuracy = 92.1%, Precision = 88.6%, Recall = 89.5%

TABLE III
POPULAR STUDIES CARRIED OUT IN THE DIRECTION OF FAKE NEWS DETECTION ON SOCIAL MEDIA.

comprises of text which has an ill effect on the community members. Severe toxicity comprises words or text which are highly toxic. Obscene refers to words or text targeting the morality of the users. Identity attack refers to text which causes the identity of social media users in jeopardy. Insult happens when disrespectful speech is given against individuals on web platforms. Threat refers to text and comments that threaten an individual. Sexually Explicit refers to text which promotes nudity and sexual activities. Toxic analysis has been carried out on E1 to prepare the toxicity-based attributes for the training of our classification models.

C. Methods

Here in this section, we will briefly explain the methods and machine learning techniques utilized in the proposed study. In order to present the data to the model in the machine-readable format, we first used the BERT (Bidirectional Encoder Representations from Transformers)[40]. BERT is a neural network-based technique (for natural language processing) designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right contexts in all layers.

BERT makes use of Transformer, to analyze the relations between words and sub-words in a paragraph. The transformer contains two components — an Encoder and a Decoder. The encoder component reads the entire user text of words at once in contrast to directional models, which read the text input either from left-to-right or right-to-left. This flexibility of BERT makes the model learn the context of a word from all directions. The decoder component performs the prediction. Since BERT's main aim is to produce a language model, therefore the main focus is on the encoder mechanism. In our case, we also utilized the BERT's encoder component to have the machine-understandable representation of the user text. BERT comes in two variants, first is known as BERT_large, having around 345 million parameters, and is considered the largest model of its kind. The second is called BERT_base, with the same architecture design (as BERT_large) but having only 110 million parameters and is considered suitable for small tasks.

Apart from the BERT models we have used traditional algorithms such as Linear Support Vector Machine (Linear SVM) [41] and Random Forest[42] for the classification task. The following section presents the specific techniques utilized in the proposed study in more detail.

IV. Methodology for the proposed System

A. System Model

Every news whether fake or non-fake has an emotion associated with it. According to Plutchik's [43] human emotions can be categorized into dual categories positive and negative comprising of seven basic emotions [44], [45]. Positive emotions include emotions such as *Joy, Trust, and Anticipation* while negative emotions comprise *Fear, Anger, Sadness, and Disgust*. However, the *surprise* emotion can be both positive and negative. The negative emotions in the context of social media can be further defined as toxic emotions. Fake news generally has these toxic and non-toxic emotions associated with it. Toxicity refers to any undesirable or unwanted behaviour shown by the users on the internet with the intent to *offend, insult*, or cause *harm* to an individual or community [46]. These toxic emotions can have several attributes such as *Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat, and Sexually Explicit.*

Furthermore, we want to clarify that we are not mixing emotions and toxicity of a user post. In Figure 1, the sub-figures a and b present the basic emotion categories and toxicity categories found in a text sentence (or a paragraph) respectively. To make the difference between the two more clear from the technical point of view that how emotion mining and toxicity analysis are different, below we are mentioning a general example.

Example: Sentence (S1): "*War between two countries always ends up with the loss of important souls. The world is Bleeding. GOD has to help everyone.*"

Table IV shows the results of toxicity analysis obtained by calculating toxicity scores for Sentence S1. These scores correspond to different toxicity classes such as *Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat, and Sexually Explicit*. The formal way of calculating the toxicity attributes, in general, is as follows:

$$TE_{i_j} = \Sigma_k frequency(ToxicLexicon_{j,k}), \forall_{j=1..7} \qquad (1)$$

Seven toxicity attributes of the s^{th} sentence (TE_{i_1}, $TE_{i_2}, \ldots, TE_{i_7}$) correspond to seven classes of Toxicity lexicon, respectively. For example, consider $ToxicLexicon_{j,k}$ represents the k^{th} toxicity term in the j^{th} class. Then, values from TE_{i_1} to TE_{i_7} can be determined by Equation (1), where $ToxicLexicon_{j,k}$ returns the toxicity word to be counted in the s^{th} sentence.

Likewise, to calculate the overall toxicity of a post (given sentence here) comprising all the seven toxicity attributes, we use the following formula given in Equation (2)

$$Toxicity(T_{Sentence}) = \sum_{n=1}^{i} X_i = [TE_i, TE_{i+1}, TE_{i+2}, ...TE_{i+n}] \qquad (2)$$

Where $Toxicity(T_{Sentence})$ is the toxicity of given sentence (S), or a user post (P). TS is the toxicity score of an attribute n and n is the toxicity attribute comprising of *Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat and Sexually Explicit.*

Table V shows the results of emotion analysis obtained by calculating emotion scores for sentence S1. These scores correspond to different emotion classes such as *Joy, Trust, Anticipation, Fear, Anger, Sadness, and Disgust*. The mathematical formula used to extract the emotions from the given

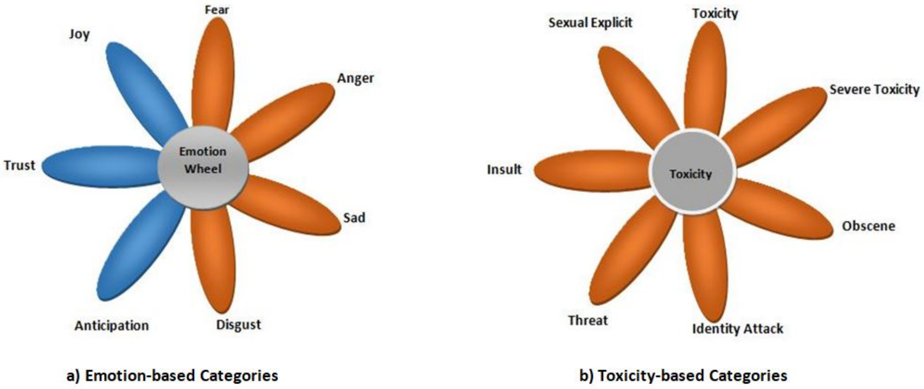


Fig. 1. Basic Emotion and Toxicity-based features (Defined in [45], and [39], respectively).

Toxicity Class	**Toxicity Score (by Detoxify Package [39])**
Toxicity	0.007462
Severe_toxicity	2.469439
Obscene	0.000129
Identity_attack	0.000138
Insult	0.000163
Threat	7.068944
Sexual_explicit	0.634636

TABLE IV
TOXICITY ANALYSIS ON EXAMPLE SENTENCE S1

sentence S is presented as equation 3. The MoodBook dictionary [45] has been utilized to obtain these emotion-based attributes. This lexicon provides a list of emotion terms for the above eight classes of emotions. For example, the fear category includes words like "*war, horror, panic, etc*" and the sad category includes terms such as "*cry, missing, painful, alone, phobia, etc.*" and "*hopeless*".

Based on the MoodBook, a total of 8 emotion-based attributes $((E_{i_1}, E_{i_2}, \ldots, E_{i_9}))$ are constructed from a user post, news article or a Tweet. Let $moodbook_{j,k}$ represents the k^{th} emotion term in the j^{th} class. Then, values from E_{i_1} to E_{i_7} can be determined by Equation (3), where $moodbook_{j,k}$ returns the emotion word to be counted in the S^{th} user post, news article or a Tweet.

$$E_{i_j} = \Sigma_k frequency(moodbook_{j,k}), \forall_{j=1..n} \qquad (3)$$

Furthermore, from this analysis, we can see in table V that the emotion score for negative emotions in sentence S1 is more in comparison to positive emotions.

Emotion Category	**Emotion Score (by Moodbook lexicon [45])**
Joy	0.00005
Trust	0.00035
Anticipation	0.14280
Fear	0.14281
Anger	0.07142
Sadness	0.21425
Disgust	0.00005

TABLE V
EMOTION ANALYSIS ON EXAMPLE SENTENCE S1

B. Architecture and Working of System

First, we started with a BERT-based representation of each sequence (Tweets, posts, and news articles) from the dataset. Secondly, we calculated the toxicity score of these sequences (Tweets, posts, and news articles) and had these two different representations stored in a file. After that, we experimented with four machine learning variants namely, Support Vector Machine (SVM), Random Forest (RF), and their BERT-based variants (bert_svm, bert_rf) to develop a toxic fake news classifier. Figure 7 presents the methodology for designing the architecture of the proposed models.

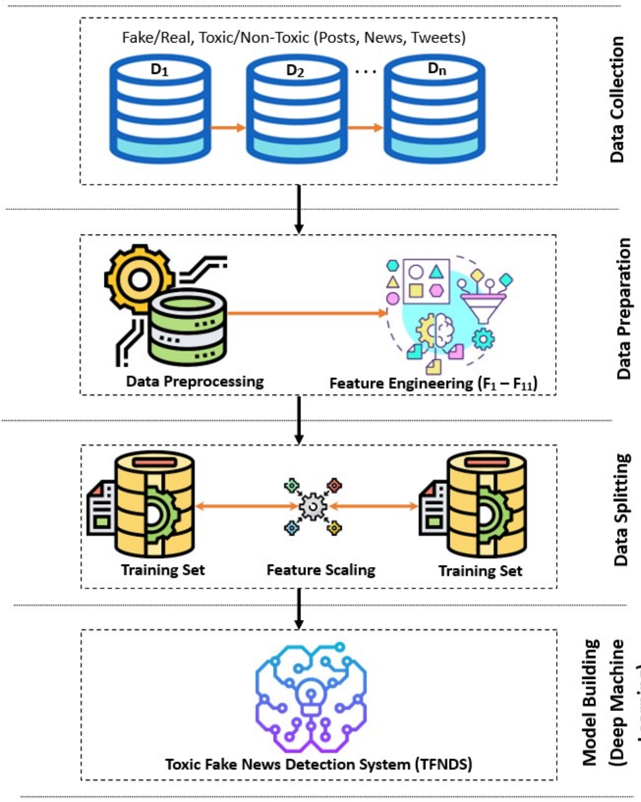

Fake/Real, Toxic/Non-Toxic (Posts, News, Tweets)
D1
D2
...
Dn
Data Collection
Data Preprocessing
Feature Engineering (F1 – F11)
Data Preparation
Training Set
Feature Scaling
Training Set
Data Splitting
Toxic Fake News Detection System (TFNDS)
Model Building (Deep Machine Learning)

Fig. 2. Methodology employed in the proposed study.

Figure 2 shows the methodology adopted in this study. As one can note from the figure it is divided into four stages. In the first stage, we start with locating the datasets for fake news and toxicity classification.

Since we could not find the optimal and feasible datasets for both problems together, therefore we decided to use an existing dataset(s) related to fake news on social media and calculate toxicity for each post in the dataset as explained in V-B section. In the second stage, we performed the data preparation which includes pre-processing of data and feature engineering. In stage third, methods used to normalize the range of independent features of data have been applied. Finally, at stage four we performed model training by employing two machine learning algorithms. In the following sections, we have further explained the overall methodologies used in this study.

V. Data collection and Feature Engineering

A. Data Collection

Our data collection process and designing of a new dataset employs multiple steps as shown in Figure 3

Each step carried out during data collection is briefly explained as under:

- **Step 1:** We first used dataset D1 from Github [47] comprising both false and real tweets, Facebook posts, and news articles about COVID-19.
- **Step 2:** In order to have the latest news articles and a larger dataset (for more accurate predictions), we generated top keywords corresponding to false and real posts from D1. Based on these keywords we crawled data from different social media platforms such as Facebook, Twitter, and news portals. Manual annotations were then performed on this dataset to label them as fake/real by three experts to generate a new dataset D2. The disagreements were resolved using majority voting and averaging wherever required to derive single ground truth labels from multiple annotations.
- **Step 3:** The datasets in Step1 and Step 2 i.e. D1 and D2 were then combined to generate a new dataset "D3". This dataset comprises of user_IDs or news URLs, Tweets or posts, and a label indicating if the news is real or false.
- **Step 4:** To perform future machine-based toxicity predictions, toxicity analysis was performed. Based on the toxicity score generated after toxicity analysis, we labelled dataset D3 as either toxic or non-toxic and obtained a new dataset E1 (experimental dataset). E1 dataset thus comprises User IDs or news URLs, Tweets, or posts, a label indicating if the data is real or false and another label indicating toxic or non-toxic.

The toxic label assigned in step 4 is based upon the summation of the toxicity scores of all the attributes as per equation4 below. It should be noted that in the instances for which the labels are not present, for example, the newly collected instances the formal annotation process is applied followed by a manual human check. Otherwise, the labels are not modified at any stage of experiments and are kept the same as were in the existing datasets.

$$Toxicity(T_P) = \sum_{n=1}^{i} X_i = [TS_i, TS_{i+1}, TS_{i+2}, ...TS_{i+n}] \tag{4}$$

Where $Toxicity(T_P)$ is the toxicity of a user post P, TS is the toxicity score of an attribute n and n is the toxicity attribute comprising of Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat and Sexually Explicit.
For finding non-toxic score ($Non\_Toxicity(NT_P)$) we use the equation below:

$$Non\_Toxicity(NT_P) = (1 - Toxicity(P)) \tag{5}$$

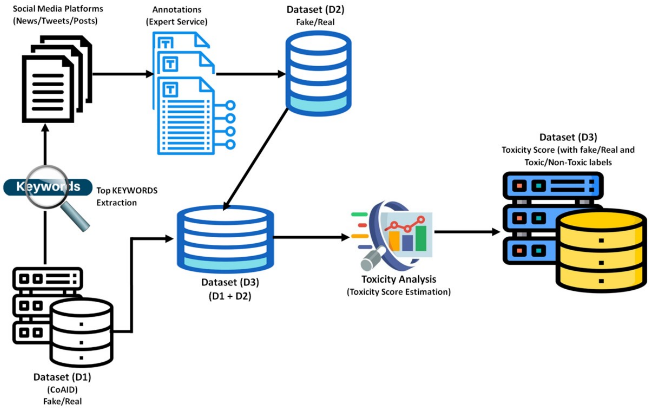


Fig. 3. Data Collection and Preparation Process

We label a post as toxic whenever $Toxicity(T_P) \geq Non-Toxicity(NT_P)$ score. We label the sequence as toxic even when the values of $Toxicity(T_P)$ and $Non-Toxicity(NT_P)$ are equal because here in this study we care and are more interested to deal with the toxic content as much as possible. Therefore, even with very poor signals of toxicity, we consider it toxic until proven wrong. Also, we need to consider the ill effects of toxicity spread on social media and safeguard the netizens against it. We labelled the data as non-toxic when $Toxcity(T_P) < Non-Toxcity(NT_P)$.

B. Data Cleaning

The data collected contains a lot of redundant and repetitive information. Therefore, in order to remove such unnecessary information following methods have been adopted:

- Exclude repetitive tweets: Dataset D2 may contain repetitive tweets. Therefore, these redundant tweets were removed.
- Remove retweets, hashtags, modified tweets, and emoticons.
- Remove punctuation
- Remove stop words: Remove stop words as they are not required for analysis.
- Lemmatization: Perform lemmatization to convert data to its root form.

In order to have the capacity to gain an accurate and deep understanding of data from both toxic and non-toxic COVID-19-related news categories after pre-processing, we made an attempt to visualize it using word clouds as shown in figure 4 and figure 5 below:

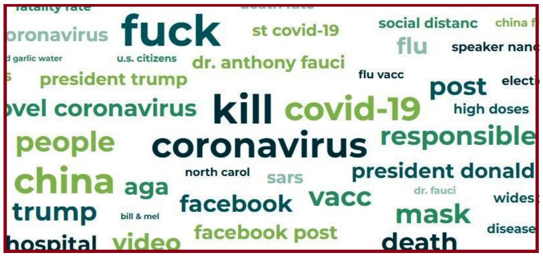


Fig. 4. Word Cloud from the Toxic Fake corpus

From the word cloud figures of both categories, we can clearly notice several distinguishing vocabulary terms used by the people and/or news organizations while posting about COVID-19. For example, we see the dictionary terms like *f***(1530), Kill(976), China(389), Responsible(441)*, etc. have been used mostly used by people from the toxic fake news category. These words with other several terms can be in the toxic fake news word cloud as well.

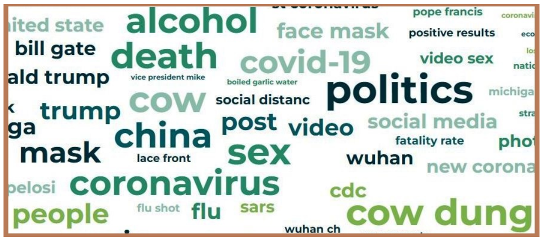


Fig. 5. Word Cloud from the Non-Toxic Fake corpus

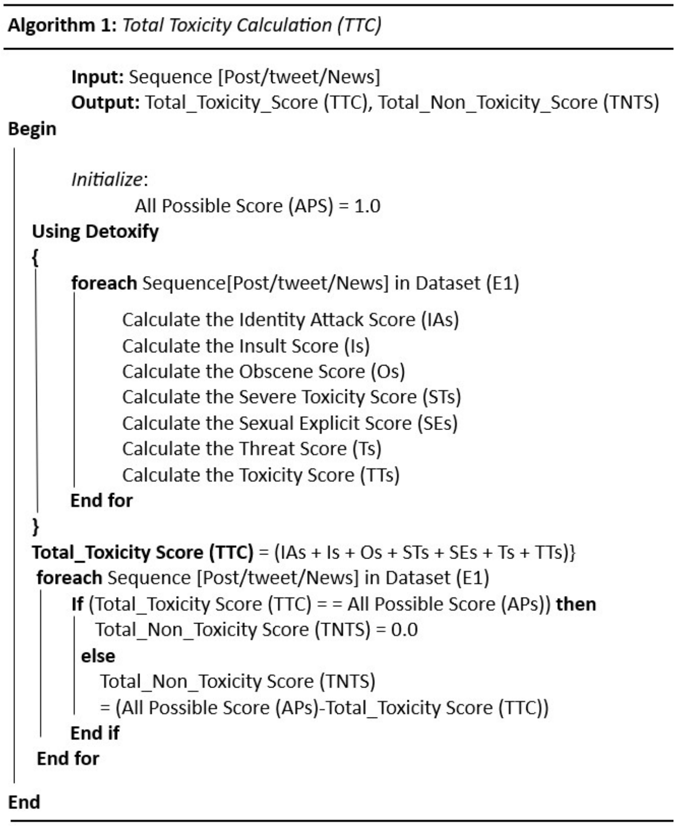
After analyzing the data we found that these are the top terms used by people in their sentences to express their hatred or other negative emotions towards people, country, and communities. The examples from the dataset include:

"*China is responsible for spreading COVID-19*"
"*Muslims are responsible for COVID19 spread in India*"
"*The only solution to stop infection is to kill infected ones*"
"*China is responsible for COVID19 spread*", etc.

On the other hand, the terms such as *Politics(970), Sex(378), cow dung(518), China(366), mask(288)*, etc have been seen as top terms used by people while sharing their opinions or expressing feelings about COVID19. Upon analyzing the non-toxic fake news corpus we found a number of sentences corresponding to these words including:

"*Coronavirus began with a man having sex with a bat*"
"*Having continuous sex with your partner will reduce the chances of COVID-19 infection*"
"*China is the only country with COVID-19 vaccine*"
"*face mask play no role in controlling infection*"
"*Cow dung cakes are the best and only cure for COVID-19 disease*", etc.

So we noticed that there are two kinds of fake news related to COVID-19, toxic-fake and non-toxic fake, as evident from the dataset and word clouds plotted in this study. Talking about toxic fake news posts and articles, they surely have a very bad impact and effect the society, for example, the sentence "*China is responsible for COVID-19 spread*" can cause hatred and discrimination in people for each other's community and can cause racist and xenophobic violence in the society. While the non-toxic sentence such as "*Having continuous sex with your partner will reduce the chances of COVID-19 infection*", and "*China is the only country with COVID-19 vaccine*", etc. will not have much effect on the public and society, therefore, can be ignored at this stage. After performing this Exploratory Data Analysis using the word cloud and manual data analysis we understood the toxicity-based features have the potential to distinguish the two fake news categories. The following section will discuss the features used in our experiments.



Algorithm 1: *Total Toxicity Calculation (TTC)*

```
        Input: Sequence [Post/tweet/News]
        Output: Total_Toxicity_Score (TTC), Total_Non_Toxicity_Score (TNTS)
Begin
        Initialize:
                All Possible Score (APS) = 1.0
   Using Detoxify
   {
        foreach Sequence[Post/tweet/News] in Dataset (E1)
               Calculate the Identity Attack Score (IAs)
               Calculate the Insult Score (Is)
               Calculate the Obscene Score (Os)
               Calculate the Severe Toxicity Score (STs)
               Calculate the Sexual Explicit Score (SEs)
               Calculate the Threat Score (Ts)
               Calculate the Toxicity Score (TTs)
        End for
   }
   Total_Toxicity Score (TTC) = (IAs + Is + Os + STs + SEs + Ts + TTs)}
    foreach Sequence [Post/tweet/News] in Dataset (E1)
        If (Total_Toxicity Score (TTC) = = All Possible Score (APs)) then
           Total_Non_Toxicity Score (TNTS) = 0.0
         else
           Total_Non_Toxicity Score (TNTS)
           = (All Possible Score (APs)-Total_Toxicity Score (TTC))
        End if
    End for
End
```

Fig. 6. Algorithm 1: Total Toxicity Calculation(TTC)

C. Feature Engineering

Feature engineering is the process of selection and transformation of the most distinguishing and meaningful variables from raw data. Its goal is to improve the performance of machine learning models to focus on data more than developing new algorithms [48]. It allows the creation and selection of important predictive variables for predictive models. Feature extraction comprises feature creation for identifying important variables, transformation i.e. manipulating predictor variables to improve the performance of models, feature extraction for automatic creation of new variables from raw data, and feature selection to identify the most useful features and remove irrelevant or redundant features [49], [50].

We performed feature engineering on our raw data and obtained seven toxicity-based features ($F_1 - F_9$) which include *Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat, and Sexually Explicit*. Their values comprise toxicity scores from a range of 0-1. We also had two additional features which are derived from features $F_1 - F_9$ namely toxicity and non-toxicity. Furthermore, we also have two class labels representing features F_{10} showing if a tweet is toxic or non-toxic and F_{11} indicating if the post is real or false. Thus, a total of 11 features were created at the phase as shown in table VI. This table shows features used for analyzing text data in terms of toxicity. F_1 to F_7 represents different types of toxicity, such as Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat, and Sexually Explicit, which can have values between 0 to 1. F_8 to F_9 are numerical/nominal features indicating the Toxicity-Score and Non-Toxicity-Score respectively, which can have values between 0 to 1. F_{10} and F_{11} are class labels representing Toxic/Non-Toxic and Fake/Non-Fake, which can have values of either 0 or 1.

The procedure to calculate the total toxicity and non-toxic score of news, post, or tweets is presented in Figure 6. The output of the algorithm is two values Total_Toxicity Score (TT_C), and Total_Non_Toxicity Score (TNT_S) representing the toxicity and non-toxic values recorded for an English sequence.

In the case of BERT variants of SVM, and RF, we supplied a vector of the BERT of size 768 to develop the BERT_SVM and BERT_RF toxic fake news detector. This is just adding a classification layer on top of the encoder output. In the case of linear SVM and conventional RF, toxicity-based scores for each toxicity category are supplied as vectors having size 9 to train toxicity classification models.

VI. Experimental Setup

The prime objective of this paper is to detect toxic fake news related to COVID-19 disease on social media and thus limit the efforts put into investing in all the fake news categories which are not otherwise toxic. For our experiments, we initially started with two datasets CoAID [47] and we referred to it as D1, we also collected the recent tweets and posts using the tweepy [51] package based on top keywords from fake news articles from social media and stored it into dataset D2. Furthermore, to have a comparatively larger fake news dataset we combined all the fake news-labeled instances of D1 and all instances from D2 into the new dataset D3. Finally, out of the D3 dataset, we created one more machine-understandable dataset E1.

A. Training and Validation

To obtain the potential of toxicity-based features in designing a ToxicFake News detection system, we performed experiments using BERT-based SVM, Random Forest, Linear_SVM, and conventional Random Forest. For Bert-based algorithms (bert_svm and bert_rf) we supplied the D1 dataset for training and validation. Similarly, for non-bert-based versions of these algorithms (SVM and Random Forest), we used dataset E1. The training data in all four experiments have been split into the ratio of 80:20 for training and validation respectively. Finally, all the algorithms have been evaluated and tested on a separate subset of the original dataset. To evaluate the performance of the model we used commonly known performance measures such as Accuracy, Precision, and Recall. Based on the confusion matrix given in Table VII, the formula for calculating each of the metrics is as under:

$$Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{6}$$

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \tag{7}$$

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{8}$$

To analyze the trade-off between precision and recall, F-measure has been used. The F-measure or F-score is a measure of the model's accuracy on a given dataset used to combine the precision and recall of the model and is defined as the harmonic mean of the model's precision and recall. The formula for calculating F-measure is as under:

$$F_\beta - score = (1 + \beta^2) * \frac{\mathrm{PR}}{\beta^2\mathrm{P} + \mathrm{R}} \tag{9}$$

The F$_\beta$ score is a generalization of the F-score which adds a configuration parameter called beta (β). The default value for β is 1.0, which is the same as the F-measure or referred to as F1-score. A smaller beta value (such as β = 0.5) gives more weight to precision and less to recall, whereas a larger beta value (such as β = 2.0) favours recall more than precision. It is helpful to use both precision and recall, but slightly more attention is needed on one or the other, such as when false negatives are more important than false positives or vice-versa.

B. Running Environment

It is now clear that the proposed study has conducted around four experiments on two datasets(D1 and E1) using two basic machine learning algorithms (SVM and RF). As per the basic data splitting rule, we have divided our final corpus into the ratio of 80:20 for training and validation respectively for all the experiments. All the experiments were conducted under the same setting with hardware features as follows: Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz, 16GB RAM, 8GB NVIDIA GEFORCE RTX 3080 GPU card, and 1TB ROM. Since BERT is a huge model with more than hundreds of millions of parameters. Therefore, a GPU is needed to fine-tune as well as in inference time.

In order to reproduce the experiments all the source code files, employed datasets, and trained models are made available in our GitHub[1] repository for the researchers working in this domain. Access to further datasets will be granted upon request under a few predefined data-sharing regulations. Furthermore, to make it more convenient for the researcher of the same domain the parameter (files) required to conduct each

[1] https://github.com/Mudasir-IIIT-Bangalore/Toxic_Fake-News-Detection-in-OSN/tree/main

Feature	Feature Type	Possible values/category	Range
$F_1 - F_7$	Toxicity_based	Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat and Sexually Explicit	0-1
F_8	Numerical/Nominal	Toxicity-Score	0-1
F_9	Numerical/Nominal	Non-Toxicity-Score	0-1
F_{10}	Class1-Label	Toxic/Non_Toxic	0 or 1
F_{11}	Class2-Label	Fake/Non_Fake	0 or 1

TABLE VI
FEATURES USED FOR TOXICITY CLASSIFICATION

		Actual	
		Positive	**Negative**
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

TABLE VII
CONFUSION MATRIX

experiment is mentioned in table VIII. The table provides a detailed overview of four different experiments aimed at testing the effectiveness of different feature types and classifiers in two different datasets. Two feature types, namely BERT-based and toxicity-based, and two classifiers, namely SVM and RF, were used in the experiments. The experiments were divided into training and testing stages, and for each stage, the relevant dataset, feature type, and parameter/file details were provided. The purpose of these experiments was to evaluate the performance of different feature types and classifiers in different datasets and to provide insights into the most effective techniques for classification tasks.

VII. RESULTS AND DISCUSSION

As of now, we understood that both the datasets (D1 and E1) hold the same data with different machine representation formats. Here, our main aim is to investigate and analyze the potential of bert representation and the toxic features to develop an efficient classification model. This experimental setup is more clearly presented in Figure 7.

Table IX and table X shows the results obtained after applying the selected machine learning techniques to the two datasets (D1 and E1) based on precision, recall, and accuracy, and F$_\beta$ evaluation metrics, respectively.

The results clearly show that all classification techniques achieved significantly good results on both datasets. It can be seen from the outcome of the results that the conventional machine learning techniques (SVM and Random Forest) performed better than their BERT-based versions.

Out of which SVM performed better on dataset E1 in terms of accuracy (92.49%), recall (94.23%), F_1-score (95%), $F_{0.5}$ (85%) and F_2-score (87 %) than Random Forest. However, the Random Forest algorithm outperformed Linear_SVM in terms of precision score. Furthermore, slightly lower performance is recorded for all the Bert-based models on dataset D1, which contains pre-trained Bert-based features of direct textual instances. However, within BERT-based techniques, the BERT_SVM performed well on all the evaluation measures than the BERT_RF approach. Thank you very much for the valuable and insightful comment on the performance of proposed approaches for toxic fake news detection. Upon comparing the results between different approaches used in our study we found that the linear SVM and its bert-based variant outperformed Random conventional forest and its bert-based version, respectively. One of the reasons why SVM performed better than RF is that the SVM is considered intrinsically better for two-class problems and Random forests are designed to provide solutions to multi-class problems. Our dataset contains two class instances i.e. toxic fake-news instances and normal news samples, therefore, SVM is a suitable choice and thus performed better. Also, Random Forest works well with a mixture of numerical and categorical features. On the other hand, SVM maximizes the margin and thus relies on the concept of distance between different points, and SVM-based models perform better on sparse data than tree-based approaches in general. Since we are using features generated using toxicity analysis most of the attributes have values near zero which also gives the chance to SVM to show better performance.

On comparing the results obtained on datasets D1 and E1 we can clearly notice the difference under all the evaluation metrics. For the dataset D1, we obtained an accuracy of 65.37% and 64.56% by BERT-based SVM and BERT-based Random forest respectively. While as for linear SVM and conventional Random forest, we recorded accuracy as 92.17%, and 89.56% respectively. Therefore it can be concluded that conventional machine learning techniques on toxicity-based features performed well than the Bert-based model on the dataset. Hence, we obtained a simple but efficient machine-learning model for toxicity classification. Since we are dealing with fake instances only, thus our main objective of this study which is to design a toxic-fake news detection system is accomplished successfully.

VIII. COMPARATIVE STUDY

This study aims to the identification of *toxic fake news* and does not pay any attention to the general fake news which is not toxic at this stage. From the literature we were not able to find a study that is directly related to the proposed one, therefore, at this stage, it seems better to compare our results with the fake news studies in the literature as in the end, it is comparing the techniques and models, not the content. Thus, we identified some popular studies carried out for identifying fake news on social media.

Table XI below presents the comparison between the results obtained by existing studies and the proposed one based on performance measures such as accuracy and F-measure.

Experiment	Stage	Feature Type	Dataset	Parameters / Files
Exp-1 (BERT-SVM)	Training	BERT_based	D1 (Training-Data)	Embedding-feature-file (*embedding_train.text*) Bert-feature-file (*bert_training.csv*)
-	Testing	BERT_based	D1 (Test-Data)	Embedding-feature-file (*embedding_test.text*) Bert-feature-file (*bert_test.csv*)
Exp-2 (Linear-SVM)	Training	Toxicity_based	E1 (Training-Data)	Toxicity-feature-file (*toxicity_train.text*) Toxicity-label(Training) (*toxicity_training.csv*)
-	Testing	Toxicity_based	E1 (Test-Data)	Toxicity-feature-file (*toxicity_test.text*) Toxicity-label(Test) (*toxicity_test.csv*)
Exp-3 (BERT-RF)	Training	BERT_based	D1 (Training-Data)	Embedding-feature-file (*embedding_train-RF.text*) Bert-feature-file (*bert_training-RF.csv*)
-	Testing	BERT_based	D1 (Test-Data)	Embedding-feature-file (*embedding_test-RF.text*) Bert-feature-file (*bert_test-RF.csv*)
Exp-4 (Linear-SVM)	Training	Toxicity_based	E1 (Training-Data)	Toxicity-feature-file (*toxicity_train-RF.text*) Toxicity-label(Training) (*toxicity_training-RF.csv*)
-	Testing	Toxicity_based	E1 (Test-Data)	Toxicity-feature-file (*toxicity_test-RF.text*) Toxicity-label(Test) (*toxicity_test-RF.csv*)

TABLE VIII
PARAMETER SETTING FOR EXPERIMENTS CONDUCTED UNDER THE SAME PLATFORM WITH TWO FEATURE REPRESENTATION TECHNIQUES.

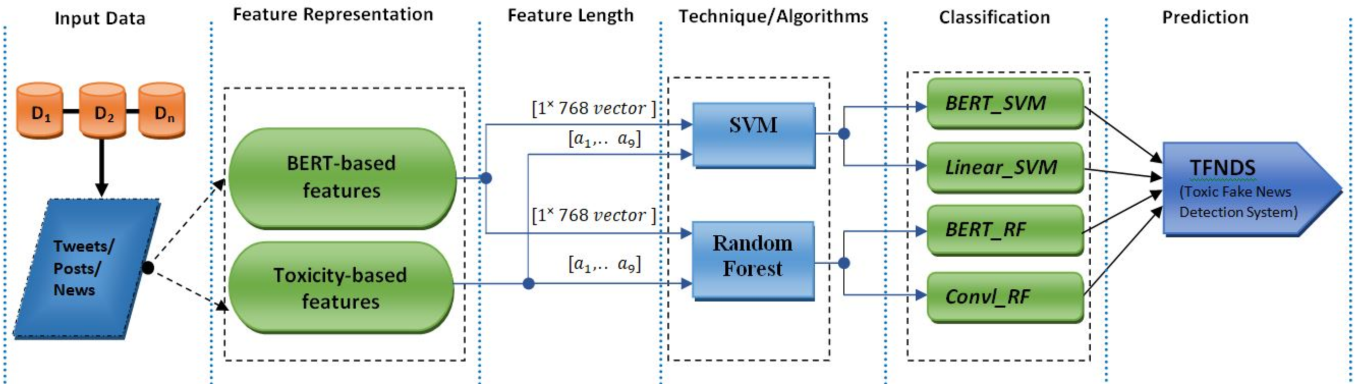


Fig. 7. Toxic-Fake News Detection (TFND)- Experimental setup.

Dataset	Technique	Precision	Recall	Accuracy
Dataset D1	BERT_SVM	0.65	1.0	0.65
Dataset D1	BERT_RF	0.62	1.0	0.64
Dataset E1	Linear_SVM	0.96	0.94	0.92
Dataset E1	Random Forest	0.98	0.88	0.89

TABLE IX
PERFORMANCE OF SEVERAL SUPERVISED LEARNING TECHNIQUES ON TOXICITY AND BERT-BASED FEATURES USING *Precision, Recall* AND *Accuracy*

Dataset	Technique	F_1	$F_{0.5}$	F_2
Dataset D1	BERT_SVM	0.79	0.65	0.65
Dataset D1	BERT_RF	0.76	0.63	0.63
Dataset E1	Linear_SVM	0.95	0.85	0.87
Dataset E1	Random Forest	0.93	0.59	0.44

TABLE X
PERFORMANCE OF SEVERAL SUPERVISED LEARNING TECHNIQUES ON TOXICITY AND BERT-BASED FEATURES USING F_βEVALUATION METRICS

As it is clear from the table we have compared our approach with seven existing approaches in this study. Before jumping into the results, let us put some light on the features used in these studies. The study [36] has employed several features including Rhetorical Structure Theory (RST) [55], Linguistic Inquiry Word Count (LIWC) [56], Castillo [52]. The RST designs a tree structure to show rhetorical or speaking or writing relationships between the words in a paragraph or a text summary and on the other hand, LIWC assists in mining the psycho-linguistic characteristics from a corpus. It includes emotional attributes, sentiments, and part-of-speech-like categories. The Castillo is mainly used to produce features from the user profile and his/her friendship network and about those users who have shared a particular news article, post, or Tweet. Pérez-Rosas et al. [38] present a linguistic model for the detection of fake news articles by employing features such as n-grams (i.e., uni-grams and bi-grams), LIWC, CFGs based on TF-IDF, etc. Zhou et al. [37] in their study used several features to distinguish fake news from non-fake news. The features used in this study include frequency of words (obtained by a Bag-Of-Words (BOW) model, Part-Of-Speech (POS) tags, and Probability Context-Free Grammar (PCFG) parsing trees to obtain the

Study/ Method	Accuracy (A)(%)	F-measure(%)
Shu et al.[36]	↓	↓
RST	60.0	63.3
LIWC	79.1	79.0
Castillo	80.0	79.7
RST + Castillo	81.6	80.5
LIWC + Castillo	82.5	82.2
TriFN	86.4	87.0
Zhou et al.[37]	89.2	89.2
Perez-Rosses et al.[38]	81.1	81.1
n-grams + TFIDF	75.5	75.5
CFG + TFIDF	74.9	74.8
Castillo et al.[52]	89	92
C. Baydogan et al. [26]	↓	↓
ANN	88.92	71.8
RNN	90.31	73.5
LSTM	87.22	65.6
CNN	89.08	69.7
GRU	87.44	66.3
NB	68.59	69.1
RF	73.91	71.4
LR	74.99	71.3
A. Wani et al. [53]	↓	↓
BERT	98.36	NA
BERT-cased	98.41	NA
M. Choudhary et al. [54]	↓	↓
BERT+CNN	97.45	97.5
ELMo+ANN	93.58	93.6
Proposed Approach	**92.1**	**95.1**

TABLE XI
PERFORMANCE COMPARISONS BETWEEN THE PROPOSED APPROACH AND THE EXISTING STUDIES ON FAKE NEWS DETECTION BASED ON ACCURACY AND F-MEASURE.

Study/ Method	Precision(P)	Recall(R)
TriFn[36]	0.849	0.893
Zhou et al.[37]	0.877	0.893
M. Choudhary et al. [54]	0.968	0.982
Perez-Rosses et al.[38]	0.742	0.751
Castillo et al.[52]	0.891	0.891
C. Baydogan et al. [26]	0.810	0.680
Proposed Approach	**0.983**	**0.884**

TABLE XII
PERFORMANCE COMPARISONS BETWEEN THE PROPOSED APPROACH AND THE EXISTING STUDIES ON FAKE NEWS DETECTION BASED ON PRECISION AND RECALL.

rewrite rules of a sentence within a news article. Furthermore, they have also used ClickBait-related Attributes (CBAs) and Disinformation-related Attributes (DIAs) to build a fake news classification model. However, the two studies [26] and [53] have learned the deep learning algorithms on BERT-based features for classification tasks and achieved better results than the rest of the approaches in the study. In [26] authors have employed Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), and BERT achieving an accuracy of 97.45% and 93.58% with BERT+CNN, and ELMo+ANN respectively while as another study [53] reported accuracy of 98.36% and 98.41% for BERT and BERT-cased, respectively. One of the main reasons for the better performance of these approaches is they have employed BERT-based approaches on a larger corpus while the proposed study has a comparatively smaller experimental dataset. It should be noted here that in table XI we have mentioned the results of best-performing models or approaches from the existing studies.

The tabulated data provides empirical evidence that the proposed methodology has outperformed most of the previously published studies in terms of accuracy and f-measure, registering scores of 92% and 95% respectively. The innovation of this approach lies in its utilization of toxicity-based features and BERT sequence representations to discriminate between toxic and non-toxic fake news posts on social media platforms. We conducted an extensive evaluation of our approach by comparing its performance against existing studies using precision and recall metrics, as shown in Table XII. Specifically, precision (P) and recall (R) metrics measure the number of true positives and the number of false negatives, respectively. A high precision score is obtained by minimizing false-positive errors, while a high recall score is obtained by minimizing false-negative errors. Our results indicate that the proposed methodology achieved a recall score of 88.4%, which is almost equivalent to the highest-scoring approach in previous studies. However, our approach outperformed all existing studies with a precision score of 98.3%, implying that it produces fewer false-positive errors than previous approaches. Therefore, we conclude that our approach offers better overall performance in terms of minimizing false-positive errors and almost equivalent false-negative errors compared to other studies.

Furthermore, we can also infer that the proposed approach has been shown to outperform several existing methods for sentiment analysis in terms of accuracy and F-measure. It has achieved an accuracy that is 52.1% higher and an F-measure that is 31.8% higher than Shu et al.'s RST method [36], 21.6% higher accuracy and 19.6% higher F-measure than n-grams + TFIDF, 23.1% higher accuracy and 27.3% higher F-measure than CFG + TFIDF [38], and 33.51% higher accuracy and 37.6% higher F-measure than C. Baydogan et al.'s NB method [26]. Furthermore, the proposed approach has demonstrated an F-measure that is 29.5% and 28.8% higher than C. Baydogan et al.'s LSTM and GRU methods, respectively. While the proposed approach still falls short compared to some state-of-the-art approaches, it shows promising results and a significant improvement over several existing methods.

As shown in the plot in Figure 8 the proposed approach in the table has an accuracy of 92.1%. When compared with the approaches defeated by the proposed approach, it can be seen that the proposed approach outperforms several methods such as RST (60%), CFG + TFIDF (74.9%), and NB (68.59%). However, the proposed approach has a lower accuracy than some of the other approaches such as BERT-cased (98.41%), BERT (98.36%), and ELMo+ANN (93.58%).

Similarly, upon comparing the proposed approach with the existing approaches based on F-score, it can be seen in Figure 9 that the proposed approach outperforms several methods such as RST (63.3%), CFG + TFIDF (74.8%), CNN (69.7%), GRU (66.3%), LSTM (65.6%), and NB (69.1%). However, the proposed approach has a lower F-measure than some of the other approaches such as BERT+CNN (97.5%) and ELMo+ANN (93.6%).

The results presented in the plot 10 are precision and recall scores for different studies and methods. Based on the results

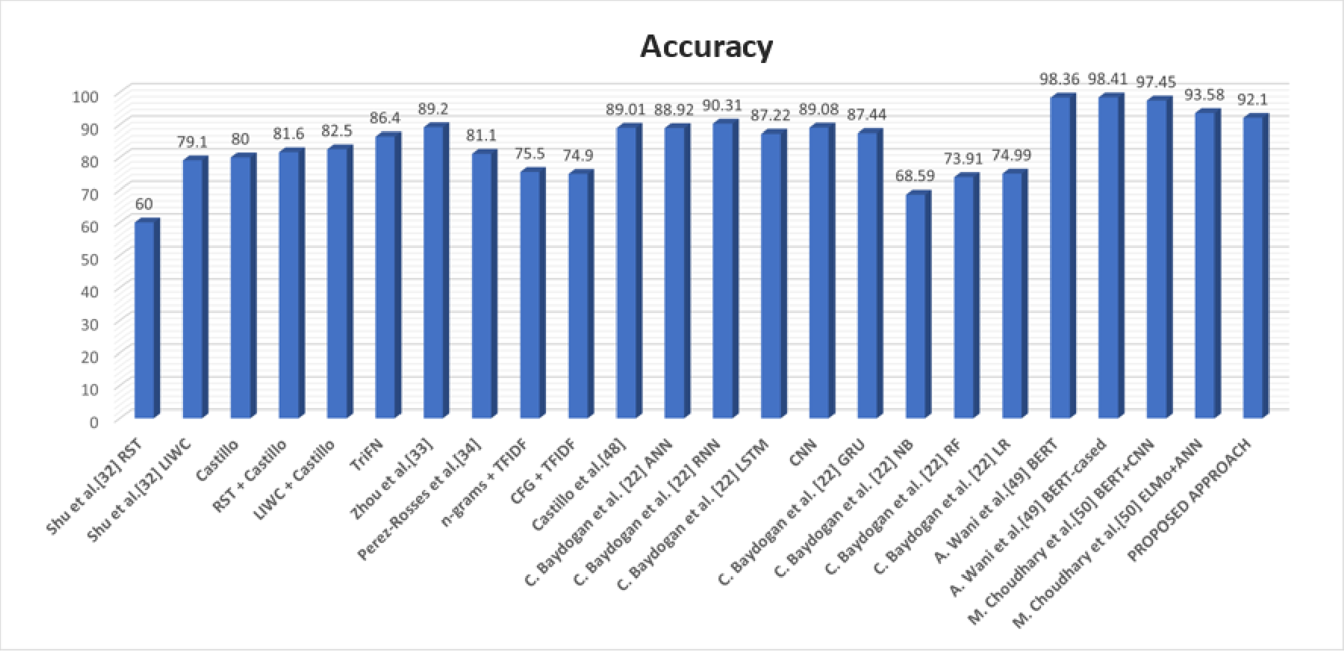


Fig. 8. Performance of proposed approaches against existing approaches on Accuracy measure

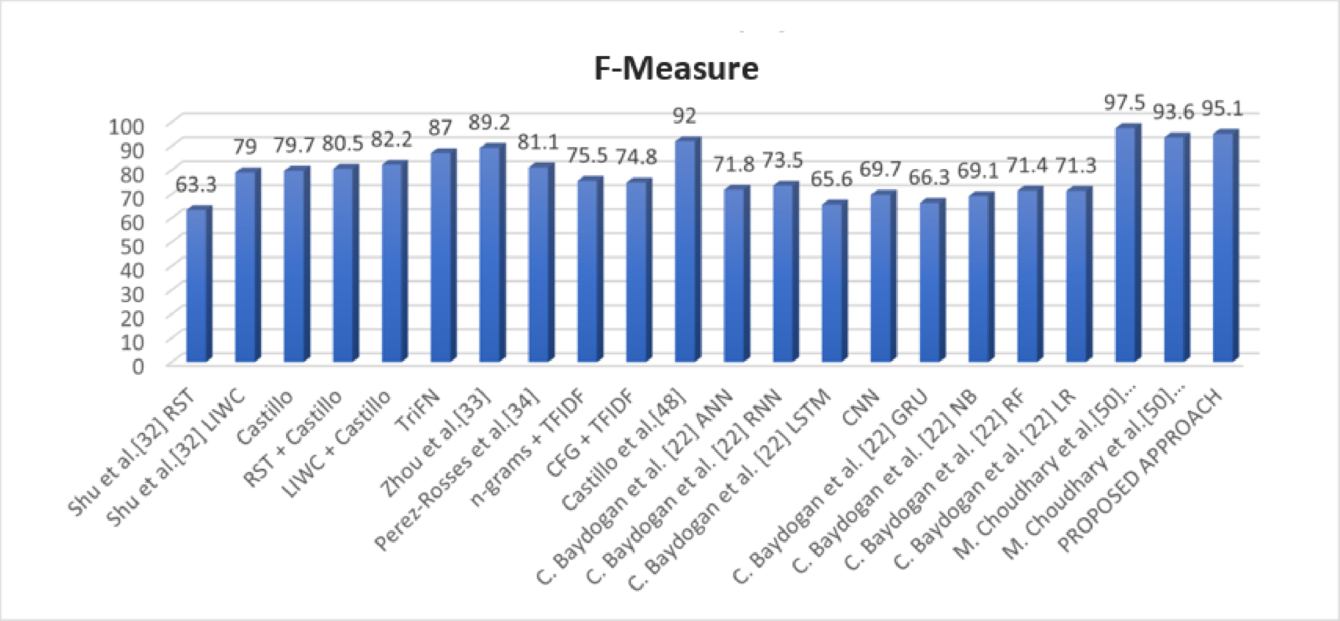


Fig. 9. Performance of proposed approaches against existing approaches on F-measure

presented in the plot, the proposed approach has the highest precision score of 0.983, which indicates that it has a low false positive rate and a high degree of accuracy in predicting positive cases. However, its recall score of 0.884 suggests that it may miss some true positive cases, meaning that it is not as comprehensive in identifying all positive cases as some of the other methods.

Among the other methods, M. Choudhary et al. [54] achieved the highest precision and recall scores of 0.968 and 0.982, respectively. This suggests that their approach is highly accurate in predicting positive cases and comprehensive in identifying all positive cases.

On the other hand, Perez-Rosses et al. [38] achieved the lowest scores for both precision (0.742) and recall (0.751), suggesting that their approach has a higher false positive rate and may miss some true positive cases.

Overall, the precision and recall scores provide insights into the performance of different approaches, and it is important to consider both scores together when evaluating the effectiveness of a method on a particular task.

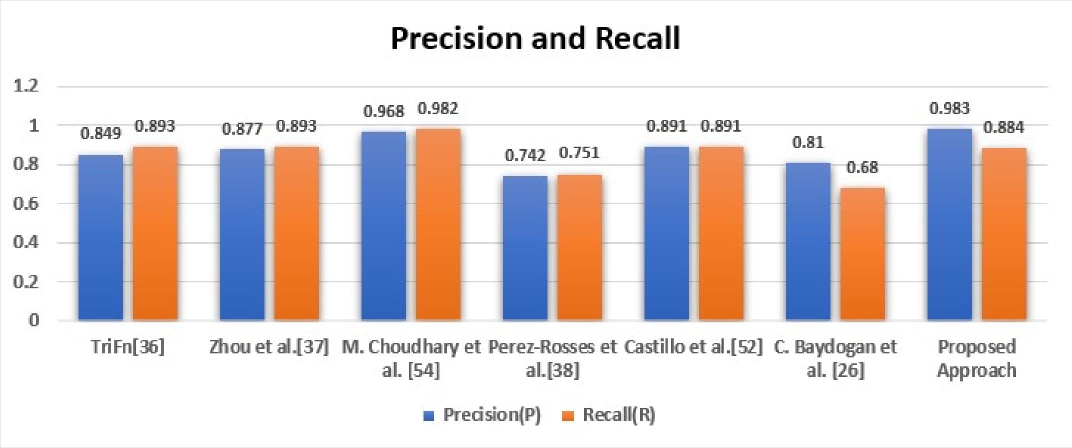


Fig. 10. Performance of proposed approach against existing approaches based on Precision and Recall.

IX. Limitations of the study

This study investigates the use of toxicity-based attributes in the detection and control of toxic fake posts or news articles on social media by utilizing the transformers (BERT) with traditional machine-learning techniques such as SVM and Random Forest. No doubt this study reported promising results from all the employed algorithms on toxicity and BERT-embedding-based features. But there are several gaps still unfilled by this study.

First, the data used in this study is limited in size and quality. AI and machine learning models are only as good as the data we use to train the model. Therefore, in order to have efficient and reliable models, we need to incorporate a large number of ground truth samples.

Second, in this study, we are only considering the toxicity-based features and word-embedding representations to train a model for the identification of toxic fake news. The use of other popular feature representation models such as Word2Vec, TF-IDF, BoW, etc. could have been employed to analyze the performance of the overall system. Also, the emotion-based features as discussed in the paper have not been used in the study. The use of emotion-based features along with other features may have enhanced the performance of the system.

Third, we have not exercised the basic deep learning techniques such as CNN(Convolutional Neural Network), LSTM (Long Short-Term Memory), etc, here at this stage for the classification task. Since CNN models employ convolutional layers and maximum pooling layers to extract higher-level features, and LSTM-based models capture long-term dependencies among word sequences, therefore these techniques are considered better for text classification. However, BERT works well for models designed to perform specific tasks. But on the other hand, the model training time is huge because of its training structure and corpus. There are also a lot of weights to update while working with such models. Furthermore, the size of the BERT model makes it expensive as it requires more computation power.

X. Conclusion and Future Scope

This paper primarily focuses on extracting the toxicity-based features and BERT-based representation of user content written in English to detect and classify *toxic-fake news* on social media, particularly in COVID-19 times. A dataset has been designed by merging two datasets (one was used by a previous study, second was crawled based on top keywords from the first dataset). It is observed that toxicity-based features such as *Toxicity, Severe Toxicity, Obscene, Identity Attack, Insult, Threat, and Sexually Explicit* have the potential to detect fake news which contains toxicity on social media.

The main motive behind this study is to detect toxic fake news on different social media platforms related to COVID-19 disease and thus save the efforts put into investing all the fake news categories which are not otherwise toxic. The Detoxify (python package) has been used to detect toxicity-based features out of the user's contents. The potential of the proposed approach is tested on the mixture of existing datasets using machine learning techniques, including SVM, and Random Forest, and also their BERT variants. The experiments are conducted on two datasets comprising toxicity-based features, and BERT representations of the same instances, respectively. The training dataset has been divided into two sets in the ratio of 80:20 for training and validation, respectively. Finally, trained models have been tested separately on the test set extracted earlier from the original dataset. It has been seen that all three algorithms achieved good results on both datasets. We recorded accuracy of 64.16%, 65.39%, 89.38%, and 92.49% by RF, linear SVM, Bert-based RF, and Bert-based SVM respectively. It is clearly seen from the results that linear SVM outperformed other employed techniques in this study. Furthermore, we recorded the performance of linear SVM on other measures as well and obtained values of 0.95, 0.85, and 0.87 for F_1, $F_{0.5}$, and F_2-score, respectively. As obvious this model outperformed other models on all the measures except the value (0.96) received for precision which is less than the value (0.98) received for Random forest. Upon comparing the results obtained by our best classifier (i,e linear SVM), the proposed approach has outperformed most of the popular methods in the literature by recording the best values on performance metrics such as accuracy, F_1-score, precision, and recall for linear SVM. Overall the proposed methods have either suppressed or achieved results very close to the state-of-the-art techniques.

The objective of this study was to demonstrate the utilization of toxicity-based attributes for toxic fake news identification tasks. As toxicity may present in any language and textual format, including social media posts, political speeches, student feedback, or customer reviews, the study aimed to explore toxicity across languages. For instance, the investigation will explore how toxicity varies between English and Arabic languages. Moreover, the study intends to examine the application of deep learning algorithms to solve this classification task. Additionally, the investigation will analyze the variation of toxicity across different types of text, such as fake news text and spam review text.

Acknowledgment

This research work is supported by the EIAS, Data Science and Blockchain Laboratory, Prince Sultan University, Riyadh, Saudi Arabia.

Key Abbreviations

Several key abbreviations utilized in this study are enumerated as follows:
FND: Fake News Detection
HSD: Hate Speech Detection
DT: Decision Tree
KNN: K-Nearest Neighbor
LR: Logistic Regression
LSVM: Linear Support Vector
MNB: Machines Multinomial Naïve Bayes
BNB: Bernoulli Naïve Bayes
NN: Neural Network
ERF: Ensemble Random Forest

References

[1] I. Gunasekara and I. Nejadgholi, "A review of standard text classification practices for multi-label toxicity identification of online content," in *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 2018, pp. 21–25.

[2] B. Zhu, X. Zheng, H. Liu, J. Li, and P. Wang, "Analysis of spatiotemporal characteristics of big data on social media sentiment with covid-19 epidemic topics," *Chaos, Solitons & Fractals*, vol. 140, p. 110123, 2020.

[3] M. A. Wani, N. Agarwal, and P. Bours, "Impact of unreliable content on social media users during covid-19 and stance detection system," *Electronics*, vol. 10, no. 1, p. 5, 2020.

[4] S. Majó-Vázquez, R. Nielsen, J. Verdú, N. Rao, N. de Domenico, and O. Papaspiliopoulos, "Volume and patterns of toxicity in social media conversations during the covid-19 pandemic," 2020.

[5] S. A. Khan, K. Shahzad, O. Shabbir, and A. Iqbal, "Developing a framework for fake news diffusion control (fndc) on digital media (dm): A systematic review 2010–2022," *Sustainability*, vol. 14, no. 22, p. 15287, 2022.

[6] A. R. Mahlous and A. Al-Laith, "Fake news detection in arabic tweets during the covid-19 pandemic," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 778–788, 2021.

[7] K. Shahzad, S. A. Khan, S. Ahmad, and A. Iqbal, "A scoping review of the relationship of big data analytics with context-based fake news detection on digital media in data age," *Sustainability*, vol. 14, no. 21, p. 14365, 2022.

[8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[9] S. Velichety and U. Shrivastava, "Quantifying the impacts of online fake news on the equity value of social media platforms–evidence from twitter," *International Journal of Information Management*, vol. 64, p. 102474, 2022.

[10] M. O. Ahmed and A. E. Msughter, "Assessment of the spread of fake news of covid-19 amongst social media users in kano state, nigeria," *Computers in Human Behavior Reports*, p. 100189, 2022.

[11] M. K. Elhadad, K. F. Li, and F. Gebali, "Detecting misleading information on covid-19," *Ieee Access*, vol. 8, pp. 165 201–165 215, 2020.

[12] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 540, p. 123174, 2020.

[13] D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein, and A. Nabil, "Coaid-deep: an optimized intelligent framework for automated detecting covid-19 misleading information on twitter," *Ieee Access*, vol. 9, pp. 27 840–27 867, 2021.

[14] M. S. Al-Rakhami and A. M. Al-Amri, "Lies kill, facts save: detecting covid-19 misinformation in twitter," *Ieee Access*, vol. 8, pp. 155 961–155 970, 2020.

[15] F. A. Ozbay and B. Alatas, "Adaptive salp swarm optimization algorithms with inertia weights for novel fake news detection model in online social media," *Multimedia Tools and Applications*, vol. 80, no. 26, pp. 34 333–34 357, 2021.

[16] B. A. Feyza Altunbey Ozbay, "A novel approach for detection of fake news on social media using metaheuristic optimization algorithms," *Elektronika ir Elektrotechnika*, vol. 25, no. 4, pp. 62–67, 2019.

[17] M. A. Wani, M. A. ELAffendi, K. A. Shakil, A. S. Imran, and A. A. Abd El-Latif, "Depression screening in humans with ai and deep learning techniques," *IEEE Transactions on Computational Social Systems*, 2022.

[18] T. Chauhan and H. Palivela, "Optimization and improvement of fake news detection using deep learning approaches for societal benefit," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100051, 2021.

[19] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Applied Soft Computing*, vol. 100, p. 106983, 2021.

[20] F. Ma and G. Tan, "Nlp in fake news detection," in *IRC-SET 2020*. Springer, 2021, pp. 71–83.

[21] J. Antony Vijay, H. Anwar Basha, and J. Arun Nehru, "A dynamic approach for detecting the fake news using random forest classifier and nlp," in *Computational methods and data engineering*. Springer, 2021, pp. 331–341.

[22] R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia tools and applications*, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[23] Y.-J. Lu and C.-T. Li, "Gcan: Graph-aware co-attention networks for explainable fake news detection on social media," *arXiv preprint arXiv:2004.11648*, 2020.

[24] M. Aldwairi and A. Alwahedi, "Detecting fake news in social media networks," *Procedia Computer Science*, vol. 141, pp. 215–222, 2018.

[25] J. Risch and R. Krestel, "Toxic comment detection in online discussions," in *Deep learning-based approaches for sentiment analysis*. Springer, 2020, pp. 85–109.

[26] C. Baydogan *et al.*, "Deep-cov19-hate: A textual-based novel approach for automatic detection of hate speech in online social networks throughout covid-19 with shallow and deep learning models," *Tehnički vjesnik*, vol. 29, no. 1, pp. 149–156, 2022.

[27] A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal, "Identifying toxicity within youtube video comment," in *International conference on social computing, Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*. Springer, 2019, pp. 214–223.

[28] B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar, "Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 90–94.

[29] C. Baydogan and B. Alatas, "Metaheuristic ant lion and moth flame optimization-based novel approach for automatic detection of hate speech in online social networks," *IEEE Access*, vol. 9, pp. 110 047–110 062, 2021.

[30] K. Varjas, J. Talley, J. Meyers, L. Parris, and H. Cutts, "High school students' perceptions of motivations for cyberbullying: An exploratory

study," *Western Journal of Emergency Medicine*, vol. 11, no. 3, p. 269, 2010.

[31] S.-H. Lee and H.-W. Kim, "Why people post benevolent and malicious comments online," *Communications of the ACM*, vol. 58, no. 11, pp. 74–79, 2015.

[32] P. Shachaf and N. Hara, "Beyond vandalism: Wikipedia trolls," *Journal of Information Science*, vol. 36, no. 3, pp. 357–370, 2010.

[33] J. W. Kim, A. Guess, B. Nyhan, and J. Reifler, "The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity," *Journal of Communication*, vol. 71, no. 6, pp. 922–946, 2021.

[34] Perspective, "Using machine learning to reduce toxicity online," Perspective.https://perspectiveapi.com/, 2022.

[35] A. Obadimu, T. Khaund, E. Mead, T. Marcoux, and N. Agarwal, "Developing a socio-computational approach to examine toxicity propagation and regulation in covid-19 discourse on youtube," *Information Processing & Management*, vol. 58, no. 5, p. 102660, 2021.

[36] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 312–320.

[37] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, 2020.

[38] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *arXiv preprint arXiv:1708.07104*, 2017.

[39] L. Hanu and Unitary team, "Detoxify," Github. https://github.com/unitaryai/detoxify, 2020.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[41] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification*. Springer, 2016, pp. 207–235.

[42] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[43] R. Plutchik and H. Kellerman, *Theories of emotion*. Academic Press, 2013, vol. 1.

[44] K. A. Shakil, K. Tabassum, F. S. Alqahtani, and M. A. Wani, "Analyzing user digital emotions from a holy versus non-pilgrimage city in saudi arabia on twitter platform," *Applied Sciences*, vol. 11, no. 15, p. 6846, 2021.

[45] M. A. Wani, N. Agarwal, S. Jabin, and S. Z. Hussain, "User emotion analysis in conflicting versus non-conflicting regions using online social networks," *Telematics and Informatics*, vol. 35, no. 8, pp. 2326–2336, 2018.

[46] D. Dementieva, D. Moskovskiy, V. Logacheva, D. Dale, O. Kozlova, N. Semenov, and A. Panchenko, "Methods for detoxification of texts for the russian language," *Multimodal Technologies and Interaction*, vol. 5, no. 9, p. 54, 2021.

[47] L. Cui and D. Lee, "Coaid: Covid-19 healthcare misinformation dataset," *arXiv preprint arXiv:2006.00885*, 2020.

[48] T. Verdonck, B. Baesens, M. Óskarsdóttir, *et al.*, "Special issue on feature engineering editorial," *Machine Learning*, pp. 1–12, 2021.

[49] OmniSCI, "What is feature engineering? definition and faqs — omnisci," OmniSCI.https://www.omnisci.com/technical-glossary/feature-engineering, 2022.

[50] A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[51] Tweepy, "Tweepy," Tweepy.org. https://docs.tweepy.org/en/stable/index.html, 2022.

[52] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[53] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, "Evaluating deep learning approaches for covid19 fake news detection," in *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, 2021, pp. 153–163.

[54] M. Choudhary, S. S. Chouhan, E. S. Pilli, and S. K. Vipparthi, "Berconvonet: A deep learning framework for fake news classification," *Applied Soft Computing*, vol. 110, p. 107614, 2021.

[55] V. L. Rubin, N. J. Conroy, and Y. Chen, "Towards news verification: Deception detection methods for news discourse," in *Hawaii International Conference on System Sciences*, 2015, pp. 5–8.

[56] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

Mudasir Ahmad Wani is currently working as a Research Scientist and Assistant Professor (Guest) at Prince Sultan University in KSA, where he specializes in NLP and Social Cybersecurity. He has served as a Lecturer and Researcher at the Department of Information Security and Communication Technology (IIK) at the Norwegian University of Science and Technology (NTNU), Norway. He pursued his postdoctoral research at the Norwegian Biometrics Laboratory, NTNU, Norway and He is the recipient of the Alain Bensoussan Fellowship award under the European Research Consortium for Informatics and Mathematics (ERCIM), Sophia Antipolis Cedex, France. He obtained his Ph.D. from Jamia Millia Islamia (A Central University), New Delhi, India in 2019 in Computer Science. He holds a master's in computer Applications (MCA) and M.Phil. (Data Mining) from the University of Kashmir (UoK) in 2012 and 2014 respectively. His research focuses on the extraction and analysis of social data, as well as the application of different statistical and machine/deep learning techniques in developing prediction models. He is an active member of the academic community, regularly organizing and reviewing international conferences, workshops, and journals.

Mohammed A ElAffendi is currently a Professor of computer science with the Department of Computer Science, Prince Sultan University, a Former Dean of CCIS, AIDE, the Rector, a Founder, and the Director of Data Science Laboratory (EIAS), a Founder and the Director of The Center of Excellence in CyberSecurity. His current research interests include data science, intelligent and cognitive systems, machine learning, and natural language processing.

Kashish Ara Shakil has received her Ph.D. degree in Computer Science from Jamia Millia Islamia, New Delhi. She is currently working as an Assistant Professor at the College of Computer and Information Sciences, Princess Nora Bint AbdulRahman University, Saudi Arabia. She has a Bachelor's degree in Computer Science from Delhi University and an MCA degree from Jamia Hamdard. She serves as the co-Editor-in-Chief of the Journal of Applied Information Science. She is on the Editorial Board of many reputed International Journals in Computer Sciences and has published several research papers. She has three books entitled as "Internet of Things (IoT): Concepts and Applications (S.M.A.R.T. Environments)", "Emerging Technologies for Sustainable and Smart Energy - Prospects in Smart Technologies", and "Green Automation for Sustainable Environment" to her credit. Her areas of interest include Cloud Computing, Big Data, Machine learning and NLP.

Ibrahem Mohammed Abuhaimed Received a bachelor's degree in software engineering from King Saud University, Saudi Arabia, 2011, and received his master's degree in Software Engineering from Prince Sultan University, Saudi Arabia, 2018 he is currently working as a research assistant in Prince Sultan University since 2018, his research interests are in Blockchain, deep learning, natural language processing, and CyberSecurity.

Anand Nayyar (Senior Member, IEEE) received his Ph.D (Computer Science) from Desh Bhagat University in 2017 in the area of Wireless Sensor Networks and Swarm Intelligence. He is currently working in the School of Computer Science-Duy Tan University, Da Nang, Vietnam as an Assistant Professor, Scientist, Vice-Chairman (Research) and Director- IoT and Intelligent Systems Lab. A Certified Professional with 80+ Professional certificates from CISCO, Microsoft, Oracle, Google, Beingcert, EXIN, GAQM, Cyberoam and many more. Published more than 125+ Research Papers in various High-Quality ISI-SCI/SCIE/SSCI Impact Factor Journals cum Scopus/ESCI indexed Journals, 50+ Papers in International Conferences indexed with Springer, IEEE Xplore and ACM Digital Library, 40+ Book Chapters in various SCOPUS, WEB OF SCIENCE Indexed Books with Springer, CRC Press, Elsevier and many more with Citations: 5500+, H-Index: 38 and I-Index: 134.Member of more than 50+ Associations as Senior and Life Member including IEEE, ACM. He has authored/co-authored cum Edited 30+ Books of Computer Science.

Amir Hussain (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in electronic and electrical engineering from the University of Strathclyde, Scotland, U.K., in 1992 and 1997, respectively.,Following Postdoctoral and Senior Academic Positions at West of Scotland (1996-1998), Dundee (1998-2000) and Stirling Universities (2000-2018), respectively, he joined Edinburgh Napier University as founding Head of the Cognitive Big Data and Cybersecurity (CogBiD) Research Lab and the Centre for AI and Data Science. His research interests include cognitive computation,machine learning and computer vision.

Ahmed A. Abd El-Latif (Senior Member, IEEE) received the B.Sc. degree with honour rank in Mathematics and Computer Science in 2005 and M.Sc. degree in Computer Science in 2010, all from Menoufia University, Egypt. He received his Ph. D. degree in Computer Science Technology at Harbin Institute of Technology (H.I.T), Harbin, P. R. China in 2013. He is an associate professor of Computer Science at Menoufia University, Egypt, and at the College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia. He is the author and co-author of more than 200+ papers, including refereed IEEE/ACM/Springer/Elsevier journals, conference papers, books and book chapters. He received many awards, State Encouragement Award in Engineering Sciences 2016, Arab Republic of Egypt; the best Ph.D. student award from Harbin Institute of Technology, China 2013; Young scientific award, Menoufia University, Egypt 2014. He is a fellow at Academy of Scientific Research and Technology, Egypt. His areas of interests are multimedia content encryption, secure wireless communication, IoT, applied cryptanalysis, perceptual cryptography, secret media sharing, information hiding, biometrics, forensic analysis in digital images, and quantum information processing. Dr. Abd El-Latif has many collaborative scientific activities with international teams in different research projects. Furthermore, he has been reviewing papers for 700+ International Journals including IEEE Communications Magazine, IEEE Internet of Things journal, Information Sciences, IEEE Transactions on Network and Service Management, IEEE Transactions on Services Computing, Scientific reports Nature, Journal of Network and Computer Applications, Signal processing, Cryptologia, Journal of Network and Systems Management, Visual Communication and Image Representation , Neurocomputing, Future Generation Computer Systems, etc. Dr. Abd El-Latif is an associate editor of Mathematical problems in Engineering, Journal of Cyber Security and Mobility, and IET Quantum Communication. Dr. Abd El-Latif also leading many special issues in several SCI/EI journals. Currently, He had many books, more than 10 books, in several publishers for process in Springer, IET, CRC press, IGI-Global, Wiley, and IEEE.