# STUDY ON WEB CONTENT FACT CHECKING WITH MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

*C. Vishnu Mohan[1]*, G. Angeline Prasanna[2]*

**Abstract**

Over the last decade, people have been widely using online platforms for sharing information and for understanding the news that has been happening around them. Classification of social media texts, tweets etc., are one of the emerging areas of research in today's world, especially when it comes to information about political and entertainment sectors. However, there are certain challenges due to the fact that most commonly used Machine Learning techniques have not proven to be optimal, when considering both textual and image data for fake content detection. We investigate the use of Gated Recurrent Unit (GRU) with Long Short-Term Memory (LSTM) Ensemble for classifying news as real or fake.

**Keywords:** Machine Learning; Fake content Detection; Gated Recurrent Unit; Long Short-Term Memory

## I INTRODUCTION

The advances in technology over time and the widespread use of the Internet have changed the nature of the digital world and the way information is shared. The Internet has become a key tool for information in research. Social media is the most popular reason for people to connect to the Internet. People's habits have altered much because of the fact that they use social media so often and have thus increased its popularity. Digital news has become most people's primary information source to know about the happenings around them. However, there are large volumes of online information that are questionable and often even meant to deceive. Also, a few false news stories are so close to the actual ones that it is challenging for people to distinguish them apart.

Due to their low cost, ease of use, and the viral nature, a number of online social media platforms, including WhatsApp, Facebook, Twitter, Instagram, YouTube, and many others, have grown in popularity. There are now a lot more people using the internet, and they utilize it for a variety of purposes. Internet-based news disseminates quickly and may be valid or invalid. People lack the intelligence to discern whether news is reliable or not. False news spreads quickly. Social media and word-of-mouth are two ways by which news can spread. News that is intentionally produced to deceive people is referred to as fake news. The term "fake news" refers to a phenomenon that has several definitions and takes many forms, ranging from exaggeration leading to fabrication [1]. This is even worse when they are even accepted by the society. False news has developed and evolved from time to time such that its frequency in online media is inappropriate and overwhelming [2].

Fake news has a negative effect on a person's, society's, or institution's reputation. The first time it was noticed was during the 2016 US presidential elections, when false information promoting one of the two candidates was accepted and spread more than 37 million times on Facebook [3]. But even though it has recently grabbed a lot of attention, identifying fake news[4,5,6,7] is a very difficult challenge. Fake news is typically produced by editing images, text, or videos, which emphasizes the importance of a multimodal detection. Section II discusses briefly on the literature where researches are conducted on applying various techniques in

[1]Department of Computer Science
[2]Department of Computer Applications
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

the classification of content as fake or real [8, 9,10]. The Problem statement (Section III), motivation towards the topic (Section IV) and the research gap (Section V) are also identified. Section VI states the objectives that are set based on the review of literature followed by the new methodology proposed in Section VII.

## II REVIEW OF LITERATURE

Academicians are now seriously considering the widespread dissemination of false information on the social media, as explained by Wu et al. [11]. Facebook, Twitter, Reddit, PolitiFact, Instagram, and other social media sites became increasingly popular, especially following the 2016 US Presidential election campaigns. Contrary to misinformation, which may be unintentional, disinformation is typically the false information that has been deliberately spread. Recently, a number of methods for spotting fake news have been created. This section includes pertinent research on spotting fake news on social media websites. The literature review that is now accessible indicates that machine learning models were frequently used to identify fake news, followed by deep learning models, and that transfer learning and pre-trained models are now also performing well in this domain.

Using n-gram analysis, Ahmed et al. [12] suggested a method to identify fake news. At first the authors decided to use two feature generation techniques and tested them against 6 various machine learning classifiers.For feature extraction, they have used two methods namely Term Frequency (TF) and Term Frequency-Inverted Document Frequency(TFIDF). Then they have compared the K-Nearest Neighbors (k-NN), Decision Trees (DT),Support Vector Machine (SVM), Logistic Regression (LR), and the Linear Support Vector Machine (LSVM), classifier's performance in order to determine which one is the most effective. With an accuracy of 92%, they achieved the best result utilizing the feature extraction method Term Frequency-Inverse Document Frequency (TF-IDF) with the classifier Linear Support Vector Machine (LSVM).Although this study showed a great accuracy, this may be due to a Population Bias or Representation Bias, as the authors focused on n-gram analysis, as explained in the study conducted by Ninareh et al. [13]. Reliance solely on n-grams could be problematic, as we can see in Cruz et al. (2019) [14], because this feature extraction method may change based on media attention over time.

For the purpose of detecting fake news, Perez et al. [15] first introduced and discussed collecting, annotation, and validation procedures of two novel datasets. Second, the authors conducted a series of experiments and exploratory data analyses utilizing the datasets indicated above to pinpoint linguistic characteristics that are predominately present in the fake content. The authors used the Linear Support Vector Machine (LSVM) classifier and employed a fivefold cross-validation technique for classification. The highlight of this research is that the best possible combination of feature variables were selected, as opposed to the research published by Ahmed et al. [12] earlier, which puts more emphasis on finding the best feature variable generation and classification methods and less emphasis on the features themselves (features generated by n-grams). The authors also performed a number of experiments with various feature combinations in order to achieve this, including n-grams, punctuation characters, psycho-linguistic features, readability, and syntax. They created a fake news detector that performed at its peak with 78% accuracy when all features were utilized. The findings point to significant discrepancies in the substance of fake and real news. Some of these variations include the employment of more social and positive phrases, the expression of greater certainty, the emphasis on the current scenario and of future, and the presence of punctuation characters, verbs, and adverbs in false news articles.

Yaqing Wang et al. [16] have made use of the Multi-Modal Feature Extractor: EANN (Event Adversarial Neural Network) to detect fake news across many media channels. As they only learn event-specific properties that cannot be applied to unobserved events, the existing models struggle to distinguish between true and false reports on recently emergent and time-critical occurrences. However, this EANN can pick up on traits that are independent of the course of an event, which gives it the ability to spot fake news reports during live events. Their proposed model consisted of a multi-modal feature extractor, the event discriminator, along with a fake-news detector that forms the model. Weibo and Twitter are only among a couple of the multimedia datasets that this study is built on. Using transferable characteristics depiction, the suggested system performs better than the current baseline methodologies.

On three large and varied datasets, [17] conducted a traditional experiment to assess the effectiveness of different machine learning algorithms. Eight of the 19 machine learning techniques use standard models, while six use typical deep learning algorithms and five use cutting-edge pre-trained language models like BERT. On all datasets, it has been found that BERT-based systems perform better than alternative techniques in terms of potential and performance. Additionally, BERT-based methods are reliable enough to work effectively with a small sample size. On acceptable huge data sets, naïve Bayes with N-Gram models have produced the same outcomes as neural network-based models.

To determine whether an event is real, Ma et al. [18] used a GRU with multiple layers and trained it using sequence of tweets based on time. A TFIDF score of 5000-dimension was fed to the model as the input from each tweet. Comparing this strategy with the non-deep learning methods, it produces an accuracy performance gain of 10%. (e.g., DT ranking, SVM, RF classification).

According to the content of the news items, Fang et al. [19] advocated using self-attention-based CNN, and they explained that the self-attention-based CNN produced greater accuracy than RNN-based models when gven the task to identify articles that contain non-factual information. Their learning approach often uses features that are derived using linguistic techniques and static network analysis. However, it does not employ dynamic network information.
Rohit et. al observed that, on the Kaggle fake news data set, FNDNet performed better than feature engineering and conventional machine learning solutions [20]. The GloVe technique was used to embed the words into a 100-dimensional vector used as an input for the FNDNet architecture, and the model in this instance only considers the features in the vector space. This deep learning architecture is built on an updated Convolutional Neural Network (CNN) network, where three convolutional layers are concatenated simultaneously and then dense layers are added on top. Using CNN and Long Short-Term Memory (LSTM) models, it outperformed both conventional machine learning and deep learning.

Detection of Fake News Deep Learning techniques were used by Hiramath & Deshpande [21]to compare the classification algorithms Logistic Regression (LR), Naïve Bayes (NB), SVM, Random Forest (RF), and Deep Neural Network (DNN). On the LIAR dataset, they conducted experiments using standard text preprocessing techniques from the natural language processing (NLP) area (such as stemming, stop word removal, etc.). They thereby validated the FNDNet findings and observed that Deep Neural Networks outperform conventional machine learning techniques.

Detection of Fake News Similar to the first article, using a Deep Neural Network also analyzed several models using Hashing Vectorizer in addition to TF-IDF as a vector space representation[22]. The authors used K-Nearest Neighbors

(KNN), Naïve Bayes (NB), Convolutional Neural Networks (CNN), Decision Tree (DT), Long Short-Term Memory (LSTM), and Random Forest (RF).The algorithms' performance accordingly declined in the order listed. Combining CNN and LSTM produced the greatest results, supporting the notion that deep learning models perform well. They combined a number of Kaggle datasets for the experiment.

Similar vector space representations and stylometric features were utilized in the Text-mining-based Fake News Detection Using Ensemble Methods[23]. Three distinct feature subsets were created from the stylometric data. The first one had a high character counts (with or without whitespace), high complexity score, Gunning-Fog index, Flesch-Kincaid readability score, and a number of unique words. The second collection is based on a dataset for lie detection, and its features may be broken down into the following groups: vocabulary, uncertainty, quantity, Flesch-Kincaid score, grammar etc. The last feature subset consisted of a write-print feature set that contain authorship attributes given in brief texts, which was divided into the following categories: Character, Word, Syntactic, Structural, and Content. They employed a variety of techniques for vector space representation, including TFIDF, the bag-of-words (BoW), TF-IDF, Continuous Bag-of-Words (CBoW), Skip-Gram (SG), which were employed to predict the next contextual word, and both Word2Vec and FastText tools. Both varieties of features were subject to feature selection. Recursive reduction of the weakest features led to the selection of stylometric features. Lemmatization, stemming, and Chi-square tests for feature selection were used to reduce the vocabulary in the word-vector space. They employed Random Forest (RF), Naïve Bayes (Gaussian and Multinomial), SVM, KNN, LR, Extra Trees Classifier,General Bagging Classifier,and Bagging with AdaBoost - Gradient boosting, as classifiers. Gradient boosting using CboWWord2Vec embeddings outperformed all other non-ensemble machine learning techniques in terms of overall accuracy. Notably, CboW representation enhanced the performance of non-ensemble algorithms.

In order to identify fake online book reviews, the usage of Rhetorical Structure Theory (RST) for Fake Online Review Detection is proposed by Olu [24]. The author used Deceptive Review as a dataset (DeRev). They created common macro-relations by grouping certain RST properties. According to the corpus analysis, the fake reviews have more macro-relations for Elaboration, Joint, and Background, whereas the genuine reviews possess relations attributed by Explanation, Evaluation, and Contrast. It was also observed that the genuine reviews have relations for better comparison. This study demonstrates that reviewers who have been paid to write fake reviews frequently use the misleading pragmatics as seen in RTS method. They tend to mention the title, author, or substance, which is against genre norm.

Sentiment analysis, sentiments, and cosine similarity scores on Naïve Bayes, Random Forest classifiers trained on LIAR dataset are used in Fake News Detection in the work by Bhutani et. al [25]. They concluded that incorporating sentimental score improves the model's accuracy.

To predict humour, irony, and satire in the news story, Victoria et. al [26] used vector space TF-IDF representation utilizing unigrams and bigrams along with the text preprocessing step of removing the stop words and extracted feature vectors. As a predictive machine learning algorithm, the SVMmodel was used. Punctuation extraction, absurdity using Part-of-Speech (PoS) tagging and Named Entity Recognition (NER), humour using knowledge-based punchline identification, grammar by counting PoS tags, and negative affect using the LIWC lexicon were among the attributes that were extracted. The detection was enhanced by each of these criteria, with humour features showing the weakest increase. The SVM was trained for a classification job using 10-fold cross-validation by the machine learning library sklearn.

Using collective user intelligence to detect fake news, Feng et. al. [27] has developed a Neural User Response Generator. The two-level CNN with User Response Generator (TCNN-URG) is employed to determine the news's credibility based on both its substance and readers' responses to similar items in the past, as well as to predict how they would react to the latest information. When real-time user reactions are unavailable, this method can be used to identify fake news early. Both the Twitter dataset and Weibo dataset were used. The conditional variational auto encoder serves as the basis for the User Response Generator.

According to the researchers conducted by Natali et. al [28], the authors have developed a hybrid model for fake news detetection, where textual information is combined with user feedback from articles as well as data of the people who posted the news. It consists of three modules: the first module uses an RNN network to process the text and response; the second module analyzes user and group information on the reliability of the source; and the third module combines these methods, tested on data sets from Weibo and Twitter.

The study conducted by Diego et. al [29] focuses on evaluating the legitimacy of the entire websites. It examines the current state of this field's research as well as recent setbacks, such as the price of external APIs and Google PageRank's discontinuation. They ignore user-based social variables due to the significant bias that these variables inject into the final model supported by the ANOVA test, and instead focuses on the online credibility model by using just content-based features. The final model was assessed using the Likert 5-star scale and two data sets, the Microsoft Dataset and the Content Credibility Corpus, both of which contained URLs. Readability, PageRank data, General Inquirer (a dictionary similar to LIWC), Vader Lexicon (sentiment), Lexical Categories (Lex Rank, LSA), Authority data (address, contact email, etc.), social tags, webpage ataset ility, and their HTML2seq feature in the form of a bag-of-tags were all content-based features they used (based on BoW). Regression and classification were the two configurations used for the credibility prediction. As a result, they put this model to test a real-world fact-checking problem and discovered that the model was able to distinguish between reputable and unreliable websites based on the assertions made in support of and opposition to each.

**Similarity-Aware:** SAFE Multi-Modal Fake News Detection, proposed by Xinyi et. al [30] uses multi-modal detection to identify fake news by using both textual and visual content. Although this has been done before, their method is innovative since it considers the similarities between textual and visual data and the technique that they implement to convert image data to text. They used the Linguistic Inquiry and Word Count (LIWC) for the textual data, and the VGG-19 – a convolutional neural network of 19 layers deep for the visual data, and the att-RNN network for the multi-modal data as baselines for their trials where all of which were outperformed.

By querying the knowledge graphs created for news stories from the knowledge base Dbpedia, it is possible to assess the credibility of the news based on the reliability of the content itself. This strategy was one of four ways listed in the survey [31] and used in [32][33][34][35]. It is regarded as an automatic fact-checking method.

Agrawal et al. [36] has considered time series into account on the Twitter news and employed a fake news classification method based on two algorithms - logistic regression, and a harmonic algorithm, and finally examined the performance. They inferred that the harmonic algorithm performed best with an accuracy of 90%. Ni et. al. [37] has proposed a model that uses attention-based neural networks to study about fake news classification that spots the clues surrounding fake news and the trend by which they spread. For this, a Multi-View Attention Network (MVAN) was being developed for detecting fake news on Twitter. This model had the ability to spot the clue words related to a particular news event.

A FACTDRIL (Fact Checking Dataset for Regional Indian Languages) dataset was introduced by Singhal et. al. [38] with a particular focus on low resource Indian languages like Marathi, Bangla, Telugu, Malayalam, Oriya, Tamil, Punjabi, Assamese, Urdu, Burmese, and Sinhala. The 22,435 samples from 11 Indian Low Resource Languages that have received IFCN (International Fact Checking Network) accreditation are included in the proposed dataset FACTDRIL. This FACTDRIL is the first large-scale multilingual dataset used to provide information on the accuracy of unverified claims for Indian languages with few resources. This work introduces a novel feature known as Investigation reasoning through manual interference. This section covers the numerous approaches fact-checkers use to decide whether or not a piece of news is reliable.

On the basis of two key elements, namely content-based and user-based, Azer et. al. [39] created a supervised machine learning strategy for credibility checking on twitter news. On the Pheme dataset, which was divided in this ratio, the authors used seven supervised machine learning techniques: Maximum Entropy (ME), Support Vector Machine (SVM),Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF),Conditional Random Forest (CRF), and Logistic Regression (LR). 80% of the dataset was used for training, 10% for testing, and 10% for validating the data set. The study's findings are as follows: Random Forest (RF) performs best, with accuracy ratings of 82.2% on user-based features and 83.4% on combined features (content-based features plus user-based features). On content-based features, Logistic Regression (LR) performs best, with an accuracy of 73.2%. In addition, it was also inferred that user-based features had a greater impact than content-based features.

Sahoo & Gupta [40] have proposed an automated method to identify fake news based on a variety of data properties of Facebook using deep learning and machine learning techniques using a chrome environment. This proposed methodology uses certain additional information tied to the user's Facebook account and its news content for identifying fake tales. The Long Short-Term Memory (LSTM) algorithm, which is a deep learning technique, has proved with an exceptional performance of 99.4% when compared to the other learning methodologies.

A Fake BERT was proposed by Kaliyar et. al. [41] combining multiple parallel blocks of a single-layer deep Convolutional Neural Network (CNN) with the BERT strategy. The BERT is a deep learning strategy that depends on bidirectional encoder representations from transformers. The most challenging component of understanding natural language is ambiguity, which is handled well by this combination.

To extract attitude representations from a post and any accompanying replies, Xie et. al. [42] suggested using the model –Stance Extraction and Reasoning Network (SERN). To accomplish binary fake news classification, they merged the posture representations and multimodal representation of both textual and visual content of a post.

The PHEME dataset and a condensed representation of the authors' own dataset from Fakeddit are used by the researchers Zubiaga et. al [43]. There are 5802 tweets in the PHEME dataset, 3830 of which are true and 1972 fake. The obtained accuracy rates are 76.53% and 96.63% respectively for Fakeddit and PHEME datasets.

The News Detection Graph (NDG), used by Kang et al. [44], is a heterogeneous graph that includes source nodes, domain nodes, review nodes and news nodes. Additionally, they suggested that implementing the Heterogeneous Deep Convolutional Network (HDCN) is beneficial in order to extract the news nodes' embeddings in the graph. Utilizing condensed versions of the Weibo and Fakeddit datasets, the authors assessed this approach. They achieved an F1 score of 96% for the Weibo dataset, 86% (three classes), 89% (binary classification),and 83% for the Fakeddit dataset (six classes).

## III PROBLEM STATEMENT

Our purpose is to research the viability of automated methods to spot fake news spread on digital channels. While fact-checking is a crucial method for spotting fake news, it is ineffective even though simple. Therefore, an automatic fake news detection system may be used to help readers to identify whether a content is more likely to be false, while ultimate final decision is left for a professional.

Formally, the fake news prediction can be defined as –"to assess whether a series of news stories from social media that contain text and image information is fake or not". However, it is not that simple to recognize fake news automatically. First, it is intrinsically difficult for people to distinguish between true and false news [4], especially when it comes to touchy themes like politics, entertainment and health. The problem of identifying fake news is made even more difficult by the fact that news items are generated by several sources, each of which has a unique style of representing the news contents and inherent biasing. In addition, they are transmitted in many ways in various platforms.

Digital media and social networking platforms presents a variety of research issues in identifying fake news. Firstly, the fact is that there are people who purposefully create fake news to confuse readers, such that the readers find it difficult to identify whether the news is real or not by just substantial reference. Thus, identifying fake news solely based on textual traits is inappropriate. Second, additional data must be provided to improve detection, such as knowledge bases and user social interactions [4]. However, the use of this supplemental data really contributes to a further significant problem with data quality. Although information from various modalities can offer hints for false identification, it can be difficult to draw out key aspects from each modality and successfully combine them.

The majority of studies are focused on unimodal data, however as information can come from various modalities, it is important to take into account both text and visual data for better fake news detection performance.

## IV MOTIVATION

Over one-third of people on the planet actively utilize digital platforms, such as social media networks and messaging platforms [5]. Through the launching of a flood of new applications and the alteration of current information ecosystems, these platforms have profoundly altered how people engage and communicate online. In particular, digital platforms have fundamentally altered how news is generated, delivered, and consumed, generating both unexpected opportunities and obstacles.

The nature of these digital platforms is partly to be blamed for this shift for the reasons: (i) online news is frequently more immediate and cheaper to generate and to consume than on similar traditional news media, like newspapers, magazines or news on television channels; and (ii) online news likely to be shared faster, and people comment on it at a rapid rate, and even involve in discussions with other readers on digital platforms.

Despite the many advantages that these platforms offer to our society, they have turned into a venue for disinformation operations that frequently aim to confuse individuals, particularly in contexts like politics and health. Online content often confuses the reader more such that the reader finds it difficult to determine what they read is trustable or not, and they are more susceptible to inaccurate or unreal information. The more incorrect information sources that people are exposed to, the harder it is for them to make a choice on reliability [6]. Therefore, the potential impact that fake news on our society can be too adverse making it a genuine issue.

False news reports existed since the past, but recently, their usage as a tool for manipulating people and controlling them has grown well because of the speed and immediacy with which they are disseminated through social media without any form of moderation or filtration. Furthermore, because of the captivating headlines, lay people are drawn to this type of news and frequently pay it more attention than accurate reporting [7].

Recently in 2018, a deceptive video about Kerala's flood-stricken region went viral on Facebook, claiming that the state's chief minister is ordering the Indian Army to stop carrying out rescue efforts in Kerala's flood inundation. Additionally, over 900,000 WhatsApp groups were created to spread false information about India's ruling party during the 2019 national elections [8]. The majority of fake stories are designed to mislead readers and inspire mistrust. The 2020 US presidential elections serve as another illustration of this consequence. According to a BBC article [10], the former president Donald Trump tried to overturn the president election's outcome by disseminating false information - according to a BBC News article [10].These issues prompted researchers to consider several automated techniques for spotting bogus news on social media networks.

Misinformation, spin, falsehoods, and dishonesty have always existed, but with the emergence of digital platforms, they may have become more widespread. As a result, the issue of fake news has become a global issue, and the absence of scalable fact-checking techniques is particularly concerning. As a result, an automated fake news detection system may be helpful in reducing the effects of widespread creation, spread, and consumption of online fake news. Additionally, photographs that are twisted, meaningless, or misrepresented accompanied with texts in fake news articles may be used to deceive viewers. This implies the necessity for a multimodal network to identify false news.

## V RESEARCH GAP

The gap identified after reviewing similar studies in the area of fake news detection is pointed out below:

- A single modality feature makes it difficult to spot fake news.
- Numerous strategies to identify fake news have been developed using linguistic approaches. However, there hasn't been much work done on visual-based verification.
- Source verification is seen as a component that is absent from the current models.
- The size of the datasets used in the literature is rather small.
- Time-sensitive and recently occurring events have received less attention from the current methodologies.
- Dataset bias is a concern because the bulk of studies are concentrated on a specific category of news (such as political news).

## VI OBJECTIVES

The following are the objectives that are finally arrived at after a detailed review of literature related to fake news detection:

- To analyze the prediction performance of fake news detection solutions in the-state-of-art.
- To propose a model for automatically detecting fake news for both long and short series of text data, such as

news articles and tweets.

- To build a system to identify fake images automatically.
- To assess the performance of the proposed approach using various news datasets and compare its effectiveness to other methods and techniques already in the literature.

## VII METHODOLOGY

There are different machine learning methods currently available for automatically detecting fake news [45]. Deep learning, one of its more recent branches, began to gain increasing significance in the discipline over time as more researches were done on it. This is because deep learning approaches, which outperform traditional machine learning techniques in a number of sectors [46], have more than one hidden layer between the input and output. Recurrent neural networks (RNN), which include simple RNN, Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM), are among the deep learning techniques that take time series into account when trying to classify news since it might be crucial to monitor changes Because language is made up of a series of words, each word depends on the words that came before it, RNN is the ideal way to represent languages. RNN also proved its competence in image processing and classification systems [47][48].

Design Science Research Methodology (DSRM) has been used in earlier works, such as Liu et al. [49], to construct deep learning-based artefacts. In DSRM, the artifact's design may be thought of as a search process that entails iterative review and artefact improvement. The suggested artefact thus combines Natural Language Processing (NLP) and ensemble LSTM and GRU-based models to categorize news information into credible and unreliable. In order to increase the solution's generalizability, the artefact will be implemented using Deep Learning, NLP, and Genetic Algorithm (GA)-based methodologies. Figure 1 represents the general high-level design of the proposed classifier that predicts and classifies the credibility of news. By comparing the model's performance on long and short text news along with images across different datasets, the effectiveness of the proposed approach will be ensured.

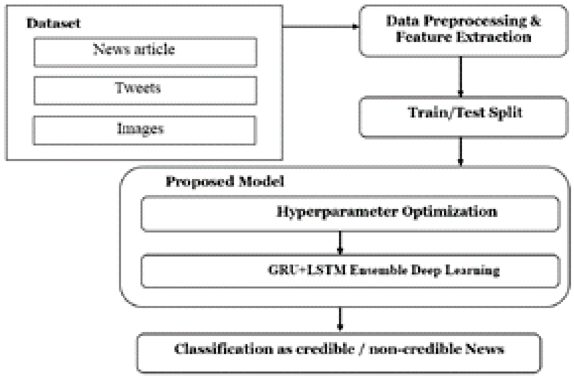


***Figure 1. High-level Representation of Proposed Methodology***

From the review of literature, we could see that LSTM models can be used for fake news detection but we prefer an ensemble model because it could combine the faster training and performance from GRU and greater volatility throughout its gradient descent from LSTM. When it comes to tasks involving the modelling of long-distance relations in language modelling challenges, LSTMs perform better than GRUs and can recall longer sequences. For language modelling issues, GRUs use less training data than LSTMs and train more quickly.

Textual data typically contains redundant information and inconsistencies; therefore, preprocessing is necessary before submitting the textual content to be learnt by the model. The textual content will be preprocessed using Natural Language Processing (NLP) methods in order to make them ready for additional analysis. Stop words, punctuation marks like "!", "&," and "$," emoticons like emojis, repetitive period marks, spaces between lines, extra spaces between words, empty rows, and additional parameters like tweetids, user ids etc., will be eliminated. Articles, pronouns, and prepositions that might be relevant in

English grammar have no semantic significance in the model's learning process. Thus they are also considered as stop words that would also be eliminated. The Porter Stemmer, created by Martin Porter, will be then used to accomplish stemming. Stemming is employed as it reduces words to their word stems, improving efficiency and minimizing data storage.

To develop the model, firstly textual features will be taken and employed. N-grams, or a series of words that frequently appear together in tweets, such as unigrams, bigrams, and trigrams, as well as popular hashtags (such as #pray4boston) and trending person mentions (such as @Barack Obama), are examples of features. The quantity of words and characters were also provided as components for the model creation.

Images will be preprocessed to extract texts written on images and the image will be converted to grayscale as the color information adds little knowledge to our domain of interest.

The next process involves tuning the hyper parameters. There exist a range of methods, including evolutionary algorithms like particle swarm optimization, the k-fold cross-validation, and genetic algorithms, to optimize the hyper parameters of model. Numerous similar strategies have been shown to have the issue of convergent in the local minima of the solution space (i.e., delivering the minimum values of the hyper parameters iteratively) [50]. The problem of local minima may be delayed by Genetic Algorithm (GA), which will eventually converge at the global maxima (i.e., provide the highest values for the hyper parameters) [51]. As a result, GA has advantages over conventional hyper parameter tuning methods, including the ability to avoid local minima and maxima and to handle complex issues and huge hyper parameter values [52]. Furthermore, previous research [53] [54] has shown that the GA has the ability to throw out the alternative cross-validation parameter tuning strategies. As a result, the GA was also included in the suggested methodology.

## VIII CONCLUSION

The viral nature of unreal news may vary and spread as it circulates on the internet and may appear on numerous online sites. The majority of the methods appear to be less effective at identifying fake news in practical situations. This study makes the case that contextual, emotive, semantic, syntactic, and semantic-syntactic linkages between real and fake news are crucial for fake news identification tasks. For automatically identifying fake news on social media, the GA-optimized ensemble LSTM -GRU-based model was proposed.

## REFERENCES


[1] Rachel R. Mourão and Craig T. Robertson,"Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information", in Journalism Studies, 20(14):2077–2095, 2019.

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu,"Fake news detection on social media: A data mining perspective", in ACM SIGKDD explorations newsletter, 19(1):22–36, 2017.

[3] Ahmed, A. A. A., Aljabouh, A., Donepudi, P. K., & Choi, M. S.,"Detecting Fake News using Machine Learning: A Systematic Literature Review". arXiv preprint arXiv:2102.04458, 2021, https://arxiv.org/abs/2102.04458.

[4] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu,"Fake news detection on social media: A data mining perspective",in ACM SIGKDD Explorations Newsletter, 19(1):22–36, 2017.

[5] Gallagher, K., "The social media demographics report: Differences in age, gender, and income at the top platforms", 2017http://www.businessinsider.com/thesocial-media-demographics-report-2017-8.

[6] Supanya Aphiwongsophon and Prabhas Chongstitvatana. "Detecting fake news with machine learning method",in 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology(ECTI-CON). Institute of Electrical and Electronics Engineers Inc., 1 2019.

[7] Ryan Browne,"'junk news' gets massive engagement on facebook ahead of eu elections", in Study Finds, CNBC, 2019, Accessed: 19-09-2022.

[8] Trends Desk, "Kerala Floods: Indian Army warns of fake video showing man wearing Army combat uniform," in The Indian Express, The Indian Express, 20-Aug-2018.

[9] Kaur, Sawinder, Parteek Kumar, and Ponnurangam Kumaraguru, "Automating fake news detection system using multi-level voting model", In Soft Computing 24, no. 12 (2020): 9049-9069.

[10] Reality Check team.,"Us election 2020: Fact-checking trump team's main fraud claims", https://www.bbc.com/news/election-us-2020-55016029, 11 2020. Accessed: 20-09-2022.

[11] Wu, L., Morstatter, F., Carley, K. M., & Liu, H.,"Misinformation in Social Media: Definition, Manipulation, and Detection",in SIGKDD Explor. Newsl., 21(2), 80–90, 2019, https://doi.org/10.1145/3373464.3373475

[12] Hadeer Ahmed, Issa Traore, and Sherif Saad,"Detection of online fake news using n-gram analysis and machine learning techniques", In International conference on intelligent, secure, and dependable systems in distributed and cloud environments. Springer, 2017.

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan,"A survey on bias and fairness in machine learning". arXiv, 8 2019.

[14] André Cruz, Gil Rocha, Rui Sousa-Silva, and Henrique Lopes Cardoso,"Team fernandopessa at SemEval-2019 task 4: Back to basics in hyperpartisan news detection", In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 999–1003, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[15] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea,"Automatic detection of fake news", arXiv preprint arXiv:1708.07104, 2017.

[16] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, GuangxuXun, KishlayJha, Lu Su, and Jing Gao.2018,"EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection",in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (London, United Kingdom) (KDD '18). Association for Computing Machinery, New York, NY, USA, 849–857. https://doi.org/10.1145/3219819.3219903

[17] Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., & Afroz, S.,"A benchmark study on machine learning methods for fake news detection", Computation and Language, 2021, https://arxiv.org/abs/1905.04749.

[18] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., and Cha, M.,“Detecting rumors from microblogs with recurrent neural networks”, In Proceedings of the ACM International Joint Conference on Artificial Intelligence, pages 3818–3824, 2016.

[19] Fang, Y., Gao, J., Huang, C., Peng, H., and Wu, R.,“Self multi-head attentionbased convolutional neural networks for fake news detection”, PloS One, 2019, 14(9):e0222713.

[20] KALIYAR, Rohit; GOSWAMI, Anurag; NARANG, Pratik; SINHA, Soumendu,“FNDNet- A Deep Convolutional Neural Network for Fake News Detection”, Cognitive Systems Research. 2020, vol. 61. Available from DOI: 10.1016/j.cogsys.2019.12.005.

[21] HIRAMATH, C. K.; DESHPANDE, G. C.,“Fake News Detection Using Deep Learning Techniques”,in: 2019 1st International Conference on Advances in Information Technology (ICAIT). 2019, pp. 411–415.

[22] KALIYAR, R. K.,“Fake News Detection Using A Deep Neural Network”, In: 2018 4th International Conference on Computing Communication and Automation (ICCCA). 2018, pp. 1–7.

[23] REDDY, Harita; RAJ M, Namratha; GALA, Manali; BASAVA, Annappa, “Text-mining-based Fake News Detection Using Ensemble Methods”, International Journal of Automation and Computing. 2020. Available from DOI: 10.1007/s11633-019-1216-5.

[24] POPOOLA, Olu,“Using Rhetorical Structure Theory for Detection of Fake Online Reviews”, In: Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms. Santiago de Compostela, Spain: Association for Computational Linguistics, 2017, pp. 58–63. Available from DOI: 10.18653/v1/W17-3608.

[25] BHUTANI, B.; RASTOGI, N.; SEHGAL, P.; PURWAR, A,“Fake News Detection Using Sentiment Analysis”, In: 2019 Twelfth International Conference on Contemporary Computing (IC3). 2019, pp. 1–5.

[26] RUBIN, Victoria; CONROY, Niall; CHEN, Yimin; CORNWELL, Sarah, “Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News”, in: Proceedings of the Second Workshop on Computational Approaches to Deception Detection. San Diego, Cali fornia: Association for Computational Linguistics, 2016, pp. 7–17. Available from DOI: 10.18653/v1/W16-0802.

[27] QIAN, Feng; GONG, Chengyue; SHARMA, Karishma; LIU, Yan,“Neural User Response Generator: Fake News Detection with Collective User Intelligence”, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 3834–3840. Available from DOI: 10.24963/ijcai.2018/533.

[28] RUCHANSKY, Natali; SEO, Sungyong; LIU, Yan, “CSI: A Hybrid Deep Model for Fake News Detection”, Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM ’17. 2017. ISBN 9781450349185. Available from DOI: 10.1145/3132847.3132877.

[29] ESTEVES, Diego; REDDY, Aniketh Janardhan; CHAWLA, Piyush; LEHMANN, Jens,“Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web”,CoRR. 2018, vol. abs/1809.00494.

Available from arXiv: 1809.00494.

[30] ZHOU, Xinyi; WU, Jindi; ZAFARANI, Reza,"SAFE: Similarity-Aware Multi-Modal Fake News Detection", 2020. Available from arXiv: 2003.04981 [cs.CL].

[31] ZHOU, Xinyi; ZAFARANI, Reza, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities", in ACM Computing Surveys. 2020. ISSN 1557-7341. Available from DOI: 10.1145/3395046.

[32] CIAMPAGLIA, Giovanni Luca; SHIRALKAR, Prashant; ROCHA, Luis M.; BOLLEN, Johan; MENCZER, Filippo; FLAMMINI, Alessandro,"Computational Fact Checking from Knowledge Networks", PLOS ONE. 2015, vol. 10, no. 6, pp. e0128193. ISSN 1932-6203. Available from DOI: 10.1371/journal.pone.0128193.

[33] PAN, Jeff; PAVLOVA, Siyana; LI, Chenxi; LI, Ningxi; LI, Yangmei; LIU, Jinshuo,"Content Based Fake News Detection Using Knowledge Graphs", in 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I. In: 2018, pp. 669–683. ISBN 978-3-030-00670-9. Available from DOI: 10.1007/978-3-030-00671-6_39.

[34] LIN, Peng; QI, Song; SHEN, Jialiang; WU, Yinghui,"Discovering Graph Patterns for Fact Checking in Knowledge Graphs" , in book: Database Systems for Advanced Applications,2018, pp. 783–801. ISBN 978-3-319-91451-0. Available from DOI: 10.1007/978-3-319-91452-7_50.

[35] WANG, Youze; QIAN, Shengsheng; HU, Jun; FANG, Quan; XU, Changsheng,"Fake News Detection via Knowledge-Driven Multimodal Graph Convolutional Networks", In: Proceedings of the 2020 International Conference on Multimedia Retrieval. Dublin, Ireland: Association for Computing Machinery, 2020, pp. 540–547. ICMR '20. ISBN 9781450370875. Available from DOI: 10.1145/3372278.3390713.

[36] Agrawal, R., de Alfaro, L., Ballarin, G., Moret, S., Di Pierro, M., Tacchini, E., & Della Vedova, M. L.,"Identifying Fake News from Twitter Sharing Data: A Large-Scale Study", 2019, ArXiv Preprint ArXiv:1902.07207. http://arxiv.org/abs/1902.07207

[37] Ni, S., Li, J., & Kao, H.-Y.,"MVAN: Multi-View Attention Networks for Fake News Detection on Social Media", IEEE Access, 2021, 9, 106907–106917. https://doi.org/10.1109/ACCESS.2021.3100245

[38] Singhal, S., Shah, R. R., & Kumaraguru, P.,"Factorization of Fact-Checks for Low Resource Indian Languages", 2021, arXiv preprint arXiv:2102.11276.

[39] Azer, M., Taha, M., Zayed, H. H., &Gadallah, M.,"Credibility Detection on Twitter News Using Machine Learning Approach",in International Journal of Intelligent Systems & Applications, 2021, 13(3).

[40] Sahoo, S. R., & Gupta, B. B.,"Multiple features based approach for automatic fake news Detection on social networks using deep learning", Applied Soft Computing, 2021, 100, 106983.

[41] Kaliyar, R. K., Goswami, A., & Narang, P., "FakeBERT: Fake news detection in Social media with a BERT-based deep learning approach", Multimedia tools and applications, 80(8), 2021, 11765-11788. 'doi: 10.1007/s11042-020-10183-2.

[42] Xie, J.; Liu, S.; Liu, R.; Zhang, Y.; Zhu, Y,“SERN: Stance Extraction and Reasoning Network for Fake News Detection”, In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Conference, 6–12 June 2021; pp. 2520–2524.

[43] Zubiaga, A.; Liakata, M.; Procter, R.,“Exploiting context for rumour detection in social media”, In Proceedings of the International Conference on Social Informatics, Oxford, UK, 13–15 September 2017; pp. 109–123.

[44] Kang, Z.; Cao, Y.; Shang, Y.; Liang, T.; Tang, H.; Tong, L,“Fake News Detection with Heterogenous Deep Graph Convolutional Network”, In Proceedings of the Advances in Knowledge Discovery and Data Mining, Virtual Event, 11–14 May 2021.

[45] Manzoor, S. I., Singla, J., Nikita,“Fake news detection using machine learning approaches: A systematic review”, In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). Presented at the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp.230–234). https://doi.org/10.1109/ICOEI.2019.8862770.

[46] Ongsulee, P.,“Artificial intelligence, machine learning and deep learning”, In 2017 15th International Conference on ICT and Knowledge Engineering (ICT KE). Presented at the 2017 15th International Conference on ICT and Knowledge Engineering (ICT KE) (pp. 1–6). https://doi.org/10.1109/ICTKE.2017.8259629.

[47] Agilandeeswari, Loganathan, Kuhoo Sharma, and Saurabh Srivastava, "Intelligent Image Captioning Approach with Novel Ensembled Recurrent Neural Network Model." In International Conference on Intelligent Systems Design and Applications, pp. 1227-1236. Springer, Cham, 2022.

[48] Das, A., Mohapatra, S. K., & Mohanty, M. N.,“Design of deep ensemble classifier with fuzzy decision method for biomedical image classification”, Applied Soft Computing, 2022, 115, 108178.

[49] Liu, X., Wang, G. A., Fan, W., & Zhang, Z.,“Finding useful solutions in online knowledge communities: A theory-driven design and multilevel analysis”, Information Systems Research, 2020, 31(3), 731–752.

[50] Ritchie, M. D., White, B. C., Parker, J. S., Hahn, L. W., & Moore, J. H.,“Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases”,in BMC Bioinformatics, 4(1), 28, 2003, https://doi.org/10.1186/1471-2105-4-28

[51] Koza, J. R., & Rice, J. P.,“Genetic generation of both the weights and architecture for a neural network”, IJCNN-91-Seattle International Joint Conference on Neural Networks, 1991, ii, 397–404 vol.2. https://doi.org/10.1109/ IJCNN.1991.155366

[52] Bakhashwain, N., & Sagheer, A.,“Online tuning of hyper parameters in deep LSTM for time series applications”, International Journal of Intelligent Engineering and Systems, 2021, 14(1), 212–220.

[53] Wicaksono, A. S., & Supianto, A. A.,“Hyper parameter optimization using genetic algorithm on machine learning methods for online news popularity prediction”, International Journal of Advanced

Computer Science and Applications, 2018, 9(12), 263–267.

[54] Liashchynskyi, P., & Liashchynskyi, P.,"Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS", 2019, https://arxiv.org/abs/1912.06059v1.