Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model

Koosha Sharifani [1], Mahyar Amini [2], Yaser Akbari [3], Javad Aghajanzadeh Godarzi [4]
[1] University of North Carolina at Charlotte, United States
[2] University Technology Malaysia (UTM), Malaysia
[3] University of Applied Science and Technology S.O.K.A.N, Iran
[4] ARYAN Institute of Science and Technology, Iran

ABSTRACT

Fake news or fabricated news, refers to false information published under the guise of being authentic news, often to influence political views. Fabricated news articles are a threat to people's trust in the government and in effect, one of the biggest threats that modern-day democracies are facing. As the menace of fake news is growing with each passing day, so is the research community getting more actively involved in curbing this issue. This paper reviews the current progress of the advancements done to solve the issue. The paper also presents various ensemble techniques to perform the binary classification of news articles. Additionally, Natural Language Processing (NLP) emerges as one of the hottest topic in field of speech and language technology and Machine Learning (ML) can comprehend how to perform important NLP tasks. This is often achievable and cost-effective where manual programming is not. This paper strives to study NLP and ML and gives insights into the essential characteristics of both. It summarizes common NLP tasks in this comprehensive field, then provides a brief description of common machine learning approaches that are being used for different NLP tasks. Also this paper presents a review on various approaches to NLP and some related topics to NLP and ML. Respectively and with regard to this research article, fake news detection research is still in the early stage as this is a relatively new phenomenon in the interest raised by society. Machine learning helps to solve complex problems and to build AI systems nowadays and especially in those cases where we have tacit knowledge or the knowledge that is not known. We used machine learning algorithms and for identification of fake news; we applied three classifiers; Passive Aggressive, Naïve Bayes, and Support Vector Machine. Simple classification is not completely correct in fake news detection because classification methods are not specialized for fake news. With the integration of machine learning and text-based processing, we can detect fake news and build classifiers that can classify the news data. Text classification mainly focuses on extracting various features of text and after that incorporating those features into classification. The big challenge in this area is the lack of an efficient way to differentiate between fake and non-fake due to the unavailability of corpora. We applied three different machine learning classifiers on two publicly available datasets. Experimental analysis based on the existing dataset indicates a very encouraging and improved performance.

KEYWORDS: Machine Learning, Natural Language Processing, Classification Techniques.

1.0 INTRODUCTION

Fake news detection topic has gained a great deal of interest from researchers around the world [1 - 3]. When some event has occurred, many people discuss it on the web through the social networking [4 - 9]. They search or retrieve and discuss the news events as the routine of daily life [10 - 13]. Some type of news such as various bad events from natural phenomenal or climate are unpredictable [11 – 16]. When the unexpected events happen, there are also fake news that are broadcasted that creates confusion due to the nature of the events [17 – 22]. Very few people know the real fact of the event while the most people believe the forwarded news from their credible friends or relatives [22 – 24]. These are difficult to detect whether to believe or not when they receive the news information. So, there is a need of an automated system to analyze truthfulness of the news [25 – 27]. Predictive active machine learning is a supervised learning method in which the learner is in control of the data from which it learns [28 – 32]. That control is used by the learner to ask an oracle, a teacher, typically a human with extensive knowledge of the domain at hand, about the classes of the instances for which the model learned so far makes unreliable predictions [33 – 36]. The active learning process takes as input a set of labelled examples, as well as a larger set of unlabelled examples, and produces a classifier and a relatively small set of newly labelled data [35 – 41]. The overall goal is to produce as good a classifier as possible, without having to mark-up and supply the learner with more data than necessary [42 – 47]. The learning process aims at keeping the human annotation effort to a minimum, only asking for advice where the training utility of the result of such a query is high [43 – 45]. Fake news is a type of yellow journalism or propaganda that consists of deliberate misinformation or hoaxes spread via traditional print and broadcast news media or online social media [40 – 44]. Fake news is as old as the news industry itself misinformation, propaganda, hoaxes and satire have long been in existence [48 – 52]. Today anybody can publish anything credible or not that can be consumed by the World Wide Web. Due to this, people can be deceived intentionally or unintentionally and do not think before sharing such types of news to the far ends of the world [50 – 54]. The counterfeited news problem can be resolved or at least overcome with machine learning and artificial intelligence. In general, fake news detection is considered as a challenging task that requires multidisciplinary efforts [51 – 54]. For deception detection, there exists a large body of research done where machine learning methods are applied [55 – 57]. Classification of online news and social media posts were the target of those methods but after the 2016 United States Presidential elections, determining fake news has also been the subject of attention in the literature [57 – 59]. Simple content related classification n-gram and part of speech (POS) tagging have proven insufficient in fake news context. Fake news detection through classification is not sufficient since it missed the important context of the information, however a deep analysis of the content that can be useful [60 – 64]. Context-free grammar (CFG) produced good results with the combination of the n-gram in deception related classification. The accuracy achieved 85%-91% when applied on news article datasets through classification [61 – 69]. We propose a hypothesis that simple classification is not enough to tackle the issue; we need to combine it with machine learning techniques. The hypothesis is proven on publicly available datasets by developing the proposed model after several experiments [70 – 74]. We observe that the relative frequency of words can also be the reason for fake and non-fake class categorization. Using word cloud visualization, we observe the corpus trend, as shown in Fig. 1. The word cloud representation reflects important word entities [75 – 79]. We use different sources of news for the testing and training datasets so that we can observe how well our models generalize to unseen data points. In the first step, we applied text extraction features covered under the text classification module [80 – 84]. Fake news can be categories in seven different types. Table 1 explains seven types of fake news.

TABLE 1 - SEVEN TYPES OF FAKE NEWS

No.	Type	Details
1	False Connection	hen headlines, visuals or captions don’t support the content.
2	False Context	When genuine content is shared with false contextual information.
3	Manipulated Content	When genuine information or imagery is manipulated to deceive.
4	Satire	No intention to cause harms but has potential to fool.
5	Misleading Content	Use of information to frame an issue.
6	Imposter Content	When genuine sources are impersonated.
7	Fabricated Content	New content that is 100% false, designed to deceive and do harm.

With advances in computer technology, we presently have the ability to store and process enormous amounts of data, and likewise to access it from physically far locations over a computer network [85 – 92]. Most data acquisition devices are digital now and record dependable data. There is a process that explains the data that is observed. Machine learning (ML), systems automatically learn models from data to make better decisions. As such, they are part of a major subfield of artificial intelligence (AI). There are 3 main approaches to learning from data: supervised, unsupervised, and reinforcement learning. In supervised learning, a target attribute is predicted, and ML algorithms infer a model from labelled input data (i.e., a training data set that provides examples described by predictive attributes and values for the target attribute). The goal is to make target predictions on new data to obtain good generalization performance [90 – 97]. In contrast, there is no target attribute in unsupervised learning, and thus no labelled data. Unsupervised learning consists of inferring a model to describe hidden patterns from unlabelled data. Under circumstances in which labelled data acquisition proves to be difficult, (e.g., costly), semi supervised ML methods can use both labelled and unlabelled data for learning [94 – 105]. The third main category of ML is reinforcement learning, in which the ML model uses feedback that acts as a reward or punishment to maximize its performance. ML is limited to certain capacities [106 – 117]. For one, it relies on collections of data that may be incomplete, noisy, or subject to systematic bias, all of which can lead to erroneous predictions. In addition, ML algorithms may introduce bias. Interesting questions to be addressed in ML are discussed in an article by Domingo's. However, when carefully conducted, ML can have great utility. AI and ML have many applications, many of which are encountered in daily life. Supervised ML, for example, is widely used for spam filtering (i.e., classifying incoming email as spam or not spam). It is also used to classify credit applicants based on their probabilities of default. Unsupervised ML, such as algorithm clustering, is able to group customers with similar characteristics and their likelihood to purchase. This is widely used by banks for market segmentation [114 – 123]. Finally, automatic document clustering that organizes similar documents into classes (for purposes of improving information retrieval, for example) is gaining importance due to the increasing number of documents on the internet. Though the details of the process underlying the generation of data are unknown, it may not be feasible to identify the process entirely, but it can construct a good and helpful approximation [124 – 127]. Though identifying the complete process may not be possible, it can still be suitable to detect specific patterns or regularities. This is the role of machine learning. Such patterns can help to comprehend the process, or use those patterns to make predictions [128 – 134]. Application of machine learning methods to large databases is called data mining. In data mining, a large volume of data is processed to construct a simple model with valuable use [135 – 142]. But machine learning is not just a database problem; it is also a part of artificial intelligence. To be intelligent, a system that is in a changing environment should have the ability to learn [143 – 149]. If the system can learn and adapt to these changes, the system designer needs no predict and provides solutions for all proper situations [150 – 156]. Machine learning also helps to find solutions for many problems in vision, speech recognition, and robotics [157 – 164]. Machine learning is programming computers to optimize a performance criterion using example data or past experience [165 – 172]. There is a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience [173 – 181]. The model may be predictive to make predictions in the future, or descriptive to obtain knowledge from data, or both [182 – 187].

Fig. 1 Word Cloud of News Articles

Natural Language Processing (NLP), is a sub discipline of computer science that emerged in the 1960s. In 1967, the first published book on the subject, Introduction to Computational Linguistics, clearly considers language from a symbolic point of view: it describes techniques such as syntax parsing using dependency trees or Chomsky transformational grammars and statistical methods (word counting) are only hinted at [188 – 195]. At that time, computing resources were sparse and had to be carefully managed; hence, a whole chapter of the book is dedicated to the storage of grammars in memory. The situation changed in the 1990s when personal computers became largely available and increasingly powerful [196 – 201]. A new approach to NLP based on statistical methods emerged [1 – 17]. The book by Manning and Schultz, Foundations of Statistical Natural Language Process, is a landmark of this evolution [14 – 23]. The 3 main sections of the book are dedicated to (1) methods at the word level (collocations, n-grams, and word sense disambiguation), (2) methods at the sentence level (morph syntactic parsing using Markov models, and probabilistic context-free grammars), and (3) clustering, classification, and information retrieval. Probabilistic context-free grammars are a typical example of the evolution of NLP methods: the symbolic approach by Chomsky—or at least a simplified version—is endowed with probabilities attached to productions, and the ambiguity of natural language is reflected in the coexistence of several syntax trees with different probabilities [24 – 39]. Natural Language Processing is a hypothetically driven range of calculative techniques for analyzing and representing naturally texts at one or more levels of linguistic analysis in order to achieve human-like language processing for a range of tasks or applications [33 – 42]. The term Natural Language Processing surrounds a wide set of techniques for automated generation, manipulation and analysis of natural or human languages [43 – 56].

TABLE 2 - SOME OF NLP CHARACTERISTICS

No.	Character	Details
1	Origins	Computer Science; Linguistic; Cognitive Psychology.
2	Divisions	Language Processing; Language Generation.
3	Approaches to NLP	Symbolic Approach; Statistical Approach; Connectionist Approach.
4	NLP Applications	Retrieval; Extraction; Question Answering; Dialogue Systems.

Machine learning uses the theory of statistics in building mathematical models, because the core task is making inference from a sample [48 – 59]. The role of computer science is divided into two parts: First, in training, it is required the effective algorithms to solve the optimization problem, and also to store and process the enormous amount of data. Second, once a model is learned, its representation and algorithmic solution for inference needs to be efficient too [60 – 69]. In particular applications, the effectiveness of the learning or inference algorithm, namely, its space and time complexity, perhaps are as significant as its predictive accuracy [64 – 71]. Natural Language Processing (NLP) deals with real text element processing [70 – 78]. The text element is transformed into machine format by NLP. A system capable of obtaining and combining the knowledge automatically is referred as machine learning [72 – 83]. Machine learning systems automatically learn programs from data [74 – 86]. The application of machine learning to natural language processing is constantly increasing. Spam filtering is one where spam generators on one side and filters on the other side keep finding more and more talented ways to surpass each other [84 – 92]. Perhaps the most striking would be machine translation. After decades of research on hand-coded translation rules, it has become apparent lately that the most favourable way is to provide a very large number of example pairs of translated texts and have a program understand automatically the rules to map one string of characters to another [92 – 117].

TABLE 3 - LEVELS OF NLP

No.	Level	Details
1	Phonetics	Knowledge About Linguistic Sounds.
2	Morphology	Knowledge Of The Meaningful Components Of Words.
3	Syntactic	Knowledge Of The Structural Relationships Between Words.
4	Semantic	Knowledge Of Meaning.
5	Pragmatics	Knowledge Of The Relationship Of Meaning To The Intentions Of The Speaker.
6	Discourse	Knowledge About Linguistic Units Larger Than A Single Utterance.

NLP techniques are affected by Linguistics and Artificial Intelligence, Machine Learning, Computational Statistics and Cognitive Science [97 – 108]. Here is introduction of some basic terminology in NLP that will be avail. A brief description about NLP, is shown in tables 2 and 3.

Table 2 illustrates some of NLP characteristics. Table 3 presents levels of NLP [109 – 118]. We observe that the relative frequency of words can also be the reason for fake and non-fake class categorization. Using word cloud visualization, we observe the corpus trend, as shown in Fig. 1. The word cloud representation reflects important word entities [119 – 127]. For example, we can easily observe the highly frequent words Political, Americas, 2016, President, Obama and Presidential Debates, respectively [128 – 133]. We use different sources of news for the testing and training datasets so that we can observe how well our models generalize to unseen data points. In the first step, we applied text extraction features covered under the text classification module [134 – 141]. Fake news can be categories in seven different types. Table 1 explains seven types of fake news [142 – 155]. The rest of this paper is organized as follows, Second II reviews the previous work, and Section III describes the Methodology. The Proposed model, Pre-processing and Machine learning are described in Sections IV-VI, Section VII describes the implementation task, Results and discussion are described in Section VIII and finally, the last section gives the Conclusion and Future Work.

2.0 LITERATURE REVIEW

With advances in computer technology, we presently have the ability to store and process enormous amounts of data, and likewise to access it from physically far locations over a computer network [156 – 161]. Most data acquisition devices are digital now and record dependable data [162 – 169]. There is a process that explains the data that is observed [170 -176]. Though the details of the process underlying the generation of data are unknown, it may not be feasible to identify the process entirely, but it can construct a good and helpful approximation [177 – 183]. Though identifying the complete process may not be possible, it can still be suitable to detect specific patterns or regularities [184 – 192]. This is the role of machine learning. Such patterns can help to comprehend the process, or use those patterns to make predictions [193 -201]. Application of machine learning methods to large databases is called data mining. In data mining, a large volume of data is processed to construct a simple model with valuable use [1- 16]. But machine learning is not just a database problem; it is also a part of artificial intelligence [17 – 25]. To be intelligent, a system that is in a changing environment should have the ability to learn. If the system can learn and adapt to these changes, the system designer needs no predict and provides solutions for all proper situations. Machine learning also helps to find solutions for many problems in vision, speech recognition, and robotics. Machine learning is programming computers to optimize a performance criterion using example data or past experience [26 – 42]. There is a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience [43 – 49]. The model may be predictive to make predictions in the future, or descriptive to obtain knowledge from data, or both. Machine learning uses the theory of statistics in building mathematical models, because the core task is making inference from a sample [50 – 58]. The role of computer science is divided into two parts: First, in training, it is required the effective algorithms to solve the optimization problem, and also to store and process the enormous amount of data. Second, once a model is learned, its representation and algorithmic solution for inference needs to be efficient too. In particular applications, the effectiveness of the learning or inference algorithm, namely, its space and time complexity, perhaps are as significant as its predictive accuracy [54 – 63]. Based on a critical and systematic review of recent research papers published over the last four years, different approaches of dealing with fake news have been analyzed in this paper [64 – 76]. This survey investigates the role of machine learning, deep learning, and natural language processing applications to detect fake news, focusing on the characteristics of the different techniques and approaches, conceptual models for detecting fake news and the role of cognitive agents in this context as they have gained great popularity in the last three years [43 – 52]. The literature review outlines the research shortcoming and gaps in current automatic fake news detection models. In Kai Shu et al. present FakeNewsTracker, a system to understand and detect fake news. FakeNewsTracker benefits researcher in identifying fake news by automatically collecting data for news and social context with a number of effective visualization techniques [1 – 11]. The dataset has been built through Politifact and twitter feed and considers article body, retweets and engagements as the features for binary classification of news article. LSTM with two layers consisting of 100 cells has been employed as their base technique to train the model and testing has been done with other Machine Learning algorithms like Support Vector Machine, Logistic Regression and Naïve Bayes Classifier. While Support Vector Machine and Logistic Regression obtained relatively close accuracies at 68.4% and 68.3 respectively, Naïve Bayes returned 62% accuracy [5 – 18]. Also, retweets were not considered for both training and testing. The experiment was performed on a crowdsourced database and not a standard dataset and accuracies obtained are at best, in high 60 percent range [8 – 23]. Hadeer

Ahmed et al. in propose a fake news detection model in this paper. They have employed n-gram analysis and machine learning techniques. They investigated and compared two different features extraction techniques and six different machine classification techniques. Using Term Frequency-Inverted Document Frequency (TF-IDF) and Linear Support Vector Machine (LSVM) as feature extraction technique and as a classifier respectively has fetched a high accuracy of 92%. Their new dataset has been obtained from real-world sources such as Reuters website. The fake news dataset has been gathered from a dataset on kaggle.com [1 – 23]. The Kaggle dataset is based on a collection of fake news articles from untrustworthy web sites which have been documented jointly by Politifact and Facebook in their initiative to weed out these websites. Their dataset consists of 12,600 truthful articles and equal number of fake news articles from kaggle.com. Their primary focus remains only on political news article since they have been the main target of fake news distributors [14 – 26]. Each article in the dataset consists of Article Text, Article Type, Article label (fake or truthful), Article Title and Article Date [8 – 22]. First, the unigram (n = 1) model was considered for their experiment, then bigram (n = 2), eventually added 1 to n until reaching tetragram. Furthermore, the experiment was performed by combining each n value with a different number of attributes or features during the testing phase. The experiments were run on a 5- fold cross validation; with an 80:20 ratio of training and testing data during each validation. From the results they obtained, Linear-based classifiers (Linear SVM, SDG, and Logistic regression) achieved better results than nonlinear ones. The lowest accuracy of 47.2% was achieved using KNN and SVM with four-gram words and 50,000 and 10,000 feature values [1 – 23]. In Veronica Perez-Rosas et al, the research aims at creating an automatic fake news detector. Their dataset is diverse, such that it covers seven different domains. FakeNewsAMT and Celebrity datasets have been employed for their research. Their feature set consists of n-grams, punctuations, psycholinguistic features, readability, and syntax. They performed several experiments with various feature sets combination to explore their predictive models both separately and jointly. A linear SVM classifier was used using five-fold cross validation, with accuracy, recall, precision and F-score as performance metrics. They used the machine learning algorithms implementation available in the e1071 packages (Meyer et al., 2015) and caret (Kuhn et al., 2016). Results show that they achieved the best accuracies, with 0.74 and 0.76 respectively when using all the features on the two datasets. Soham Mone et al in aim to emulate the popular chrome extension, BS Detector [8 – 33]. Their dataset consists of 8071 true news sample from Kaggle and 4094 samples of fake news data (headlines + body). The features used are Article Title and Article Body. Two models were developed independently in order to achieve their desired objective: An Average-Hypothesis model, and a Neural Network. They ran Naïve Bayes, SVM, and Logistic Regression model and obtained an average accuracy of 83%. Building on previous work in detecting satire, Victoria Rubin et al in proposed an SVM-based algorithm, utilizing 5 predictive features (Absurdity, Humor, Grammar, Negative Affect, and Punctuation) and tested their combinations on 360 news articles. They have used Linear SVM with 10-fold cross validation. Their best predicting feature combination (Absurdity, Grammar and Punctuation) detects satirical news with 82% accuracy. However, this research only focuses on sarcasm detection which is the lowest level of fake news. Ting Su et al in their research paper examines recurrent neural networks-based language representations (e.g., BERT, BiLSTM) and the advantages that they possess to build ensemble classifiers. These classifiers can predict if one news title is either related to, or even, additionally disagrees with an earlier news title [17 – 42]. The dataset consists of 321,000 news article titles created during the WSDM 2019 challenge. The experiments on this dataset show that the BERT-based models outperform BiLSTM substantially, which in-turn significantly outdoes a simpler representation based on embeddings. Furthermore, even BERT approach can be improved by combining it with a simple BM25 feature. The experiment was able to reach an accuracy of 88.5% on an ensembled BERT and BM25. However, the research focuses only on article title to judge the veracity of the news article [13 – 26]. In Aravinder Singh Bali et al have performed a comparative analysis of seven different machine learning algorithms namely Support Vector Classifier, Random Forest, Gaussian Naïve Bayes, k-Nearest Neighbor, AdaBoost, Gradient Boosting and Multilayer Perceptron. The dataset is a mixture of Open Source (11161 articles), Kaggle dataset (20800 articles) and George McIntire dataset (6335 articles). N-gram, Sentiment, Readability, Cosine Similarity, and Word Embedding were the features studied for each article. The accuracies for the three datasets were 86.2%, 91.05% and 87.3% respectively. Dataset is not standard and there is no analysis of psychometric features. In the research by Pranav Bhardwaj et al in their research paper implemented Naive Bayes classifier, Recurrent Neural Networks, and Random Forest classifiers using five groups of linguistic features. The model was evaluated with real or fake dataset obtained from Kaggle. The experiment saw the highest accuracy of 95.66% achieved using bigram features with the Random

Forest classifier. However, semantic features may be additionally combined with other linguistic cues and meta data to improve the performance of the classifiers [12 – 32]. The study by Christian Janze et al examines visual, affective, cognitive, and behavioral cues of the news posts and its usage by machine learning classifiers to identify fake news automatically. They used Support Vector Machine, Logistic Regression, Decision Tree, Random Forest and XGBoost separately to build their model. The best performing configurations, i.e. with SVM with a stratified 10-fold cross validation achieved an average accuracy around 80%, and a recall of around 90% while Logistic Regression giving the lowest accuracy at 76%. The "balanced" dataset used for this research is solely based on Facebook data that is directly available. Mykhailo Granik et al proposed a simple technique based on the Naïve Bayesian classifier for fake news detection. The experiment uses BuzzFeed news dataset and performs the Naïve Bayes classification over the dataset. The technique used in this research achieved an accuracy up to 74% on the testing set. Marco L. Della et al in her paper propose a novel technique to recognize how social networks and gadget studying strategies can be utilized for fabricated news detection. A novel ML fake news detection method - Content-based (CB), Logistic regression (LR) on social signals and Harmonic Boolean label crowdsourcing (HC) on social signals has been used. This novel approach is carried out on a Facebook Messenger chatbot. The experiment achieves an accuracy of 81.7% for faux news detection [17 – 34]. In Shloka Gilda presented concept approximately how NLP can be helpful in detecting fake information. Time period frequency-inverse record frequency (TFIDF) of bi- grams and probabilistic context free grammar (PCFG) detection has been used. They have evaluated the data over multiple class algorithms to find out the best model. They conclude that TF-IDF of bi-grams ran on a Stochastic Gradient Descent model recognizes non-credible articles with an accuracy of 77.2%. Priya S. Gadekar used two different classifiers namely SVM classifier and Naive Bayes Classifier. With these classifiers, she achieved 60.97% accuracy with the SVM classifier and a 59.76% with the Naïve Bayes classifier. QIN Yumeng et al discuss a more advanced topic - to counteract misinformation and rumor detection in real time in their paper. It uses novelty-based feature. The dataset is obtained from Kaggle. The model achieves an accuracy of 74.5%. Clickbait and unreliable sources are not considered in these experiments which led to lower accuracy [23 – 42]. Arushi Gupta et al in their research paper aim to differentiate between spammers and non-spammers in Twitter. The various models used are clustering, Naïve Bayesian classifier, and decision tree. Accuracy rate to detect spammers are at 70% and non-spammers are at 71.2%. The models that were used attained a low average accuracy to segregate spammer and non-spammer. The problem addressed is very relevant in this information age, several previous works have been carried out from different perspectives, focused on different ways and using various techniques, but ultimately all seek to combat misinformation; some of these studies will be presented below. Traditional approaches based on verification by humans and expert journalists do not scale the volume of the news content that is generated online. Text classification is the fundamental task in Natural Language Processing (NLP) and researchers have addressed this problem quite extensively. Researchers proposed a model that can check the real-time credibility within 35 seconds after combining user-based, propagation-based, and content-based text. The basic idea of Naïve Bayes is that all features are independent of each other. Naïve Bayes needs a smaller data set and less storage space. Facebook post prediction through real or fake labeling can be done through naïve Bayes and it performs well. A proposed method can separate fake contents in three categories: serious fabrication, large scale hoaxes and humorous fake. It can also provide a way to filter, vet and verify the news. PHEME was a three-year research project funded by the European Commission from 2014-2017, studying NLP techniques for dealing rumor detection, stance detection, contradiction detection and analysis of social media rumors [27 – 48]. Fake news stories can be easily shared on social media platforms but it is difficult to identify fake content automatically. Using information sources (Visual cues & Cognitive cues) and social judgment (Cognitive, Behavioral & Affective), Facebook examines that machine learning classifiers can be helpful to detect fake news. We preferred Support Vector Machine for fake news detection as it is a more researched algorithm nowadays. It is difficult to say that it is the best classifier in fake news because the selection of classifiers totally depends on the organizational requirements. Stance detection of the headline for binary classification through n-gram matching can also be assessed after comparing "related" vs. "unrelated" pairs [1 – 36]. This approach can be applied in the detection of fake news, especially clickbait detection. They used a dataset released by the organization Fake News Challenge (FNC1) on stance detection for experiments. The dataset is publicly available and can be downloaded from the corresponding GitHub page along with baseline implementation. Deep learning using NLP for the detection of fake news and applied different models are presented, an assessment is made of which may be the option to obtain good results [31 – 54].

TABLE 4 - SEVEN TYPES OF FAKE NEWS

Title of Article	**Methodology**
Fake News Tracker: A Tool for Fake News Collection, Detection, and Visualization	• LSTM (2 layers of 100 cells each) • SVM • Logistic Regression • Naïve Bayes
Detection of Online fake News Using N-Gram Analysis and Machine Learning Techniques	• N-Gram • SVM • kNN • Logistic Regression
Automatic Detection of Fake News	• Linear SVM with 5-fold cross validation
Fake News Identification	• Naïve Bayes • SVM • Logistic Regression
Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News	• Linear SVM with 10-fold cross validation
Ensembles of Recurrent Networks for Classifying the Relationship of Fake News Titles	• BERT (Bidirectional Encoder Representations for Transformers) • BiLSTM
Comparative Performance of Machine Learning Algorithms for Fake News Detection	• Random Forest • Support Vector Classifier • Gaussian Naïve Bayes • AdaBoost • Multilayer Perceptron • Gradient Boosting
Fake News Detection with Semantic Features and Text Mining	• Bi-grams • Random Forest • Naïve Bayes • RNN
Automatic Detection of Fake News on Social Media Platforms	• Support Vector Machine • Logistic Regression • Decision Tree • Random Forest • XGBoost
Fake News Detection Using Naive Bayes Classifier	• Naive Bayes Classifier
Automatic Online Fake News Detection Combining Content and Social Signals	• Content-based (CB) • Logistic regression • Harmonic Boolean label crowdsourcing (HC)
Evaluating Machine Learning Algorithms for Fake News Detection	• TFIDF • Bigram • Stochastic Gradient Descent
Fake News Identification using Machine Learning	• Naïve Bayes • SVM
Predicting Future Rumours	• Liu • Yang
Improving Spam Detection in Online Social Networks	• Naïve Bayes Classifier • Clustering • Decision Tree

3.0 METHODOLOGY

The rest of this paper is organized as follows, Section 2.0 reviews the previous work, and Section 3.0 describes the Methodology. The Proposed model, Preprocessing and Machine learning are described in Sections 4.0, Section 5.0 describes the implementation task, Results and discussion are described in Section 6.0 and finally, the last section gives the Conclusion in Section 7.0. Our proposed model starts with the extraction phase and then we have four main steps. The first step is related to the NLP models where we measure the frequency of words and build the vocabulary of known words in fake news datasets. Next, fake news is detected using NB, SVM and PA classifiers. Finally, we test our models with several experiments and some other datasets and propose the final fake news detection model. Fig. 2 shows the flowchart of our model. The objective of this phase is to reduce the size of the data by removing irrelevant information that is not necessary for classification. Subsequently, for processing, the data were changed so that the first half of the data with the fake label set and the second half with a true label were not simply what would cause impartiality when applying the machine learning methods. One common task in NLP is tokenization that takes a text or set of texts and breaks it up into individual words [11 – 49]. We converted words to their base form for better understanding. Then we applied stemming that decreases the number of words on the bases of word type and class. Let us suppose we have three similar words in the dataset like running, ran and runner; it will be reduced and changed to the word, run. There are different stemming algorithms, but we used Porter due to its high accuracy rate. We used stop word removal as it removes common words used in articles, prepositions and conjunctions. Fake news is increasing every second without proper checks and balances, so there is a need for computational tools that can handle this problem [55 – 89]. Machine learning algorithms like “CountVectorizer”, “TFIDFVectorizer”, naïve Bayes, Support Vector Machine, Passive Aggressive Classifier and NLP for the identification of false news in public data sets are proposed. This is purely a text-based classification problem but our actual goal is the combination of text-based classification with machine-based text transformation and then choosing which type of text is to be used, e.g. single news or the full body of the news [90 – 124]. The overall data cleaning process is shown in Fig. 2.

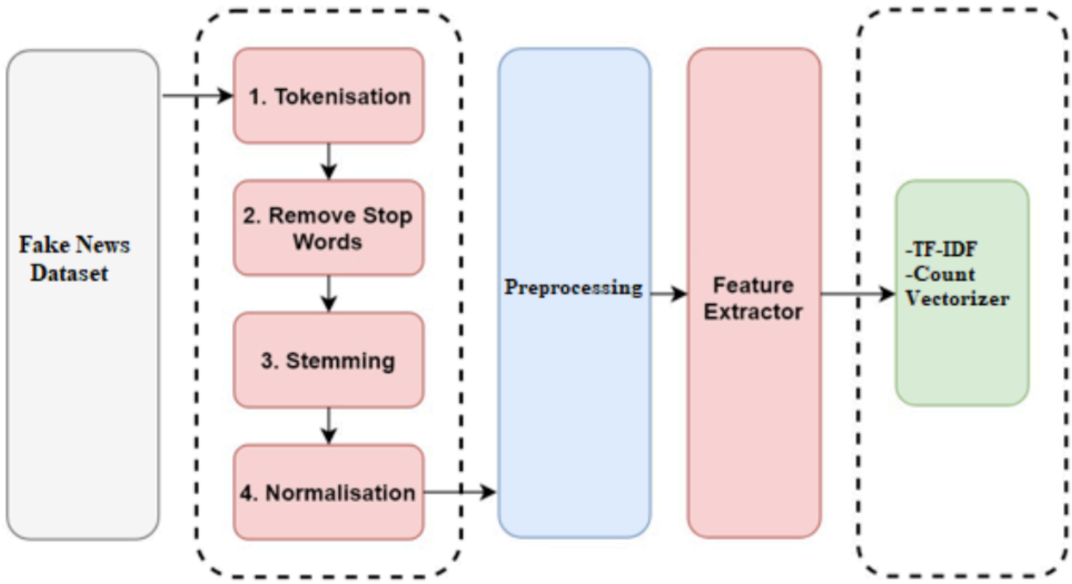


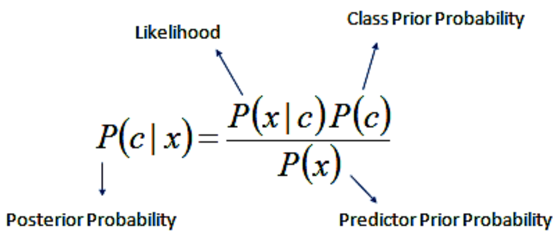
Fig. 2 Overall Flowchart of Our Model

In the study carried out, NLP is used as a Python computational tool, which uses different libraries and platforms. We applied PANDAS (Python Data Analysis Library) which is an open-source library with BSD license that provides data structures and data analysis tools during classification [125 – 149]. We applied NLTK in the extraction and characterization phase. Numpy and Scipy libraries are applied for programming but our main program is run on Jupyter Notebook. Keeping in mind the training and testing data, we further attached test data with tokenization algorithms. The main objective is to develop a model based on the count vectorization and TF-IDF [150 – 174]. Fake news detection is a binary classification task that the news is fake or not fake. Classification is not completely correct in fake news detection because classification methods are not specialized for fake news detection. So, keeping in mind that classification can separate fake text from non-fake, the goal is to develop a model that is specialized for fake news detection [175 – 192]. To develop a classification method that is specialized for fake news detection we need to identify relevant features before classification. We applied different features to extract optimal features in the text that help us for better text classification [193 – 201].

4.0 MODEL DEVELOPMENT

Different classification models can be applied in this case, but to choose the most adequate one and to tune its parameters we run several experiments on different models. We started experimenting with classification models that have proven to be effective and give good results in related sentence classification tasks. Some of the models did not give good results and were discarded, one of them was Logistics Regression, while Support Vector Machines, naïve Bayes and Passive Aggressive gave promising results and we continued to experiment on them [1 – 23]. To check the accuracy, we compare our results with other datasets through performance metrics.

a) Naïve Bayes: It is a powerful classification model that performs well when we have a small dataset and it requires less storage space. It does not produce good results if words are co related between each other [17 – 36]. Fig. 3 contains the Naïve Bayes formula that explains the probability of an attribute that belongs to a class independent from other classes.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig. 3 NaïVe Bayes Formula

b) Support Vector Machine: It performs supervised learning on data for regression and classification. The SVM computes the data and converts it into different categories. The advantages of Support Vector Machine are learning speed, accuracy, classification and tolerance to irrelevant features. Support Vector Machine is one of the most researched classifiers nowadays and it performs well in the fake news detection problem [24 – 49].

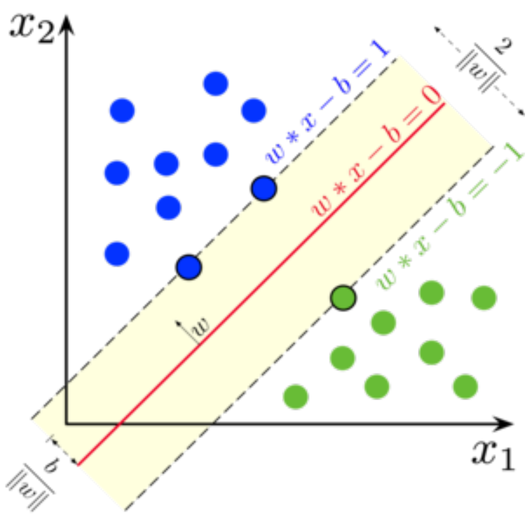


Fig. 4 Support Vector Machine

c) Passive Aggressive: These algorithms are mainly used for classification. The idea is very simple and the performance has been proven with many other alternative methods like Online Perceptron and MIRA [47 – 58].

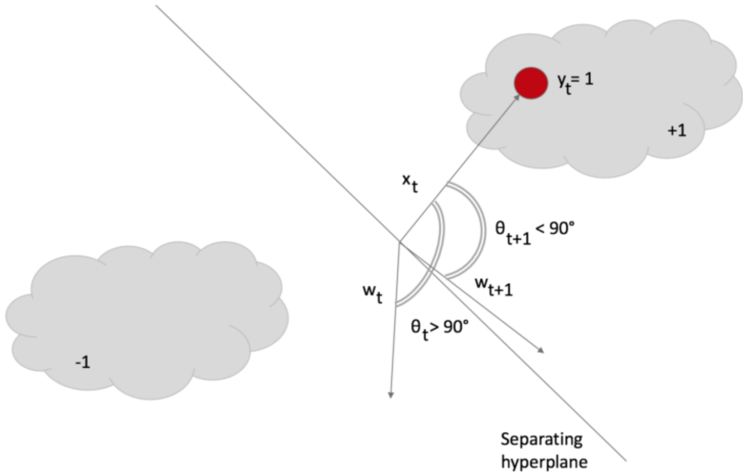


Fig. 5 Passive Aggressive

d) Logistic Regression: It is used to estimate the relationship between variables after using statistical methods. It performs well in binary classification problems because it deals with classes and requires a large sample size for initial classification [59 – 73].

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Fig. 6 Logistic Regression

e) NLP Models: Irrelevant and redundant features in a dataset have a negative impact on the accuracy and performance of the classifier. So, in those cases, we perform feature reduction to reduce the text feature size that limited the words like "the", "and", "there", "when" and focus only on those words which appear a given number of times. This is done by using n- number of use words, lower casing and stop word removal since the sensitivity of the problem, which is increasing every second without check and balance, is understood. It is essential to use machine learning algorithms like CountVectorizer and TF-IDF to speed up the task and improve performance [68 – 92].

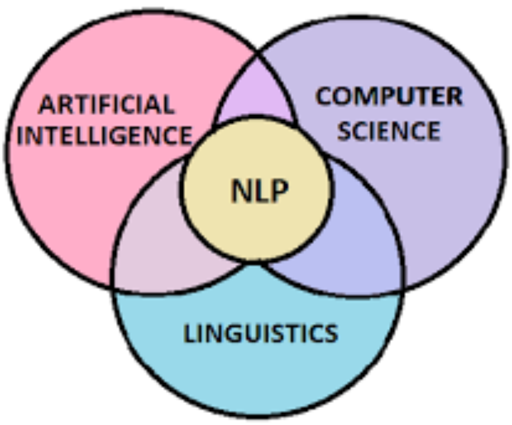


Fig. 7 NLP Models

f) Count Vectorization: It provides a simple way to collect text documents and to help build the vocabulary of known distinctive words and also to encode new documents using that vocabulary. Given a collection of text documents, S to Count Vectorizer and it will generate a sparse matrix of size A where m = total number of documents, n = total number of distinct words used in S. With the Count Vectorizer, we can produce a table for each word and occurrence of each class [93 – 128].

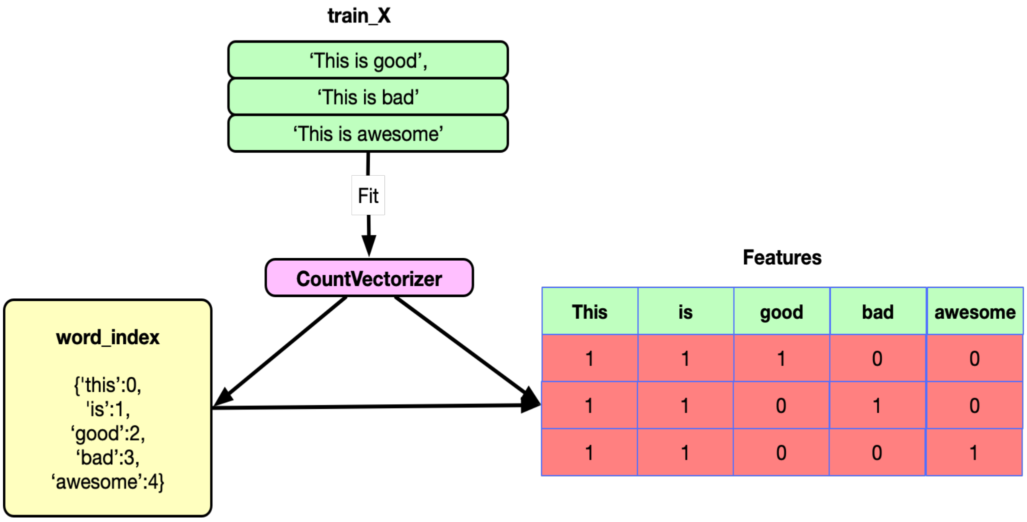


Fig. 8 Count Vectorization

g) Term Frequency - Inverse Document Frequency: To measure a term in documents over a dataset, we used the term frequency-inverted document frequency. A term's importance increases in the document which appears in the dataset and also the frequency of the words. So, with the help of this method, we can weigh the metric that is used for information retrieval [129 – 141]. TF-IDF for the word with respect to document d and corpus D is calculated as follows:

$$TF(i) = \frac{log_2(Freq(i,j) + 1}{log_2(L)}$$

Fig. 9 Term Frequency - Inverse Document Frequency

h) Decision Tree Classifier: The main objective of Decision Tree classifier is to optimally partition a space of possible observations by subsequent recursive splits. Decision Trees mimic human thinking unlike the other algorithms like SVM and neural network which are essentially, black box algorithms. We have used the CART model since we are dealing with a binary classification problem. The CART model uses Gini index as cost function to evaluate the split in partitioning the features. Gini index is a measure of inequality in the data sample. It is essentially the sum of squares of the probabilities of each class and is calculated and to find which attribute classifies the dataset in the best manner, we have to calculate the information gain of each attribute, for which we first calculate the entropy and The attribute with the smallest entropy value is used to split the set on the respective iterations [142 – 178]. Information Gain is the change in entropy when a set is split on attribute. The attribute with the highest Information Gain value is used to split the set on that particular iteration. as follows:

$$Gini\ Index = 1 - \sum_{i=1}^{n} p_i^2$$

$$Entropy\ H(S) = \sum_{c \in C} -p(c) \log_2 p(c)$$

Fig. 10 Decision Tree Classifier

i) Logistic Regression is a commonly used classification algorithm and it is used to label an observation to a discrete set of class. Since the problem in hand is a binary classification problem, Logistic regression has been used, and successfully indeed. Logistic Regression function is basically a sigmoid function and assigns a probability value which, is then assigned to a class in a discrete set of two or more classes. In regression analysis, logistic regression (or logic regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). In statistics, the logistic model is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression is estimating the parameters of a logistic model. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1, [179 – 189].

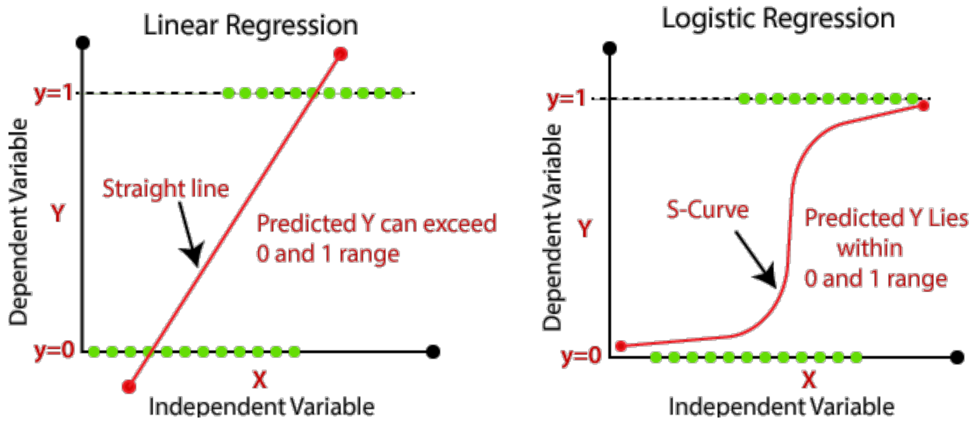


Fig. 11 Logistic Regression & Linear Regression

j) Bagging Classifier: Bootstrap aggregating classifier, commonly known as Bagging classifier is as an ensemble meta-estimator. It fits the base algorithm and create subsets of the sample data, and aggregates the individual prediction through techniques like voting and averaging to output a final prediction. This estimator is commonly used to reduce variance when a black-box algorithm such as decision tree is used which has a tendency to produce high variance in the predictions [190 – 201].

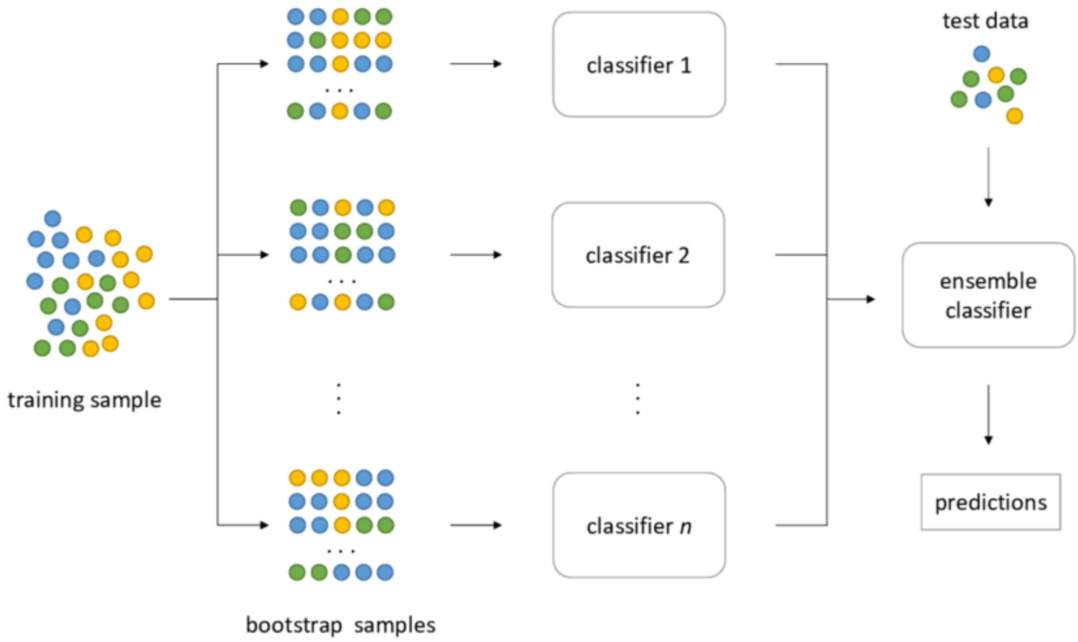


Fig. 12 Bagging Classifier

5.0 IMPLEMENTATION

Fake news detection topic has gained a great deal of interest from researchers around the world [1 – 17]. When some event has occurred, many people discuss it on the web through the social networking [18 – 27]. They search or retrieve and discuss the news events as the routine of daily life [28 – 37]. Some type of news such as various bad events from natural phenomenal or climate are unpredictable. When the unexpected events happen, there are also fake news that are broadcasted that creates confusion due to the nature of the events [38 – 49]. Very few people know the real fact of the event while the most people believe the forwarded news from their credible friends or relatives [50 – 64]. These are difficult to detect whether to believe or not when they receive the news information [65 – 78]. So, there is a need of an automated system to analyze truthfulness of the news. In the study carried out, NLP is used as a Python computational tool, which uses different libraries and platforms [79 – 96]. We applied PANDAS (Python Data Analysis Library) which is an open-source library with BSD license that provides data structures and data analysis tools during classification [97 – 116]. We applied NLTK in the extraction and characterization phase [117 – 124]. Numpy and Scipy libraries are applied for programming but our main program is run on Jupyter Notebook [125 – 137]. Keeping in mind the training and testing data, we further attached test data with tokenization algorithms [138 – 142]. The main objective is to develop a model based on the count vectorization and TF-IDF [143 – 157]. Fake news detection is a binary classification task that the news is fake or not fake. Classification is not completely correct in fake news detection because classification methods are not specialized for fake news detection [158 – 169]. So, keeping in mind that classification can separate fake text from non-fake, the goal is to develop a model that is specialized for fake news detection [170 – 184]. To develop a classification method that is specialized for fake news detection we need to identify relevant features before classification [185 – 194]. We applied different features to extract optimal features in the text that help us for better text classification [195 – 201]. Different classification models can be applied in this case, but to choose the most adequate one and to tune its parameters we run several experiments on different models [1 – 17]. We started experimenting with classification models that have proven to be effective and give good results in related sentence classification tasks [18 – 27]. Some of the models did not give good results and were discarded, one of them was Logistics Regression, while Support Vector Machines, naïve Bayes and Passive Aggressive gave promising results and we continued to experiment on them [1 – 44]. To check the accuracy, we compare our results with other datasets through performance metrics [16 – 56]. Fake news is increasing every second without proper checks and balances, so there is a need for computational tools that can handle this problem. Machine learning algorithms like "CountVectorizer", "TFIDFVectorizer", naïve Bayes, Support Vector Machine, Passive Aggressive Classifier and NLP for the identification of false news in public data sets are proposed [1 – 36]. This is purely a text-based classification problem but our actual goal is the combination of text-based classification with machine-based text transformation and then choosing which type of text is to be used, e.g. single news or the full body of the news [37 – 75]. The overall data cleaning process is shown in Fig. 13.

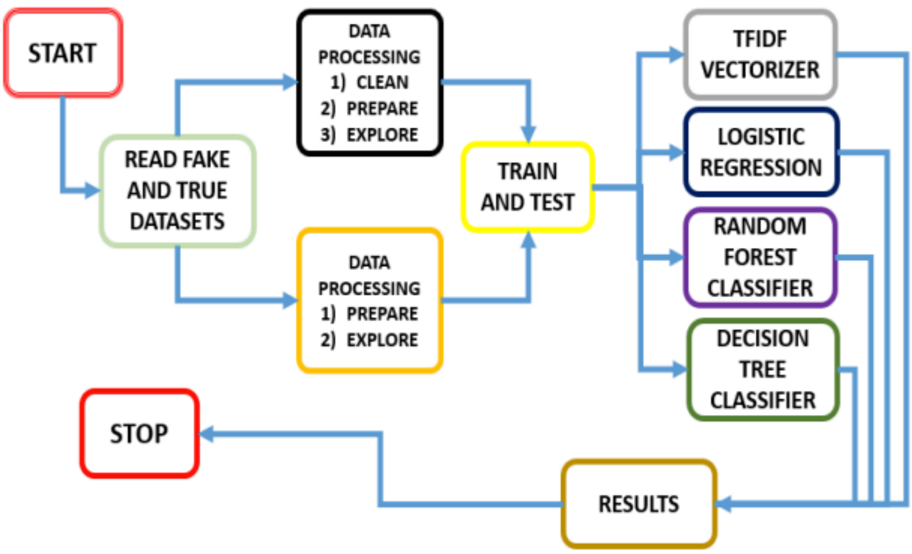


Fig. 13 Overall Data Cleaning Process

6.0 RESULTS & DISCUSSION

The dataset used for classification was drawn from a public domain. Fake news articles were collected from an open source Kaggle dataset that was published during the 2016 election cycle. The collection is made up of 18000 news articles highlighted in Fig. 14. These articles were collected from news organizations NYT, Guardian, and Bloomberg during the election period. Articles are separated through binary labels 0 and 1. The dataset is already sorted qualitatively with fake, non-fake and not clear labels. This division can be seen in Fig. 14 where we have 15,115 articles from the false category and 1,846 from the true category. The remaining articles are classified as not clear due to some other reasons like unique ID missing, source not clear etc. [1 – 26]. The task itself leads to a quite imbalanced dataset, as shown in Fig. 15, wherefrom the total articles, roughly 12% are in the true category. This imbalance is typical in this task, and also seen in previous similar works. The second dataset contains 5000 real news articles collected from the Signal Media News dataset, in which 2,541 belong to the false class and 299 to the true class, as shown in Fig. 15. We skipped the unclear class due to the missing values. Articles were collected from major news media organizations e.g. the Guardian, Bloomberg, New York Times, NPR, etc. The dataset was published in 2016 before and after the United States presidential elections [27 – 48].

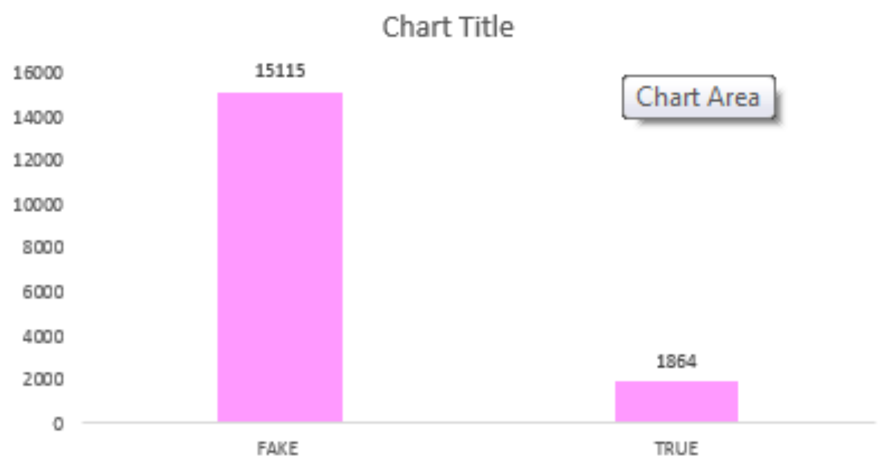


Fig. 14 Class Distribution Kaggle Dataset

For the purpose of this project, the dataset was acquired from Kaggle. The dataset itself is known as "Fake News" dataset. The training set consists of 20800 rows built through various articles obtained from internet and other news sources [49 – 67]. A lot of preprocessing has been done in order to train the data for our models, which we will discuss in the next section. In addition, the dataset contains around 5200 rows for testing purposes [68 – 91].

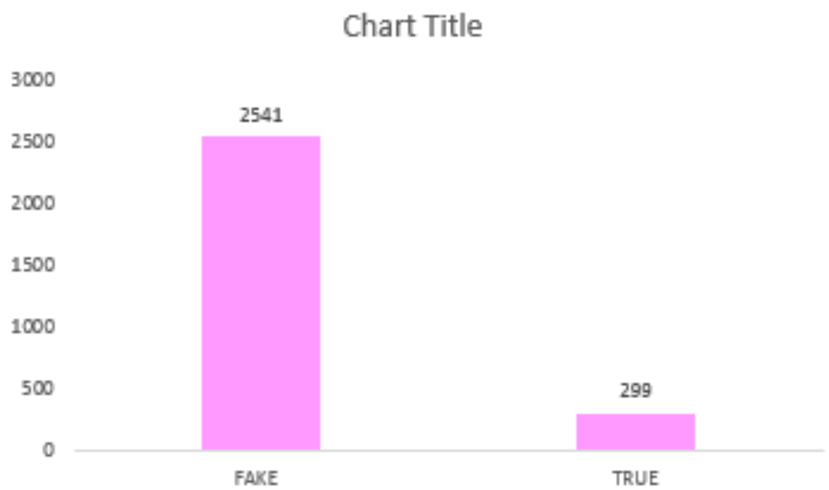
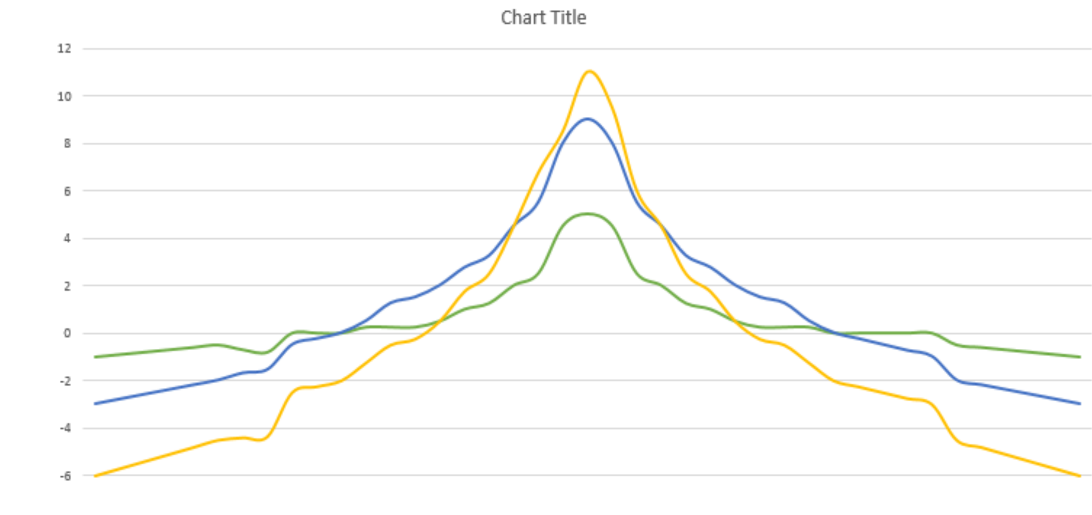


Fig. 15 Signal Media News Dataset

Initially, preprocessing tasks such as data cleaning, removing English stopwords, punctuations and special characters is done on the dataset. Stopwords are words which do not necessarily lend any additional semantical meaning to a sentence and are considered to be useless for any natural language processing tasks and hence, removed from the textual dataset before processing. However, in the absence of stopwords, a sentence may not grammatically make sense for humans. Later, a comma-separated lists of words is produced from the cleaned data. These lists are further fed into the doc2vec algorithm in order to produce 300 length embedding vector for every article in the dataset. Doc2Vec is an extension of the pre-existing word2vec algorithm. It came exactly a year after word2vec in 2014. The doc2vec algorithm aims at creating a numeric representation of documents which is an analogous concept to word2vec. This numeric representation is independent of its length. However, documents do not come in logical structures such as words, so the authors of word2vec model Mikilov and Le came up with a simple yet clever solution and added another vector. This new 'document vector' contains in itself information about the document as a whole. It can contain unique paragraph ids which will help track the context of the information paragraph-wise and other features based on the application in hand. The doc2vec is called as an extended version of word2vec because like word2vec, it also allows the model to learn about the word order. The fact that the word order remains preserved in the doc2vec model as well as the whole document information is learnt makes it very useful for the project's purpose.We used RapidMiner, a powerful machine learning tool for data exploration, preparation, information extraction, result visualization and result optimization. We analyzed the fake and true sentences through RapidMiner and initial results can be seen in Fig. 16.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Series3	-3	-3	-3	-3	-3	-3	-3	-3	-2	-2	-2	-2	-2	-2	-2	-1	-1	0	1.3	0.5
Series2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-0	0	0.3	1	1.3	1.5	1.8	2	2.5	3	3.5
Series1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0.3	0.3	0.3	0.5	1	1.3	2	2.5	4.5

	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
Series3	2	1.5	0.5	0	-1	-1	-2	-2	-2	-2	-2	-2	-2	-2	-2	-3	-3	-3	-3	-3	-3
Series2	4	3.5	3	2.5	2	1.8	1.5	1.3	1	0.3	0	-0	-1	-1	-1	-2	-2	-2	-2	-2	-2
Series1	5	4.5	2.5	2	1.3	1	0.5	0.3	0.3	0.3	0	0	0	0	0	-1	-1	-1	-1	-1	-1

Fig. 16 Dataset class labeling chart

We conducted several experiments with different feature set combinations as discussed in Section three C and the model selection in Section three B. Our proposed combination works well and obtains performance above the baseline 0.50. The best performing classifier is PA when we check the performance through accuracy and precision. However, somehow in the recall it reduced. Table 5 shows the performance of our proposed classifiers.

TABLE 5 - RESULTS

Classifier	Accuracy	Precision	Recall
Naïve Bayes	0.85%	0.89%	0.87%
Passive Aggressive	0.93%	0.92%	0.89%
Support Vector Machine	0.84%	0.82%	0.87%

Confusion matrix tells the overall performance of the model on the testing dataset when true values are known. It provides us with the summary of the performance of the model and provides us with valuable insights like true positive, true negative, false positive and false negative results of our classifier model. Accuracy Score, referred to classification accuracy rate is defined as the ratio of number of correct predictions and total number of predictions that the model has made. In other words, the number of true positives and true negatives when divided over the total number of predictions gives us the accuracy score. Precision is defined as the fraction of total number of correct positive outcomes out of number of positive outcomes predicted by the classification model. Recall is defined as the total number of correct positive outcomes over total relevant results as predicted by the model. F1-score describes about the preciseness and the robustness of your model. It is mathematically, defined as the harmonic mean of precision and recall. F1 score is directly proportional to the performance of the model [17 – 42]. The precision recall curve is the plot between two basic evaluation parameters – precision and recall. The receiver operator characteristic curve, generally known as the ROC Curve, is a graph representing the trade- off between specificity and sensitivity of a model. Specificity measures the entire negative part of the results while sensitivity deals with the positive spectrum of the results obtained by the model. We compare our results with the same model but different datasets and different features, as highlighted in Table III. It is observed that the proposed models perform well and achieved the highest accuracy up to 93% with Passive Aggressive, 85% with naïve Bayes and 84% with SVM. Ott et al. applied SVM with features LIWC+ Bigrams and achieved an accuracy level of up to 89%. Similarly, when they changed the Stylometric features, it achieved 84% accuracy. On the other side, Horrne and Adali achieved 71% accuracy when they applied text-based features [43 – 76]. The results show that the proposed combination improves the existing performance in some categories. For further analysis, we applied different combinations to check the accuracy of the proposed model with other models. Accuracy comparison of Passive Aggressive and Support Vector Machine (a), Passive Aggressive and Logistic Regression (b), Passive Aggressive and Support Vector Machine (c) with a different dataset, Passive Aggressive and Naïve Bayes (d), Support Vector Machine and Naïve Bayes (e), Naïve Bayes and Support Vector Machine (f), Support Vector Machine and Logistic Regression (g) and Support Vector Machine and Naïve Bayes (h) can be seen in Fig. 17. Through our experiment, we find that a Hard Voting Ensemble model of Decision Tree Classifier and Logistic Regression performs the best with over 88% accuracy. Decision Trees tend to be more preferred while making any ensemble model because while Decision Trees are simple yet powerful, they tend to exhibit high variance and low variance. Since the problem in hand is essentially a binary classification problem, logistic regression proved to be a good algorithm to aggregate the decision tree model [1 – 42]. However, it still gave an accuracy of 0.785. The low accuracy was of course, due to the nature of decision tree. So we ensemble a Bagging Classifier which is Bootstrap Aggregating technique which is known to be very good reducing variance at the cost of more computation and little bias. The accuracy improved to 0.88 with this ensemble model. Voting ensemble technique was used on this aggregate model to make the final prediction. Hard voting gave us better results compared to soft voting which is obvious since in soft voting, prediction resutlts are averaged out from the models in the ensemble whereas in hard voting, model is selected from the ensemble to make final predictions by majority vote.. We further investigated and compared our results with when they applied a combination of CFG and n-gram accuracy in deception detection where they achieved 85%-91% accuracy. Still, our presented results are better in the context of fake news detection and our proposed classifiers achieved maximum accuracy level. Figs. 17 (a)-(h) show the results of the classification for the PA, SVM and NB classifiers. The values are the maximum accuracy level achieved by the classifier after combining it with others [77 – 95]. For further understanding of the results, we changed the classifier and fake news dataset proposed by others. These experiments highlighted some important features that we still want to investigate further.

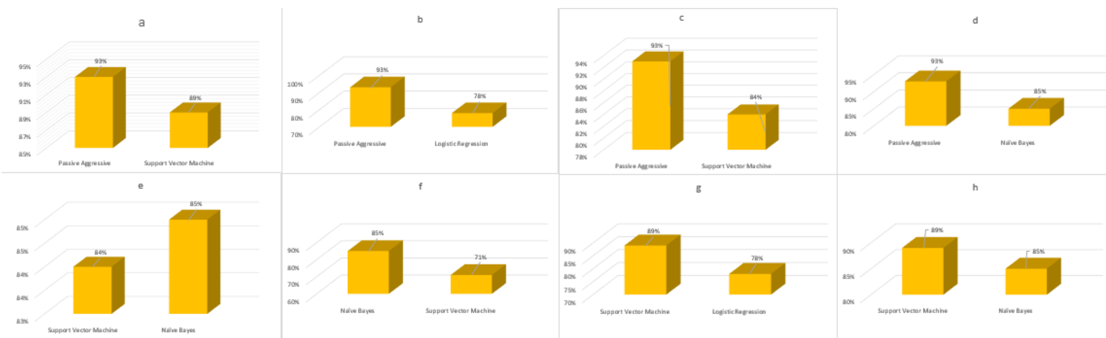


Fig. 17 Word Cloud of News Articles

7.0 CONCLUSION

Natural Language Processing is widely considered to be the area of research and application, as compared to other information technology approaches. There have been adequate successes that propose NLP-based technologies will continue to be a great area of research and development in information systems now and in the future. Also Machine Learning is a significant application in NLP that can never be ignored. ML is truly a very important and elaborate, however a necessary task in the NLP development applications. While NLP is a relatively recent area of research and application, as compared to other information technology approaches, there have been sufficient successes to date that suggest that NLP-based information access technologies will continue to be a major area of research and development in information systems now and far into the future. The state-of- the-art Natural Language Processing techniques applied to speech technologies, specifically to Text-To-Speech synthesis and Automatic Speech Recognition. In 3TTS. The importance of NLP in processing the input text to be synthesized is reflected. The naturalness of the speech utterances produced by the signal-processing modules are tightly bound to the performance of the previous text-processing modules. In ASR the use of NLP particularly is complementary. It simplifies the recognition task by assuming that the input speech utterances must be produced according to a predefined set of grammatical rules. Its capabilities can though be enhanced through the usage of NLP aiming at more natural interfaces with a certain degree of knowledge. Reviews the major approaches proposed in language model adaptation in order to profit from this specific knowledge. The number of people consuming news from social media, internet, micro-blogging website, blogs etc., instead of traditional news media, are increased exponentially. In the recent past, the role of social media in spreading fake news and its negative impacts on our society, from personal level to a global level, have been well documented. One of the ways to tackle the challenge presented by the rising menace of fabricated news is through the applications of Machine Learning and Natural Language Processing techniques as described in this paper. In future, we aim to incorporate a lot more features such as the medium of publication, URL if any, topic and additional linguistic features which are not part of this paper. We would also like to exploit deep learning algorithms to create ensemble models which are even more accurate. The results suggested that the approach is highly favorable since this application helps in classifying fake news and identifying key features that can be used for fake news detection. Our proposed technique suggests that to differentiate fake and non-fake news articles, it is worthwhile to look at machine learning methods. The developed system with accuracy up to 93% proves the importance of the combination; next, we need to look into other methods for fake news detection except for simple text classification. The producers of fake news are using different techniques to hide their identity, so they can easily mislead readers. As we are aware that every single news has different characteristics so there is a need for a system that can check the content of the news in depth. Our future work includes building an automated fact-checking system that combines data and knowledge to help non-experts and checks the content of the news thoroughly after comparing it with known facts. We want to look into the issue of fake news from different angles like known facts, source, topic, associated URLs, geographical location, year of publication, and credibility of the source for a better understanding of the problem. The open issues and challenges are also presented in this paper with potential research tasks that can facilitate further development in fake news research.

REFRENCES

[1] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

[2] Arega, Kedir Lemma, and S. KotherMohideen. "Grouping and Detection of Fake News via web-based media Using Machine Learning in Amharic Language." CENTRAL ASIAN JOURNAL OF THEORETICAL & APPLIED SCIENCES 3.5 (2022): 89-110.

[3] Arushi Gupta, Rishabh Kaushal, "Improving Spam Detection in Online Social Networks", 978-1-4799-7171-8/15/$31.00 ©2015 IEEE.

[4] Mladenova, Tsvetelina, and Irena Valova. "Research on the Ability to Detect Fake News in Users of Social Networks." 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2022.

[5] Arvinder Pal Singh Bali, Mexson Fernandes, Sourabh Choubey, Mahime Goel, "Comparative Performance of Machine Learning Algorithms for Fake News Detection", Part of the Communications in Computer and Information Science book series (CCIS, volume 1046) (2022).

[6] Kumar, Abhinav, Jyoti Prakash Singh, and Amit Kumar Singh. "COVID-19 Fake News Detection Using Ensemble-Based Deep Learning Model." IT Professional 24.2 (2022): 32-37.

[7] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," Lecture Notes in Computer Science Intelligent, Secure, and Dependable Systems in Distributed and Cloud

Environments, pp. 127–138, Oct. 2017

[8] Khalil, Ashwaq, et al. "AFND: Arabic fake news dataset for the detection and classification of articles credibility." Data in Brief 42 (2022): 108141.

[9] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in Proceedings of the 50thAnnual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 171–175.

[10] Varlamis, Iraklis, et al. "A Survey on the Use of Graph Convolutional Networks for Combating Fake News." Future Internet 14.3 (2022): 70.

[11] Tandoc, E. C., & Lim, Z. W., & Ling, R. (2018). Defining "fake news." Digital Journalism, 6, 137-153.

[12] Piya, Fahmida Liza, Rezaul Karim, and Mohammad Shamsul Arefin. "BDFN: A Bilingual Model to Detect Online Fake News Using Machine Learning Technique." Soft Computing for Security Applications. Springer, Singapore, 2022. 799-816.

[13] Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018). Fake News Detection in Social Networks via Crowd Signals, 517–524. https://doi.org/10.1145/3184558.3188722

[14] Bhartiya, Priyanka, et al. "Fake News Predictor Model-Based on Machine Learning and Natural Language Processing." Handbook of Research on Machine Learning. Apple Academic Press 479-501, (2022).

[15] Dubey, Yogita, et al. "Framework for Fake News Classification Using Vectorization and Machine Learning." Combating Fake News with Computational Intelligence Techniques. Springer, Cham, 2022. 327-343.

[16] Lorent, S. (2019). Master thesis: Fake news detection using machine learning.

[17] Janze, C., Risius, M., "Automatic detection of fake news on social media platforms." Pacific Asia Conference on Information Systems (2017)

[18] Agarwal, Swati, and Adithya Samavedhi. "Profiling fake news: Learning the semantics and characterisation of misinformation." International Conference on Advanced Data Mining and Applications. Springer, Cham, 2022.

[19] Kozik, Rafał, et al. "Technical solution to counter potential crime: Text analysis to detect fake news and disinformation." Journal of Computational Science 60 (2022): 101576.

[20] Kai Shu, Deepak Mahudeswaran, Huan Liu, "FakeNewsTracker: a tool for fake news collection, detection, and visualization", Computational & Mathematical Organization Theory, vol. 25 issue 1, pp. 60-71, (2022).

[21] De Magistris, Giorgio, et al. "An explainable fake news detector based on named entity recognition and stance classification applied to covid-19." Information 13.3 (2022): 137.

[22] Sedik, Ahmed, et al. "Deep fake news detection system based on concatenated and recurrent modalities." Expert Systems with Applications 208 (2022): 117953.

[23] Marco L. Della Vedova, Eugenio Tacchini, Stefano Moret, Gabriele Ballarin, Massimo DiPierro, Luca de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals", ISSN 2305- 7254,2017, (2022).

[24] Mykhailo Granik, Volodymyr Mesyura, "Fake News Detection Using Naive Bayes Classifier", 2017 IEEEFirst Ukraine Conference on Electrical and Computer Engineering (UKRCON)

[25] Megıas, David, et al. "Architecture of a fake news detection system combining digital watermarking, signal processing, and machine learning."

[26] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, Nov. 2015.

[27] Ahmad, Tahir, et al. "Efficient Fake News Detection Mechanism Using Enhanced Deep Learning Model." Applied Sciences 12.3 (2022): 1743.

[28] Pranav Bhardwaj, Zongru Shao, "Fake News Detection with Semantic Features and Text Mining", International Journal on Natural Language Computing (IJNLC) Vol.8, No.3, June 2019

[29] Chen, Mu-Yen, Yi-Wei Lai, and Jiunn-Woei Lian. "Using Deep Learning Models to Detect Fake News About COVID-19." ACM Transactions on Internet Technology (2022).

[30] Jain, A., & Kasbe, A. (2018, February). Fake news detection. In 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-5). IEEE.

[31] Al-Asadi, Mustafa A., and Sakir Tasdemir. "Using artificial intelligence against the phenomenon of fake news: a systematic literature review." Combating Fake News with Computational Intelligence Techniques (2022): 39-54.

[32] Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection (pp. 7-17).

[33] Patil, Dharmaraj R. "Fake News Detection Using Majority Voting Technique." arXiv preprint arXiv:2203.09936 (2022).

[34] Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015a. Classifying tweet level judgements of rumours in social media. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP‟15). 2590–2595.

[35] Bozuyla, Mehmet, and AKIN ÖZÇİFT. "Developing a fake news identification model with advanced deep languagetransformers for Turkish COVID-19 misinformation data." Turkish Journal of Electrical Engineering and Computer Sciences 30.3 (2022): 908-926.

[36] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. arXiv preprint arXiv:1704.00656, 2017

[37] Kumar, Sudhanshu, and Thoudam Doren Singh. "Fake News Detection on Hindi News Dataset." Global Transitions Proceedings (2022).

[38] Zhou, X., Cao, J., Jin, Z., Xie, F., Su, Y., Chu, D. ... & Zhang, J. (2015, May). Real-Time News Cer tification System on Sina Weibo. In Proceedings of the 24th International Conference on World Wide Web (pp. 983-988). ACM.

[39] Nassif, Ali Bou, et al. "Arabic fake news detection based on deep contextualized embedding models." Neural Computing and Applications (2022): 1-14.

[40] Rubin, V. L., Chen, Y., & Conroy, N. J. (2015, November). Deception detection for news: three types of fakes. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (p. 83). American Society for Information Science.

[41] Mohtaj, Salar, and Sebastian Möller. "The Impact of Pre-processing on the Performance of Automated Fake News Detection." International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, Cham, 2022.

[42] Ruchansky, N., Seo, S., & Liu, Y. (2017, November). Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 797- 806). ACM.

[43] Sciucca, Laura Della, et al. "FakeNED: A Deep Learning Based-System for Fake News Detection from Social Media." International Conference on Image Analysis and Processing. Springer, Cham, 2022.

[44] Janze, C., & Risius, M. (2017). Automatic Detection of Fake News on Social Media Platforms.

[45] Nassif, Ali Bou, et al. "Arabic fake news detection based on deep contextualized embedding models." Neural Computing and Applications (2022): 1-14.

[46] Hiramath,C.K.,&Deshpande,G.C.(2019,July).FakeNewsDetection Using Deep Learning Techniques. In 2019 1st International Conference on Advances in Information Technology (ICAIT) (pp. 411-415). IEEE.

[47] Sciucca, Laura Della, et al. "FakeNED: A Deep Learning Based-System for Fake News Detection from Social Media." International Conference on Image Analysis and Processing. Springer, Cham, 2022.

[48] Bourgonje, P., Schneider, J. M., & Rehm, G. (2017). From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism (pp. 84-89).

[49] Mondal, Subrota Kumar, et al. "Fake News Detection Exploiting TF-IDF Vectorization with Ensemble Learning Models." Advances in Distributed Computing and Machine Learning. Springer, Singapore, 2022. 261-270.

[50] Priya S. Gadekar, "Fake News Identification using Machine Learning", Volume 7 Issue V, May 2019, International Journal for Research in Applied Science & Engineering Technology (IJRASET)

[51] Dong, Xishuang, and Lijun Qian. "Integrating Human-in-the-loop into Swarm Learning for Decentralized Fake News Detection." arXiv preprint arXiv:2201.02048 (2022).

[52] Shloka Gilda,"Evaluating Machine Learning Algorithms for Fake News Detection" ,2017 IEEE 15th Student Conference on Research and Development (SCOReD).

[53] Vardhan, K. Vishnu, B. Manjula Josephine, and KVSN Rama Rao. "Fake News Detection in Social Media Using Supervised Learning Techniques." 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE, 2022.

[54] Sohan Mone, Devyani Choudhary, Ayush Singhania, "FAKE NEWS IDENTIFICATION", CS229: Machine Learning: Group 21, Stanford University, (2022).

[55] Ting Su, Craig McDonald, Iadh Ounis, "Ensembles of Recurrent Networks for Classifying the Relationship of Fake News Titles", Proceedings of the 42nd International ACM SIGIR Conference, pp. 893-896, (2022).

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

[57] Kulkarni, Chaitanya, et al. "COVID-19 Fake News Detection Using GloVe and Bi-LSTM." Proceedings of Second International Conference on Sustainable Expert Systems. Springer, Singapore, 2022.

[58] Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, June). Finding deceptive opinion spam by any stretch of the imagination. InProceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 (pp. 309-319). Association for Computational Linguistics.

[59] Das, Sourya Dipta, Ayan Basak, and Saikat Dutta. "A heuristic-driven uncertainty based ensemble framework for fake news detection in tweets and news articles." Neurocomputing 491 (2022): 607-620.

[60] Pratiwi, I. Y. R., Asmara, R. A., & Rahutomo, F. (2017, October). Study of hoax news detection using naïve Bayes classifier in Indonesian language. In2017 11th International Conference on Information & Communication Technology and System (ICTS) (pp. 73-78). IEEE.

[61] Del Ser, Javier, et al. "Efficient Fake News Detection using Bagging Ensembles of Bidirectional Echo State Networks." 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022.

[62] Davuth, N., & Kim, S. R. (2013). Classification of malicious domain names using support vector machine and bi-gram method. International Journal of Security and Its Applications7(51-58)

[63] Madani, Mirmorsal, Homayun Motameni, and Hosein Mohamadi. "Fake news detection using deep learning integrating feature extraction, natural language processing, and statistical descriptors." Security and Privacy: e264.

[64] Banerjee, S., Chua, A. Y., & Kim, J. J. (2015, January). Using supervised learning to classify authentic and fake online reviews. InProceedings of the 9th International Conference on Ubiquitous Information Management and Communication (pp. 1-7).

[65] da Silva, Fernando Cardoso Durier, Ana Cristina Bicharra Garcia, and Sean Wolfgand Matsui Siqueira. "A Systematic Literature Mapping on Profile Trustworthiness in Fake News Spread." 2022 IEEE 25th International Conference on

Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2022.

[66] Torunoğlu, D., Çakirman, E., Ganiz, M. C., Akyokuş, S., & Gürbüz, M. Z. (2011, June). Analysis of preprocessing methods on classification of Turkish texts. In2011 International Symposium on Innovations in Intelligent Systems and Applications (pp. 112-117). IEEE.

[67] Jain, Mayank Kumar, et al. "Review on Analysis of Classifiers for Fake News Detection." International Conference on Emerging Technologies in Computer Engineering. Springer, Cham, 2022.

[68] Raulji, J. K., & Saini, J. R. (2016). Stop-word removal algorithm and its implementation for Sanskrit language. International Journal of Computer Applications, 150(2), 15-17.

[69] Epstein, Ziv, et al. "Do explanations increase the effectiveness of AI-crowd generated fake news warnings?." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 16. 2022.

[70] Vijayaraghavan, S., Wang, Y., Guo, Z., Voong, J., Xu, W., Nasseri, A. ... & Wadhwa, E. (2018). Fake News Detection with Different Models. arXiv preprint arXiv:2003.04978.

[71] Shushkevich, Elena, Mikhail Alexandrov, and John Cardiff. "BERT-based Classifiers for Fake News Detection on Short and Long Texts with Noisy Data: A Comparative Analysis." International Conference on Text, Speech, and Dialogue. Springer, Cham, 2022.

[72] Gilda, S. (2017, December). Evaluating machine learning algorithms for fake news detection. In2017 IEEE 15th Student Conference on Research and Development (SCOReD) (pp. 110-115). IEEE.

[73] Michail, Dimitrios, Nikos Kanakaris, and Iraklis Varlamis. "Detection of fake news campaigns using graph convolutional networks." International Journal of Information Management Data Insights 2.2 (2022): 100104.

[74] Nørregaard, J., Horne, B. D., & Adalı, S. (2019, July). NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 13, No. 01, pp. 630-638).

[75] Shahid, Wajiha, et al. "Are You a Cyborg, Bot or Human?—A Survey on Detecting Fake News Spreaders." IEEE Access 10 (2022): 27069-27083.

[76]] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F. ... Zittrain, J. L. (2018a). The science of fake news. Science, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

[77] Koloski, Boshko, et al. "Knowledge graph informed fake news classification via heterogeneous representation ensembles." Neurocomputing (2022).

[78] Kostakos, P., Nykanen, M., Martinviita, M., Pandya, A., & Oussalah, M. (2018, August). Meta-terrorism: identifying linguistic patterns in public discourse after an attack. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1079-1083). IEEE.

[79] Verma, Pawan Kumar, et al. "UCred: fusion of machine learning and deep learning methods for user credibility on social media." Social Network Analysis and Mining 12.1 (2022): 1-10.

[80] Samonte, M. J. C. (2018). Polarity analysis of editorial articles towards fake news detection. ACM International Conference Proceeding Series, 108–112. https://doi.org/10.1145/3230348.3230354.

[81] Bhatt, Shaily, et al. "Fake News Detection: Experiments and Approaches beyond Linguistic Features." Data Management, Analytics and Innovation. Springer, Singapore, 2022. 113-128.

[82] Gencheva, P. et al. (2017) 'A context-aware approach for detecting worth-checking claims in political debates', in International Conference Recent Advances in Natural Language Processing, RANLP. doi: 10.26615/978-954-452-049-6-037.

[83] Pranto, Protik Bose, et al. "Are You Misinformed? A Study of Covid-Related Fake News in Bengali on Facebook." arXiv preprint arXiv:2203.11669 (2022).

[84] Patwari, A., Goldwasser, D. and Bagchi, S. (2017) 'TATHYA: A Multi- Classifier System for Detecting Check-Worthy Statements in Political Debates', in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. doi: 10.1145/3132847.3133150.

[85] Verma, Pawan Kumar, et al. "MCred: multi-modal message credibility for fake news detection using BERT and CNN." Journal of Ambient Intelligence and Humanized Computing (2022): 1-13.

[86] Gruppi, M., Horne, B. D., & Adalı, S. (2020). NELA-GT-2019: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles. arXiv preprint arXiv:2003.08444.

[87] Singhal, Shivangi, et al. "Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection." Companion Proceedings of the Web Conference 2022. 2022.

[88] Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. In Proceedings of the 2008 international conference on web search and data mining (pp. 219-230).

[89] Najadat, Hassan, Mais Tawalbeh, and Rasha Awawdeh. "Fake news detection for Arabic headlines-articles news data using deep learning." International Journal of Electrical & Computer Engineering (2088-8708) 12.4 (2022).

[90] Pathak, A., & Srihari, R. K. (2019, July). BREAKING! Presenting Fake News Corpus for Automated Fact Checking. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 357-362).

[91] Garcia, Gabriel L., Luis Afonso, and João P. Papa. "FakeRecogna: A New Brazilian Corpus for Fake News Detection." International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2022.

[92] Singh, Kavinder, et al. "A Comprehensive Study on Data-Driven Fake News Detection Methods." 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2022.

[93] Yumeng Qin, Dominik Wurzer and Cunchen Tang. “Predicting future rumours”, Chinese Journal of Electronics. 2018

[94] Sharma, Aniket, Ishita Singh, and Vipin Rai. "Fake News Detection on Social Media." 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2022.

[95] Liddy, E. D, "Natural language processing",In Encyclopedia of Library and Information Science, 2nd Ed,NY, Marcel Decker, Inc., 2001.

[96] Nistor, Andreea, and Eduard Zadobrischi. "The Influence of Fake News on Social Media: Analysis and Verification of Web Content during the COVID-19 Pandemic by Advanced Machine Learning Methods and Natural Language Processing." Sustainability 14.17 (2022): 10466.

[97] Amini, Mahyar, et al. "MAHAMGOSTAR.COM as a Case Study for Adoption of Laravel Framework As the Best Programming Tool for PHP Based Web Development for Small and Medium Enterprises." Journal of Innovation & Knowledge, ISSN (2021): 100-110.

[98] Batailler, Cédric, et al. "A signal detection approach to understanding the identification of fake news." Perspectives on Psychological Science 17.1 (2022): 78-98.

[99] Pérez-Almendros, Carla, Luis Espinosa Anke, and Steven Schockaert. "SemEval-2022 Task 4: Patronizing and Condescending Language Detection." Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). 2022.

[100] M. Alrubaian, M. Al-Qurishi, M. Mehedi Hassan, and A. Alamri, “A credibility analysis system for assessing information on Twitter,” IEEE Trans. Depend. Sec. Comput., vol. 15, no. 4, pp. 661–674, Aug. 2018.

[101] Madnani, Nitin,"Getting started on natural language processing with python". Vol 13, pp. 1–3, 2013.

[102] M. Glenski, T. Weninger, and S. Volkova, “Propagation from deceptive news sources who shares, how much, how evenly, and how quickly?” IEEE Trans. Comput. Social Syst., vol. 5, no. 4, pp. 1071–1082, Dec. 2018.

[103] Nanath, Krishnadas, et al. "Examination of fake news from a viral perspective: an interplay of emotions, resonance, and sentiments." Journal of Systems and Information Technology (2022).

[104] Zervopoulos, Alexandros, et al. "Deep learning for fake news detection on Twitter regarding the 2019 Hong Kong protests." Neural Computing and Applications 34.2 (2022): 969-982.

[105] Amini, Mahyar, and Aryati Bakri. "Cloud computing adoption by SMEs in the Malaysia: A multi-perspective framework based on DOI theory and TOE framework." Journal of Information Technology & Information Systems Research (JITISR) 9.2 (2015): 121-135.

[106] Shu, Kai, Ahmadreza Mosallanezhad, and Huan Liu. "Cross-Domain Fake News Detection on Social Media: A Context-Aware Adversarial Approach." Frontiers in Fake Media Generation and Detection. Springer, Singapore, 2022. 215-232.

[107] Nadkarni, Prakash M, Ohno-Machado, Lucila, and Chapman, Wendy W. "Natural language processing: an introduction", Published by group.bmj.com, pp. 545-547, 2011.

[108] E. Lancaster, T. Chakraborty, and V. S. Subrahmanian, “M AL T P : Parallel prediction of malicious tweets,” IEEE Trans. Comput. Social Syst., vol. 5, no. 4, pp. 1096–1108, Dec. 2018.

[109] Vinothkumar, S., et al. "Fake News Detection Using SVM Algorithm in Machine Learning." 2022 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2022.

[110] Amini, Mahyar, and Nazli Sadat Safavi. "A Dynamic SLA Aware Heuristic Solution For IaaS Cloud Placement Problem Without Migration." International Journal of Computer Science and Information Technologies 6.11 (2014): 25-30.

[111] Chung, Wingyan, Yinqiang Zhang, and Jia Pan. "A Theory-based Deep-Learning Approach to Detecting Disinformation in Financial Social Media." Information Systems Frontiers (2022): 1-20.

[112] Jurafsky, D. and Martin, J. H., "Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics", 2nd edition. Prentice-Hall, Upper Saddle River, NJ, 2008.

[113] P. K. Verma and P. Agrawal, “Study and detection of fake news: P 2 C 2 -based machine learning approach,” in Proc. Int. Conf. Data Manage., Anal. Innov., vol. 1175. Singapore: Springer, Sep. 2020, pp. 261–278.

[114] Blanc, Olivier, et al. "CODE at CheckThat! 2022: multi-class fake news detection of news articles with BERT." Working Notes of CLEF (2022).

[115] Amini, Mahyar, and Nazli Sadat Safavi. "A Dynamic SLA Aware Solution For IaaS Cloud Placement Problem Using Simulated Annealing." International Journal of Computer Science and Information Technologies 6.11 (2014): 52-57.

[116] Tomas, Frédéric, Olivier Dodier, and Samuel Demarchi. "Computational measures of deceptive language: prospects and issues." Frontiers in Communication 7 (2022): 792378.

[117] Cambria, Erik and White, Bebo. "Jumping nlp curves: A review of natural language processing research" . IEEE Computational Intelligence Magazine, pp. 51-55, 2014.

[118] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, “Novel visual and statistical image features for microblogs news verification,” IEEE Trans. Multimedia, vol. 19, no. 3, pp. 598–608, Mar. 2017.

[119] Low, Jwen Fai, et al. "Distinguishing between fake news and satire with transformers." Expert Systems with Applications 187 (2022): 115824.

[120] Amini, Mahyar, et al. "Development of an instrument for assessing the impact of environmental context on adoption of cloud computing for small and medium enterprises." Australian Journal of Basic and Applied Sciences (AJBAS) 8.10 (2014): 129-135.

[121] Tandoc, Edson C., and Seth Kai Seet. "War of the Words: How Individuals Respond to “Fake News,“ Misinformation,“ Disinformation,” and “Online Falsehoods”." Journalism Practice (2022): 1-17.

[122] Alpaydın, Ethem., "Introduction to Machine Learning", 2nd edition, The MIT Press, Massachusetts Institute of Technology, 2010.

[123] B. Ratner, "The correlation coefficient: Its values range between +1/ − 1, or do they?" J. Targeting, Meas. Anal. Marketing, vol. 17, no. 2, pp. 139–142, Jun. 2009.

[124] Truică, Ciprian-Octavian, Elena-Simona Apostol, and Adrian Paschke. "Awakened at CheckThat! 2022: fake news detection using BiLSTM and sentence transformer." Working Notes of CLEF (2022).

[125] Amini, Mahyar, et al. "Types of cloud computing (public and private) that transform the organization more effectively." International Journal of Engineering Research & Technology (IJERT) 2.5 (2013): 1263-1269.

[126] Arrese, Ángel. "Cultural Dimensions of Fake News Exposure: A Cross-National Analysis Among European Union Countries." Mass Communication and Society just-accepted (2022).

[127] Buche, Arti., Chandak, Dr. M. B., and Zadgaonkar, Akshay. "Opinion mining and analysis: A survey",International Journal on Natural Language Computing (IJNLC), Vol. 2, No.3,pp. 41, 2013.

[128] A. De Salve, P. Mori, B. Guidi, and L. Ricci, "An analysis of the internal organization of Facebook groups," IEEE Trans. Comput. Social Syst., vol. 6, no. 6, pp. 1245–1256, Dec. 2019.

[129] Amer, Eslam, Kyung-Sup Kwak, and Shaker El-Sappagh. "Context-Based Fake News Detection Model Relying on Deep Learning Models." Electronics 11.8 (2022): 1255.

[130] Amini, Mahyar. "The factors that influence on adoption of cloud computing for small and medium enterprises." (2014).

[131] Azhar, Wardah, and Syed Kazim Shah. "A Corpus Driven Text Analysis of" Fake Social Media News": A case study of The Indian Chronicles." Journal of English Language, Literature and Education 4.1 (2022): 60-80.

[132] Domingos, Pedro. "A few useful things to know about machine learning", communications of the ACM magazine vol.55, issue.10, pp 78- 79,2012.

[133] P. K. Verma, P. Agrawal, and R. Prodan, WELFake Dataset for Fake News Detection in Text Data (Version: 0.1) [Data Set]. Genéve, Switzer- land: Zenodo, 2021.

[134] Siddikk, Abu Bakkar, et al. "FakeTouch: Machine Learning Based Framework for Detecting Fake News." Big Data Intelligence for Smart Applications. Springer, Cham, 2022. 317-334.

[135] Amini, Mahyar, et al. "Types of cloud computing (public and private) that transform the organization more effectively." International Journal of Engineering Research & Technology (IJERT) 2.5 (2013): 1263-1269.

[136] Pereira, Fernando. "Machine Learning in Natural Language Processing",University of Pennsylvania, pp. 3, 2002.

[137] Indurkhya, Nitin and Damerau, Fred J. , "Handbook Of Natural Language Processing", second edition, Chapman & Hall/CRC press, 2010.

[138] V. Madaan and A. Goyal, "Predicting ayurveda-based constituent bal- ancing in human body using machine learning methods," IEEE Access, vol. 8, pp. 65060–65070, 2020.

[139] Himdi, Hanen, et al. "Arabic fake news detection based on textual analysis." Arabian Journal for Science and Engineering (2022): 1-17.

[140] Amini, Mahyar, et al. "The role of top manager behaviours on adoption of cloud computing for small and medium enterprises." Australian Journal of Basic and Applied Sciences (AJBAS) 8.1 (2014): 490-498.

[141] Verma, Sabu, and S. Rajagopal. "NLP Based Fake News Detection Using Hybrid Machine Learning Techniques." 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2022.

[142] Daumé III, Hal ,"A Course in Machine Learning", Published by TODO,2014

[143] M. Li, G. Clinton, Y. Miao, and F. Gao, "Short text classification via knowledge powered attention with similarity matrix based CNN," 2020, arXiv:2002.03350.

[144] Loey, Mohamed, Mohamed Hamed N. Taha, and Nour Eldeen M. Khalifa. "Blockchain Technology and Machine Learning for Fake News Detection." Implementing and Leveraging Blockchain Programming. Springer, Singapore, 2022. 161-173.

[145] Amini, Mahyar, et al. "Agricultural development in IRAN base on cloud computing theory." International Journal of Engineering Research & Technology (IJERT) 2.6 (2013): 796-801.

[146] Galli, Antonio, et al. "A comprehensive Benchmark for fake news detection." Journal of Intelligent Information Systems (2022): 1-25.

[147] Li, Qi." Literature survey:domain adaptation algorithms for natural language processing", Department of Computer Science The Graduate Center, The City University of New York, pp. 8-10, 2012.

[148] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification," in Proc. China Nat. Conf. Chin. Comput. Linguistics, vol. 11856. Cham, Switzerland: Springer, Feb. 2020, pp. 194–206.

[149] Patel, Ritik H., et al. "Detecting Fake News Using Machine Learning." Intelligent Data Communication Technologies and Internet of Things. Springer, Singapore, 2022. 613-625.

[150] Amini, Mahyar, and Nazli Sadat Safavi. "Cloud Computing Transform the Way of IT Delivers Services to the Organizations." International Journal of Innovation & Management Science Research 1.61 (2013): 1-5.

[151] Meddeb, Paul, et al. "Counteracting French Fake News on Climate Change Using Language Models." Sustainability 14.18 (2022): 11724.

[152] Mihalcea, Rada, Liu, Hugo, and Lieberman, Henry. "Nlp (natural language processing) for nlp (natural language programming)", pp. 323– 325, Springer-Verlag Berlin Heidelberg, 2006.

[153] K. Dzmitry Bahdanau, Y. Cho, and Bengio, "Neural machine translation by jointly learning to align and translate," in Proc.

3rd Int. Conf. Learn. Represent., 2015, pp. 1–15.

[154] Lai, Chun-Ming, et al. "Fake News Classification Based on Content Level Features." Applied Sciences 12.3 (2022): 1116.

[155] Amini, Mahyar, and Nazli Sadat Safavi. "Critical success factors for ERP implementation." International Journal of Information Technology & Information Systems 5.15 (2013): 1-23.

[156] Rohera, Dhiren, et al. "A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects." IEEE Access 10 (2022): 30367-30394.

[157] Mishra, Shubha, Piyush Shukla, and Ratish Agarwal. "Analyzing machine learning enabled fake news detection techniques for diversified datasets." Wireless Communications and Mobile Computing 2022 (2022).

[158] W. Jiang, J. Wu, F. Li, G. Wang, and H. Zheng, "Trust evaluation in online social networks using generalized network flow," IEEE Trans. Comput., vol. 65, no. 3, pp. 952–963, Mar. 2016.

[159] M. Alrubaian, M. Al-Qurishi, A. Alamri, M. Al-Rakhami, M. M. Hassan, and G. Fortino, "Credibility in online social networks: A survey," IEEE Access, vol. 7, pp. 2828–2855, 2019.

[160] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as 'false news,"' in Proc. ACM Workshop Multi- modal Deception Detection, Nov. 2015, pp. 15–19.

[161] Biradar, Shankar, Sunil Saumya, and Arun Chauhan. "Combating the infodemic: COVID-19 induced fake news recognition in social media networks." Complex & Intelligent Systems (2022): 1-13.

[162] Sadat Safavi, Nazli, et al. "An effective model for evaluating organizational risk and cost in ERP implementation by SME." IOSR Journal of Business and Management (IOSR-JBM) 10.6 (2013): 70-75.

[163] Nistor, Andreea, and Eduard Zadobrischi. "The Influence of Fake News on Social Media: Analysis and Verification of Web Content during the COVID-19 Pandemic by Advanced Machine Learning Methods and Natural Language Processing." Sustainability 14.17 (2022): 10466.

[164] S. Ranganath, S. Wang, X. Hu, J. Tang, and H. Liu, "Facilitating time critical information seeking in social media," IEEE Trans. Knowl. Data Eng., vol. 29, no. 10, pp. 2197–2209, Oct. 2017.

[165] Z. Zhang, R. Sun, X. Wang, and C. Zhao, "A situational analytic method for user behavior pattern in multimedia social networks," IEEE Trans. Big Data, vol. 5, no. 4, pp. 520–528, Dec. 2019.

[166] M. Schudson and B. Zelizer, "Fake news in context," in Understanding and Addressing the Disinformation Ecosystem. Philadelphia, PA, USA: Annenberg School for Communication, Apr. 2017, pp. 1–4.

[167] P. Bourgonje, J. Moreno Schneider, and G. Rehm, "From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles," in Proc. EMNLP Workshop, Natural Lang. Process. Meets Journalism, 2017, pp. 84–89.

[168] Raja, M. Senthil, and L. Arun Raj. "Fake news detection on social networks using Machine learning techniques." Materials Today: Proceedings (2022).

[169] Sadat Safavi, Nazli, Nor Hidayati Zakaria, and Mahyar Amini. "The risk analysis of system selection and business process re-engineering towards the success of enterprise resource planning project for small and medium enterprise." World Applied Sciences Journal (WASJ) 31.9 (2014): 1669-1676.

[170] Lahby, Mohamed, et al. "Online fake news detection using machine learning techniques: A systematic mapping study." Combating Fake News with Computational Intelligence Techniques (2022): 3-37.

[171] S. Zaryan, "Truth and trust: How audiences are making sense of fake news," M.S. thesis, Media Commun. Studies, Lund Univ. Publications Student Papers, Stockholm, Sweden, Jun. 2017.

[172] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact- checking," in Proc. Conf. Empirical Methods Natural Lang. Process., 2017, pp. 2931–2937.

[173] Hossain, Fahima, Mohammed Nasir Uddin, and Rajib Kumar Halder. "An Ensemble Method-Based Machine Learning Approach Using Text Mining to Identify Semantic Fake News." Proceedings of the International Conference on Big Data, IoT, and Machine Learning. Springer, Singapore, 2022.

[174] Sadat Safavi, Nazli, Mahyar Amini, and Seyyed AmirAli Javadinia. "The determinant of adoption of enterprise resource planning for small and medium enterprises in Iran." International Journal of Advanced Research in IT and Engineering (IJARIE) 3.1 (2014): 1-8.

[175] Althabiti, S., Mohammad Ammar Alsalka, and E. Atwell. "SCUoL at CheckThat! 2022: fake news detection using transformer-based models." Working Notes of CLEF (2022).

[176] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[177] Surekha, T. Lakshmi, et al. "Digital Misinformation And Fake News Detection Using WoT Integration With Asian Social Networks Fusion Based Feature Extraction With Text And Image Classification By Machine Learning Architectures." Theoretical Computer Science (2022).

[178] Safavi, Nazli Sadat, et al. "An effective model for evaluating organizational risk and cost in ERP implementation by SME." IOSR Journal of Business and Management (IOSR-JBM) 10.6 (2013): 61-66.

[179] Humayoun, Muhammad. "The 2021 Urdu Fake News Detection Task using Supervised Machine Learning and Feature Combinations." arXiv preprint arXiv:2204.03064 (2022).

[180] R. Sequeira, A. Gayen, N. Ganguly, S. K. Dandapat, and J. Chandra, "A large-scale study of the Twitter follower network to characterize the spread of prescription drug abuse tweets," IEEE Trans. Comput. Social Syst., vol. 6, no. 6, pp. 1232–1244, Dec. 2019.

[181] Arif, Muhammad, et al. "CIC at CheckThat! 2022: multi-class and cross-lingual fake news detection." Working Notes of CLEF (2022).

[182] Khoshraftar, Alireza, et al. "Improving The CRM System In Healthcare Organization." International Journal of Computer Engineering & Sciences (IJCES) 1.2 (2011): 28-35.

[183] Seetharaman, R., et al. "Analysis of fake news detection using machine learning technique." Materials Today: Proceedings 51 (2022): 2218-2223.

[184] Alhakami, Hosam, et al. "Evaluating Intelligent Methods for Detecting COVID-19 Fake News on Social Media Platforms." Electronics 11.15 (2022): 2417.

[185] Choraś, Michał, et al. "How Machine Learning May Prevent the Breakdown of Democracy by Contributing to Fake News Detection." IT Professional 24.2 (2022): 25-31.

[186] Taskin, Suleyman Gokhan, Ecir Ugur Kucuksille, and Kamil Topal. "Detection of Turkish Fake News in Twitter with Machine Learning Algorithms." Arabian Journal for Science and Engineering 47.2 (2022): 2359-2379.

[187] Verma, Sabu, and S. Rajagopal. "NLP Based Fake News Detection Using Hybrid Machine Learning Techniques." 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2022.

[188] Lee, Colin Jingwei, and Hui Na Chua. "Using Linguistics and Psycholinguistics Features in Machine Learning for Fake News ClassificationThrough Twitter." Proceedings of International Conference on Data Science and Applications. Springer, Singapore, 2022.

[189] Patra, Debasish, and Biswapati Jana. "Fake News Identification Through Natural Language Processing and Machine Learning Approach." International Conference on Computational Intelligence in Communications and Business Analytics. Springer, Cham, 2022.

[190] Garg, Sonal, and Dilip Kumar Sharma. "Linguistic features based framework for automatic fake news detection." Computers & Industrial Engineering 172 (2022): 108432.

[191] Nanath, Krishnadas, et al. "Examination of fake news from a viral perspective: an interplay of emotions, resonance, and sentiments." Journal of Systems and Information Technology (2022).

[192] Ghafoor, Hafiz Yasir, et al. "Fake News Identification on Social Media Using Machine Learning Techniques." Proceedings of International Conference on Information Technology and Applications. Springer, Singapore, 2022.

[193] Rahman, Shagoto, et al. "Efficient Machine Learning Approaches to Detect Fake News of Covid-19." Machine Intelligence and Data Science Applications. Springer, Singapore, 2022. 513-525.

[194] L.-L. Shi et al., “Human-centric cyber social computing model for hot- event detection and propagation,” IEEE Trans. Comput. Social Syst., vol. 6, no. 5, pp. 1042–1050, Oct. 2019.

[195] Hangloo, Sakshini, and Bhavna Arora. "Combating multimodal fake news on social media: methods, datasets, and future perspective." Multimedia Systems (2022): 1-32.

[196] Gupta, Anchal, et al. "Empirical Framework for Automatic Detection of Neural and Human Authored Fake News." 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2022.

[197] Baria, Reeya, et al. "Theoretical Evaluation of Machine And Deep Learning For Detecting Fake News." 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). IEEE, 2022.

[198] Shahid, Wajiha, et al. "Detecting and Mitigating the Dissemination of Fake News: Challenges and Future Research Opportunities." IEEE Transactions on Computational Social Systems (2022).

[199] Bozkir, Efe, et al. "Regressive Saccadic Eye Movements on Fake News." 2022 Symposium on Eye Tracking Research and Applications. 2022.

[200] Rodriguez, Guillermo Romera, Sanjana Gautam, and Andrea Tapia. "Understanding Twitters behavior during the pandemic: Fake News and Fear." arXiv preprint arXiv:2202.05134 (2022).

[201] Waghmare, Akash Dnyandeo, and Girish Kumar Patnaik. "Social Media Fake News Detection using mNB in Blockchain." 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE, 2022.