Fake News Detection with Topic Modeling Using LDA and BERTopic

Asha Latha Nama, Harshini Badam, Sathwik Vattikuti, Rushith Challa, Sai Subash Edara

Abstract

Social media at its high rise in the recent world, there are many pros and cons with its utilization. As users can post anything in the social media without any consent, spread of fake news has become one of the most challenging problems in the recent years. Fake news is a type of news which misleads people to believe in false information. Fake news content has the potential to make some serious issues in the society. Although it is a problem since many decades but spreading fake news has increased with the recent rise of social media. There are many negative impacts on the individuals due to fake news and cyber criminals are targeting users with the help of fake news. Segregating fake news from the vast data available is one of the most challenging problems. In this survey, we are using two different datasets to analyze how LDA and BERTopic works in topic modeling and compare the topics from each model. Also, to develop a model with supervised learning, we are using the python libraries Scikit-Learn, Pandas, NumPy, and NLP for textual analysis to explore machine learning models. The overall results generated by BERTopic showed better results when compared to LDA. Also, we have got 90% and 92% accuracy when trained and tested the model with logistic regression and passive aggressive classifier respectively.

1.Introduction

The development of online social media in recent times has greatly facilitated how people communicate with one another. As more of our lives are spent conversing online through social media platforms, more people are opting to seek out and consume news through social media channels rather than conventional news organizations. However, because of the growing popularity of online social media, the Internet has become a perfect breeding ground for the transmission of fake news, such as misleading information, fake reviews, fraudulent ads, rumours, fake political comments, satires, and so on. For many people, social media has suddenly become a significant source of news. According to Statistical, about 50 percent of the world's population use social media. In the spread of news, social media websites and networks offer evident advantages, including quick access to information, free distribution, no time restriction, and diversity. However, most of these locations remain unregulated. As a result, determining whether given news is accurate, or fake is usually impossible.

Fake news is increasingly more common and widely spread on social media than in conventional media. False news may have a harmful impact on individuals and society. First and foremost, fake news has the ability to mislead people. Second, bogus news may influence how people react to legitimate news. Third, the broad availability of erroneous material may undermine public faith in the news profession as a whole. As a result, it's critical to recognize bogus news on social media. Google and Facebook are testing new tools to make it simpler for users to detect and report fake news websites. Facebook is building a new system that will allow users and fact-checkers to submit dubious articles, in addition to testing fact-checking labels in Google News and banning hoax websites from its ad platform.

Fraudulent news detection is a subtask of text classification and is commonly characterized as the task of determining whether news is authentic or fake. The advancements in big data and machine learning have been key among the tools and approaches employed by researchers. Among machine learning approaches, supervised machine learning techniques have seen widespread application and outstanding results. To extract themes from news, topic modeling machine learning techniques are utilized. For example, Latent Dirichlet Allocation (LDA) and BERTopic performed well in extracting themes, but Decision Tree and Random Forest algorithms will be utilized to forecast if the news is true or false. The comparison is based on many assessment criteria such as accuracy, precision, recall, F-measures, and the models' ROC curves.

Large volumes of textual user-generated content are created every day in the form of comments, reviews, and short-text messages, thanks to the expansion of online social network platforms and applications. As a result, people frequently find it difficult to glean relevant information or learn more about the issue at hand from such content. Machine learning and natural language processing algorithms, as well as topic modeling approaches that have gained prominence in recent years, are utilized to examine the huge quantity of textual social media data available online.

2. Literature Review

Abeer Abuzayed and Hend Al-Khalifa) [1], analyzed BERTopic with pre-trained arabic language models as its embeddings and compared its accuracy against LDA and NMF topic modelling techniques. LDA uses probabilistic approach and NMF uses the matrix factorization approach the dataset taken for this paper contains,111,728 documents where the data was collected from three arabic newspapers. Using Normalized pointwise mutual information (NPMI) evaluated the results of the 3 topic modeling techniques where -1 indicates never occurring together, 0 for independence and +1 for complete co-occurrence. Out of all three models BERTopic excelled.

Roman Egger and Joanne Yu [2], analyzed the topic modeling algorithms like LDA, NMF, Top2Vec, and BERTopic and their performances in analyzing the twitter data and compare its features against each other to understand which algorithm excels. The author compared the models by dividing the results into two parts, in the 1st part of the result, LDA and NMF, Top2vec and Bert. Bert produced better insights and this research suggests Top2Vec compared to LDA, as the results produced by LDA are not interesting.

Rania Albalawi, Tet Hin Yeap1, and Morad Benyoucef[3] performed topic modelling on short-text data with LSA,LDA,NMF,PCA and RP algorithms, results were compared using the performance metrics and overall results of LDA and NMF were most valuable outputs with their meaningful topic extractions. It produces higher quality topics and the authors said that LDA is slower and for the real time NMF will be a better choice.

Marina Danchovsky Ibrishimova and Kin Fun Li [4], in this paper, they analyzed the previous researchers in detecting the fake news. By examining the previous efforts, a new definition for fake news is derived in this paper in terms of bias and accuracy. Here the incident classification model consists of 5 NLP features that are combined with 3 knowledge verification features. The system implemented here using google NLP API for the NLP analysis portion and News API for knowledge verification part. The results showed that, approach for spotting false news combines source and fact checking with NLP analysis.

Julio C. S. Reis, Andre Correia, Fabricio Murai, Adriano Veloso, and Fabricio Benevenuto[5], the aim of this paper to explore the features from the fake news, sources and posts in social media. In addition to the existing features, externally they have added new set of features to predict the performance for detection of fake news. Different machine learning algorithms are used to train the model and the results showed that random forest and XGBoost has achieved a better accuracy. They are also trying to find that, what features are useful and important for detecting fake news.

Iftikhar Ahmad, Muhammad Yousaf,Suhail Yousaf , and Muhammad Ovais Ahmad[6], proposed various methods for automated classification of a text. Using a different textual property that helps in differentiating the fake news from real, some machine learning algorithm used ensemble method and applied on 4 real world datasets. The different ensemble learners used are Random Forest, Bagging Ensemble classifier and boosting ensemble classifier.From the results they concluded that, ensemble learner approach has higher performance in comparison to individual learners.

Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq [7], In order to computerize the identification of fake news in Twitter datasets, this research suggests a model for identifying fabricated news messages from tweets by learning how to anticipate precision appraisals. Then the data is trained using different well known machine learning algorithms Support vector machine, naïve bayes classifier, logistic regression and recurrent neural network. The results showed that, svm and naïve bayes has a better accuracy with 89% when compared with the other models.

Jia Ding, Yongjun Hu, Huiyou Chang [8], proposed a method for building a model using BERT to capture the features of fake news as they analyzed RNN, and CNN based model have less performance. For all the languages, a pattern was formed in each text to classify the features. With all the features that are formed from a text we tune the bert model to adjust the features. The paper suggested 3 methods to input the information of feature, the three methods are group method,pair method, news text method. The only difference between the methods is how we input the data. They showed that, self-attention mechanisms will help to improve the accuracy. The main aim is to detect a mental feature and build a bert model, according to that, bert will keep the mental feature for detecting the fake news.

Maryam Heidar, Samira Zad, Parisa Hajibabaee, Masoud Malekzadeh, SeyyedPooya HekmatiAthar, Ozlem Uzuner, James H Jr Jones [9], they analyzed the previous research papers on bot detection and fake news detection, where it raised a new question for fake news and bot detection on online platforms. For creating a model, we are using bert method. They used the covid19 dataset to predict the fake news and transfer learning model was used to detect the bot accounts. To improve the fake news detection model, then we applied new features on the covid19 dataset. The results and research stated that, for improving the fake news detection model a characteristic from the bot accounts is used as it claims less fake news.

Linguistic Feature Based Learning Model for Fake News Detection and Classification

Anshika Choudhary, Anuja Arora proposed a linguistic model to find out the properties of content that will generate language-driven features. This linguistic model is used to extract syntactic, grammatical, sentimental, and readability features of particular news. The sequential learning technique based on neural networks was utilized to get the best outcomes for fake news identification. Finally, the integrated linguistic feature-driven model can classify and detect bogus news with an average accuracy of 86%. Results indicate that a sequential model based on features can complete an evaluation with a comparable level of performance in a noticeably shorter amount of time. Syntactic, grammatical, emotive, and readability features were collected for the news dataset using a variety of computational techniques. They also carried out preliminary experiments to confirm the effectiveness of base learners and ensemble algorithms for machine learning. The ensemble algorithm has a maximum accuracy of 72% when all algorithms were measured. They also looked into sequential neural networks based on linguistic features and neural networks to classify news.

Aswini Thota, Priyanka Tilak ,Simrat Ahluwalia and Nibrat Lohia presents the solution to the task of fake news detection by using Deep Learning architectures. Additionally, they provided a neural network architecture that can precisely predict the relationship between a given title and article body. Their model performs 2.5% better than current model architectures, and they can attain a 94.21% accuracy on test data. They have research. Researchers that looked at how quickly misleading information spread on Twitter came to the conclusion that false information spreads on Twitter six times more quickly than true information. They also suggested a method for computing the Tf-Idf vectors based on bigrams and unigrams, which was quite successful. They also noted that we were able to produce a very smooth and consistent learning process by using regularization techniques including Dropout, L2 regularization, crossvalidation, and early stopping.

Srijan Malhotra, Amitrajit Sarkar, Vasu Agarwal, and H. Parveen Sultana discussed about the use of NLP and ML to address fake news. To determine which of five classifiers performs best for this dataset of labelled news statements, bag-of-words, n-grams, count vectorizer, and TF-IDF were used. The data was then trained on the classifiers. Which model performs best was determined using the precision, recall, and f1 scores.

Reza Mansouri, Mahmood Naderan-Tahan and Mohammad Javad Rashti proposed a method that is based on a semi supervised learning framework targeting both labelled and unlabelled data using convolutional neural network. Using CNN, different aspects of text and visual data are retrieved.

To predict the classes of unclassified data, they next used linear discrimination analysis (LDA), which was modified such that it could estimate classes in a semi-supervised fashion. The unlabelled data is subsequently labelled using a CNN. Additionally, they changed the fitness function to enhance the effect of the estimated class at each stage. The evaluations' findings indicate that the method's precision is 95.6% and its recall is 96.7%. With an accuracy value of 95.6%, the proposed method performs better than previous methods in terms of recall, specificity, and sensitivity.

3. Methodology

In this study, we conducted experiments in google Collaboratory(executing the code) with different topic modeling algorithms like LDA and BERTopic to analyze which model predicts the topics accurately from the vast data. We have used genism implementation of LDA(Latent Dirichlet Allocation) multicore and BERTopic with word embedding representations.

Steps taken to Validate the expected outcome are pre-processing data, training the LDA model on the pre-processed data, training BERTopic, comparing the results of each topic modeling algorithm then visualizing the clusters of each topic that these models produced.

3.1 Dataset

we have taken two datasets from a website called Kaggle. One dataset has a labels column i.e., whether the news is Real or Fake (let it be Dataset 1) and the other dataset (let it be Dataset 2) has a categories column that contains the published about around 24 different topics. A clear picture of dataset details is provided below.

3.1a. Dataset1 attributes:

Title: Title of the article published

Text: a short description of the article

Label: Contains whether the article is Real or Fake

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

FIG 3.1a: Figure shows the first 5 rows of Dataset1

3.1b. Dataset2 attributes:

Link: the link to the article that is been published

Headline: the headline of the article

Category: Type of category that news belongs

Short description: Outline of the article

Authors: Author of the article that is published

Date: The date on which the article is published online

	link	headline	category	short_description	authors	date
0	https://www.huffpost.com/entry/covid-boosters-...	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS	Health experts said it is too early to predict...	Carla K. Johnson, AP	2022-09-23
1	https://www.huffpost.com/entry/american-airlin...	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS	He was subdued by passengers and crew when he ...	Mary Papenfuss	2022-09-23
2	https://www.huffpost.com/entry/funniest-tweets...	23 Of The Funniest Tweets About Cats And Dogs ...	COMEDY	"Until you have a dog you don't understand wha...	Elyse Wanshel	2022-09-23
3	https://www.huffpost.com/entry/funniest-parent...	The Funniest Tweets From Parents This Week (Se...	PARENTING	"Accidentally put grown-up toothpaste on my to...	Caroline Bologna	2022-09-23
4	https://www.huffpost.com/entry/amy-cooper-lose...	Woman Who Called Cops On Black Bird-Watcher Lo...	U.S. NEWS	Amy Cooper accused investment firm Franklin Te...	Nina Golgowski	2022-09-22

FIG 3.1b: Figure shows the sample data of Dataset2

Firstly, we have done some pre-processing steps on both the datasets like checking for null values, removing duplicates, performing stemming(reducing words to their root forms), tokenization, and removing stop words from the datasets. Now on both Dataset1 and Dataset2, we have applied the LDA topic modeling technique for the first 1000 rows on the bag of words corpus and the model segregated the topics. Now that we have topics available to us, the labelling must be done for the available topics using the word cloud, where we can visually interpret each word i.e. based on its frequency, word with highest frequency will be shown large and in such a manner other words are shown in a decreasing order of frequenct, based on that we can interpret and give the labelling manually. The main drawback with LDA is it cannot be performed on a very less amount of data and had to assign the topics manually which can cause errors sometimes. We cannot verify if the topic assigned is correct or wrong. Sometimes the data will not have any relation to the cluster with which it is assigned. In this case, here we are trying to evaluate the results with the help of BERTopic and to analyze which results are best in topic modeling. Before sending the data to a BERTopic, we need to do some adjustments for our column, i.e. converting the words in a text into tokens and also removing the numbers, special characters, as bert only trains the string type and it gives incorrect results if there is any unusual text.

For BERTopic, we are using sentence transformers embedding i.e.," sentence-transformers/all-MiniLM-L6-v2" which is a very high-performing model with 384-dimensional sentence embeddings then the BERTopic converts them into 2/3 dimensional vectors. Then the topics are generated based on frequency and using the wordcloud, we visually interpret the words from high frequency to lower frequency. Topic (-1) will have outliners of the topics which needs to be ignored and the rest will have topics related to other issues.

To validate the results, we used to visualization methods in both LDA and BERTopic and couldn't conclude anything with Dataset 1 since we didn't get predicted results and the results in both seems almost the same. To evaluate this, we have chosen a dataset with category labels and performed both LDA and BERTopic on Dataset 2 and found that in topic modeling BERTopic outperformed LDA. Also, we have plotted how the topic has changed throughout the years with the help of dynamic Bert to visualize the results.

Finally, we have used logistic regression and Passive aggressive classifier models to predict the accuracy of the models in predicting fake news. We are also using the confusion matrix, we can evaluate the performance of classification models using this matrix.

4. Machine Learning Algorithms

We have used all the below mentioned machine learning algorithms to predict the topics and understand which model is performing well in deriving them. Also, to differentiate the news whether it is real or fake.

- LDA (Latent Dirichlet Allocation)
- BERT (Bidirectional Encoder Representations from Transformers)
- Dynamic BERT
- Logistic Regression
- Passive-Aggressive Classifier
- Confusion Matrix

LDA

Latent Dirichlet Allocation (LDA) is a common approach for extracting themes from a corpus. The phrase latent refers to anything that exists but has not yet manifested itself. The themes we wish to extract from the data are now "hidden topics." It has not yet been discovered. As a result, the term "latent" is used in LDA. Following the Dirichlet distribution and method comes the Dirichlet allocation. Because this method is a distribution among distributions, every draw from a Dirichlet procedure is a distribution of its own. A Dirichlet process is therefore a probability distribution with a spectrum that is a group of probabilities. The Dirichlet process is closely connected to clustering, a family of algorithms in which the size of the clusters is determined by the process and its parameters. LDA is very effective for locating relatively accurate topic combinations within a given material. The Dirichlet model depicts the pattern of words that recur together, appear often, and are related to one another. In the instance of topic modelling, the approach aids in evaluating the likelihood that the words scattered across the page will appear again. This allows the model to generate data points and estimate probabilities. LDA creates odds for the terms used to build themes, and the topics are subsequently categorized as documents. The LDA is based on two essential assumptions. One is that documents are a combination of themes, and the other is that topics are a combination of tokens (or words), and these topics create the words using a probability distribution. The papers are defined as the range of themes in statistical terms, and the concepts are the probability concentration of words.

How does LDA work:

To begin, LDA has two fundamental assumptions: one is that documents are a combination of concepts, and the second is that topics are a combination of tokens. These are the two significant hypotheses applied to the supplied corpus. The initial stage with text data is cleaning, which includes tokenizing (reducing a content to its atomic constituents), stopping (removing useless words), and stemming (merging words that are equal in sense to the text to words). The document word matrix is obtained after preprocessing the documents. The corpus is now mostly a preprocessed document-word matrix, with each row representing a document and each column representing tokens or words. The document-word matrix is converted by LDA into two further matrices: the Document Term grid and the Topic Word matrix. The Document-Topic grid currently includes all of the available themes for the documents. In this case, we have 5 subjects and 5 documents, thus the matrix has a dimension of 5*6, and the Topic-Word matrix contains the words that those topics might include. For example, because the vocabulary contains 5 subjects and 8 unique tokens, the matrix has a shape of 6*8. Every word, according to LDA, is connected with a latent theme. These documents now provide a subject term breakdown available in the dataset. The LDA model contains two parameters that regulate the distributions: one controls the per-document topic distribution and the other controls the per-topic distribution. Because LDA thinks that papers are a mix of themes and topics are a mix of words, LDA goes back to the document level to determine which subjects originated these papers and which phrases generated those topics. LDA is an iterative technique that assigns subjects to each phrase in the document at random in the first iteration. As a result, the terms in the texts will be connected with certain random subjects in our corpus. This produces Documents having Topics composed of words and Topics composing of Topics. Following the initial iteration of LDA, the objective is to strengthen the acquired results, which LDA accomplishes by iterating through all of the articles and all of the words. LDA also assumes that all of the associated are accurate except for the present term. So, based on previous already-correct subject-word assignments, LDA attempts to correct and amend the current word's topic assignment with a new assignment. It is completed by computing two probability. One is the percentage of the document's words that are presently allocated to the

subject, while the other is the fraction of allocations to the topic that arise from this word across all papers for each topic. Reassignment of document words to a new subject using unit chances, the LDA is run for a number of times for the phase of choosing a brand-new topic until a stable state is attained. LDA converges at the point when it provides the best approximation of the document-term grid and topic-word mixture. This accomplishes the operation and procedure of LDA Allocation.

Comparison between LDA and BERTopic:

LDA does not require any knowledge for converting the text to feature vectors and it can deal with the large number of inputs. Number of topics here will be limited, so that it is easier to interpret. A document can contain several topics. It generates a generalized model where different vectors are generated for the same word in different sentence. Within the topics it shows both adjectives and nouns. LDA doesnot support for multilingual analysis and is not stable across different domains. This is where BERTopic came into the picture, its stability is high across all the domains and it is a pre trained model, it doesnot require any preprocessing of the data. As in LDA, there is no need for declaring number of topics that need to be produced for the bert, it automatically gives the different number of topics based on the data. This has a builtin modules which helps in searching topics in a document. There are also many disadvantages for the BERTopic, the approach used in bert will result in many number of topics which will be difficult for interpreting. In this the text should be converted into tokenizers, as the bert has a disadvantage where it assigns each document to single topic. Also it generates many outliers. Detailed assumptions are not produced in LDA and the output of this can give overlapping of topics. The other advantage is the hyperparameters need to be adjusted before it is trained. The main drawback is user needs to give the number of topics. In LDA, as the topics are independent of each other, for similar words same frequency will be used which gives a less accurate results. The correlations between words are ignored, so there is no relationship between the topics. Both LDA and BERTopic failed to give objective evaluation metrics. By comparing both the models, we can say that BERTopic is more accurate in topic modelling for producing the topics, also it generates the topics automatically and are all the topics produced are unique and not repeated, it just needs a proper text converted into tokens.

BERT:

Similarity between two words can be found by deriving the features and each feature is converted to vectors. These feature vectors for each word is called as word embeddings. One of the method for word embedding is word2vec. It is a two layer neural network, its input is a text corpus and the output of this model is the vector. Word2vec converts vectors to neural network understandable form. The usefulness of word2vec to group the vectors of similar words. But word2vec has an disadvantage, when we have a similar word with different meanings, word2vec generates a fixed vector even if the two words are of different sentences. Even if the word is same, its not the correct way of generating a fixed vector. We need a model which can generate a contextualized meaning of a word. We can achieve this using bert embedding. It captures the meaning of a word and generates a meaningful number based on the context of a given sentence. Not only for one word, it can also generate a embedding for the whole sentence, usually it generates around 768 size of the vector. Even google uses bert where search becomes better with this method. Bert is a topic modelling technique it creates clusters using transformers and class-based TF-IDF. Also it allows for visualizing topics that are generated.It is a pre trained model of 50+ languages.

Bert is a bidirectional transformer,it can read the entire text in both directions simultaneously, the learning phase in this model is divided into two phases, where one phase in a pre-trainining phase and the second phase is fine-tuning phase. The first phase is a time consuming and requires a lot of computation. After the first phase, to have a general understanding of the language a network will be created. Then to adjust the first phase, fine tuning phase tunes the learned knowledge to make it applicable for the target. The input of bert is a single text or a pair of sentence. This input should be represented in a linear sequence and for each word input presentation if performed by three parts of embedding, token embedding, segment embedding and position embedding. Here the token embedding represents the word vector, where classification task can use this and non-classification task can ignore this vector. Segment embeddings used for distinguishing between two sentences. Position embedding represents position information.

For the pre training phase, it is combined of two steps i.e. masked LM and next sentence prediction. The multiple layers in the bert forms a "Transformer", learns a relationship between different words of a text. The main aim of the transformer is to analyze the words of a complex sentence and to relate them for a better understanding for the overall meaning. These transformers makes it faster and efficient. The two transformers used in the bert are paraphrase-MiniLM-L6v2, it is an English bert based model used specifically for the similarity task and the other one is paraphrase-multilingual-MiniLM-L12-v2, but the main difference between the two is this can work for many number of languages. Core architecture of the bert consists of two components, the first component uses an encoder for reading the input text, and a vector representation of words is generated. The second component is bert uses a decoder to perform the task.

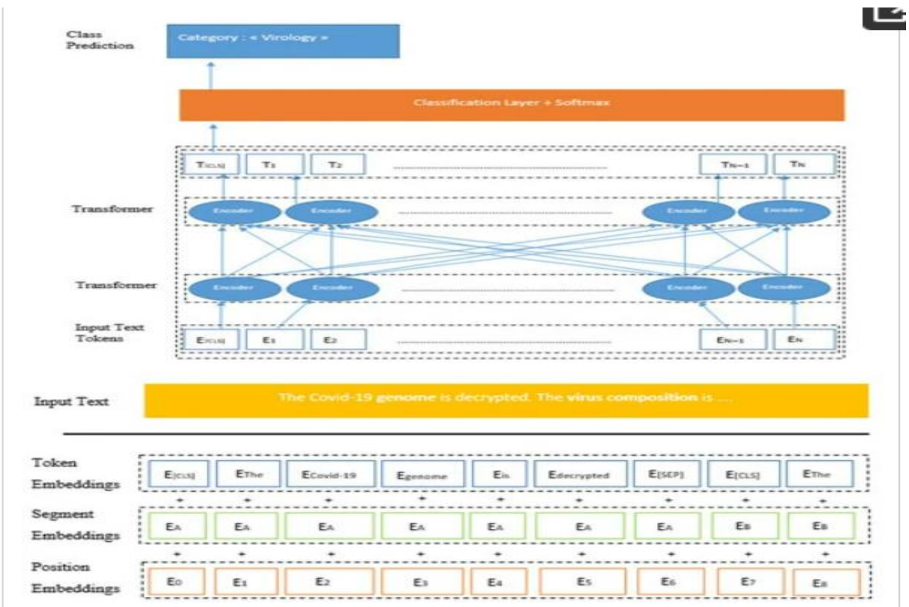

FIG 4: Figure shows the visual representation of Bert model

Dynamic bert:

After analyzing the topics with bert, using the dynamic bert it helps to evolute the topics over time which allows to analyze how a topic is represented in different timestamps. With no need to generate clusters from embeddings because they were already created, this will lead to a distinct topic representation at each timestep. The dynamic bert will be more interpretable if it is represented visually how topics are changed over time, where the topics and the date or time of that topic are considered for the visualization.

Logistic regression:

It is a supervised classification algorithm where it is used for predicting a categorical variable, where it calculates a probability of an event. The dependent variable contains discrete values. We cannot use linear regression for discrete values, it doesn't classify the data correctly, if there are outliers in the data and logistic regression removes these disturbing outliers using the sigmoid function. Logistic regression has 3 types i.e. binary, multinomial and ordinal. This type of regression can be used to classify if we have the target variable as the binary value 0 or 1 types. The loss function of logistic regression can be calculated using the maximum likelihood estimation.

Passive Aggressive Classifier:

It is a type of machine learning and a high-level algorithm, one of the category of machine learning algorithm, as supervised and unsupervised algorithms, there is another category in this machine learning algorithms i.e. online learning algorithm. In this algorithm, we train a model incrementally in the form of small group of data called mini batches. It is useful in situations where there is a huge amount of data, training the total data will computationally take a lot of time. Every time a new data is added, it continues to learn and updates the model. It doesnot require any learning rate. In the passive aggressive classifier, passive means, if the prediction is correct, there is no need to make any changes and the data is perfect, aggressive indicates that if the prediction is incorrect, changes are required for the model. The online learning algorithm can be used in real world for detecting fake news in social media, where every time new data adds, it can easily learn and update. The passive aggressive classifier works the same as other learning algorithms, i.e., splitting the data into training and testing sets. Then this training set is given for the model, to learn how correctly can it classify the two categories and accuracy is measured. It is found to have a better accuracy than the other machine learning algorithms, also in the text classification it tends to perform well.

Confusion matrix:

It is used for evaluating the performance of a classification models. A matrix is produced between the actual values as column and predicted values as rows. The values in each row of the matrix are, True positive, False positive, False negative and True negative. The false positive is also called as Type1 error, and the False negative is called as Type 2 error. Using this confusion matrix we can calculate different performance metrics like Accuracy, precision, recall and F1-score, which are generally used for evaluating a model. Accuracy is the total positives by the total number of predictions. Precision tells, how many are actually predicted are turns out to be positive. Recall tells, how many positive values are used for predicting correctly with our model.

5. Results

In this study, we performed analysis on BERTopic (Bidirectional Encoder Representations) and LDA (Latent Dirichlet Allocation) topic modelling techniques to predict its performance using two datasets with unsupervised clustering. BERTopic used default embedding model i.e., all-MiniLM-L6-v2 while selecting language =" English" in which embedding for each topic is calculated by taking the weighted average of the word embeddings in a topic based on their c-TF-IDF values. On the other hand, LDA works by decomposing the corpus documents words into lower dimension matrices and mainly LDA does 2 tasks i.e., first it finds the topics from the corpus, also assigns the topics to the documents within the same corpus at the same time. We observed that BERTopic excelled LDA in deriving the topics from the available two datasets. Also, we have visualised the topics over the recent months using dynamic BERT (FIG 5). Finally, we have trained and tested on classification models for predicting the fake news using passive aggressive classifier and logistic regression to analyse which performs better in differentiating the news. Passive-aggressive classifier's results were most convincing when compared in predicting the fake news.

5.1 Dataset1 Results

5.1.1 Result of LDA topic modeling

```
Topic: 0
Words: 0.005*"said" + 0.005*"biden" + 0.004*"gop" + 0.004*"trump" + 0.004*"abbott" + 0.004*"republican" + 0.004*"presid" + 0.003*"hous" + 0.003*"new'


Topic: 1
Words: 0.006*"water" + 0.005*"former" + 0.005*"trump" + 0.004*"twitter" + 0.004*"week" + 0.003*"one" + 0.003*"first" + 0.003*"texas" + 0.003*"leader'


Topic: 2
Words: 0.006*"trump" + 0.005*"man" + 0.004*"said" + 0.004*"jackson" + 0.004*"new" + 0.004*"court" + 0.003*"show" + 0.003*"black" + 0.003*"want" + 0.0


Topic: 3
Words: 0.011*"trump" + 0.006*"russian" + 0.005*"ukraine" + 0.005*"presid" + 0.005*"new" + 0.004*"former" + 0.004*"donald" + 0.004*"said" + 0.004*"hou


Topic: 4
Words: 0.008*"said" + 0.005*"time" + 0.004*"gop" + 0.004*"get" + 0.004*"russian" + 0.004*"new" + 0.004*"trump" + 0.003*"r" + 0.003*"democrat" + 0.00


Topic: 5
Words: 0.007*"star" + 0.004*"kardashian" + 0.004*"children" + 0.004*"home" + 0.004*"gun" + 0.003*"said" + 0.003*"lawsuit" + 0.003*"million" + 0.002*'


Topic: 6
Words: 0.005*"said" + 0.004*"trump" + 0.004*"russian" + 0.004*"fight" + 0.004*"million" + 0.003*"new" + 0.003*"make" + 0.003*"work" + 0.003*"host" +
```

FIG 5.1.1: The above figure displays the topics derived by using LDA for dataset1

5.1.2 Result of BERTopic modeling

	Topic	Count
0	-1	426
1	0	74
2	1	67
3	2	64
4	3	39
5	4	36
6	5	35
7	6	34
8	7	28
9	8	24
10	9	23
11	10	19
12	11	19
13	12	17
14	13	17
15	14	16
16	15	16
17	16	13
18	17	12
19	18	11
20	19	10

FIG 5.1.2 : The above figure displays the topics derived from BERTopic

5.1.3 Words clouds of both LDA and BERTopic

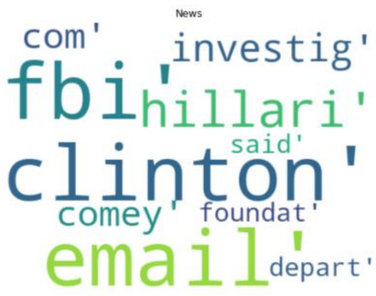

FIG 5.1.3a : cluster1 topics word cloud of LDA **FIG 5.1.3b: cluster1 topics word cloud of BERTopic**

5.2 Dataset2 Results

5.2.1 Result of LDA Topic modeling

```
Topic: 0
Words: 0.005*"said" + 0.005*"biden" + 0.004*"gop" + 0.004*"trump" + 0.004*"abbott" + 0.004*"republican" + 0.004*"presid" + 0.003*"hous" + 0.003*"new

Topic: 1
Words: 0.006*"water" + 0.005*"former" + 0.005*"trump" + 0.004*"twitter" + 0.004*"week" + 0.003*"one" + 0.003*"first" + 0.003*"texas" + 0.003*"leader

Topic: 2
Words: 0.006*"trump" + 0.005*"man" + 0.004*"said" + 0.004*"jackson" + 0.004*"new" + 0.004*"court" + 0.003*"show" + 0.003*"black" + 0.003*"want" + 0.

Topic: 3
Words: 0.011*"trump" + 0.006*"russian" + 0.005*"ukraine" + 0.005*"presid" + 0.005*"new" + 0.004*"former" + 0.004*"donald" + 0.004*"said" + 0.004*"hc

Topic: 4
Words: 0.008*"said" + 0.005*"time" + 0.004*"gop" + 0.004*"get" + 0.004*"russian" + 0.004*"new" + 0.004*"trump" + 0.003*"r" + 0.003*"democrat" + 0.00

Topic: 5
Words: 0.007*"star" + 0.004*"kardashian" + 0.004*"children" + 0.004*"home" + 0.004*"gun" + 0.003*"said" + 0.003*"lawsuit" + 0.003*"million" + 0.002*

Topic: 6
Words: 0.005*"said" + 0.004*"trump" + 0.004*"russian" + 0.004*"fight" + 0.004*"million" + 0.003*"new" + 0.003*"make" + 0.003*"work" + 0.003*"host" +
```

FIG 5.2.1 : The above figure displays the topics derived from LDA

5.2.2 Result of BERTopic

	Topic	Count
0	-1	367
1	0	109
2	1	80
3	2	78
4	3	49
5	4	47
6	5	42
7	6	42
8	7	27
10	8	22
9	9	22
11	10	20
12	11	20
13	12	16
14	13	13
15	14	12
16	15	12
17	16	11
18	17	11

FIG 5.2.2 : The above figures displays the topics derived from BERTopic

5.2.3 Words clouds of both LDA and BERTopic

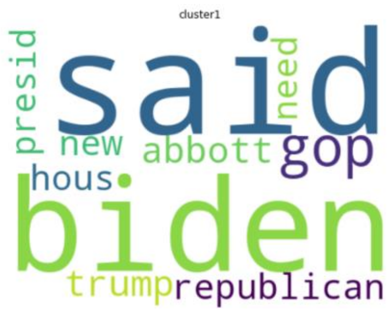

FIG 5.2.3a : cluster1 topics word cloud of LDA **FIG 5.2.3b: cluster1 topics word cloud of BERTopic**

5.3 Results of Dynamic Bert

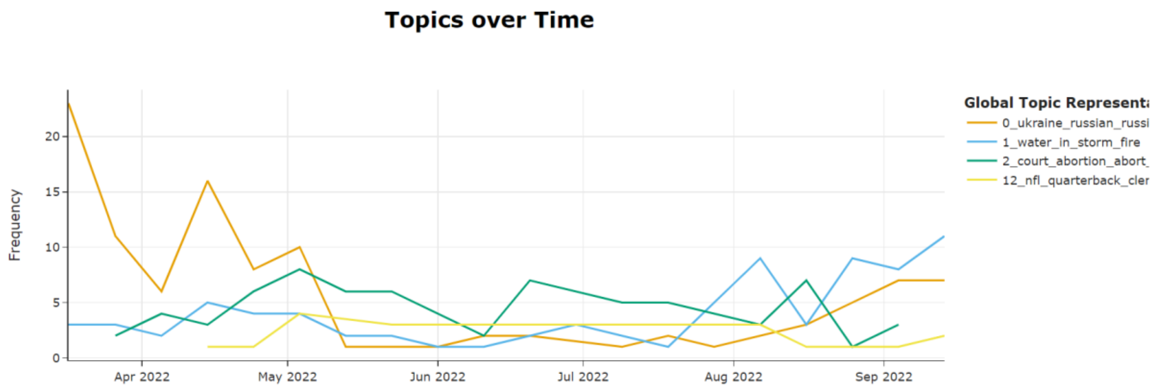

FIG 5.3: Visualized results of topics derived from BERTopic using Dataset2

6. Discussion:

As we mentioned before we started our experiments with two topic modeling techniques i.e., LDA and BERTopic. We sliced the dataset and experimented with the first 1000 rows of both Dataset1 and Dataset2 this is because when we ran the dataset with large amounts of data LDA and BERTopic stopped working due to the limitations of the computational power and on the other hand we ran the same for 1000 rows we were able to get the topics using both LDA and BERTopic. We noticed when comparing the topics from each model, LDA derived fewer topic modeling results and BERTopic outperformed. Also, we have experimented with two machine learning models i.e., passive-aggressive classifier and linear regression model to check if the news is real or fake, and segregated the data into 80 and 20 for training and testing the model trained it. We noticed that passive aggressive model outperformed in accuracy when compared to the other model.

7. Conclusion and Future Work:

Out of LDA and BERTopic, for smaller amounts of data BERTopic derives more relevant topic clusters when compared with LDA. Also, generates topics without human intervention, allows for multilingual analysis, contains built in functions for different purposes and uses embeddings where there is no requirement of pre-processing the data. With all the advantages that we have with BERTopic, this model can be used for short text data thereby working as a tool for topic modelling. Future work should continue in finding the best topic modelling tools across all platforms and can work on reducing the outliers that are being generated in BERTopic.

8. References:

[1]. Marina Danchovsky Ibrishimova and Kin Fun Li. A Machine Learning Approach to Fake News Detection Using Knowledge Verification and Natural Language Processing. L. Barolli et al. (Eds.): INCoS 2019, AISC 1035, pp. 223–234, 2020.

[2]. Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita. Fake News Detection Using Machine Learning approaches: A systematic Review. Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8.

[3]. Julio C. S. Reis, Andre Correia, Fabricio Murai, Adriano Veloso, and Fabricio Benevenuto. Supervised Learning for Fake News Detection. Erik Cambria, Nanyang Technological University, Singapore.

[4]. Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake News Detection Using Machine Learning Ensemble Methods. https://doi.org/10.1155/2020/8885861.

[5]. Junaed Younus Khan, Md. Tawkat Islam Khondaker, Anindya Iqbal1 and Sadia Afroz. A Benchmark Study on Machine Learning Methods for Fake News Detection. https://www.snopes.com/fact-check/soros-caravan-refugees/.

[6]. Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq. Detecting Fake News using Machine Learning and Deep Learning Algorithms. 2019 7th International Conference on Smart Computing & Communications (ICSCC).

[7]. Jia Ding, Yongjun Hu, Huiyou Chang. BERT-Based Mental Model, a Better Fake News Detector | Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence (acm.org).

[8]. Uma Sharma, Sidarth Saran, Shankar M. Patil. Fake News Detection using Machine Learning Algorithms. International Journal of Engineering Research & Technology (IJERT).

[9]. Mykhailo Granik, Volodymyr Mesyura. Fake News Detection Using Naive Bayes Classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON).

[10]. Wenlin Han and Varshil Mehta. Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation. 2019 IEEE International Conference on Industrial Internet (ICII).

[11]. Stefan Helmstetter, Heiko Paulheim. Weakly Supervised Learning for Fake News Detection on Twitter. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[12]. Vaishali Vaibhav Hirlekar, Dr.Arun Kumar. Natural Language Processing based Online Fake News Detection Challenges – A Detailed Review. Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020) IEEE Conference Record # 48766; IEEE Xplore ISBN: 978-1-7281-5371-1.

[13]. Anjali Jain , Avinash Shakya , Harsh Khatter , Amit Kumar Gupta. A SMART SYSTEM FOR FAKE NEWS DETECTION USING MACHINE LEARNING. 2019 2nd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT).

[14]. Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. https://doi.org/10.1007/s11042-020-10183-2.

[15]. Z Khanam, B N Alwasel, H Sirafi and M Rashid. Fake News Detection Using Machine Learning Approaches. IOP Conf. Series: Materials Science and Engineering.

[16]. Reza Mansouri, Mahmood Naderan-Tahan, Mohammad Javad Rashti. A Semi-supervised Learning Method for Fake News Detection in Social Media. 2020 28th Iranian Conference on Electrical Engineering (ICEE).

[17]. Arup Baruah, Kaushik Amar Das, Ferdous Ahmed Barbhuiya, and Kuntal Dey. Automatic Detection of Fake News Spreaders Using BERT. Notebook for PAN at CLEF 2020.

[18]. Jasmine Shaikh, Rupali Patil. Fake Detection Using Machine Learning. 978-1-7281-8880-5/20/$31.00 ©2020 IEEE.

[19]. Wesam Shishah. Fake News Detection Using BERT Model with Joint Learning. https://doi.org/10.1007/s13369-021-05780-8.

[20]. Shlok Gilda. Evaluating machine learning algorithms for fake news detection. In Research and Development (SCOReD), 2017 IEEE 15th Student Conference on, pages 110–115. IEEE, 2017.

[21]. Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, page 82. American Society for Information Science, 2015.

[22]. Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2):211–36, 2017.

[23]. Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fake news. [Online]. Available: https://www.wired.com/ 2016/12/bittersweet-sweepstakes-build-ai-destroys-fake-news/.

[24]. Y. Seo, D. Seo, and C. S. Jeong, FaNDeR: Fake News Detection Model Using Media Reliability. IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2018–October, no. October, pp. 1834–1838, 2019.

[25]. Zhang, J., Cui, L., Fu, Y., & Gouza, F. B. (2018). Fake news detection with deep diffusive network model. arXiv preprint arXiv:1805.08751.

[26]. Yang, K.; Niven, T.; Kao, H.: Fake news detection as natural language inference. arxiv:1907.07347.

[27]. Okano, Y.E.; Liu, Z.; Ji, D.; Ruiz, S.E.: Fake news detection on Fake.Br using hierarchical attention networks. Int. Conf. Comput. Process. Portug. Lang. pp. 143–152 (2020).

[28]. Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H.: dEFEND: explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 395–405 (2019). https://doi.org/10.1145/3292500. 3330935.

[29]. Singhania, S.; Fernandez, N.; Rao, S.: 3han: a deep neural network for fake news detection. Int. Conf. Neural Inform. Process. (2017).

[30]. Umer, M.; Imtiaz, Z.; Ullah, S.; Mehmood, A.; Choi, S.G.; On, B.W.: Fake news stance detection using deep learning architecture (CNN-LSTM). IEEE Access. 8, 156695–156706 (2020). https:// doi.org/10.1109/ACCESS.2020.3019735.

[31]. Liu, Y.; Wu, Y.B: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. Proceedings of the thirty-second AAAI conference on artificial intelligence, pp. 354–361 (2018).

[32]. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B.: A stylometric inquiry into hyperpartisan and fake news. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 231–240 (2018).

[33]. Rubin., Victoria, L., et al.: Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of NAACLHLT (2016).

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

[35]. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. SIGKDD Explorations 19(1), 22–36 (2017), https://doi.org/10.1145/3137597.3137600.

[36]. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020).

[37]. Cardoso Durier da Silva, F., Vieira, R., & Garcia, A. C. (2019, January). Can machines learn to detect fake news? a survey focused on social media. In Proceedings of the 52nd Hawaii International Conference on System Sciences.

[38]. Sharma, Karishma, et al. "Combating fake news: A survey on identification and mitigation techniques." ACM Transactions on Intelligent Systems and Technology (TIST) 10.3 (2019): 1-42.

[39]. I. Traore et al. "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques." International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (pp. 127–138). Springer International Publishing . 2017.

[40]. Looijenga, M. S. "The Detection of Fake Messages using Machine Learning." 29 Twente Student Conference on IT, Jun. 6th, 2018, Enschede, The Netherlands. Netherlands: essay.utwente.nl. 2018.

[41]. Prabhjot Kaur et al. " Hybrid Text Classification Method for Fake News Detection." International Journal of Engineering and Advanced Technology (IJEAT) , 2388-2392. 2019.

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

[43]. MykhailoGranik and VolodymyrMesyura. “Fake news detection using naive Bayes classifier.” First Ukraine Conference on Electrical and Computer Engineering (UKRCON). Ukraine : IEEE. 2017.

[44]. Rampersad G, Althiyabi T 2020 “Fake news: Acceptance by demographics and culture on social media” J. Inf. Technol. Politics 2020, 17, 1–11.

[45]. Alkhodair S A, Ding S H.H, Fung B C M, Liu J 2020 “Detecting breaking news rumors of emerging topics in social media” Inf. Process. Manag. 2020, 57, 102018.

[46]. Shivam B. Parikh ; Saurin R. Khedia ; Pradeep K. Atrey, “A Framework to Detect Fake Tweet Images on Social Media,” in the proceeding of IEEE Fifth International Conference on Multimedia Big data (BigMM), 2019.

[47]. Zhixuan Zhou; Huankang Guan1; Meghana Moorthy Bhat; Justin Hsu; “Fake News Detection via NLP is Vulnerable to Adversarial Attacks,” in the proceeding of 11th International Conference on Agents and Artificial Intelligence (ICAART, 2019.

[48]. Amitabha Dey ; Rafsan Zani Rafi ; Shahriar Hasan Parash ; Sauvik Kundu Arko ; Amitabha Chakrabarty, “Fake News Pattern Recognition using Linguistic Analysis,” in the Proceeding of IEEE Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2018.

[49]. Kuai Xu; Feng Wang; Haiyan Wang; and Bo Yang , “Detecting Fake News Over Online Social Media via Domain Reputations and Content Understanding,” in the proceeding of Tsinghua Science and Technology, Vol.25 , Issue: 1 , Feb. 2020.

[50]. S. Gilda, ”Evaluating machine learning algorithms for fake news detection,” 2017 IEEE 15th Student Conference on Research and Development (SCOReD), Putrajaya, 2017, pp. 110 – 115.

[51]. M. Granik and V. Mesyura, ”Fake news detection using naive Bayes classifier,” 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900 - 903.

[52]. M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, ”Automatic Online Fake News Detection Combining Content and Social Signals,” 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, Finland, 2018, pp. 272 - 279.

[53]. Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake News Detection on Social Media using Geometric Deep Learning.. arXiv: Social and Information Networks,.

[54]. D. M. J. Lazer et al., “The science of fake news,” Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

[55]. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” ACM SIGKDD Explorations Newslett., vol. 19, no. 1, pp. 22–36, 2017.

[56]. Choudhary, A., Arora, A., Linguistic Feature Based Learning Model for Fake News Detection and Classification, Expert Systems with Applications (2020), doi: https://doi.org/10.1016/j.eswa. 2020.114171

[57]. Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach," SMU Data Science Review: Vol. 1: No. 3, Article 10. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/10

[58]. Vasu Agarwal, H. Parveen Sultana, Srijan Malhotra, Amitrajit Sarkar, Analysis of Classifiers for Fake News Detection, Procedia Computer Science, Volume 165, 2019, Pages 377-383, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.035.

[59]. R. Mansouri, M. Naderan-Tahan and M. J. Rashti, "A Semi-supervised Learning Method for Fake News Detection in Social Media," 2020 28th Iranian Conference on Electrical Engineering (ICEE), 2020, pp. 1-5, doi: 10.1109/ICEE50131.2020.9261053.