





Table of Contents

**[Executive Summary](#_Toc102176867)** [3](#_Toc102176867)

[**Project motivation/background** 4](#_Toc102176868)

[**Importance** 7](#_Toc102176869)

[**Problem statement** 7](#_Toc102176870)

[**Data statistics** 8](#_Toc102176871)

[**Data preparation activities** 11](#_Toc102176872)

[**Data exploration** 11](#_Toc102176873)

[**Exploring missing values in dataset** 13](#_Toc102176874)

[**Visualizing dataset for distribution of class labels** 14](#_Toc102176875)

[**Exploring the dataset for total number of true and fake articles** 15](#_Toc102176876)

[**Exploring dataset for variety of subjects** 15](#_Toc102176877)

[**Plotting a count of all subjects in the news** 15](#_Toc102176878)

[**Visualizing the news coverage of each topic** 16](#_Toc102176879)

[**Plotting a count of all subjects in the fake news** 17](#_Toc102176880)

[**Exploring for imbalanced data** 18](#_Toc102176881)

[**Data pre-processing** 19](#_Toc102176882)

[**Data Cleaning** 19](#_Toc102176883)

[**Converting the data into lowercase** 20](#_Toc102176884)

[**Removal of punctuation** 21](#_Toc102176885)

[**Removal of stop words** 21](#_Toc102176886)

[**Tokenization** 22](#_Toc102176887)

[**Data Analytics** 23](#_Toc102176888)

[**Visualization Technique** 24](#_Toc102176889)

[**Algorithms Used** 26](#_Toc102176890)

[**Naïve Bayes** 26](#_Toc102176891)

[**Logistic Regression** 27](#_Toc102176892)

[**Support Vector Machine** 27](#_Toc102176893)

[**Decision Tree** 27](#_Toc102176894)

[**Random Forest** 28](#_Toc102176895)

[**Performance Metrics** 29](#_Toc102176896)

[**Analysis of Results** 30](#_Toc102176897)

[**Splitting the dataset** 30](#_Toc102176898)

[**Applying Naïve Bayes** 31](#_Toc102176899)

[**Applying Logistic Regression** 31](#_Toc102176900)

[**Applying SVM** 32](#_Toc102176901)

[**Applying Decision Tree** 32](#_Toc102176902)

[**Applying Random Forest** 33](#_Toc102176903)

[**Findings** 33](#_Toc102176904)

[**Plotting Results** 33](#_Toc102176905)

[**Comparing results using confusion matrix** 35](#_Toc102176906)

[**Business Implications/Intelligence** 37](#_Toc102176907)

[**Conclusion** 38](#_Toc102176908)

[**References** 38](#_Toc102176909)

# **Executive Summary**

We have witnessed a dramatic change in our society as a result of the emergence and use of digital platforms for the production, dissemination, and consumption of news. The challenge for digital platforms in the age of disinformation today is the fact that they have become a forum for campaigns of misinformation that have direct and indirect effects on the credibility of all of the news ecosystems. The emergence of fake news has fast become an increasingly troubling worldwide phenomenon, particularly in the absence of scalable platforms for fact-checking that are affordable and scalable. The use of automatic solutions for detecting fake news can thus be used as a supplementary tool to determine which article is likely to be fake, or which article merits further investigation by fact-checkers. There are many fake news articles written on purpose in an attempt to fool the public into believing the propaganda, making it difficult to detect since it is written with the intent of mislead the public (Mahir et al., 2019). There's no denying that fake news has been detrimental to the general public's mindset. In the modern age of fake news in the media it is important to check whether the information is authentic, due to the widespread spread of fake news on the internet. There has been a lot of speculation that fake news can have to significantly affect society. To tackle this problem, we propose a method that uses machine learning to detect fake news. The purpose of this project is to explore practical approaches for discovering fake news in the context of the digital platform context, by investigating how these can be automated and detected. The first step in our research is to make a survey of a large number of articles on fake news that have recently been published as an attempt to implement all possible detection mechanisms. The purpose of our study is to analyse labelled datasets and propose two novel features that are useful in assessing prediction performance in supervised machine learning approaches. Our dataset is then used to vectorize the title of the news article and to analyse the words using the tokens of the news title. We are using a dataset that comprises a predefined set of curated news events that are either fake news or they are not fake news. It is our goal to create a model which distinguishes between genuine and fake articles based on a given article.

# **Project motivation/background**

In this digital age, more than one-third of the global population uses digital platforms, such as social media networks and messaging programs to stay connected (Gallagher, 2017). These platforms have drastically changed the way in which users interact with each other and communicate over the Internet, providing the opportunity to create entirely new types of applications, as well as modifying the existing information eco system. A particularly important change in how news is produced, distributed, and consumed has been the growth of digital platforms, which offers many opportunities, but also introduces many challenges as well.

As a consequence of the characteristics of these digital platforms, this change has been spurred by a number of factors: (i) The cost and time constraints for producing and consuming news on these platforms are often less and faster than those for traditional news sources, as they are for newspapers and telecommunications; additionally, viewers are able to share news, comment on it, and participate in discussions on digital platforms, which allows them to interact and communicate with their friends and fellow readers through the use of social media (Shu et al., 2017). Hence, we can say that digital platforms are creating a paradigm shift in how information is consumed by people. About 62% of the American population and 66% of the Brazilian population nowadays receive their news through digital channels (Mitchell, 2016). Even though these systems have the capability to provide numerous benefits to our society, they have with time become a source of a lot of misinformation that is both inaccurate as well as deceptive, which is especially apparent in various contexts, including health and politics. There is no denying that the flood of fake news, which is being spread on digital platforms, has severely harmed the health of the world. An example is that an online advertisement for an experimental cancer treatment was interpreted as medically valid information by a cancer patient, and ultimately the cancer patient died (Dai et al., 2020). As a result of the COVID-19 pandemic, we have seen an increase in the occurrence of rumours and conspiracy theories spreading across the internet.

Over three thousand false claims have been detected by the International Fact-Checking Network in less than two months about COVID-19 (Poynter, 2020). It is estimated that there were at least 800 deaths worldwide as a result of misinformation about the Coronavirus during the first three months of 2020. During the last election cycle, we have seen the appearance of conflicts of interest, political corruption and the dissemination of false information as a means to influence public opinion. It has been more than a year since the US presidential elections of 2016. There was a huge "misinformation war" during the elections that took place mainly via Facebook and Twitter. There was a case in Russia in which Russians targeted advertisements with the intention of influencing public opinion (Ribeiro et al., 2019). On the eve of the Brazilian elections in 2018, a similar campaign was carried out, in which WhatsApp was exploited for the purpose of sending misinformation campaigns, which made use of heavily edited images and memes containing egregious political attacks. It was recently found that 88% of the images that were shared in Brazil in the days leading up to the presidential election in that country were fabricated or misleading in some fashion (Tardáguila et al., 2018). The use of WhatsApp was also responsible for spreading fake rumours through the online service in India that caused multiple lynching and social disorder. In the context of digital news platforms, one of the things that further contributes to the emergence of fake news is that almost anyone can set up/appoint themselves as a news publisher free of cost (for example, an individual who believes they are a news media organization or newspaper can create a Facebook page to spread news, or they can start a WhatsApp group for spreading fake news).

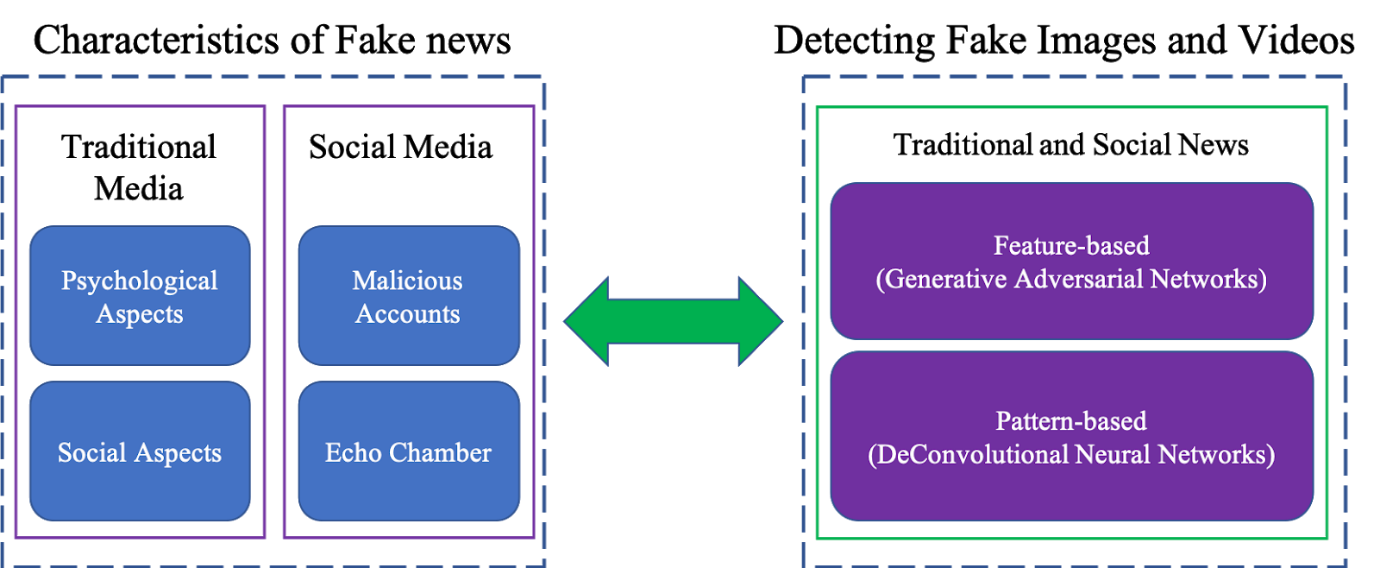
Because of this, it is not only traditional news organisations that are increasingly moving towards digital platforms, but many other news outlets are also establishing themselves on these platforms in the process. It should be noted, for instance, that previous studies have found that during the course of 2018, Facebook in the United States had grown to more than 20,000 pages that publish and share news stories (Ribeiro et al., 2018), a figure which continues to rise steadily. Furthermore, there is concern that fake news publishers are creating fake news stories and posting them on the internet, and disseminating them widely via social media platforms. In another case, a study funded by Avaaz6 included a survey that attempted to measure the spread of fake news in Brazil by asking Brazilian citizens how many of the most popular false stories were spread over social media within the last few weeks of the 2018 Brazilian elections. A surprising number of findings were revealed in the survey, where it was shown that over 98% of interviewees had been exposed to at least one fake news account, and over 90% of the respondents believed that the news stories were true.

The consequence of these factors is that both the results of the 2018 elections in Brazil and the democratic process in Brazil suffered as a result. The number of false allegations, deceptions, spins, and lies has existed for a very long time, but with the emergence of digital platforms, misinformation has become more and more widespread, causing the issue of fake news to become a concern that affects the whole world, especially when a lack of research-based fact-checking strategies is a cause for concern.

Thus, in the scenario presented here, we hypothesize that by detecting fake news automatically, we can be able to use a useful level of discrimination in the process of identifying what is more likely to be fake content, thereby supporting the process of fact-checking while also minimizing the unintended consequences of fake information widely distributed through digital channels.

The most important piece of information to understand when working on fake news detection is the concept of what fake news is and how it can be distinguished from real news. It is important to be able to differentiate the two in one of two ways: either characterization or detection. To create a detection model, characterization is critically important. In other words, in order to detect fake news, one must first be able to identify what is fake news, before trying to build a detection model.

The definition of fake news is based on two components: the legitimacy of the news and the intention behind it. By authenticity, we mean that fake news contains false information which can be verified to be false, so this implies that conspiracy theories are not part of fake news since in most cases it is difficult to prove whether an assertion is true or false. As for the second part, intent, assumes that the false information is written with the intent of mislead the reader.



## **Importance**

There is no doubt that social media offers a great platform for broadcasting news to the network at an exponential rate. This is a great source for information sharing. The fact that we see the news and the social media today does not imply that everything we see is accurate. It has been shown that words, photos, and videos are altered so as to appear as if they were subjective to   propaganda.   Since Twitter stands out as one of the most well-known and well-known continuing news sources, it also shows up as one of the most prominent news dissemination mediums.

We are living in a social era where understanding the difference between right and wrong, When, what, and why of any news is important to understand. Awareness of fake news detection and classification is slowly becoming a fundamental necessity for society in order to avoid the phenomenon of reality vertigo, as well as to protect less educated members of society. There have been numerous approaches proposed to address the issue of machine learning.

Our socio-economic life is affected by fake news in a significant way, especially our political environment. The detection of fake news is an emerging area of research, and it has gained considerable interest, but it has also been hampered by a limited amount of available resources (for example, datasets and published literature).

## **Problem statement**

The term fake news refers to an informational practice or propaganda that is based on deliberate misinformation and hoaxes which have spread through traditional print news media as well as television or radio news broadcasts or online social media. There has long been a problem of fake news in the news industry - the production of misinformation and propaganda as well as hoaxes and satire have always been a concern. There are no boundaries anymore when it comes to what can be published on the Internet. This includes anything credible or not credible. Therefore, this exposes people to the possibility of being deceived intentionally or unintentionally, and they are more likely to share such types of news without reason to the ends of the earth. With machine learning and artificial intelligence, we may be able to overcome the counterfeiting of news problem or at least make it less common. It has been concluded that the detection of fake news is usually considered a challenging task (Manzoor et al., 2019), requiring the application of a multidisciplinary approach. For deception detection, there exists a large body of research (Khanam et al., 2021) done where machine learning methods are applied.

It was thought that these methods would categorize news reported online and posts made on social media sites, but, after United States Presidential elections in 2016, researchers have also been examining methods for determining fake news from legitimate news sources. Content-based classification and tagging of n-grams and parts of speech (POS) tagging has proven to be inadequate when used in the context of fake news (Kansal, 2021). Although the classification of fake news is not sufficient to detect them as it overlooks the important context of information in the news, a thorough examination of the content may be useful to detect false news. With the combination of the context-free grammar software (CFG) and the N-gram program, context-free grammar produced good predictions for deception classification. On the basis of a classification algorithm, 85%-91% accuracy was achieved when applied to news article datasets (Sharma et al., 2019). Taking the subject of classification in this context, it is our hypothesis that simple classification alone will not be sufficient to solve the problem; we need to employ machine learning techniques together. This hypothesis will be proven by developing a model on publicly available datasets, after several experiments, based on the hypothesis.

## **Data statistics**

The dataset we are using is downloaded from Kaggle(*Fake and Real News Dataset | Kaggle*, n.d.)

We have a data set which contains all of the news stories around Real or Fake News or Text Datasets.  
There are only four columns in this data set.

1. title
2. text
3. subject
4. date

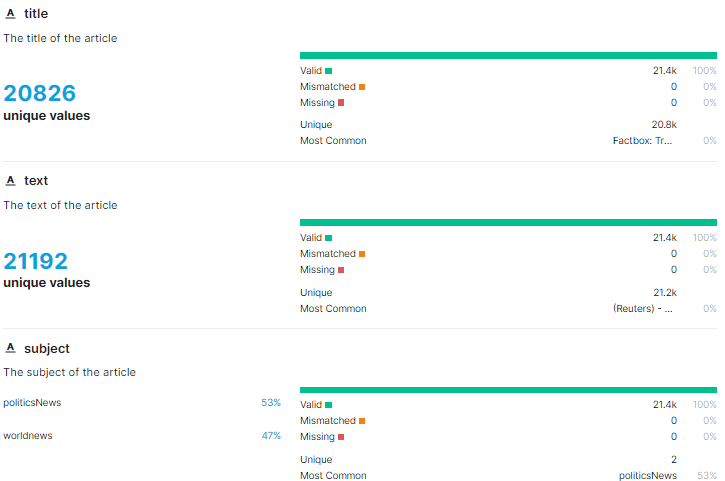
there are two csv files in the dataset.

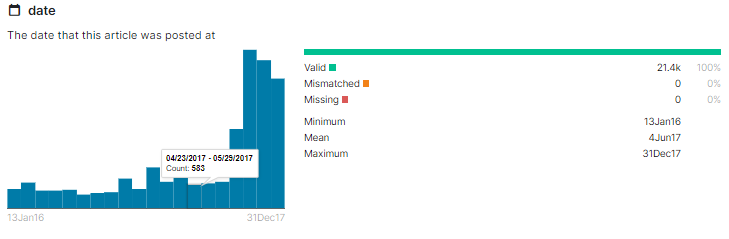
1. True.csv
2. Fake.csv

The statistics of both files are shown below:

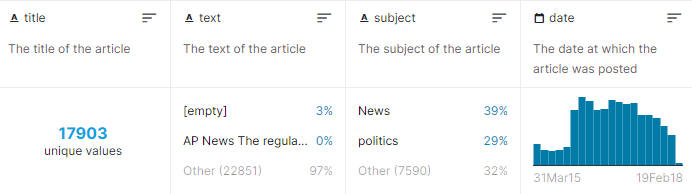
**True.csv**

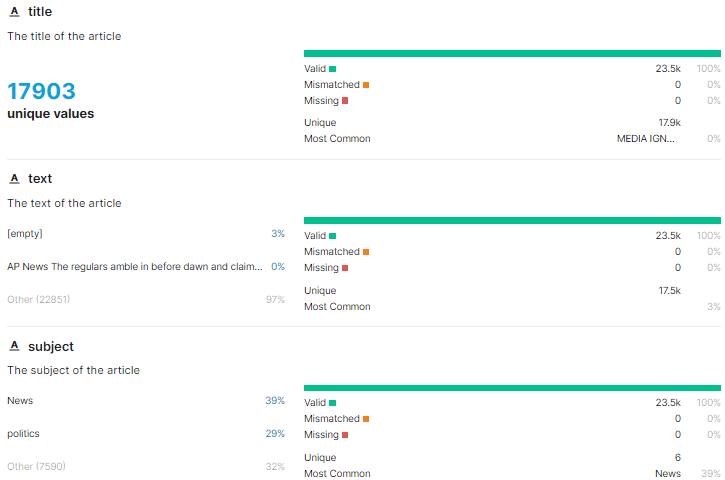


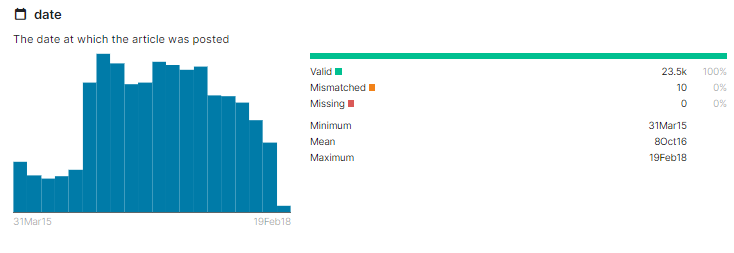




**Fake.csv**







# **Data preparation activities**

The machine learning algorithm can only process numerical features so the text data must be transformed into numerical columns so that the algorithm will work. The text must be pre-processed in order to make sense of it, and this process is known as Natural Language Processing.

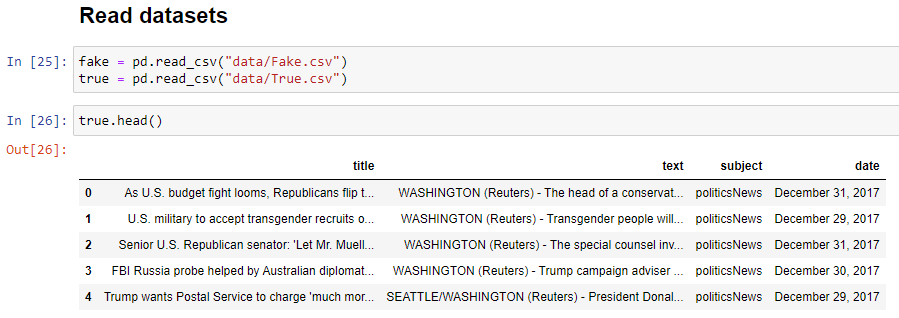
As part of text pre-processing, we will clean our text using steaming, lemmatization, removing stop words, removing special symbols and numbers from the text, etc. The data has to be cleaned first and then we have to feed these text data into a vector evaluator that will turn it into a numerical feature.

The datasets are downloaded from Kaggle and then we analyse them. One dataset deal with fake news, while the other will deal with true news. there are 21417 news articles in true news, and there are 23481 in fake news. In both datasets, there is a label column with 0 representing true news and 1 representing fake news. These two datasets are then combined using a pandas built-in library of python.

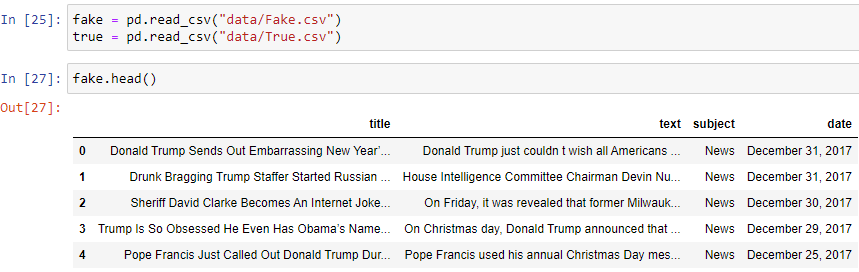
## **Data exploration**

First, we read the datasets. As there are two datasets. We will look into them one by one.

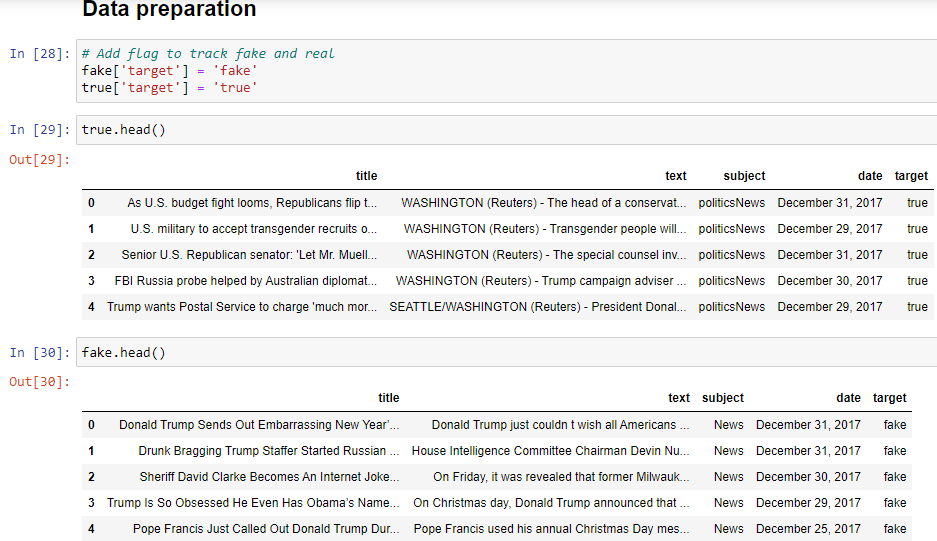
**Dataset with true news**



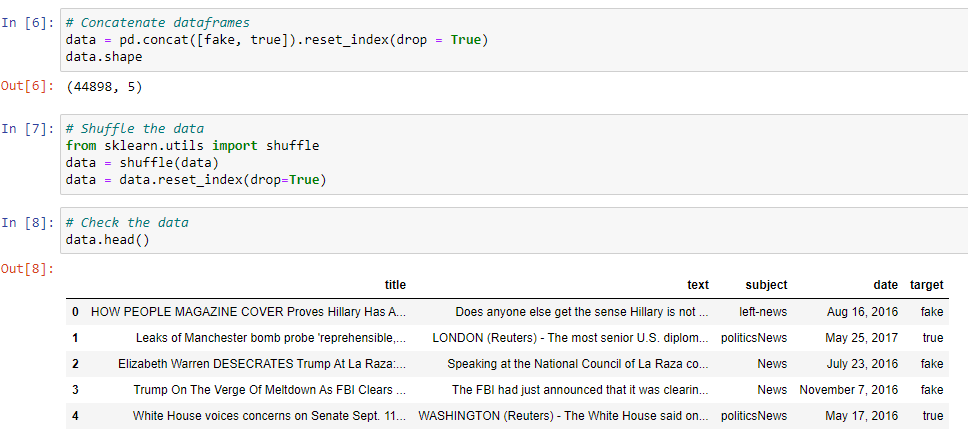
**Dataset with fake news**



As for classification in machine learning, we need a target variable. In these datasets we will create a variable and name it as “target”. For true news it will be set as true. For fake news, it will be set as “false”.

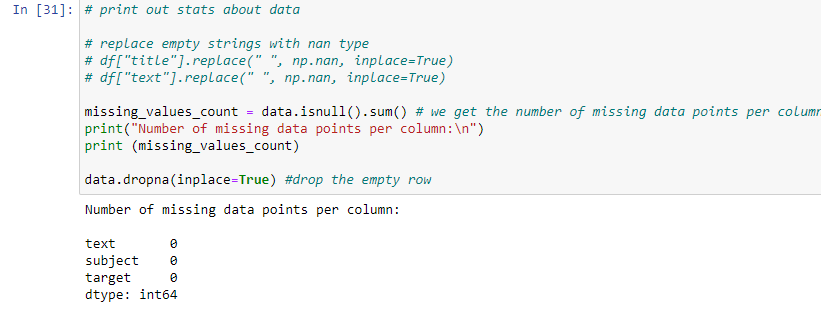


After setting the values in target variable in both datasets, we will combine them in one file as follows:



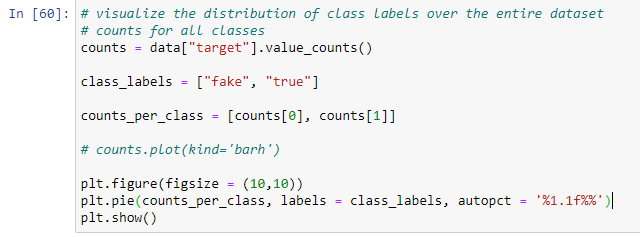
### **Exploring missing values in dataset**

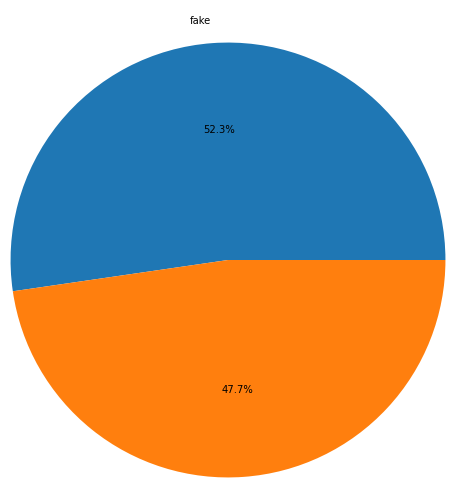
We will now explore if there are any missing values in the dataset.



From output, it can be seen that there are no missing values.

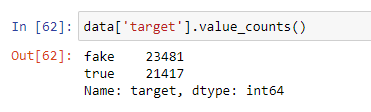
### **Visualizing dataset for distribution of class labels**





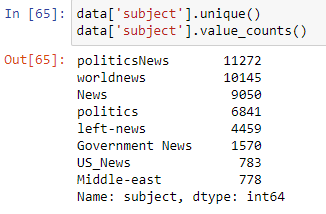
Based on the chart above, we see that the data is fairly evenly distributed amongst the two classes.

### **Exploring the dataset for total number of true and fake articles**

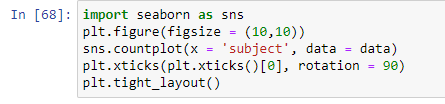


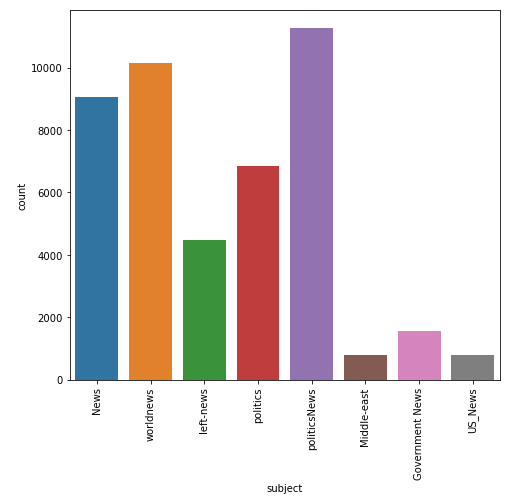
There are 23481 fake news and 21417 true news in the dataset.

### **Exploring dataset for variety of subjects**

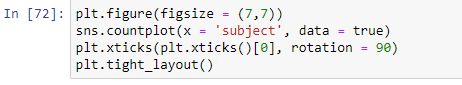


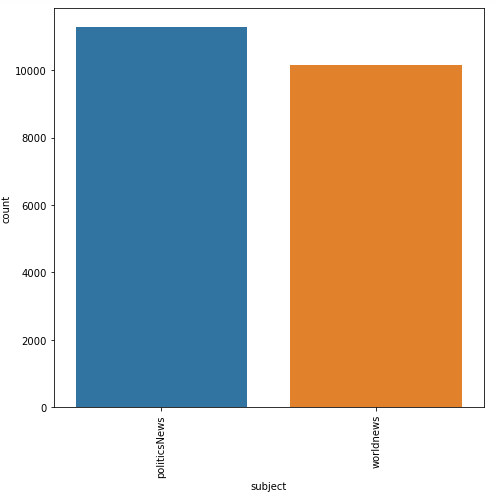
### **Plotting a count of all subjects in the news**



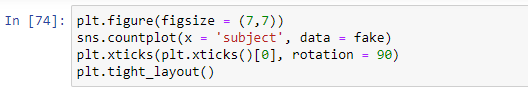


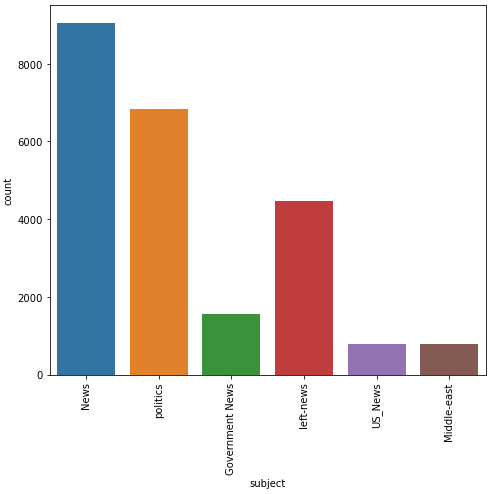
### **Visualizing the news coverage of each topic**





### **Plotting a count of all subjects in the fake news**





### **Exploring for imbalanced data**



It can be seen from graph that data is well balanced.

## **Data pre-processing**

When it comes to data processing, we are mostly going to concentrate on the text column on this data, which is actually what holds the news. In the future, we are going to change this text column in order to extract more information and make the model more accurate. This whole procedure will be carried out with the help of a tool called nltk, a library that enables us to extract information from the text in a column.

Using the nltk library in this case, we will use the functions related to removing stopwords, tokenizing, and lemmatizing.

## **Data Cleaning**

First step is to perform data cleaning which involves following steps;

1. ***Converting Upper case to Lower case***

This step involves converting the upper case to lower case so that each sentence follows same pattern and does not affect the accuracy.

1. ***Removal of punctuations***

This step involves removing all special symbols so that they do not affect accuracy.

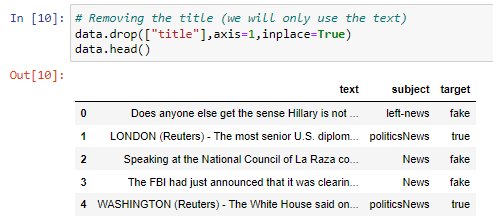
1. ***Removal of stop words***

In almost any language, there are words, pronouncements, and phrases that are used to make connections between words and to indicate the tense of sentences. The consequence for us as users of these words is that when we use them in sentences they don't really add anything to the meaning of the sentences, which means that even when we remove the stop words, we can still understand what is being said.

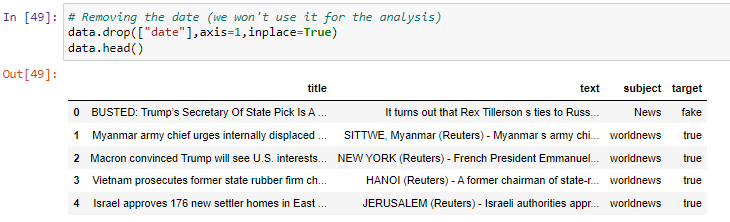
1. ***Tokenization***

By tokenization we mean breaking a text into smaller pieces which are referred to as Tokens. A token in Natural Language Processing may be a word, a special character, or a number found within a sentence.

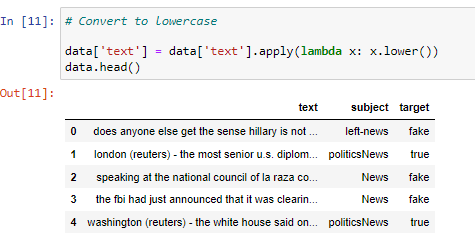
Before data cleaning, we will remove the title feature from dataset as we do not need it. We only use text feature.



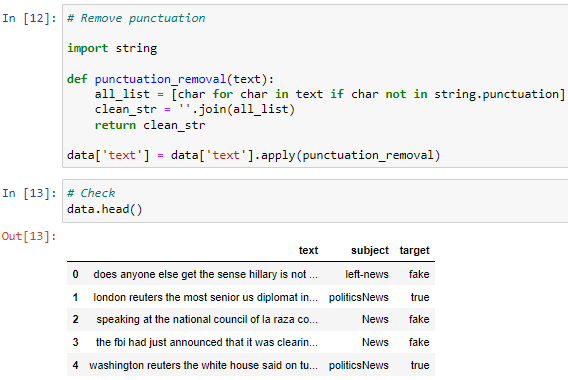
We will remove the date feature from dataset as we do not need it for analysis.



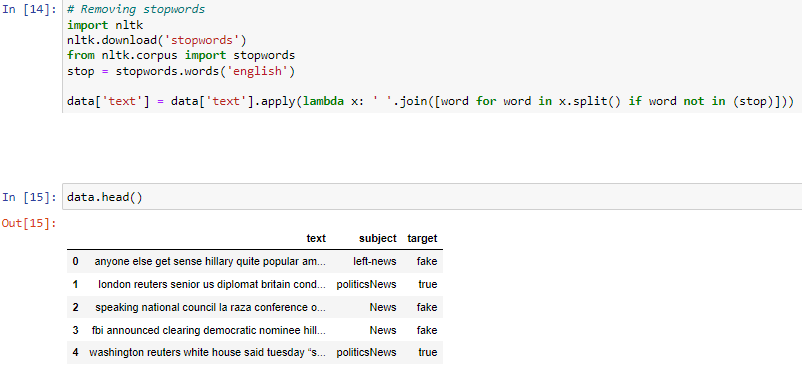
### **Converting the data into lowercase**



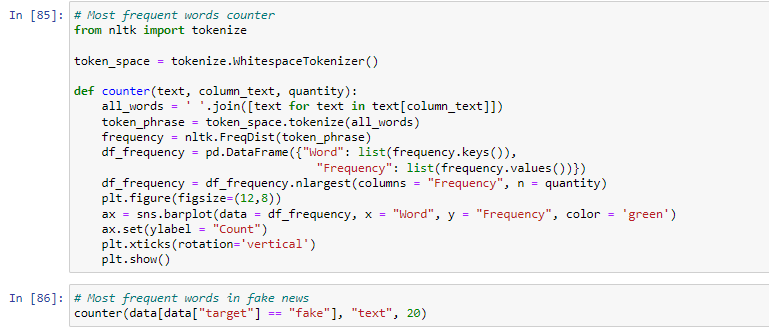
### **Removal of punctuation**

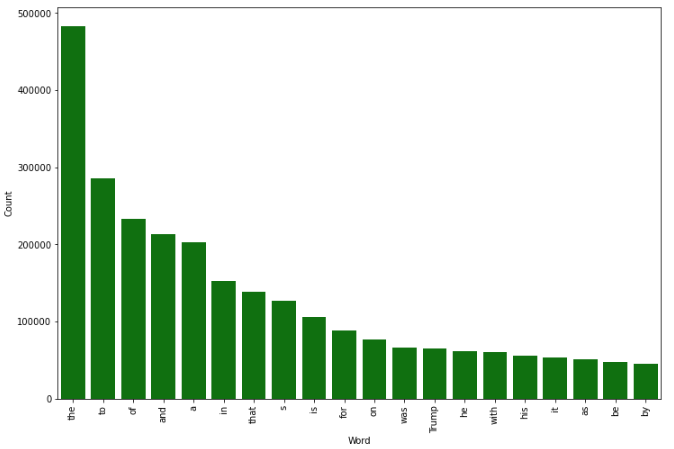


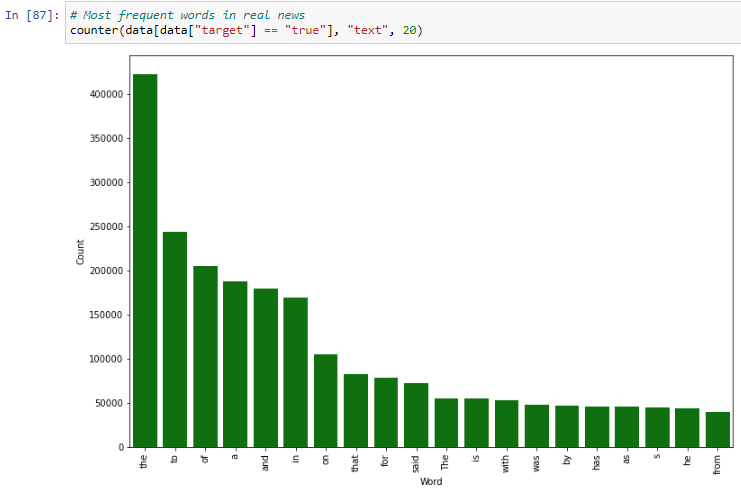
### **Removal of stop words**



### **Tokenization**



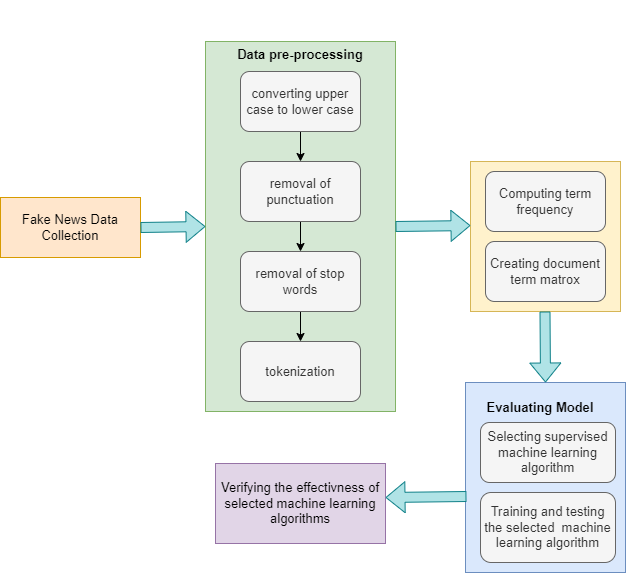




# **Data Analytics**

In our proposed framework, we intend to consolidate on the current literature by combining ensemble techniques with a variety of linguistic features sets in order to classify news articles from multiple domains as true or false. An innovative approach that makes use of the ensemble techniques is the one we have suggested in this research. An ensemble method is a technique that creates multiple models and then combines them in a way that produces improved results at the end. These methods are generally more accurate than trying to solve the problem with a single model.

There are many legitimate news websites that publish full-text news articles that can be used as a source for fact-checking. In addition, there is an open repository that is maintained by researchers so that they can always have access to a detailed list of all the datasets currently available, as well as hyperlinks to potentially useful fact-checking websites that can assist in combating the spread of false news. It is for this reason that we chose three datasets which were collected from a variety of different categories of news. These datasets are free to access online and was extracted from Kaggle. The flow diagram of proposed approach for detection of fake news is presented below:



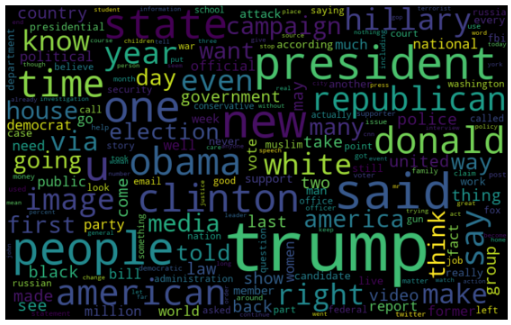
## **Visualization Technique**

Businesses often discover that visualizing data (like graphs, charts, infographics, and so on) is a valuable alternative to presenting important information at a glance.

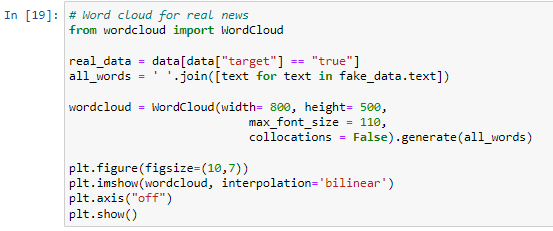
A word cloud can convey crucial information in a visually stunning way and make dull textual data seem more interesting, highlighting important textual data points in a captivating manner. Essentially, word clouds (also referred to as tag clouds or weighted lists) are visual representations of text data. Each of the words that are displayed is normally a single word, and the size and colour of the font indicate their significance. There is a Python library which permits us to create these word clouds.

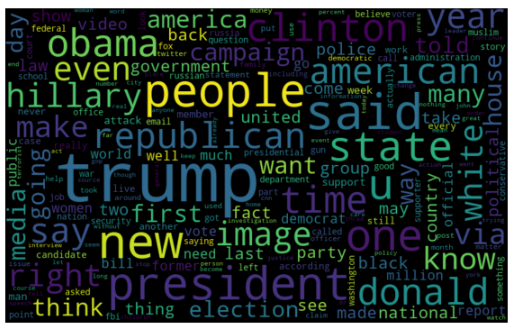
**Word cloud for fake news**





**Word cloud for true news**



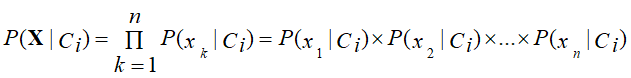


## **Algorithms Used**

We evaluated the performance of fake news detection classifiers based on the following learning algorithms in our proposed methodology.

### **Naïve Bayes**

The Naive Bayes algorithm works according to the Bayes' Theorem, which is based on the conditional probability.



A classifier can be defined by calculating the maximal posterior, which is calculated by taking the maximum P (Ci, X) generated by the Bayesian search with the assumption that it is applied to Bayesian framework. With this assumption, the cost of computation is greatly reduced by counting only the distribution of classes. In spite of the fact that this assumption is not valid in the majority of cases because attributes are not independent, it is nonetheless remarkable the amount of performance that Naive Bayes has managed to achieve.

### **Logistic Regression**

In supervised learning, a logistic regression algorithm predicts the probability of a target variable by using a classification algorithm. It would only be possible to have two classes for a target or dependent variable if it is dichotomous.

To put it simply, the dependent variable is a binary variable, containing data that is coded either as 1 (represents success/yes) or 0 (represents failure/no).

### **Support Vector Machine**

There is a machine learning approach to solving two-group classification problems called support vector machine (SVM). SVM models are capable of categorizing new text by using labels that have been given to them for each category of training data.

it is mostly used in classification problems. As part of the SVM algorithm, we are plotting each item of data as a point in a n-dimensional space. (N is the number of features you have) Where the value of each feature corresponds to the coordinates of the item of data. Then, we conduct classification based on finding the hyperplane that best distinguishes between two classes as follows.



### **Decision Tree**

The Decision Tree algorithm is a supervised learning algorithm.  An effective way to make use of Decision Trees is to create a training model which can then be used with a variety of learning algorithms to predict the target variable by detecting simple decision rules based on prior information (training data).

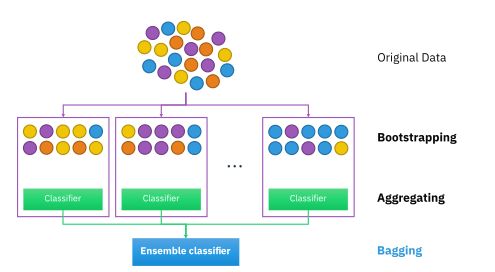
The decision tree method involves starting from the root of the tree to determine the class label for a record. The root attribute values of the record are compared with the values of the root attribute. Then we compare the two values and follow the branch that corresponds to this value and go to the next node in the tree.

### **Random Forest**

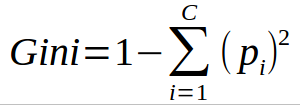
In the field of classification and regression, random forests are one of the most widely used supervised machine learning algorithms. In the case of regression, the decision tree is based on different samples and the majority vote of them will be used to classify the data and determine the average.

The Random Forest algorithm is one of the most important features of the theory that it can perform a classification and regression analysis with data sets consisting of continuous variables in the case of regression and categorical data in the case of classification. The Random Forest algorithm has a better performance when classifying data. This is an ensemble machine learning method called ***Bagging*** or ***Bootstrap Aggregation.***

The ensemble technique of random forests is called bagging. This technique randomly samples data. As a result, each model is generated from them (Bootstrap Samples). Each model is then trained independently, generating results. Based on combining all results, the final result is a majority vote. In an aggregation process, all votes are combined and the result is based on a majority of votes.



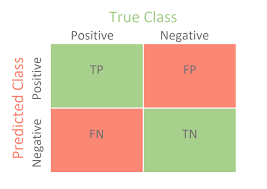
There are different parameters used in training our random forest model; this is based on the fact that a number of estimators were used in order to create a model that could predict the outcome with a high degree of accuracy. There are a number of algorithms that can determine the split in a tree structure based on the type of problem being dealt with, whether it is regression or classification. It is an important fact to note that for the classification problem, a cost function called Gini Index was applied in our model to estimate what amount a dataset will split into.



A Gini Index is calculated by squaring the probabilities of each class and then subtracting that sum from one.

## **Performance Metrics**

The confusion matrix has been used as a tool to evaluate the performance of the algorithms. Confusion matrix is a table that represents the performance of a classification model with respect to a given dataset, which is divided into four categories: true positives, false positives, true negatives, and false negatives.



* **True Positives (TP)**

These are the values that are a correct prediction of positive results. The predicted value must also be true if the actual value of the class is true.

* **True Negatives (TN)**

These are the values which indicate that a prediction was correct, meaning that both the actual and predicted classes are negative.

* **False Positives (FP)**

This refers to situations where a class is accurate, but an incorrect prediction for the same class is produced.

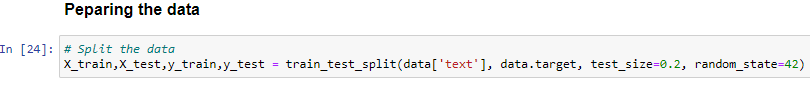
* **False Negatives (FN)**

In this case, the actual class is “yes”, but the predicted class is “no”.

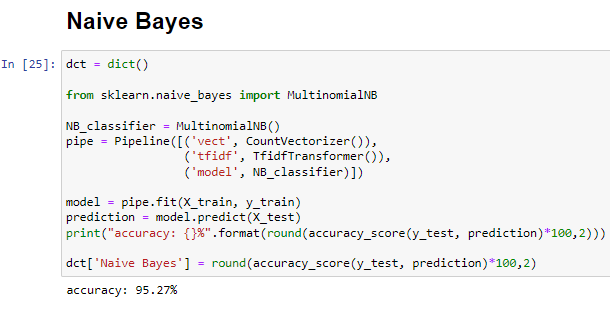
## **Analysis of Results**

Data retrieval is the first step of the process in which data is retrieved. This particular data was downloaded from Kaggle. After this, we split up the data samples into training and testing groups. In our approach we divide data using 80.20 rule. Which means 80 % data is for training and 20% of data is for testing purpose. Then the selected machine learning algorithms are trained on the dataset. In the last step, we evaluate the performance of each algorithm through confusion matrix.

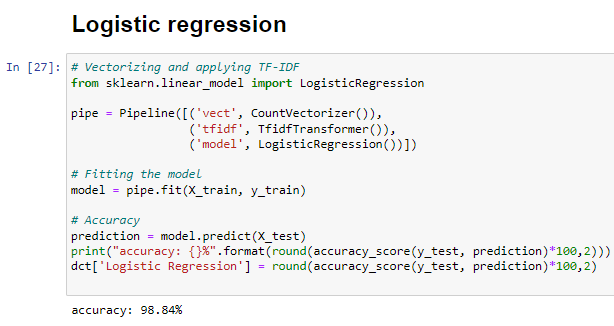
### **Splitting the dataset**



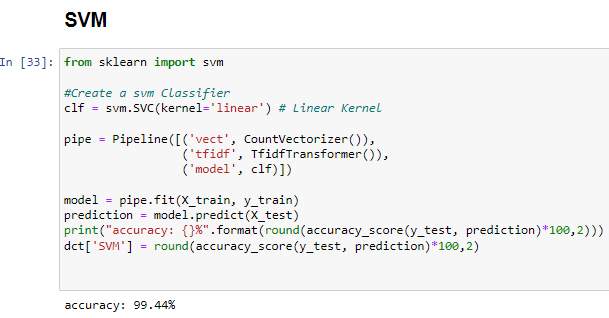
### **Applying Naïve Bayes**



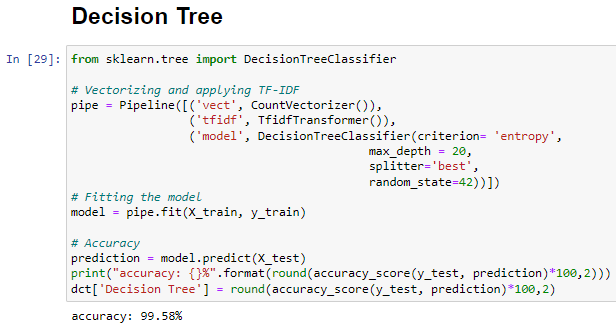
### **Applying Logistic Regression**



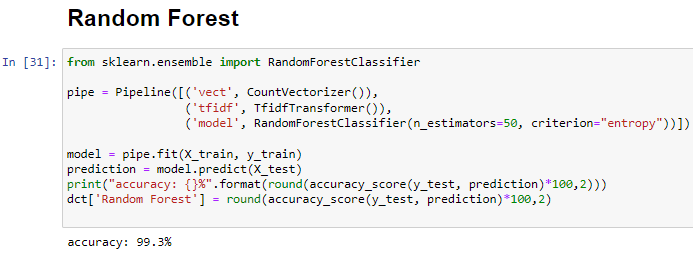
### **Applying SVM**



### **Applying Decision Tree**



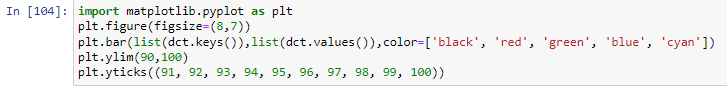
### **Applying Random Forest**

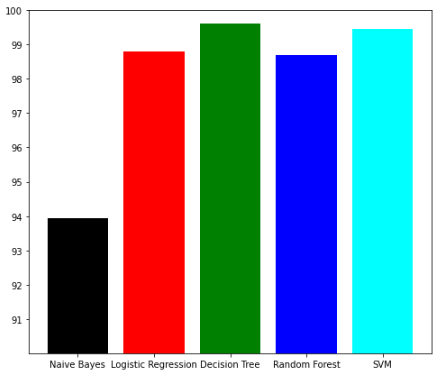


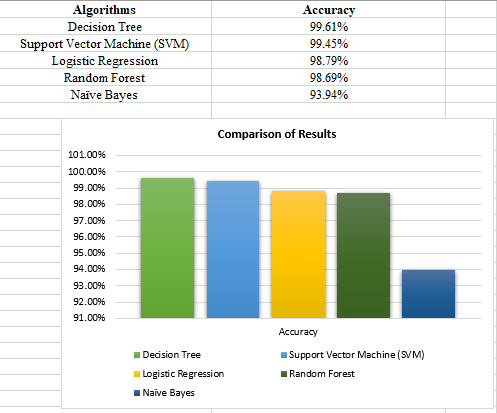
# **Findings**

In this section, we will analyze the results after performing machine learning algorithms on the dataset and will present findings at end of section.

### **Plotting Results**







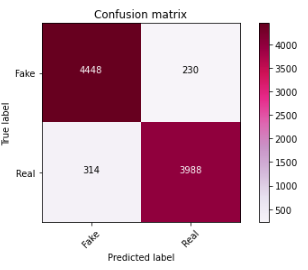
A summary of the accuracy assessed by each algorithm on the final dataset can be seen in the above graph. On the other hand, we can conclude that the Decision Tree has achieved the highest accuracy on a scale of 99.61%. On the other hand, Support Vector Machines (SVM) achieve some of the highest accuracy which is 99.45%. However, Random Forest achieves a 99.69% accuracy which is the next highest accuracy. On the contrary, the level of accuracy of Logistic Regression is seen as the next highest in the class at 98.79%. Among all the simulated models, the Naive Bayes model has the lowest accuracy at 93.94%.

A clear understanding of the subject and expertise in identifying anomalies in the text of news articles is required to classify them manually. This research presents a study that examines the problem of detecting fake news articles by combining machine learning models and ensemble learning techniques. The purpose of the study is to identify patterns in language that distinguish fake news articles from those that are true. We have pre-processed the data and remove the features that were not required. Then we pass those features as an input to the selected machine learning models. We trained and fine-tuned the learning models to obtain optimal performance from them. We have proven that certain models have a higher degree of accuracy than others. We compared the results for each algorithm using a variety of performance metrics.

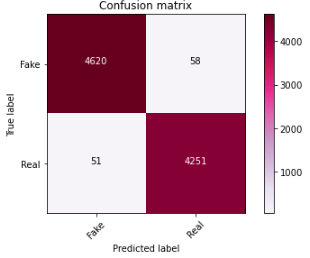
Researchers need to pay attention to many open issues concerning the detection of fake news. If fake news is to be reduced, the identification of key components involved in the spread of fake news should be a priority. It is possible to identify the key sources involved in spreading fake news by applying graph theory and machine learning techniques. Furthermore, the ability to identify fake news in real time, using video, is also one of the possible future directions that can be pursued.

### **Comparing results using confusion matrix**

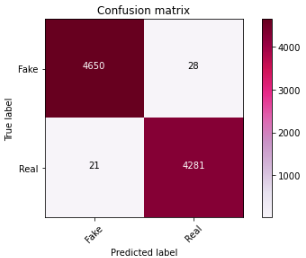
**Naïve Bayes**



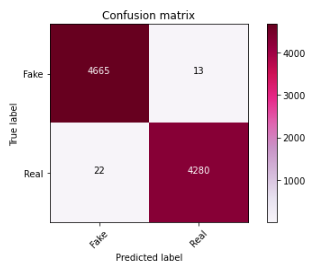
**Logistic Regression**



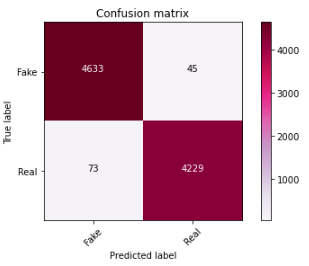
**SVM**



**Decision Tree**



**Random Forest**



# **Business Implications/Intelligence**

As the "fake news" issue has become a hot topic in the current political climate, it is an issue the business world has been facing for a lot longer than that. As a result of bad news, stocks may fall dramatically, business reputations could be damaged, and customers' expectations may be unrealistic. Business entities that do not adhere to ethical practices can also generate real news or reviews with the aim of boosting their own position or sales.

Therefore, it is not surprising at all that this is a popular and controversial topic in the current climate and it has been the subject of many studies in this area. There is substantial disagreement within the field of research and study on the topic of fake news, since there are only a limited number of studies and research reports available.

Regarding suggested solutions or ways to determine how companies can reduce the risk of fake news circulating through the internet, companies must take preventative measures in place in order to ensure that these companies have a presence on social media platforms and other news platforms that can protect them against the spread of fake news. Data storage capacity is expanding rapidly as companies are able to store more and more data on their business because of various technological advances that assist in the spreading of fake news.

Sources of news must become stronger and more credible in their ability to verify the news they are sharing on their own platforms, and consumers, meanwhile, must be more diligent in their checking of news and taking appropriate action based on their research. Fake news is becoming a major part of the conversation on social media. A potential solution to this problem would be for each individual to become more aware of the fact that there are misleading sources of information available through which they could make their investment decisions and to be more selective about what sources they rely upon for information (*Fake News and Business Consequences - \* Business and Economics - Research Guides at SUNY Geneseo*, n.d.).

It is our hope that the results from this study will contribute towards a more pragmatic and usable fake news detection system that is able to be implemented in real-world situations. In addition, the results can provide some information on how to develop applications that can be used to evaluate social media posts from a user's perspective, where hey can take advantage of machine learning algorithms' ability to determine whether it contains fake news or not. Furthermore, these models can also be used as third-party approaches auxiliary to those algorithms for the purpose of identifying malicious accounts in social networking sites.

# **Conclusion**

In this project, we sought to investigate a new approach to finding fake news quickly by analyzing only the title of a news story without reading the entire news text. The whole text analysis of news article was also conducted as a comparison measure to the proposed method. Using natural language processing, we analyzed the title and text of the news article and eventually we used machine learning algorithms, a combination of single-classifier models and ensemble models, to predict the outcome of the news report. The findings that were obtained from the experiments conducted were able to confirm the hypothesis that the authors put forward. After the initial analysis of the news titles, the application of the model proposed by us allowed you to obtain good results. Although, by analyzing the content of articles we can nonetheless very easily discern whether certain news articles are fake or not. Currently, the SVM classifier has a serious issue with running times, but we think the use of the random forest algorithm at the beginning of the experiments represents an excellent trade-off between quality of classification and resource consumption.

It may be necessary in the future to develop a two-step analysis model with a multi-criteria approach. Initial verification of fake news can be done based on the title, followed by a more detailed analysis of selected news pieces based on the text-only version. Moreover, the multi-criteria model may be selected based on a selection of suitable features, such as the classification quality and the running time. This capability can be designed to provide fast and accurate detection of fake news by studying only the title of the article or reading the article and further analysing it based on its content. Further, the questions can be posed which of the proposed classification algorithms would be most appropriate to use in an online learning environment where new fake or real news may be collected from users.

# **References**

Dai, E., Sun, Y., & Wang, S. (2020). Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 853–862.

*Fake and real news dataset | Kaggle*. (n.d.). Retrieved April 29, 2022, from https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

*Fake News and Business Consequences - \* Business and Economics - Research Guides at SUNY Geneseo*. (n.d.). Retrieved April 29, 2022, from https://libguides.geneseo.edu/business/fake-news-business-consequences

Gallagher, K. (2017). The Social Media Demographics Report: Differences in age, gender, and income at the top platforms. *Business Insider*.

Kansal, A. (2021). Fake News Detection Using Pos Tagging and Machine Learning. *Journal of Applied Security Research*, 1–16.

Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake news detection using machine learning approaches. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 12040.

Mahir, E. M., Akhter, S., & Huq, M. R. (2019). Detecting fake news using machine learning and deep learning algorithms. *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 1–5.

Manzoor, S. I., Singla, J., & Nikita. (2019). Fake news detection using machine learning approaches: A systematic review. *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, 230–234. https://doi.org/10.1109/ICOEI.2019.8862770

Mitchell, A. (2016). *Key findings on the traits and habits of the modern news consumer*.

Poynter. (2020). *CoronaVirusFacts Alliance - Poynter*. https://www.poynter.org/coronavirusfactsalliance/

Ribeiro, F. N., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., & Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. *Twelfth International AAAI Conference on Web and Social Media*.

Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., Goga, O., Gummadi, K. P., & Redmiles, E. M. (2019). On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 140–149.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1–42.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Tardáguila, C., Benevenuto, F., & Ortellado, P. (2018). Fake news is poisoning Brazilian politics. WhatsApp can stop it. *The New York Times*, *17*(10).