Fake News Detector

Manisha K. Ahirrao[1], Deekshita Bharat Nirmal[2], Soham JayantPrabhu[3], PrathameshMhaluSanap[4], Jhanvi Hemal Shah[5]
[1,2,3,4,5] *Rajiv Gandhi Institute of Technology*

***Abstract*— Fake News has become one of the biggest problems in the current society. Fraudulent stories are high the power to change ideas, facts and can be a very dangerous weapon to influence society. A proposed project strategy to detect 'false stories', that is, future misleading news from sources not reputable. By creating a Passive Aggressive Classifier model, false news can seen. The data science community has responded by taking action against the problem. Icon it is not possible to describe the stories as real or fake. The proposed project therefore uses data sets trained using a vectorizer calculation method to detect false stories and their accuracy tested using machine learning algorithms. Reduced the authenticity of the news ecosystem as is more widely distributed on social media than popular real-life stories. It is one of his major problems with the ability to change ideas and to influence decisions and disrupt them how people react to real stories.**

***Index Terms:* fake news, passive aggressive, vectorizer calculated, confusion matrix.**

I.INTRODUCTION

As more and more of our lives are spent online through social media, more and more many people tend to hunt and eat stories from social media instead of traditional news organizations. The descriptions of these behavioural changes are in line with those social context's media forums: it is often timely and very expensive to use news on social media in comparison traditional journalism, such as newspapers or television; and it is easy to share, discuss, and discuss this topic stories with friends or other readers on social media. For example, 62 percent of adults in the U.S. received news on social media in 2016, while in 2012; only 49 percent reported seeing news on social media. It had also been discovered that social media is now surpassing television because it is a major media outlet, the level of news on social media is lower thantraditional media organizations. However, because it is not expensive to deliver news online and very quicklyand very easy to spread it on social media, large falsenews journals, that is, those deliberately false news headlines, are produced online for the spread of goals, such as financial and political gain. it was estimated that there were more than a million tweets associated with the false news "Pizza gate" at the top of the presidential election. Considering the prevalence of this innovation. "False Stories" was even named Macquarie Dictionary of the Year 2016 widespread dissemination of false information can have a significant negative impact on individuals and communities. First, fake news can disrupt the authenticity of the news ecosystem for example; obviously very much Popular false stories were more widespread on Facebook than the most widely accepted real stories during the U.S. presidential election 2016. Second, false news deliberately entices consumers to simply admit biased or false beliefs. False stories are often used by propagandists to spread political messages or influence for example, another report shows that Russia has created fake accounts and public bots to spread false news. Third, false stories change the way people interpret and respond to real stories, for example, some lies. stories were created simply to arouse the imagination of the people and to confuse them; to impair their skills distinguish true from false. Helping to reduce the negative effects of false news (both to benefit the general public and therefore the news ecosystem). It is important that we build ways of saying that automatically detect false stories being spread on social media. The Internet and social media have made access to news information is very simple and comfortable. Internet users can often follow the events that concern them the online form, and the growing number of mobile devices make this process

even easier. But very much so opportunities come with great challenges. Social media has a huge impact on society, and because of that it usuallyhappens, someone wants to take advantage of this fact. Sometimes achieving certain goals Many media outlets can use information in a variety of ways. This results in the production of headlines that not entirely true or perhaps entirely false. There are even more websites that produce fake news almost only. They deliberately publish false information, half-truths, propaganda, and false information claiming to be authentic news - often using social media to drive web traffic and maximize its impact. Lots of goals for false news websites should affect the general public's opinion on certain issues (especially political ones). Samples of such websites can also be found in Ukraine, the United States of America, Germany, China and many other countries. Therefore, false news can be a global problem and a global challenge

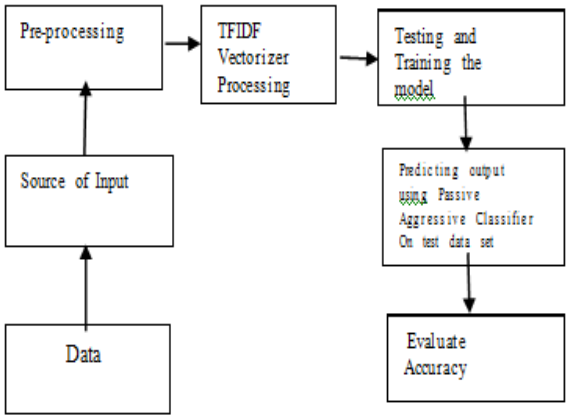
II. PROBLEM STATEMENT

News consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news. It enables the wide spread of “fake news”, i.e., low quality news with intentionally false information. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Therefore, fake news detection has recently become an emerging research that is attracting tremendous attention. First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content. To develop a FAKE NEWS DETECTION system using natural language processing and its accuracy will be tested using machine learning algorithms. The algorithm must be able to detect fake news in a given scenario.

III. METHODOLOGY

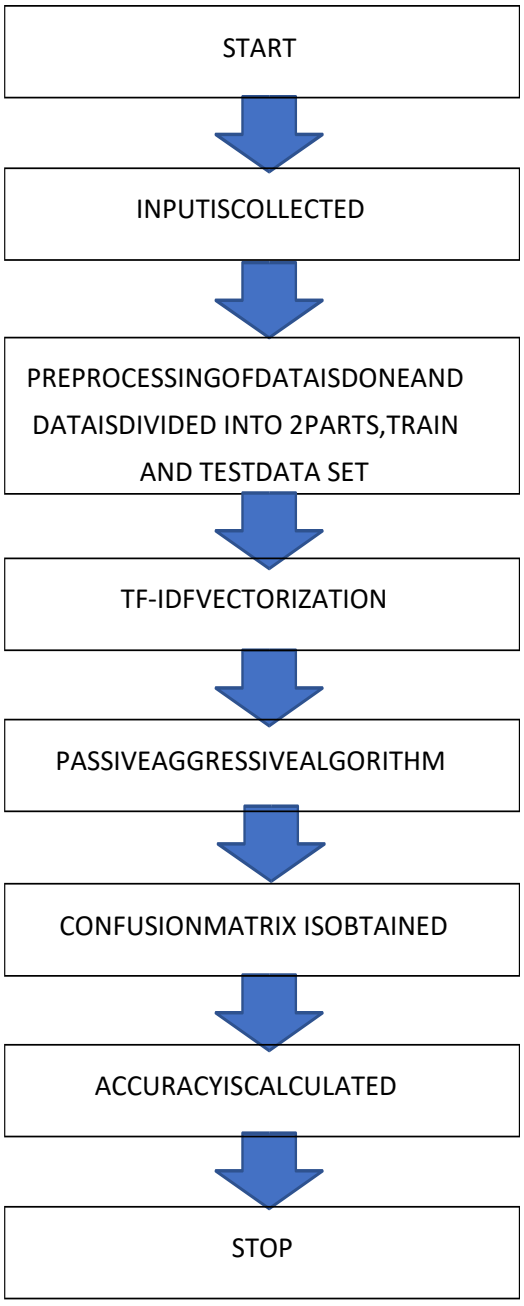
The proposed system when subjected to a scenario of a set of news articles, the new articles are categorized as true or fake by the existing data available. This prediction is done by using the relationship between the words used in the article with one another. The proposed system contains a Word to Vec model for finding the relationship between the words and with the obtained information of the existing relations, the new articles are categorized into fake and real news.

IV. SYSTEM ARCHITECHTURE



Input is collected from various sources such as newspapers social media and stored in datasets. System will take input from datasets. The datasets undergo preprocessing and the unnecessary information is removed from it and the data types of the columns are changed if required. Jupyter notebook and python libraries are used in the above step. Count vectorizer technique is used in the initial step. For fake news. detection , we have to train the system using dataset. Before entering to the detection of fake news , entire dataset is divide into two datasets . 80% is used for training and 20% is used for testing. During training , Passive Aggressive algorithm is used to train the model using the train dataset. In testing , the test dataset is given as input and the output is predicted After the testing time , The predicted output and the actual output are compared using confusion matrix obtained .The confusion matrix gives the information regarding the number of correct and wrong predictions in the case of real and fake news. The accuracy is calculated by the equation No Of Correct Predictions/Total Test Dataset
Input Size

V. ALGORITHM FOR THE PROPOSED SYSTEM



VI.DRAWBACK OF ALGORITHM

Drawback of SVM Algorithm:- final model is difficult to understand and interpret the It is hard to visualize their impact .We can not do small calibrations to the model.

Drawbacks of Naïve Bayes :- The main limitation of Naive Bayes is the assumption of independent predictor features. In real life, it's almost impossible that we get a set of predictors that are completely independent or one another. To overcome this problem we used these algorithm

VII.CONFUSION MATRIX

A confusion matrix is a summary of prediction results on a classification problem. Confusion matrices are used to visualize important predictive analytics like recall, specificity, accuracy, and precision. Confusion matrices are useful because they give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives.

1267		PREDICTED DATA		
		FAKE	REAL	
ACTUAL DATA	FAKE	567	48	615
		(TN)	(FP)	
	REAL	38	614	652
		(FN)	(TP)	
		605	662	

Accuracy = TP + TN/total
= 614+567/1267 = 0.932 = 93.21%
Error Rate = 1-accuracy or FP+FN/total
=0.09 = 9%
Precision = TP/Predicted REAL
=614/662 = 0.927
Recall = TP/Actual REAL
=614/652 = 0.941

VIII. PASSIVE AGGRESSIVE CLASSIFIER

Passive-Aggressive algorithms are called so because :
Passive: If the prediction is correct, keep the model and do not make any changes. i.e., the Data in the example is not enough to cause any changes in the model.
Aggressive: If the prediction is incorrect, make changes to the model. i.e., some change to the Model may correct it.
Passive-Aggressive algorithms are generally used for large-scale learning. It is one of the few ‘onlinelearningalgorithms‘. In online machine learning algorithms, the input data comes in sequential order and the machine learning model is updated step- bystep, as opposed to batch learning, where the entire training dataset is used at once. This is very useful in situations where there is a huge amount of data and it is computationally infeasible to train the entire dataset because of the sheer size of the data. We can simply say that an online- learning

algorithm will get a training example, update the classifier, and then throw away the example. A very good example of this would be to detect fake news on a social media website like Twitter, where new data is being added every second.

TF-IDF VECTORIZER

The term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. TF-IDF was invented for document search and information retrieval. It works by increasing proportionally to the Number of times a word appears in a document, but is offset by the number of documents that contain the word. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they don't mean much to that document in particular. However, if the word Bug appears many times in a document, while not appearing many times in others, it probably means that it's very relevant. For example, if what we're doing is trying to find out which topics some NPS responses belong to, the word Bug would probably end up being tied to the topic Reliability, since most responses containing that word would be about that topic. Term frequency of a word in a document. There are several ways of calculating this frequency, with the simplest being a raw count of instances a word appears in a document. Then, there are ways to adjust the frequency, by length of a document, or by the raw frequency of the most frequent word in a document. The inverse document frequency of the word across a set of documents. This means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the algorithm. So, if the word is very common and appears in many documents, this number will approach 0.

Otherwise, it will approach

IX. TECHNOLOGIES USED

HTML, CSS -Hypertext Markup Language, a standardized system for tagging text files, to achieve font, color, graphic, and hyperlink effects of World Wide Web pages

Cascading style sheet is a style sheet language used for describing the presentation of a document written in markup language such as HTML We have used HTML for developing the front end of our project.

Sklearn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

Flask is a micro web framework written in Python.

it allows us to build up web applications.

Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

X. CONCLUSION

Manual classification of news articles requires indepth knowledge and expertise in identifying deviation in the text. It takes a lot of time to verify a single article manually that's why we discussed the use of machine learning models and ensemble methods to classify fake news articles .In these project implementation and the accompanying experiment, it was shown that the combination of TFIDF and Passive Aggressive model shows the best performance and can detect over 9 out of 10 fake news articles correctly. This paper concludes that using Passive Aggressive and TF-IDF vectorizer is efficient as we obtained 94% of accuracy from this model.

REFERENCES

[1] Fan Yang XFake: Explainable Fake News Detector with Visualizations
[2] Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator, IEEE
[3] Fake News Detection from Data Streams, IEEE
[4] Koby Crammer, Online Passive-Aggressive Algorithms
[5] Taiwo Oladipupo Ayodele, Types of Machine Learning Algorithms,
[6] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake News Detection on Social
[7] 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
[8] Nandhin, Subhasree, Rooba-Fake news detection using machine learning algorithm
[9] Detecting Fake news with python and machine learning ,Data Fair
[10] Piero Paialunga, Fake news Detection with Machine Learning,using Python, Towards Data Science
[11] Detecting Fake News with Natural Language Processing, Analytics Vidhya