**Fake News Detection Using Machine Learning**

**SUBMITTED**

**TO**



**VISHWAKARMA INSTITUTE OF INFORMATION**

**TECHNOLOGY, PUNE.**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**ENGINEERING, VIIT PUNE.**

**UNDER GUIDANCE OF:  DR. PRAVIN FUTANE SIR**

**Head of Departments and professor, Information Technology**

**Nikita Dara (331071), Stuti Wali (331063), Rushika Solankar (331054), Mangesh Myadamwar (331032), Prajwal Suryawanshi (331066)**

**UNDER GUIDENCE OF: DR. PRAVIN FUTANE SIR**

**Head of Department and professor, Information Technology**

**Abstract -**

                This Project comes up with the applications of NLP (Natural Language Processing) techniques for detecting the 'fake news', that is, misleading news stories that come from non-reputable sources. Only by building a model based on a count vectorizer (using word tallies) or a (Term Frequency Inverse Document Frequency) TF IDF matrix, (word tallies relative to how often they’re used in other articles in your dataset) can only get you so far. But these models do not consider the important qualities like word ordering and context. It is very possible that two articles that are similar in their word count will be completely different in their meaning. The data science community has responded by taking actions against the problem. There is a Kaggle competition called the “Fake News Challenge'' and Facebook is employing AI to filter fake news stories out of users’ feeds. Combatting the fake news is a classic text classification project with a straightforward proposition. Is it possible for you to build a model that can differentiate between “Real “news and “Fake” news? A proposed work on assembling a dataset of both fake and real news and employing a Naive Bayes classifier in order to create a model to classify an article into fake or real based on its words and phrases.

1. **INTRODUCTION**

These days’ fake news is creating different issues from sarcastic articles to fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is “fake news “but lately blathering social media’s discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints. The importance of disinformation within American political discourse was the subject of weighty attention, particularly following the American president election. The term 'fake news' became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views. In this paper, it is seeked to produce a model that can accurately predict the likelihood that a given article is fake news.

              Facebook has been at the epicentre of much critique following media attention. They have already implemented a feature to flag fake news on the site when a user sees it; they have also said publicly they are working on to distinguish these articles *in* an automated way. Certainly, it is not an easy task. A given algorithm must be politically unbiased – since fake news exists on both ends of the spectrum – and also give equal balance to legitimate news sources on either end of the spectrum. In addition, the question of legitimacy is a difficult one. However, in order to solve this problem, it is necessary to have an understanding on what Fake News is. Later, it is needed to look into how the techniques in the fields of machine learning, natural language processing helps us to detect fake news.

1. **LITERATURE SURVEY**

In Shloka Gilda presented concept approximately how NLP is relevant to stumble on fake information. They have used time period frequency-inverse record frequency (TFIDF) of bi-grams and probabilistic context free grammar (PCFG) detection. They have examined their dataset over more than one class algorithms to find out the great model. They locate that TF-IDF of bi-grams fed right into a Stochastic Gradient Descent model identifies non-credible resources with an accuracy of seventy-seven.2%.

In, Mykhailo Granik proposed simple technique for fake news detection the usage of naive Bayes classifier. They used BuzzFeed news for getting to know and trying out the Naïve Bayes classifier. The dataset is taken from Facebook news publish and completed accuracy up to seventy-four% on test set. In, Cody Buntain advanced a method for automating fake news detection on Twitter. They applied this method to Twitter content sourced from BuzzFeed’s fake news dataset. Furthermore, leveraging non-professional, crowdsourced people instead of journalists presents a beneficial and much less costly way to classify proper and fake memories on Twitter rapidly. In Marco L. Della offered a paper which allows us to recognize how social networks and gadget studying (ML) strategies may be used for faux news detection. They have used novel ML fake news detection method and carried out this approach inside a Facebook Messenger chatbot and established it with a actual-world application, acquiring a fake information detection accuracy of eighty one.7%. In Rishabh Kaushal carried out 3 getting to know algorithms specifically Naive Bayes, Clustering and Decision

bushes on some of features such astweet-degree and consumer-level like Followers/Followers, URLs, Spam Words, Replies and Hashtag’s. Improvement of unsolicited mail detection is measured on the premise of general Accuracy, Spammers Detection Accuracy and No Spammers Detection Accuracy. In Saranya Krishnan used superior framework to identify faux information contents. Initially, they've extracted content material capabilities and consumer functions via Twitter API. Then functions together with statistical analysis of twitter user accounts, reverse picture searching, verification of fake news assets is used by facts mining algorithms for class and analysis

1. **PROPOSED METHODOLOGY**

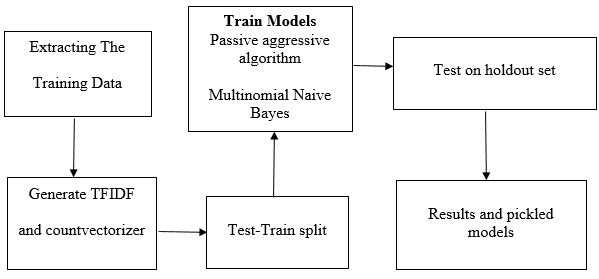
**Problem Statement:**

In this paper a model is built based on the count vectorizer or a TFIDFmatrix (i.e.) word tallies relative to how often they are used in other articles in your dataset) can help. Since this problem is a kind of text classification, implementing a Naive Bayes classifier will be best as this is standard for text-based processing. The actual goal is in developing a model which was the text transformation (count vectorizer vs tfidf vectorizer) and choosing which type of text to use (headlines vs full text). Now the next step is to extract the most optimal features for count vectorizer or tfidf-vectorizer, this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as “the”, “when”, and “there” and only using those words that appear at least a given number of times in a given text dataset.

**Display result:**

Result that was found after data matching is now displayed over here.

**Basic Working:**



1. **IMPLEMENTATION**
2. Confusion Matrix:

           It is a table that describes performance of classification model on set of test data for which the true values are known.

1. Text feature extraction:

1. Count Vectorizer 2.HashingVectorizer                            3. Tfidf Vectorizer

1. Naive Bayes Multinomial Algorithm:

1. Naive Bayes are mostly used in Natural Language problem (NLP) problem. Naive Bayes predict the tag of a text.

2.Stopword -There are common word that don’t really add anything to classification like able, either, and so on.

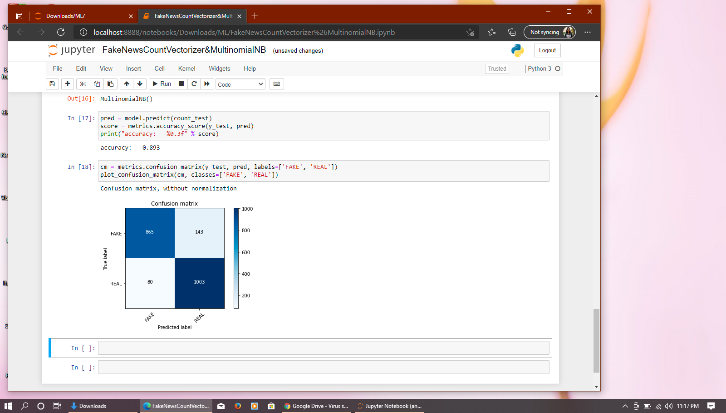
3. Stemming - stemming to take out the root of word.

Example-

**<>** “ilikethemovie” - POSITIVE

**<>**  “nicesongbutsadlyboringend” - NEGATIVE

1. **RESULTS**

****

The chatbot results into a platform for feasible interaction of doctors and patients.

1. **CONCLUSION**

From this project we are able to find accuracy of news that we provided as input. Means you have to submit your news along with label in proper format and based on algorithm and confusion matrix concept we get output as accuracy and confusion matrix diagram.

So, we can give only one news and document containing many news too.

1. **REFERENCES**

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

2) S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 171–175.

3)Shlok Gilda, Department of Computer Engineering, Evaluating Machine Learning Algorithms for Fake News Detection,2017 IEEE 15th Student Conference on Research and Development (Scored).