**Synopsis**

**on**

**Fake News Detection**

**Submitted by-**

**Muskan Agrawal-2100290140091**

**Shivansh Srivastava-2100290140131**

**Zainab Zubair-2100290140157**

**Group No: GC11**



**Submitted to-**

**Mrs Vidushi**

**(Assistant Professor)**

**Department of Computer Applications,**

**KIET Group of Institutions,**

**Delhi-NCR, Ghaziabad**

**10th September’ 2022**

**ABSTRACT**

In our modern era where the internet is ubiquitous, everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook , Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click-baits. In this project, we aim to perform binary classification of various news articles available online with the help of concepts pertaining to Artificial Intelligence, Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website publishing the news.

Approved by-

Mrs Vidushi

**TABLE OF CONTENTS**

1. Introduction
2. Technologies / Software Requirements
3. Hardware requirement / Hardware Used
4. Modules Description
5. Reports / Outputs
6. Conclusion
7. How to serve the society (If applicable)
8. Gantt Chart (In terms of weeks)

**INTRODUCTION**

Fake News contains misleading information that could be checked. This maintains lie about a certain statistic in a country or exaggerated cost of certain services for a country, which may arise unrest for some countries like in Arabic spring. There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues as confirming authors are accountable. However, their scope is so limited because they depend on human manual detection, in a globe with millions of articles either removed or being published every minute, this cannot be accountable or feasible manually. A solution could be, by the development of a system to provide a credible automated index scoring, or rating for credibility of different publishers, and news context.

This proposes a methodology to create a model that will detect if an article is authentic or fake based on its words, phrases, sources and titles, by applying supervised machine learning algorithms on an annotated (labeled) dataset, that are manually classified and guaranteed. Then, feature selection methods are applied to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results. We propose to create the model using different classification algorithms. The product model will test the unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and can be used and integrated with any system for future use.

**RELATED WORK**

**Social Media and Fake News**

Social media includes websites and programs that are devoted to forums, social websites, microblogging, social bookmarking. On the other side, some researchers consider the fake news as a result of accidental issues such as educational shock or unwitting actions like what happened in Nepal Earthquake case. In 2020, there was widespread fake news concerning health that had exposed global health at risk. The WHO released a warning during early February 2020 that the COVID-19 outbreak has caused massive ‘infodemic’, or a spurt of real and fake news—which included lots of misinformation.

**Natural Language Processing**

The main reason for utilizing Natural Language Processing is to consider one or more specializations of system or an algorithm. The Natural Language Processing (NLP) rating of an algorithmic system enables the combination of speech understanding and speech generation. In addition, it could be utilized to detect actions with various languages. suggested a new ideal system for extraction actions from languages of English, Italian and Dutch speeches through utilizing various pipelines of various languages such as Emotion Analyzer and Detection, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Chunking, and Semantic Role Labeling made NLP good Subject of the search . The Sentiment analysis extracts emotions on a particular subject. Sentiment analysis is composed of extracting a specific term for a subject, extracting the sentiment, and pairing with connection analysis. The Sentiment analysis uses dual languages Resources for analysis: Glossary of meaning and Sentiment models database. for constructive and Destructive words and attempts to give classifications on a level of -5 to 5. Parts of speech taggers tools for languages such as European languages are being explored to produce parts of language taggers tools in different languages such as Sanskrit [8], Hindi [9] and Arabic. Can be efficient Mark and categorize words as names, adjectives, verbs, and so on. Most part of speech techniques can be performed effectively in European languages, but not in Asian or Arabic languages. Part of the Sanskrit word "speak" specifically uses the tree-bank method. The Arabic utilizes Vector Machine (SVM) uses a method to automatically identify symbols and parts of speech and automatically expose basic sentences in Arabic text.

**Data Mining**

Data mining techniques are categorized into two main methods, which is; supervised and unsupervised. The supervised method utilizes the training information in order to foresee the hidden activities. Unsupervised Data Mining is a try to recognize hidden data models provided without providing training data for example, pairs of input labels and categories. A model example for unsupervised data mining is aggregate mines and a syndicate base

**Machine Learning (ML)**

Classification Machine Learning (ML) is a class of algorithms that help software systems achieve more accurate results without having to reprogram them directly. Data scientists characterize changes or characteristics that the model needs to analyze and utilize to develop predictions. When the training is completed, the algorithm splits the learned levels into new data [11]. There are six algorithms that are adopted in this paper for classifying the fake news.

**Decision Tree**

The decision tree is an important tool that works based on flow chart like structure that is mainly used for classification problems. Each internal node of the decision tree specifies a condition or a “test” on an attribute and the branching is done on the basis of the test conditions and result. Finally the leaf node bears a class label that is obtained after computing all attributes. The distance from the root to leaf represents the classification rule. The amazing thing is that it can work with category and dependent variable.

**TECHNOLOGIES**

The Technologies used in this project are-

* Python
* Numpy
* Pandas
* Matplotlib
* Seaborn

**SOFTWARE REQUIREMENTS**

|  |  |  |
| --- | --- | --- |
| **Number** | **Description** | **Type** |
| 1 | Operating System | Windows XP / Windows |
| 2 | Language | Python, and Python libraries |
| 3 | Database | IBM Db2 Warehouse ,Vertica |
| 4 | IDE | Visual Code, Anaconda |
| 5 | Browser | Google Chrome |

**HARDWARE REQUIREMENT**

|  |  |  |
| --- | --- | --- |
| **Number** | **Description** | **Type** |
| 1 | Hardware | Processor Intel dual core and above XP / Windows |
| 2 | Clock speed | 3.0 GHz |
| 3 | RAM size | 2048 MB |
| 4 | Hard Disk capacity | 32 GB |
| 5 | Monitor type | 15 inch color monitor |

**MODULE DESCRIPTION**

* Data collection from heterogeneous platforms and we will create the dataset.
* We will analyze the data
* Visualization of findings
* Data Storage

Data collection from heterogeneous platforms and we will create the dataset.

We can get online news from different sources like social media websites, search engine, homepage of news agency websites or the fact-checking websites. On the Internet, there are a few publicly available datasets for Fake news classification like-

* Buzzfeed News,
* LIAR BS Detector etc.

These datasets have been widely used in different research papers for determining the veracity of news. Online news can be collected from different sources, such as news agency homepages, search engines, and social media websites.

* Analysis

Naïve Bayes Classifier:

This classification technique is based on Bayes theorem, which assumes that the presence of a particular feature in a class is independent

of the presence of any other feature. It provides way for calculating the posterior probability.

P(x) = (P(c)\*P(c))/P(x)

P(c |x)= posterior probability of class given predictor

P(c)= prior probability of class

P(x |c)= likelihood (probability of predictor given class)

P(x) =

**REPORT/OUTPUT**

We can evaluate machine learning algorithms using various metrics like:

1. Accuracy

2. Precision

3. Recall

4. F1-Score

Hence we evaluate and analyse the result based on these metrics for different datasets, classifiers and different methods of feature extraction methodology.

**Accuracy = (TP+TN)/(TP+FP+FN+TN)**

**Recall = TP/(TP+FN)**

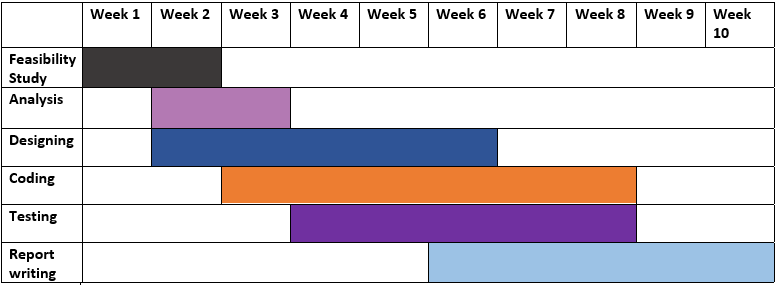
**Precision = TP/(TP+FP)**

**F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)**

**FUTURE SCOPE**

In the future, a web-based GUI can be created for the proposed fake news detection system to classify the news as fake or real on real-time social media platforms such as Facebook, Instagram, Twitter, WhatsApp, etc. Also, the annotated dataset in the sequence of images (with textual content written on them) will be collected and maintained from Facebook and Reddit platforms. The annotated dataset is often used for detecting fake images within the future as no such dataset is out there at the present. The proposed system has the potential to provide an impulse to various emerging applications such as controlling the spread of fake news during elections, terrorism, natural calamities, crimes for the betterment of society. In the future, the efficiency and accuracy of the prototype can be enhanced to a certain level, and also enhance the user interface of thep roposed model.

**GANTT CHART**

****

**CONCLUSION**

In the 21st century, the majority of the tasks are done online. Newspapers that were earlier preferred as hard-copies are now being substituted by applications like Facebook, Twitter, and news articles to be read online. Whatsapp forwards are also a major source. The growing problem of fake news only makes things more complicated and tries to change or hamper the opinion and attitude of people towards use of digital technology. When a person is deceived by the real news two possible things happen- People start believing that their perceptions about a particular topic are true as assumed. Thus, in order to curb the phenomenon, we have developed our Fake news Detection system that takes input from the user and classify it to be true or fake. To implement this, various NLP and Machine Learning Techniques have to be used. The model is trained using an appropriate dataset and performance evaluation is also done using various performance measures. The best model, i.e. the model with highest accuracy is used to classify the news headlines or articles. As evident above for static search, our best model came out to be Logistic Regression with an accuracy of 65%. Hence we then used grid search parameter optimization to increase the performance of logistic regression which then gave us the accuracy of 75%. Hence we can say that if a user feed a particular news article or its headline in our model, there are 75% chances that it will be classified to its true nature.

The user can check the news article or keywords online; he can also check the authenticity of the website. The accuracy for dynamic system is 93% and it increases with every iteration.

We intend to build our own dataset which will be kept up to date according to the latest news. All the live news and latest data will be kept in a database using Web Crawler and online database.