

**An Internship Project Report on**

“Fake News Detection using Natural Language Processing”

Submitted

In partial fulfilment of the requirement for the

Internship in Machine Learning & Artificial Intelligence during the academic year 2022-2023

**Submitted By**

**YASHWANTH SAGAR KOLUGURI (18E41A05E8)**

**(Sree dattha institute of Engineering and science)**

**Under the guidance of**

Mr. Shashank G

Technical Lead

QuantMasters Pvt. Ltd.

**Quant Masters**

#812, 6th cross 3rd main,

Rajajinagar, Bengaluru - 560021

**CONTENTS**

|  |  |
| --- | --- |
| **Chapters** | **Page no.** |
| **ABSTRACT** | **3** |
| **Chapter-1 Introduction** | **4** |
| **Chapter-2 Related Work** | **6** |
| 2.1 Machine learning | 6 |
| 2.2 Natural language processing | 7 |
| 2.3 Decision Tree | 8 |
| 2.4 Passive Aggressive Classifier | 9 |
| 2.5 Pre processing | 10 |
| 1. **Proposed Methodology** | **10** |
| 3.1 Existing system | 10 |
| 3.1.1 Need of New System | 11 |
| 3.1.2 Problem defination | 11 |
| 3.2 Proposed System | 11 |
| 1. **System Design** | **12** |
| 4.1 System Architecture | 12 |
| 4.2 UML Diagrams | 13 |
| 4.2.1 Use case diagrams | 13 |
| 4.2.2 Sequence diagram | 14 |
| 4.2.3 Class diagrams | 15 |
| 1. **Results** | **16** |
| 1. **Conclution** | **19** |
| 1. **Future scope** | **20** |

**GOOGLE COLAB LINK : [https://colab.research.google.com/drive/1KOXJ\_d66A40mQ\_VBf5Wn8DrrhOtZpK8D?usp=sharing](INTERNSHIP FINAL report.docx)**

https://colab.research.google.com/drive/1KOXJ\_d66A40mQ\_VBf5Wn8DrrhOtZpK8D?usp=sharing

**ABSTRACT**

Fake News has become one of the major problem in the existing society. Fake News has high

potential to change opinions, facts and can be the most dangerous weapon in influencing society.The recent growth of the online social media fake news has great impact to the society. There is a huge information from disparate sources among various users around the world. Social media platforms like Facebook, WhatsApp and Twitter are one of the most popular applications that are able to deliver appealing data in timely manner. Developing a technique that can detect fake news from these platforms is becoming a necessary and challenging task.

The proposed project uses NLP techniques for detecting the 'fake news', that is, misleading

news stories which come from the non-reputable sources. By building a model based on five

algorithms Logistics Regression, NAVIE BAYES, Decision Tree, Passive-Aggressive Classifier, the fake news can be detected . The data science community has responded by taking actions against the problem. It is impossible to determine a news as real or fake accurately. So the proposed project uses the datasets that are trained using count vectorizer method for the detection of fake news and its accuracy will be tested using machine learning algorithms

.

**1.Introduction**

Fake news or junk news or pseudo-news is a type of yellow journalism or propaganda that

consists of deliberate disinformation or hoaxes spread via traditional print and broadcast news

media or online social media. The news is often reverberated as misinformation in social media

but occasionally finds its way to the mainstream media as well. We plan to build a web-based

application or browser extension to help users identify if a news source is reliable or fake.

Any news whose authenticity and source cannot be validated by the reader is termed as a fake

news. The issue of fake news has become a serious problem in India because of high digital

illiteracy and low digital penetration. Like any other social phenomenon fake also has its own

pros and cons.

This project will contribute to the start of a new revolution against one of the most prevalent

hazard i.e. spread of the Fake News. It will serve as root and branch eradication of the same.

This project will help to create a next level of awareness and make the citizens more responsible.

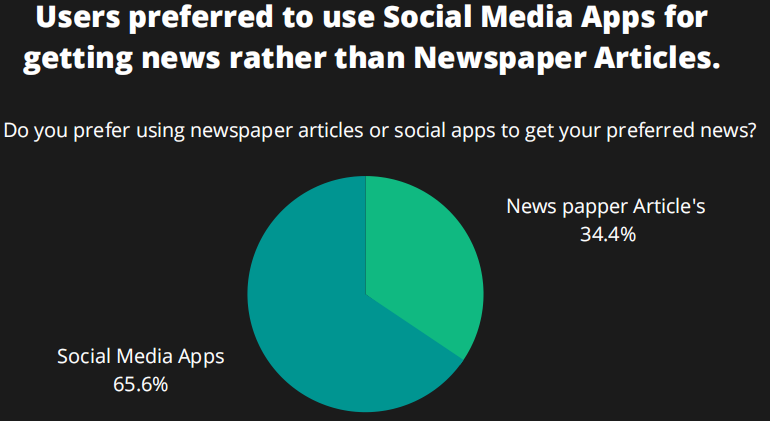
This project will help the people of a nation to take meaningful and informed decisions

**1.1 Problem Statement:**

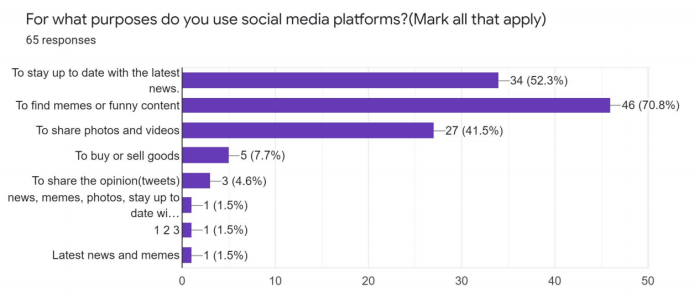
Fake news posted by famous figures on social media spreads faster because people do not fact-check what is shared with them and blindly follow the crowd. Thus , people cannot differentiate between facts and online propaganda.

It often leads to real-life consequences regardless of being good or bad. People need a transparent medium for fact-checking news to make informed decisions without contributing to a social problem





**A Major portion of the users used social media platforms to keep track of the latest news after searching content for entertainment**



**CHAPTER-2 Related Work**

**2.1 Machine Learning**

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions,relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.” This is Alan Turing’s definition of machine learning.

Deep learning is a class of machine learning algorithms that utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. The artificial neural networks are built like the human brain, with neuron nodes connected together like a web. While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach.

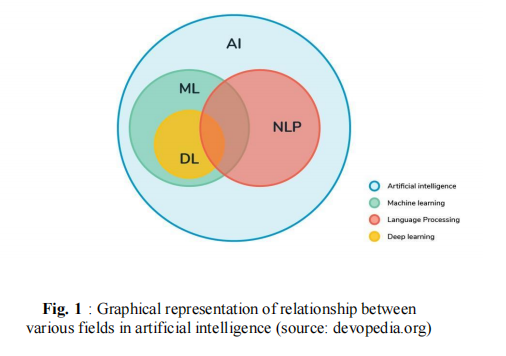
The word "deep" in "deep learning" refers to the number of layers through which the data is

transformed. More precisely, deep learning systems have a substantial credit assignment path (CAP) depth. The CAP is the chain of transformations from input to output. CAP ’s describe potentially causal connections between input and output.

For a feed forward neural network, the depth of the CAP ’s is that of the network and is the

number of hidden layers plus one (as the output layer is also parameterized). For recurrent neural networks, in which a signal may propagate through a layer more than once, the CAP depth is potentially unlimited.

Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts.



**2.2 NATURAL LANGUAGE PROCESSING**

NLP is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program

computers to fruitfully process large amounts of natural language data.

Natural language processing (NLP) is a sub field of linguistics, computer science,information

engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyse large amounts of natural language data.

**2.3 DECISION TREE**

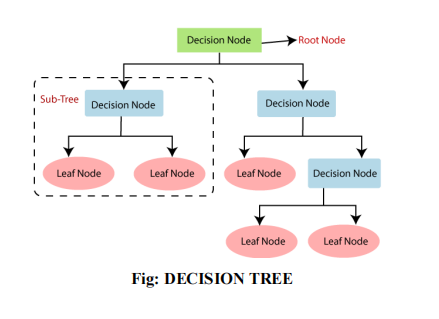
A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels.The paths from root to leaf represent classification rules. Decision tree one of the predictive modelling approaches used in statistics, data mining and machine learning.

Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions.It is one of the most widely used and practical methods for supervised learning. Decision Trees area non-parametric supervised learning method used for both classification and regression tasks.

Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is general term for this.

Data comes in records of forms i.e., (x, Y) =(x1, x2, x3, xk,Y). The dependent variable,

Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the features, x1, x2, x3 etc., that are used for that task. While making decision tree, at each node of tree we ask different type of questions. Based on the asked question we will calculate the information gain corresponding to it.



**Advantage of Decision Tree**

 Easy to use and understand.

 Can handle both categorical and numerical data.

 Resistant outliers, hence require little data preprocessing.

**Disadvantage of Decision Tree**

 Prone to overfitting.

 Require some kind of measurement as to how well they are doing.

 Need to be careful with parameter tuning.

 Can create biased learned trees if some classes dominate.

**2.4 PASSIVE AGGRESSIVE CLASSIFIER**

Passive-Aggressive Classifier belong to the family of online learning algorithms for

binary classification. It is similar to support vector machine classifier and can be considered as the online version of SVM. It finds a hyperplane to separate the instance into two halves. The margin of an example is proportional to the distance of the hyperplane. The errors in predicting examples can be corrected by using margin of the hyperplane. The update of the classifier follows the constraints passive update and aggressive update. The advantage is that it follows an online learning pattern and update the separating hyperplane for the next example and ensure the performance of the algorithm[13]. Because of the theoretical loss bound, the performance of the classifier can be easily predicted.

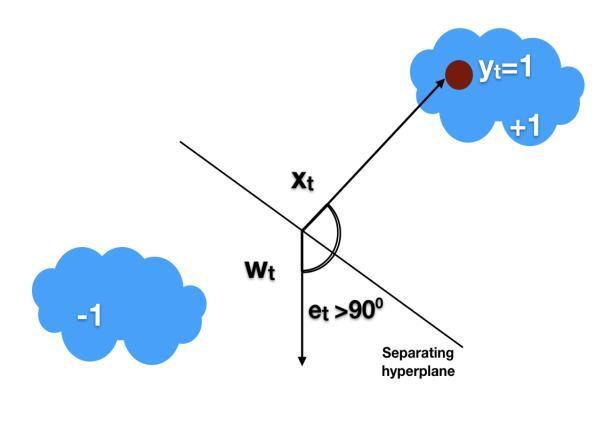


Figure shows that when the angle Ɵ>0 the sample is classified as -1 and its label is +1. In such cases the update rule is aggressive and looks for new weight wt for the sample xt.

**2.5 PRE PROCESSING**

In any Machine Learning process, Data Pre processing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

In this fake news detection, pre processing is the major thing that should be done . Firstly as the data data set is collected from various sources unnecessary information should be removed ,converted to lower case , remove punctuation , symbols , stop words.

**3.Proposed Methodology**

**3.1 EXISTING SYSTEM**

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 buzz feed News, LIAR [15], BS Detector etc. These datasets have been widely used in different research papers for determining the veracity of news. In the following sections, I have discussed in brief about the sources of the data set used in this work.This Existing system can help us to trained our model using machine learning technique.

**3.1.1 NEED OF NEW SYSTEM**

Currently, many people are using the internet as a central platform to

find the information about reality in world and need to be continue.Hence I has mention above we will create fake news and message detection model which detect the reality of the news and message.

Also whose use our website can see the up to date about main source or keyword are

getting most fake news and message and mapoed up with chart. After and all everyone want to know how to prevent this hence we are giving some important tips to avoid this fake news of spreading rumor in the world.

**3.1.2 PROBLEM DEFINATION**

The system is an Web application which help user to detect the fake news. We have

given the text box where the user has the option to paste the message or paste the url link of the news and other message link and after that it gives the reality of it. All the user gives data to detector may save for further use in order to update the statue of model, data analysis in future. We also help user by giving some guidance of how to prevent from such false event and how to stop with such event from spreading it.

**3.2 PROPOSED SYSTEM**

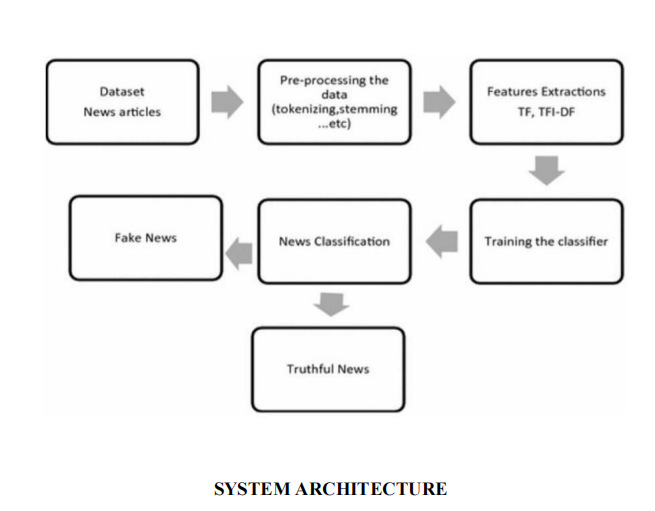
The system is an Web application which help user to detect the fake news. We have

given the text box where the user has the option to paste the message or paste the url link of the news and other message link and after that it gives the reality of it. All the user gives data to detector may save for further use in order to update the statue of model, data analysis in future. We also help user by giving some guidance of how to prevent from such false event and how to stop with such event from spreading it.

**4.SYSTEM DESIGN**

**4.1 SYSTEM ARCHITECTURE**

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 Word2Vec 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.



Input is collected from various sources such as newspapers , social media and stored

in datasets. System will take input from datasets. The datasets undergo pre processing 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 data set. Before entering to the detection of fake news , entire data set is divide into two datasets . 80% is used for training and 20% is used for testing.

During training , K-Means algorithm is used to train the model using the train data set.

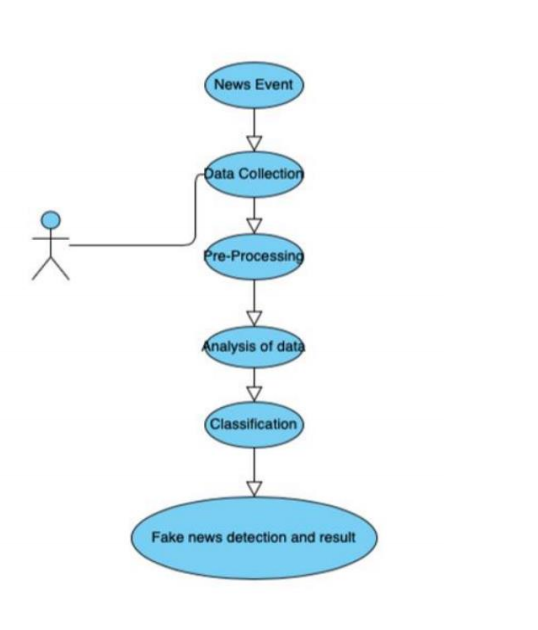
In testing , the test data set is given as input and the output is predicted.After the testing time , The predicted output and the actual output are compared using confusion matrix obtained.

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 data set Input Size.

**4.2 UML DIAGRAMS**

**4.2.1 USE CASE DIAGRAM**

Use case diagram is used to show which operations are performed by the user and which operation are performed by the system.



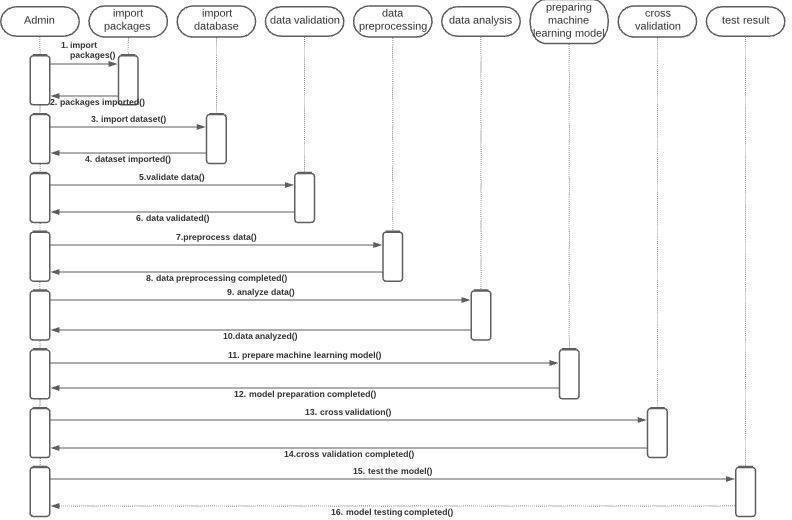
**Fig : USE CASE DIAGRAM**

**4.2.2 SEQUENCE DIAGRAM**

Sequence diagrams describe interactions among classes in terms of an exchange of

messages over time. They're also called event diagrams. A sequence diagram is a good way to visualize and validate various runtime scenarios. These can help to predict how a system will behave and to discover responsibilities a class may need to have in the process of modelling a new system.

The aim of a sequence diagram is to define event sequences, which would have a desired outcome. The focus is more on the order in which messages occur than on the message per se.However, the majority ofsequence diagrams will communicate what messages are sent and the order in which they tend to occur.



**Fig : SEQUENCE DIAGRAM**

**4.2.3 CLASS DIAGRAM**

Class diagrams are the main building blocks of every object oriented methods. The class diagram can be used to show the classes, relationships, interface, association, and collaboration. UML is standardized in class diagrams. Since classes are the building block of an application that is based on OOPs, so as the class diagram has appropriate structure to represent the classes, inheritance, relationships, and everything that OOPs have in its context. It describes various kinds of objects and the static relationship in between them. The main purpose to use class diagrams are:

1.This is the only UML which can appropriately depict various aspects of OOPs concept.

2.Proper design and analysis of application can be faster and efficient.

3.It is base for deployment and component diagram. Each class is represented by a rectangle having a subdivision of three compartments name, attributes and operation.

**Dependency:** A dependency is a semantic relationship between two or more classes where a

change in one class cause changes in another class. It forms a weaker relationship.

**Generalization:** A generalization is a relationship between a parent class (superclass) and a child class (subclass). In this, the child class is inherited from the parent class.

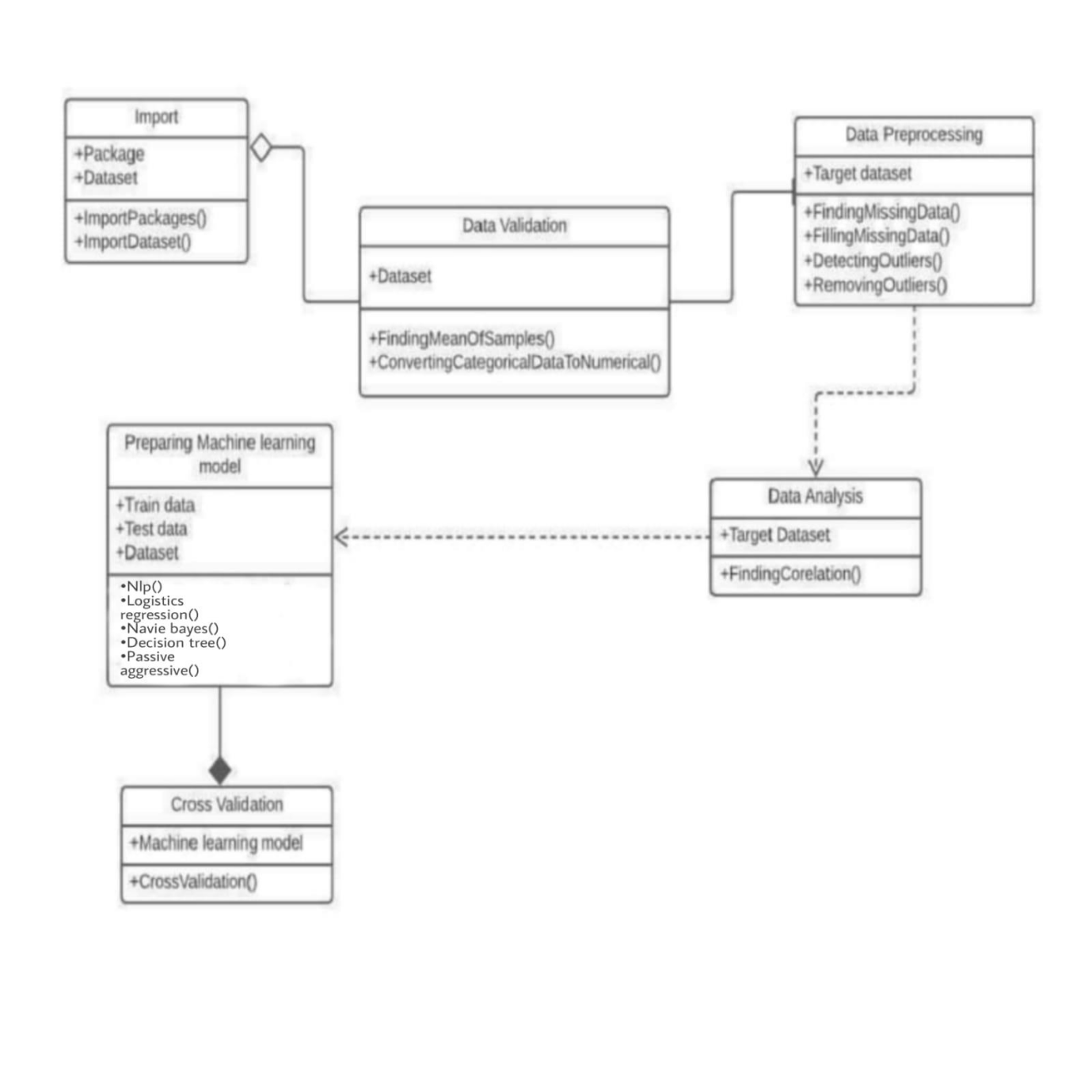
**Association:** It describes a static or physical connection between two or more objects. It depicts how many objects are there in the relationship.

**Multiplicity:** It defines a specific range of allowable instances of attributes. In case if a range is not specified, one is considered as a default multiplicity.

**Aggregation:** An aggregation is a subset of association, which represents has a relationship. It is more specific then association. It defines a part-whole or part-of relationship. In this kind

of relationship, the child class can exist independently of its parent class.

**Composition:** The composition is a subset of aggregation. It portrays the dependency between the parent and its child, which means if one part is deleted, then the other part also gets discarded. It represents a whole-part relationship.

 **Fig : CLASS DIAGRAM**

1. **RESULTS**

**5.1 IMPORTING PACKAGES AND DATA SET**

**1. Introduction**

We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic.

Do you trust all the news you consume from online media?

Every news that we consume is not real. If you listen to fake news it means you are collecting the wrong information from the world which can affect society because a person’s views or thoughts can change after consuming fake news which the user perceives to be true.

Since all the news we encounter in our day-to-day life is not authentic, how do we categorize if the news is fake or real?

In this article, we will focus on text-based news and try to build a model that will help us to identify if a piece of given news is fake or real.

Before moving to the practical things let’s get aware of few terminologies

**2. Terminologies**

**2.1 Fake News**

A sort of sensationalist reporting, counterfeit news embodies bits of information that might be lies and is, for the most part, spread through web-based media and other online media.

This is regularly done to further or force certain kinds of thoughts or for false promotion of products and is frequently accomplished with political plans.

Such news things may contain bogus and additionally misrepresented cases and may wind up being virtualized by calculations, and clients may wind up in a channel bubble.

**2.2 Tfidf Vectorizer**

**TF (Term Frequency):** In the document, words are present so many times that is called term

frequency. In this section, if you get the largest values it means that word is present so many times with respect to other words. when you get word is parts of speech word that means the document is a very nice match.

**IDF (Inverse Document Frequency):** in a single document, words are present so many times, but also available so many times in another document also which is not relevant. IDF is a proportion of how critical a term is in the whole corpus.

collection of word Documents will convert into the matrix which contains TF-IDF features using Tf idf Vectorizer.

**3. Project**

To get the accurately classified collection of news as real or fake we have to build a machine learning model.

To deals with the detection of fake or real news, we will develop the project in python with the help of ‘sklearn’, we will use ‘TfidfVectorizer’ in our news data which we will gather from online media.

After the first step is done, we will initialize the classifier, transform and fit the model. In the end, we will calculate the performance of the model using the appropriate performance matrix/matrices. Once will calculate the performance matrices we will be able to see how well our model performs.

The practical implementation of these tools is very simple and will be explained step by step in this

article.

Let’s start

**3.1 Data Analysis**

Here I will explain the dataset. In this python project, we have used the CSV data set. The data set contains 7796 rows and 4 columns.68 This data set has four columns,

1. **title**: this represents the title of the news.

2. **author**: this represents the name of the author who has written the news.

3. **text**: this column has the news itself.

4. **label**: this is a binary column representing if the news is fake (1) or real (0)

**6.CONCLUSION**



We have classified our news data using three classification models. We have analysed the

performance of the models using accuracy and confusion matrix. But this is only a beginning point for the problem. There are advanced techniques like BERT, GloVe and ELMo which are popularly used in the field of NLP. If you are interested in NLP, you can work forward with these techniques.

The passive-aggressive classifier performed the best here and gave an accuracy of 93.12%.We can print a confusion matrix to gain insight into the number of false and true negatives and positives Fake news detection techniques can be divided into those based on style and those based on content, or fact-checking. Too often it is assumed that bad style (bad spelling, bad punctuation, limited vocabulary, using terms of abuse, ungrammaticality, etc.) is a safe indicator of fake news. More than ever, this is a case where the machine’s opinion must be backed up by clear and fully verifiable indications for the basis of its decision, in terms of the facts checked and the authority by which the truth of each fact was determined.

Collecting the data once isn’t going to cut it given how quickly information spreads in today’s connected world and the number of articles being churned out.

**7.FUTURE SCOPE**

The fake news challenge is perilous and is spreading rapidly like a wildfire as it becomes easier for information to reach the mass in various flavors. Reports have shown that, just like in the last US presidential elections, fake news can have a huge impact in politics and thereafter on the people like a domino effect. With the help of artificial intelligence, we can control and limit the spread of such misinformation more quickly and efficiently as compared to manual efforts.

The work in this project proposes a stacked model which fine tunes the informational insight gained from the data at each step and then tries to make a prediction. Although many attempts have been made to solve the problem of fake news, any significant success is yet to be seen. With huge amountsof data collected from social media websites like Facebook, Twitter, etc., the best models improve every day. With the use of deep neural networks, the future work in this field seems a lot more promising. The limitations that come packaged with this problem is that, the data is erratic and this means that any type of prediction model can have anomalies and can make mistakes. For future improvements, concepts like POS tagging, word2vec and topic modelling can be utilized. These will give the model a lot more depth in terms of feature extraction and fine-tuned classification. Word2Vec: The Word2Vec technique converts text to features while maintaining the original

relationships between words in a corpus. It is a combination of techniques and is one of the best feature extraction techniques in NLP

It generally uses a model of pretrained vectors (like GloVe) and then transfer learning can be used to obtain a superior model. Topic Modelling: News can contain a vast range of topics. Just the classification based on labels is not enough if realistic results are desired. For this reason, an advanced technique called topic modelling can come in handy. Topic modelling categories each piece of text into topics and using this one can make more accurate predictions. The most popular topic modelling technique used.