



A Dissertation Report on

**“EVALUATION OF MACHINE LEARNING METHODS FOR  
DETECTING LEGITIMACY OF NEWS ON EXTENDED DATASET”**

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD

OF THE DEGREE OF

**INTEGRATED DUAL DEGREE B. TECH (Computer Science & Engineering)  
+ M. TECH (Artificial Intelligence & Robotics)**

**By**

Jagriti Shahi(16/ICS/025)

**Under the guidance of:**

Dr. Anurag Singh Baghel

SCHOOL OF INFORMATION, COMMUNICATION & TECHNOLOGY  
GAUTAM BUDDHA UNIVERSITY,  
GREATER NOIDA  
Session: 2016-2021



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY  
GAUTAM BUDDHA UNIVERSITY, GREATER NOIDA, 201312, U. P., (INDIA)

### **Candidate's Declaration**

I, hereby, declare that the work embodied in this dissertation(Part-II) entitled “Evaluation of Machine Learning Methods for Detecting Legitimacy of News on Extended Dataset” submitted in partial fulfilment of the requirements of the award of the degree of Integrated B.Tech (Computer Science &Engineering) + M.Tech with Specialization in ARTIFICIAL INTELLIGENCE & ROBOTICS submitted to the School of Information and Communication Technology, Gautam Buddha University, Greater Noida is an authentic record of my own bonafide work carried out under the supervision of Dr. Anurag Singh Baghel, School of ICT and is correct to the best of my knowledge and belief. This work has been undertaken taking care of engineering ethics. It contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of any university or other institute of higher learning, except where due acknowledgement has been made in the text. Responsibility for any plagiarism related issue stands solely with me.

Name and Signature of the Student: Jagriti Shahi

This is to certify that the above statement made by the candidates is correct to the best of my knowledge and belief. However, responsibility for any plagiarism related issue solely stands with the student.

Signature of the Supervisor.....

Name with Designation.....

Date: 30-06-2021

Place: Greater Noida

## **ACKNOWLEDGEMENT**

I would like to extend my sincere thanks to Dr. Anurag Singh Baghel for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project.

By -

Jagriti Shahi

(16/ICS/025)

## **ABSTRACT**

Recently, fake news have been creating many problems to our society. Fake news is widely spread everywhere that it is sometimes very difficult to determine if they are legitimate or not. Social media is a major source for the spread of false information, since it is trouble-free, low-cost and can be easily accessed by any individual. Millions of news are floating on social network whose sources are not verified. This type of misinformation can be both good as well as harmful depending on how it is presented. This false information are spread with an intention to influence beliefs and decision of people or to affect the major events such as political elections.

In this dissertation work, we have used a larger dataset, performed some pre-processing technique and evaluated various machine learning and natural language processing algorithms such as Random Forest classifier, K-nearest neighbours and LSTM techniques that categorizes the news as genuine or fake with a better accuracy.

## TABLE OF CONTENTS

S.no	Topics			Page no.
1.	Introduction			8-9
2.	Theoretical background			10-19
	2.1	Machine Learning		10-16
		2.1.1	Classification of Machine Learning on the basis of nature of learning	10-12
		2.1.2	Classification according to Output required	12-15
		2.1.3	Classification Vs Regression	15-16
	2.2	Natural Language Processing		16-19
		2.2.1	Applications of NLP	17-18
		2.2.2	Future of NLP	18
		2.2.3	Desirable features of NLP	19
		2.2.4	Pitfalls of NLP	19
3.	Literature Survey			20-21
4.	Problem Identification			22
5.	Research Methodology			23-40
	5.1	Data Preprocessing		23-26
		5.1.1	Stemming	23
		5.1.2	Tokenizing	24-25
		5.1.3	Stop word removal	25-26
	5.2	Feature Extraction		26
		5.2.1	Doc2Vec Model	26
	5.3	Algorithms Used		27-35
		5.3.1	K-Nearest Neighbor	27-28
		5.3.2	Random Forest Classifier	29-30
		5.3.3	Long Short-Term Memory	30-35
	5.4	Language Used		35
	5.5	Tools & Libraries Used		36-40

6.	Implementation	41-48
	6.1 Dataset Description	41-43
	6.2 Data Pre-processing	44-47
	6.3 Algorithms	47-48
7.	Results & Performance Evaluation	49-50
8.	Conclusion	51
9.	References	52-53

## LIST OF FIGURES

<b>S.no</b>	<b>Figures</b>	<b>Page no.</b>
1.	Fig 1. Types of Machine learning	10
2.	Fig 2. Classification	13
3.	Fig 3. Regression	14
4.	Fig 4. Clustering	15
5.	Fig 5. Natural Language Processing	16
6.	Fig 6. Applications of NLP	17
7.	Fig 7. Stemming	24
8.	Fig 8. Before KNN	27
9.	Fig 9. After KNN	28
10.	Fig 10. Random Forest Classifier	30
11.	Fig 11. Life Cereal Review	31
12.	Fig 12. LSTM	32-33
13.	Fig 13. Flow of Implementation	41
14.	Fig 14. train.csv	42
15.	Fig 15. train_news.csv	43
16.	Fig 16. Comparison of Algorithms	50

# **CHAPTER 1: INTRODUCTION**

## **1.1 Introduction**

Counterfeit news can be characterized as the bogus or untruthful data distributed as a certified news to deceive per users of the substance and spread fake data by means of interpersonal organizations and informal. The motive behind this kind of act can be for harming the reputation of any individual or organization or making profit by advertising manipulated news. It can be expressed as: the data which is in reality bogus, yet appears to be valid.

Because the goal of misinformation is to disseminate false claims, the clearest method of identifying it is to examine the validity of key examples in stories to determine the originality of content.

In the age of the innovation, a gigantic measure of information is created online consistently. Nonetheless, an amazing measure of information overwhelmed is phony information to draw in crowd. Web-based media have raised it amazingly simple and trouble-free to get and spread any false or misdirecting data which brings about impacting the convictions and choice of individuals or to influence the significant events like political decisions. Fraudulent news detection is an active area of research which is constantly expanding, but somehow it faces several challenges considering the limited number of entities available.

In the recent years, online substance has been assuming a critical part in influencing client's choices and suppositions. Sentiments, for example, online surveys are the fundamental wellspring of data for e-commerce clients to assist with acquiring understanding into the items they are arranging to purchase.

As of late it has become obvious that assessment spam doesn't just exist in item audits and client's criticism. Indeed, counterfeit news and deluding articles is another type of assessment spam, which has acquired attention. The absolute greatest wellsprings of



getting out counterfeit word or bits of gossip are web-based media sites like Google Plus, Facebook, Twitters, and other web-based media outlet.

Despite the fact that the issue of phony news is certainly not another issue, recognizing counterfeit news is accepted to be an intricate assignment given that people will in general think deceiving data and the absence of control of the spread of phony substance. Counterfeit news has been standing out enough to be noticed over the most recent few years, particularly since the US political decision in 2016. It is extreme for people to distinguish counterfeit news. It tends to be contended that the solitary path for an individual to physically distinguish counterfeit news is to have an immense information on the covered subject. Indeed, even with the information, it is significantly difficult to effectively recognize if the data in the article is genuine or counterfeit.

The open thought of the web and online media notwithstanding the new improvement in computer science simplify the way toward making and getting out counterfeit word. While it is clearer and follow the goal and the effect of phony surveys, the expectation, and the effect of making purposeful publicity by getting out counterfeit word can't be estimated or seen without any problem. For instance, clearly counterfeit survey impacts the thing proprietor, client and online stores; then again, it is difficult to identify the components influenced by the phony news. This is on the grounds that distinguishing these substances require estimating the news proliferation, which has demonstrated to be unpredictable and asset serious.

Distinguishing counterfeit news is accepted to be an unpredictable undertaking and a lot harder than recognizing counterfeit item audits given that they spread effectively utilizing web-based media and word of mouth.

## **CHAPTER 2: THEORETICAL BACKGROUND**

### **2.1 Machine Learning**

Machine Learning can be defined as a subfield of AI that provides the ability to machines to improve their performance by learning from their own experience, without being expressly modified. It revolves around the improvement of computer programs by gaining information and utilizing it to learn on their own.

#### **2.1.1. Categorization of Machine Learning on the basis of nature of learning**

Machine Learning can be categorized into three significant types on the basis of its nature of learning:

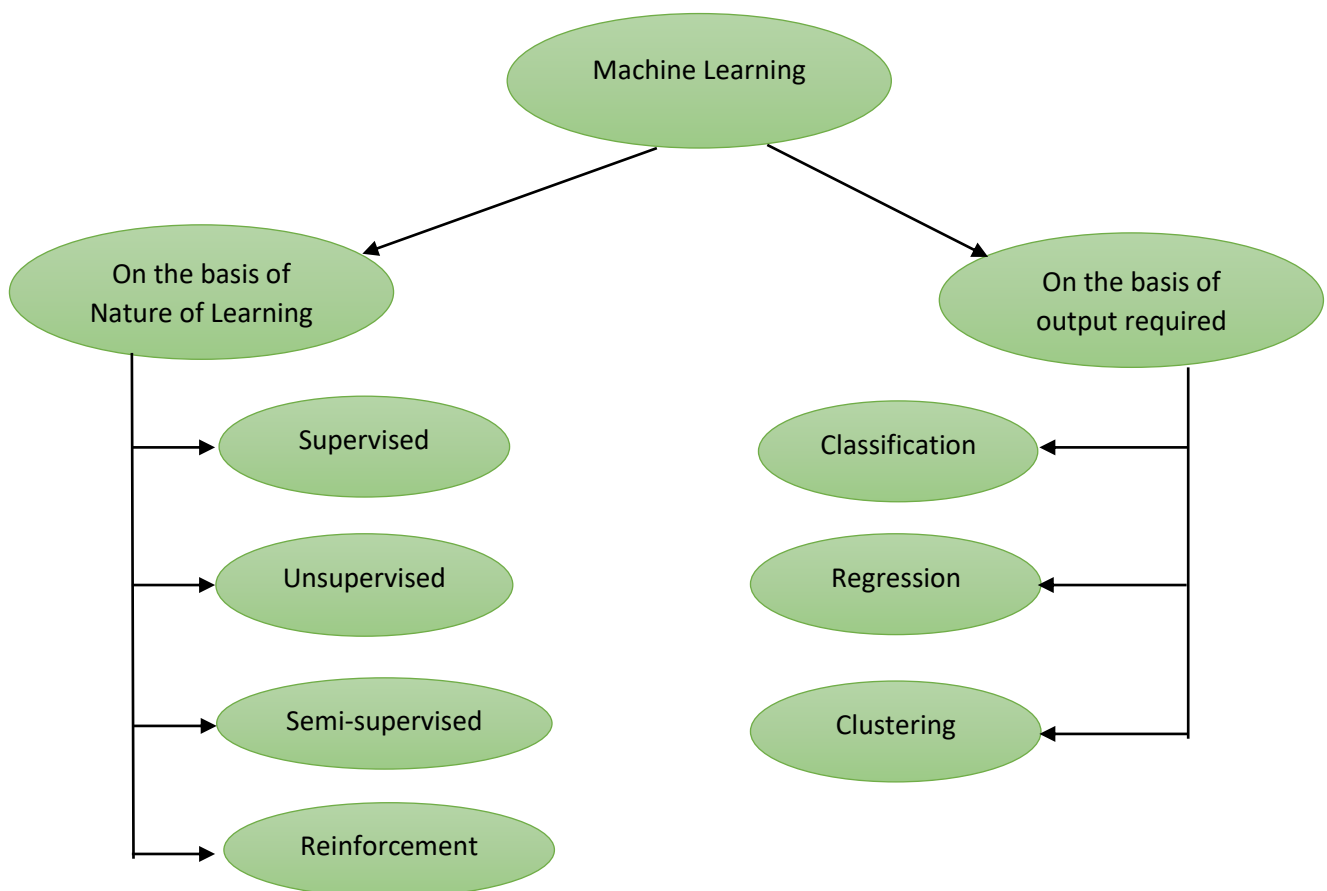


Fig1: Types of Machine Learning

- **Supervised learning:**

When a machine is trained with labelled data and on premise of that information, anticipations are done, this type of learning is categorized as supervised. The labelled data means that some of the input information will be already mentioned with right yield. As the name “supervised” indicates, this methodology is undoubtedly same as the learning under the supervision of teacher.

In this type of learning, the data given for training is paired with the its right yield too. Then the model analyses for the pattern in input that coordinate with the output while training the data. After the training, the model is fed with new input data and the task is to determine its correct output on the basis of previous training.

- **Unsupervised learning:**

This learning is unaided learning that dissect and group untagged datasets utilizing AI calculations. These algorithms help in detecting the hidden patterns or information groupings without the requirement for human mediation. It provides the capacity to find likeness and contrasts in data and thus making it the best answer for exploratory information inspection, strategically pitching techniques, client division, and picture recognition.

This algorithm can be utilized for performing more complex task as compared to that of supervised algorithms.

- **Reinforcement learning:**

At the point when you present the methods with models that does not have labels, as in unsupervised learning. It is the development of AI techniques that make decisions on a variety of options. The expert determines how to complete a task during an uncertain, perhaps difficult system. A man-made rationale considers a game-like situation to aid learning. To fix the issues, the computer use experimentation. To have the system to accomplish whatever we want, we use

computerized reasoning, which receives either incentives or penalties for the actions it does. It is very certain that doing so will raise the entire reward.

- **Semi-supervised learning:**

In this kind of learning, the algorithm is prepared upon a mix of both tagged and untagged informational records.

Routinely, this mix will contain a little amount of named information and a ton of unnamed information. The crucial framework included is that first, the programmer will bunch relative information using a unsupervised learning calculation, and subsequently utilize the current named information to name the rest of the untagged information.

The conventional use examples of such kind of calculation have a regular property among them – The obtainment of untagged information is by and large not unreasonably much while marking the referenced information is very costly.

Semi administered AI calculations are applied in an assortment of organizations starting from fintech and finishing with diversion applications. In banking, ML systems expect a key part since they assist relationship with building information security.

### **2.1.2. Classification according to Output required**

When the optimal yield of a machine-learned framework is considered, another categorization of machine learning emerges:

- i. **Classification:**

It may be described as the interaction of classifying a given arrangement of data into categories. It may be used to both structured and chaotic data. The process starts by anticipating the kind of provided content's emphasis. The categories are commonly referred to as the goal, the name, or the categorizations.

For example: speech recognition, spam detection, face recognition etc. are all classification issues.

The methodology that carries out the categorization on a dataset is known as a classifier. There are two sorts of Classifications:

- **Binary Classifier:**

When the characterization issue has just two probable results, then, at that point it is called as Binary Classifier. For instance: YES or NO, 0 or 1, WHITE or BLACK, CAT or DOG, and so on.

- **Multi-class Classifier:**

If a characterization issue has multiple results, then, at that point it is called as Multi-class Classifier. For instance: Classifications of types of fruits, Classification of sorts of music.

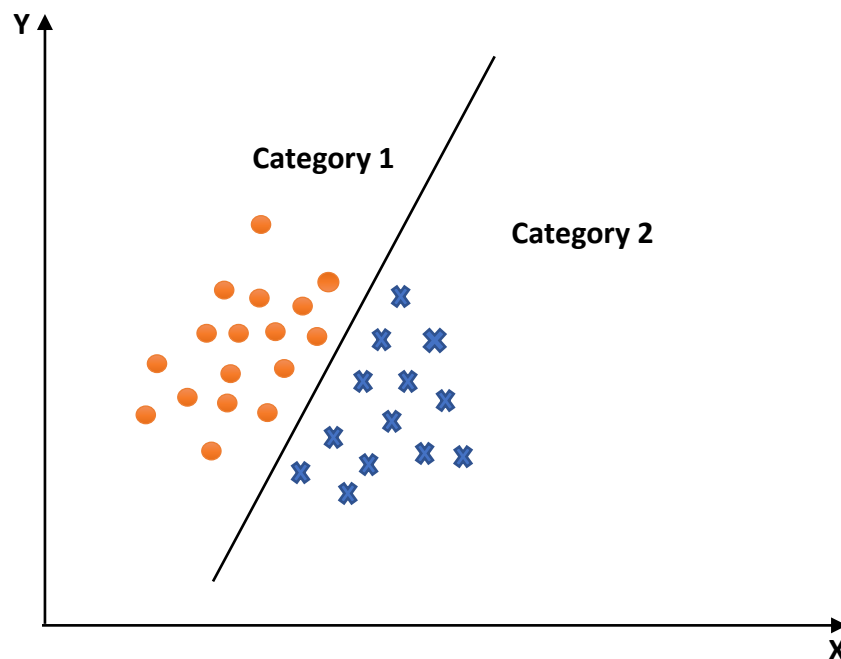


Fig2: Classification

## ii. **Regression:**

Which is additionally a supervised issue, a situation when the yields are continuous instead of discrete.

Regression analysis is an approach to discover patterns in information. This approach is mostly used for assessing and identifying situations as well as logical outcomes connections between elements. Relapse techniques vary depending on the number of free variables and the type of interaction seen between self-governing and confined elements. Such a well relapse concerns are, for example, predicting home costs or salary of a person, and so on.

All the more explicitly, Regression analysis assists us with seeing how the worth of the reliant variable is changing comparing to a free factor when other autonomous factors are held fixed. It predicts continuous/genuine qualities like temperature, age, salary, cost, and so forth.

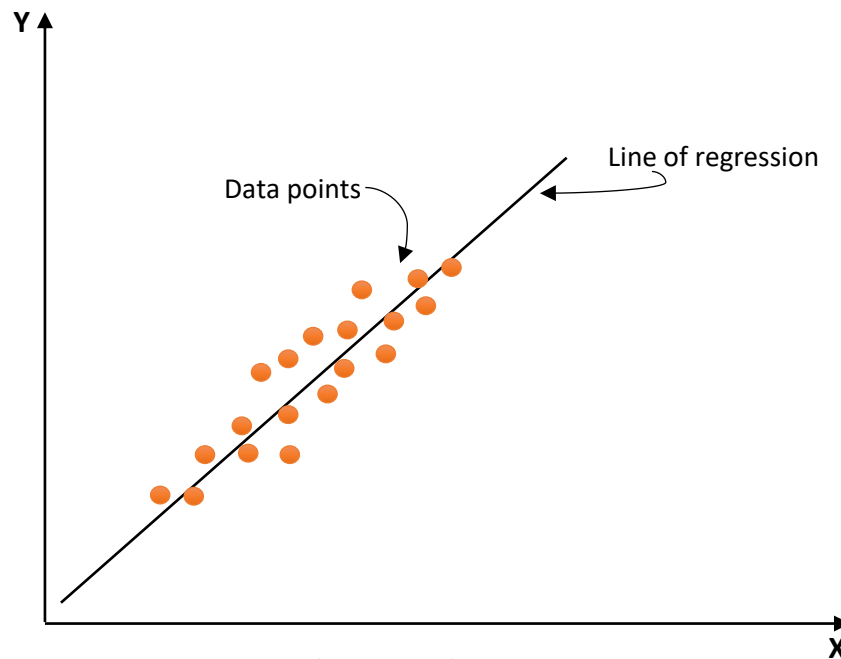


Fig3: Regression

### iii. Clustering:

As the name recommends, bunching includes separating information focuses into various groups of comparative qualities. At the end of the day, the goal of bunching is to isolate bunches with comparable qualities and pack them together

into various groups. It is preferably the execution of human psychological ability in machines empowering them to perceive various items and separate between them dependent on their regular properties. In contrast to people, it is hard for a machine to recognize from an apple or an orange except if appropriately prepared on an enormous pertinent dataset. This preparation is accomplished by solo learning calculations, explicitly bunching.

We can recognize the groups, and we can distinguish that there are 2 bunches in the underneath picture.

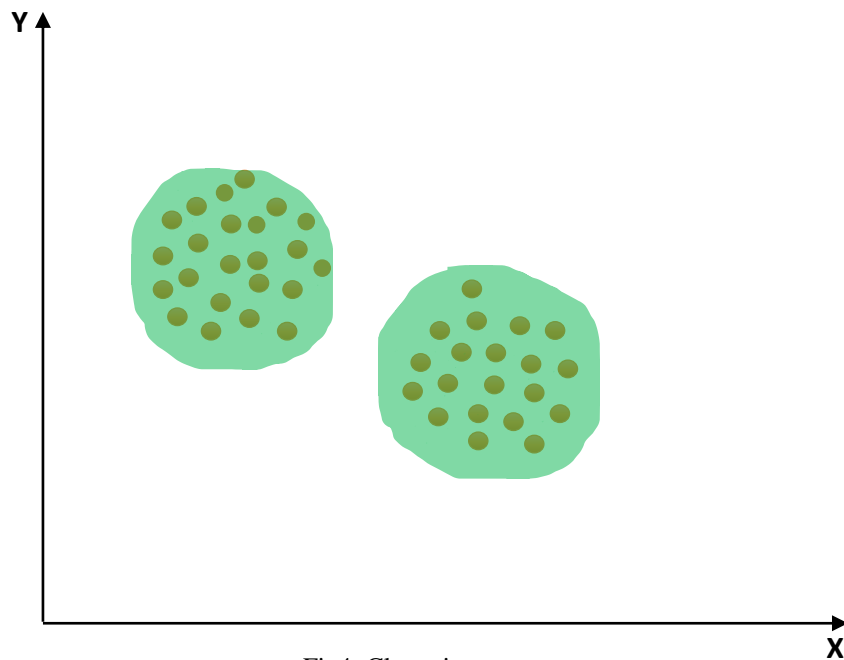


Fig4: Clustering

### 2.1.3. Classification vs Regression

Classification can be defined as the process of discovering a model or capacity which assists in isolating the information into numerous unmitigated classes for example discrete qualities. In order, information is classified under various names as per a few boundaries given in information and afterward the marks are anticipated for the information. The determined planning capacity could be exhibited as "IF-THEN" rules.

The characterization cycle manages the issues where the information can be isolated into paired or various discrete names.

Regression is the technique of locating a model or potential for relevant details into continuous real features instead of just using categories or discrete qualities.

It can likewise recognize the dissemination development relying upon the recorded information. Since a regression forecasting model predicts an amount, consequently, the ability of the model should be accounted for as a blunder in those expectations.

## 2.2 Natural Language Processing

It is viewed as a subdiscipline of AI technology & linguistics that handles the communication of machines and humans. Natural language is a language through which we humans communicate to each other. For example: text, speech. We are all surrounded by text. A definitive objective of NLP is to empower computers to understand language in the same way humans do.

NLP is computerized approach to comprehend and investigate common human language and concentrate data from such information by applying AI calculation.

For example: Google Assistant Application: it takes inquiries from human that can be composed and spoken and answer them as needs be.

NLP is one of the fields that intensely profited with the new advances in Machine Learning, particularly from Deep Learning strategies.

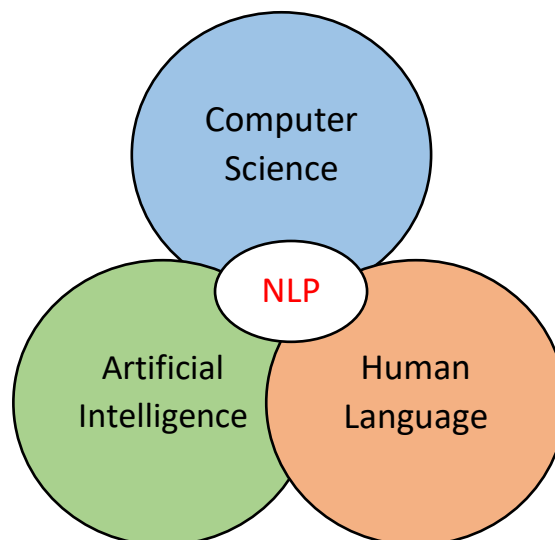


Fig5: Natural Language Processing

Source: <https://static.javatpoint.com/tutorial/nlp/images/what-is-nlp.png>



### 2.2.1 Applications of NLP

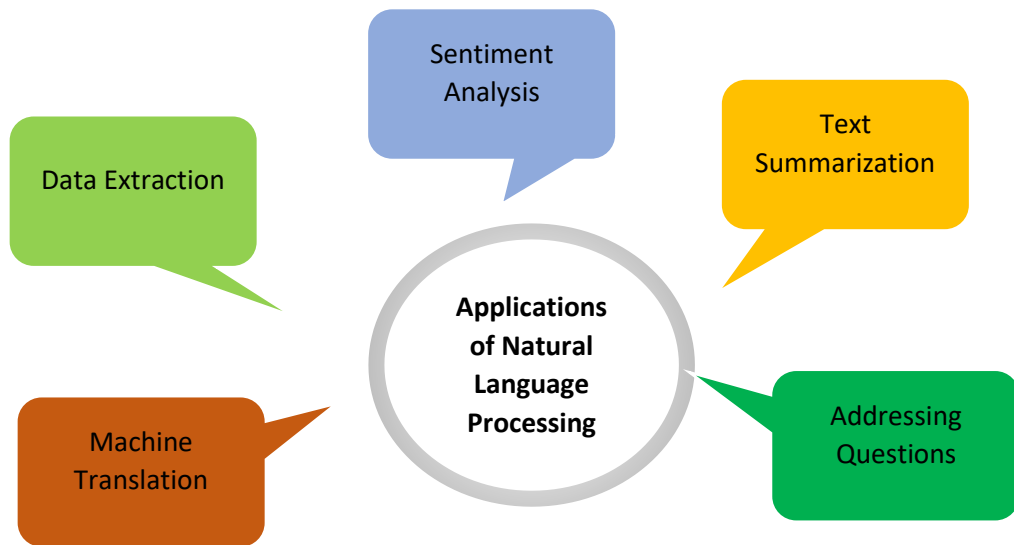


Fig6 : Application of NLP

- **Addressing Questions**

Question replying is a software engineering discipline inside the fields of data recovery and normal language handling, which deals with building frameworks that consequently answer questions presented by people in a characteristic language.

- **Data Extraction**

This procedure is the task of naturally distinguishing structured data from chaotic and semi-arranged machine-meaningful entries and some other digitally referenced sources. In the great majority of situations, this trend is concerned with the preparation of human language communications using natural language processing (NLP). Recent operations in emerging technology record management, such as automated explanation and component extraction from images/sound/video/reports, might be considered data extraction.

- **Machine Translation**

Machine Translation (MT) is the assignment of consequently changing over one common language into another, saving the importance of the information text, and delivering familiar content in the yield language.

- **Text Summarization**

It is a method that abbreviates a long piece of substance with central matters laid out that gives a thought of the entire substance. It becomes basic when somebody needs a fast and precise synopsis of extremely long substance. Summing up text can be costly and tedious whenever done manually.

- **Sentiment Analysis**

This can be defined as the process of translating and categorizing the feelings into text information availing text investigation strategies. Feeling examination apparatuses permit organizations to recognize client supposition toward items, brands or administrations in online criticism.

### **2.2.2 Future of NLP**

- Comprehensible regular language handling is the greatest AI-issue. It is all generally same as tackling the focal man-made consciousness issue and having machines as intellectual as human beings.
- With assistance of NLP, future machines will be able to gain from data available on internet and will apply in reality, in any case, a load of efforts is needed.
- Natural language toolbox or nltk turned more viable.

- Joined with normal language generation, machines will turn out to be more equipped for getting and giving valuable and creative data or information.

### **2.2.3 Desirable features of NLP**

- It provides immediate response to customers queries regardless of the time or topic the question has been asked.
- These systems provide responses in a normal language that can be easily understandable by customers.
- It can provide more accurate answers of customer queries as compared to humans.
- It also assists machines to deal with customers speaking in their language and in other work related to specific language.
- Structuring an exceptionally unstructured information source.

### **2.2.4 Pitfalls of NLP**

- Complex Query Language-the framework will be unable to give the right answer if the inquiry is ineffectively phrased or vague.
- The framework is worked for a solitary and explicit assignment in particular; it can't adjust to new spaces and issues due to restricted capacities.
- NLP framework lacks a UI which does not have feature that permit clients to additionally communicate with the framework.

### **CHAPTER 3: LITERATURE SURVEY**

A literature study portrays the different investigations and examination made in the field of revenue of the venture and the outcomes previously distributed. It is a significant part as it provides a guidance in the space of your exploration. It assists with defining an objective for your examination along these lines inferring at the difficult assertion.

In the wake of considering different boundaries and the degree of the venture, the accompanying papers were analyzed:

- i. In [1], authors have utilized three distinct methods to group the authenticity of illegitimate news. The initial two approaches carry out various blends of AI calculations, for both unsupervised clustering and classification, which anticipate the kind of information via training just the genuine news. Similarly, the third method develops the recognizable proof proposition for a verifiable circumstance, considering the hypothesis that legitimate and illegitimate news have unmistakable possible scatterings while taking into account module of their depiction in the vector space of repetition of words.

And the main technique "The Reduction Methodology with Training" utilized one-class SVM to defeat the impediment of having a marked base with genuine and counterfeit news in supervised learning. As opposed to run of the typical SVM executions, the one-class considers a bunch of training samples for a single class.

In evaluating the nature of recognition, it brings about 86% accuracy, and 94% precision even subsequent to utilizing a dimensional reduction.

- ii. In the paper [2] by Fathima Nada, Bariya Firdous Khan, Aroofa Maryam, Nooruz-Zuha, Zameer Ahmed, a detection model for counterfeit news is introduced utilizing tf-idf and diverse feature extraction methods. They had obtained the exactness of around 72% when utilizing tf-idf and calculated relapse classifier.

- iii. In the paper [3] by Shu A et al., a point-by-point study regarding distinguishing counterfeit news via online media have been done that includes enactment of counterfeit news on social hypotheses and psychology, & review of existing approaches used in false news identification from a viewpoint of an information mining, methods of extracting features, and creation of models. They additionally talk about other research regions related to it, metrics of evaluations, representative datasets, future research direction in this field, open issues for identification of counterfeit news via web-based media.
- iv. In the paper [4] by Mykhailo Granik et al, a simplified and easy way has been analyzed for recognizing the counterfeit news utilizing Naïve bayes approach. Instead of news stories on the internet, this methodology was directed mainly for news post of Facebook. They have accomplished an accuracy of 74%. And here, it have been also expressed that the simple algorithms like Naïve bayes can obtain the moderate accuracy so as to increase it further other AI technique can be utilize to handle the hazard of phony news. It likewise proceeded to express that basic AI models like the Naïve Bayes can accomplish a moderate exactness and in future, more computerized reasoning procedures could be utilized to handle the hazard of phony news.
- v. In paper [5] by Akshay Jain, they had proposed a system for detecting "false information," with plans to deploy this on Facebook, a well-known social network. And to make prediction on Facebook that whether the post is genuine or not by using Naïve bayes model. Also, they have recommended numerous strategies in the paper for improving the precision of arrangement.

## **CHAPTER 4: PROBLEM IDENTIFICATION**

In online media, it is a two-sided deal for news, utilization. From one of the perspectives, it is of minimal expense, easy access, and a quick outspread of data which entices certain people to seek out and acquire news from web-based media. On the other hand, it allows for the widespread dissemination of "falsehoods," i.e., poor quality news with deliberately misleading facts. The widespread dissemination of fake news has the ability to have serious implications for individuals and communities.

As a result, phony news acknowledgement over social media platforms has recently been an emerging research field that is attracting a lot of attention. Therefore, the problem is to distinguish the legitimacy of news and online content.

This report discusses about the methodology of machines learnings and natural language processing to tackle the counterfeit news identification problem. The purpose of this research is:

- To detect illegitimate news, which is a classic text classification problem by using existing machine learning models on extended dataset.
- To compare some of predictions method.
- To report a method that has achieved highest accuracy among all the compared models.

## **CHAPTER 5: RESEARCH METHODOLOGIES**

The theoretical ideas and algorithms utilized in the advancement of the framework are clarified in the resulting sections.

### **5.1 Data pre-processing**

Information preprocessing is that step of information preparation that includes some data mining technique where the raw information is converted into a format that could be understandable by machine or the machine can undoubtedly parse it. All in all, now the algorithms would be able to effectively elucidate the highlights of information.

Prior to addressing the information utilizing the different models, the information should be exposed to particular refinements like stop-word elimination, tokenization, conversion of letters into lower case, and removal of punctuations. This will assist us with lessening the length of real information by eliminating the unessential data that occurs in the information. A conventional preparing function was made to take out punctuation and non-letter characters for each article. Then, in the letters in the archive were lowercased prior to eliminating the stop words.

#### **5.1.1 Stemming:**

Stemming is a procedure used to separate the base type of the words by eliminating attaches from them. It is actually similar to chopping down the parts of a tree to its stems. Comprehending, searching for, and retrieving more sorts of words yields more results. When a certain sort of term is recalled, it may be possible to retrieve query items that were previously missed. That additional data acquired is the purpose stemming is essential for searching queries and data recovery.

For example: the stem of word “waiting”, “waited” and “waits” is wait.

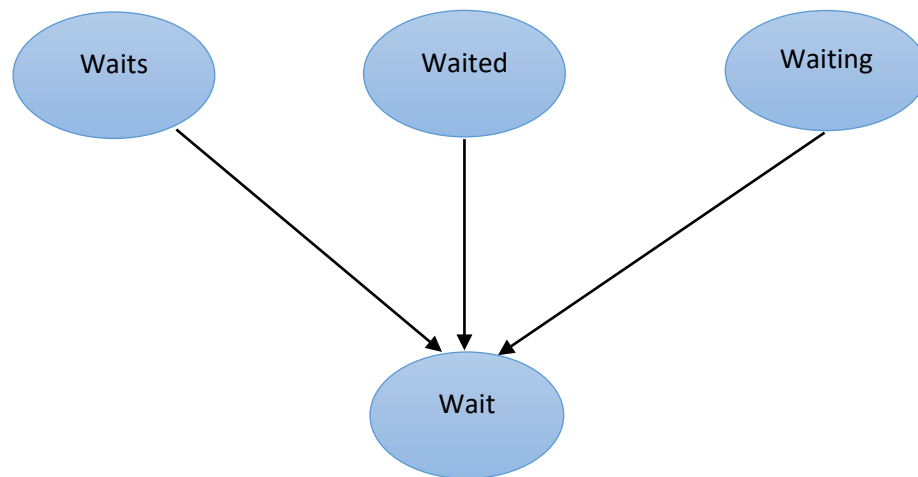


Fig7: Stemming

### 5.1.2 Tokenizing:

Tokenization is the process of separating a piece of text into more modest parts, like sentences and words. Tokens can be particular words, phrase or even full sentences. While tokenizing, some characters like punctuation marks are discarded. The tokens become the contribution for another interaction like parsing and text mining.

Tokenizing sentence into words using `word_tokenize ()` method:

```
from nltk.tokenize import sent_tokenize, word_tokenize

data = "All work and no play makes jack a dull boy, all work and no play"
print(word_tokenize(data))
```

Output:

```
['All', 'work', 'and', 'no', 'play', 'makes', 'jack', 'dull', 'boy', ',', 'all', 'work', 'and', 'no', 'play']
```

This same process can be used for tokenizing a paragraph into sentences by using `sent_tokenize ()` method:



```
from nltk.tokenize import sent_tokenize, word_tokenize

data = "All work and no play makes jack dull boy. All work and no play makes  
jack a dull boy."  
print(sent_tokenize(data))
```

Output:

```
['All work and no play makes jack dull boy.', 'All work and no play makes jack  
dull boy.']
```

### 5.1.3 Stop word Removal:

A stop word can be defined as most appearing or usual word that does not add much importance to the meaning of sentence, (for instance, "to", "this", "of", "in") For these words, search engine is modified to overlook, in both cases while input is entered as search keyword and while fetching the results of inquiry question.

Since these words are meaningless so we can safely eliminate them and save the memory space & significant handling time required by these words. They could be deleted by storing a list of terms that are considered stop words. NLTK provides a list of stop words stored in 16 different languages in Python (Natural Language Toolkit). They may be found with in nltk data section.

By typing the following command, we can check the list of stop words in a python shell:

```
import nltk  
nltk.download("stopwords")  
from nltk.corpus import stopwords  
print(stopwords.words('english'))
```

Output:

---

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

---

## 5.2 Feature Extraction

The Doc2Vec concept is used to produce the majority of the embeddings used in modelling. The goal is to build a similarity metric of each document. We undertake many required information pre-preparation steps prior to deploying Doc2Vec. This entailed deleting stop words, odd symbols, and punctuation, and then transforming all content to lowercases. This generates a comma-separated series of items that may be used in the Doc2Vec computation to generate a 300-length embedded variable with each content.

### 5.2.1. Doc2Vec Model:

Doc2Vec models were constructed in 2014, based on the actual Word2Vec model, that generates vector depiction of words. Word2Vec handles files by combining the vectors of particular words, but it destroys most word sequence information in the process. Doc2Vec extends Word2Vec via inserting a "document vector" to the output representation, that provides certain information more about file as a whole as well enables the algorithms to become familiar with some data about word sequence.

## 5.3 Algorithms Used:

### 5.3.1 K-Nearest Neighbor

The KNN computations are predicted on the presumption that similar items occur in close enough proximity. Overall, similar items are nearby to each other.

K-Nearest Neighbors is one of the most fundamental classification machine learning Models. It is integrated into the administered learning environment and uncovers amazing applications in pattern identification, mining techniques, and interrupted identification.

### Intuition

Imagine a scenario, where we have two categories already present in our dataset, so we have to identify two categories. For simplicity, taking into consideration the two column or fields (Let's say,  $X_1$  &  $X_2$ ) from our dataset, grouping is done.

Now, suppose we have to add a new data point in dataset, so the question arises that should it fall in red category or green category, how to decide that?



Fig8: Before K-NN

That's where the K-nearest neighbor algorithm will assist us. At the end of performing this algorithm we will be able to identify whether this new data point falls into red data point category or green data point category.

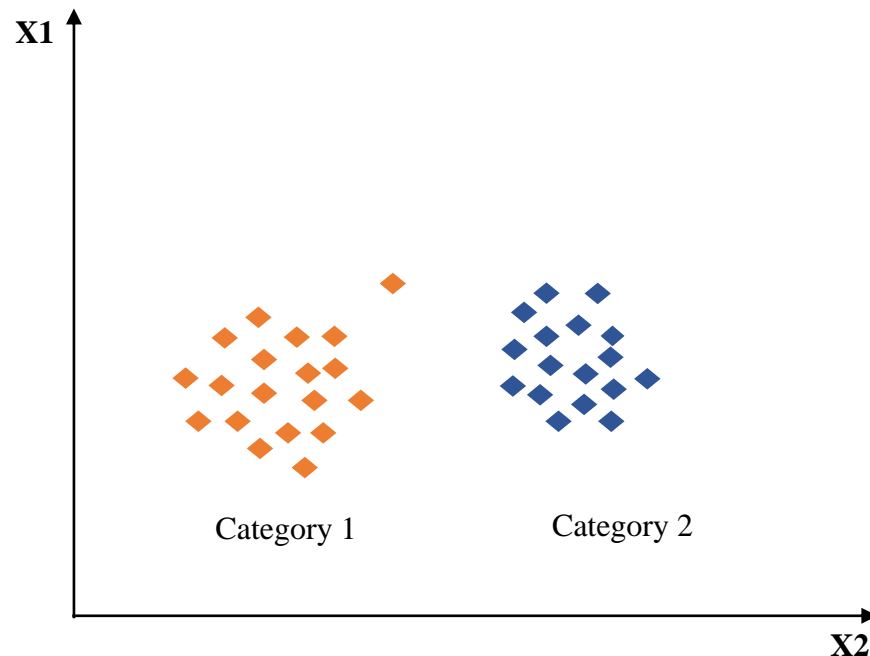


Fig9: After K-NN

## How does K-NN work?

**Step 1:** Choose the K, number of neighbors.

**Step 2:** As per Euclidean distance calculated, choose the neighbors of K that are closest to the new data point.

**Step 3:** Out of all these neighbors, check the total number of data points focuses in each category.

**Step 4:** On the basis of maximum closest neighbors in category, assign the new data point to that category.

**Step 5:** Now model is prepared!

### 5.3.2 Random Forest Classifier

Random Forest is a technique that is used in conjunction with the learning approach that is being used. It has the potential to be used in ML for both regression and classification problems. This is built on the notion of ensemble learning, that is a process of combining multiple assessors to deal with a puzzling difficulty and increase the model's representation.

Random Forest Model, as the term suggests, is a grouping of a number of decision trees on different subgroups of supplied datasets, with the aggregate used to enhance the forecasting accuracy of that dataset. Instead of relying on a single decision tree, Random Forest evaluates the performance of a number of decision trees by taking forecasting from each tree and forecasting the final outcome based on the highest number of votes.

The accuracy depends on no. of decision trees in random forest, more prominent number of trees prompts better accuracy and also prevents the problem of overfitting.

#### **Suppositions for Random Forest**

Because the random forest connects several trees to forecast the category of the dataset, it is probable that certain decision trees will anticipate the correct yield whilst others will not.

However, when all of the trees are combined, they anticipate the correct yield. As a result, below are two assumptions for an improved Random Forest technique:

- For a classifier to foresee exact outputs in comparison to speculated results, there ought to be few genuine qualities column of dataset.
- There should be very low connection for forecasts from each tree.

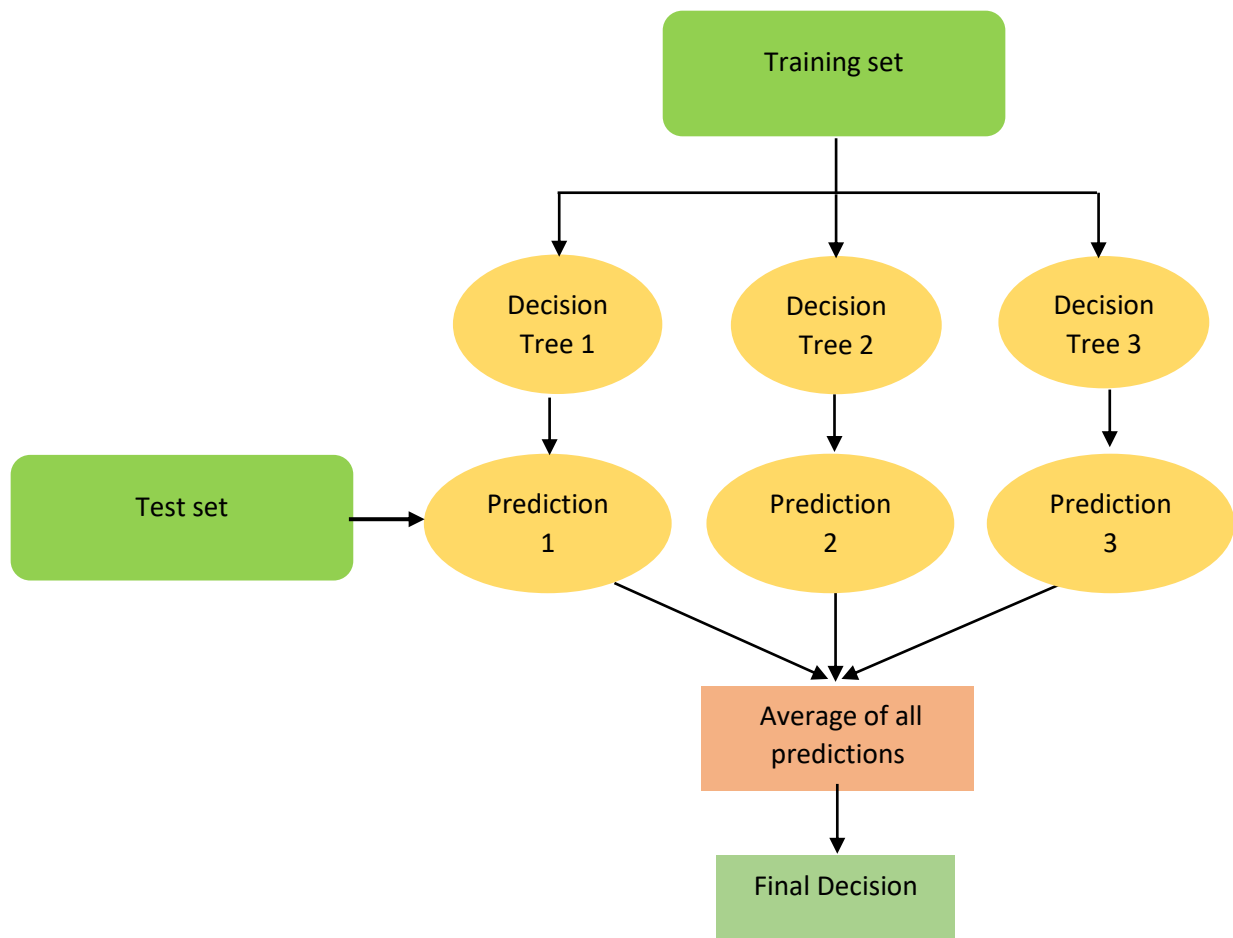


Fig10: Random Forest Classifier

### 5.3.3 Long Short-Term Memory (LSTM)

Since these networks broaden the memory of recurrent neural network so it can be referred as the expansion for RNN. The structure units for the layers of a RNN are the units of a LSTM, which is then regularly called a LSTM network. To recollect LSTM's contributions throughout a significant stretch of time, the LSTM's empower RNN's. Since the data is available in the memory for LSTMs, that is a similar to the memory of the computer since even the LSTM can peruse, compose and erase data from its memory. The LSTM is as yet a neural network. In any case, not the same as the completely associated neural network, it has cycled in the neuron connections.

RNNs are susceptible to the effects of temporary memory. If a series is relatively long, they will have difficulty transferring data from earlier time stages to newer ones. So, if you're attempting to predict anything from a stretch of text, RNNs might miss off material that is crucial all along.

LSTM 's was developed to overcome the problem of transient memory. They have internal system known as gates that can direct the progression of data.

These gates can realize which information in an arrangement is imperative to keep or discard. By doing that, it can pass pertinent data down the long chain of sequence to make forecasts.

## Intuition

Suppose we have to buy a life cereal and for that we are looking at it reviews online. We will initially peruse the survey then, at that point decide whether somebody thought it was acceptable or on the off chance that it was bad.

When we read the review, our brain subconsciously only recalls the significant words like “amazing”, “perfectly balanced breakfast” and so forth. We don’t probably recall it word by word, since we don’t care much about words like “gave”, “ate”, “this” etc.

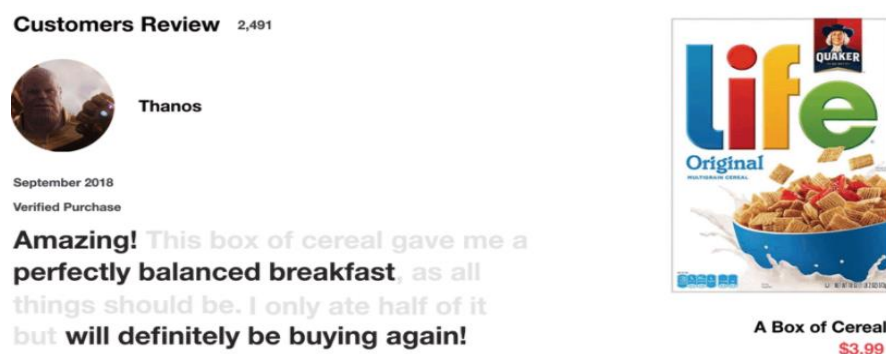
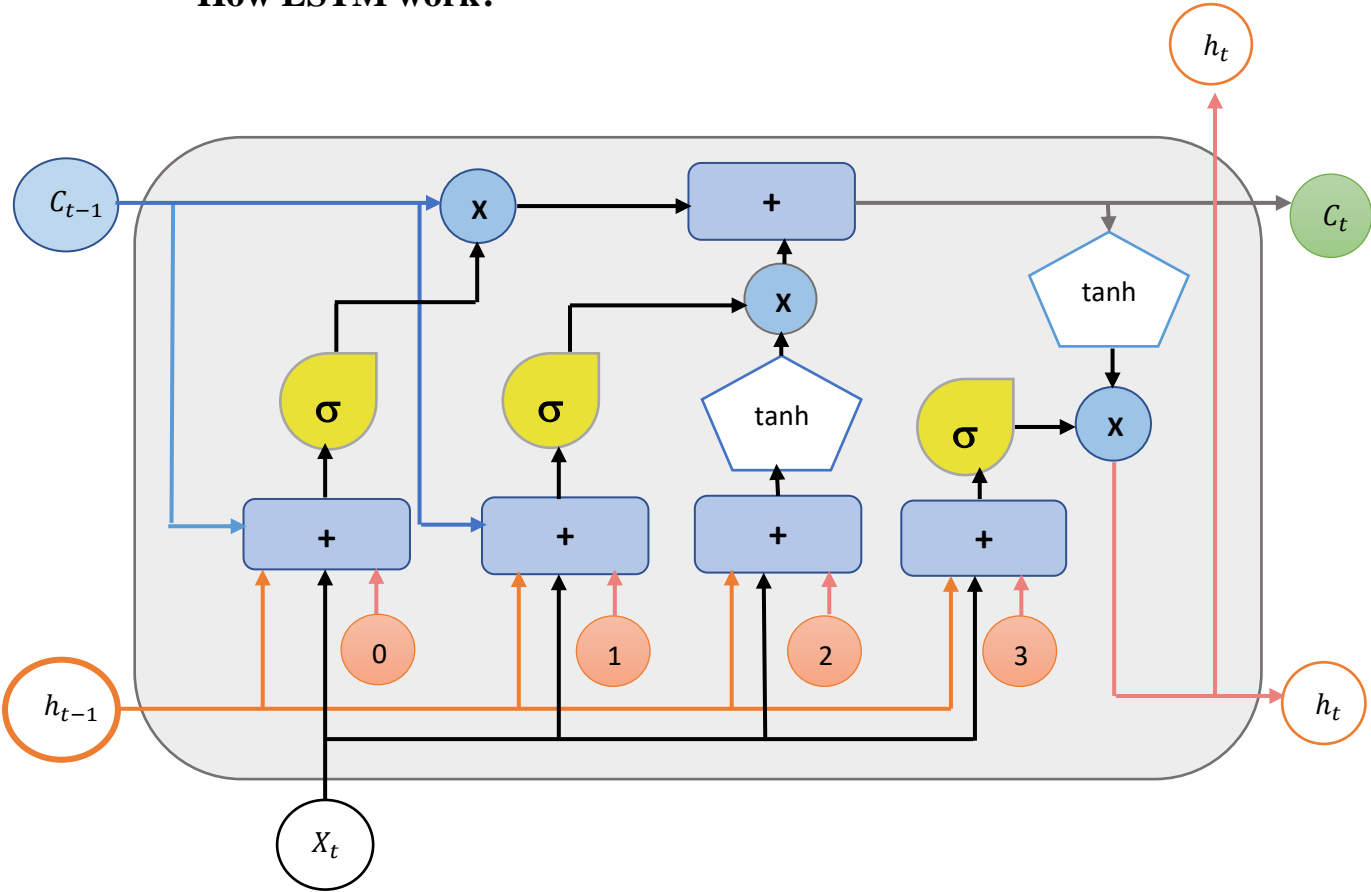


Fig11: Life Cereal Review

Source: [https://miro.medium.com/max/875/1\\*YHjfAgozQaghcsEvsBEu2g.png](https://miro.medium.com/max/875/1*YHjfAgozQaghcsEvsBEu2g.png)

Furthermore, that is basically what a LSTM does. It can figure out how to keep just applicable data to make forecasts, and does not remember non-significant information. For this situation, the words we recalled made us judge that it was acceptable.

### How LSTM work?



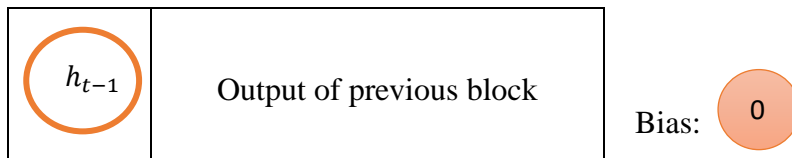
Inputs:

$X_t$	Input Vector
$C_{t-1}$	Memory from previous block

Outputs:

$C_t$	Memory from current block
$h_t$	Output of current block





Non-linearities:

Vector Operation:

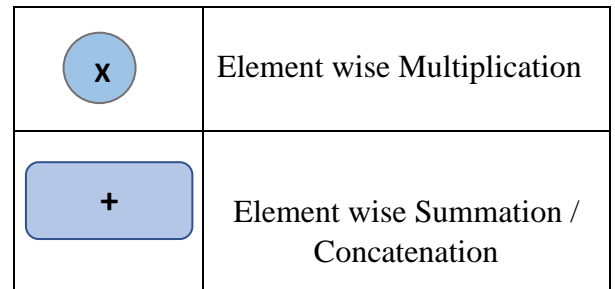
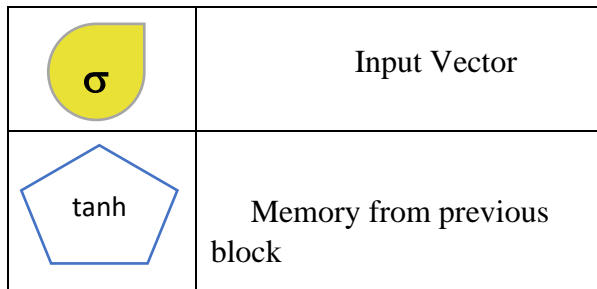


Fig12: LSTM

Source: [https://miro.medium.com/max/1400/1\\*laH0\\_xXEKFE0IKJu54gkFQ.png](https://miro.medium.com/max/1400/1*laH0_xXEKFE0IKJu54gkFQ.png)

A typical LSTM network is made up of several memory segments known as cells. The cell state and the concealed state are both being transferred to the subsequent cell. Memory cells are responsible for recalling information, and access to this memory is controlled by three crucial characteristics known as gates. Each of them is explained below:

#### a) Forget gate:

To commence, we got the forget gate. This gate determines whether data should be deleted or saved.

This gate receives two data sources;  $h_{t-1}$  and  $x_t$ .

$h_{t-1}$  is the concealed state from the past cell or the yield of the past cell and  $x_t$  is the contribution at that specific time step. The weight matrices are multiplied by the provided data sources, and a bias is applied. Then after this, the sigmoid method is applied to this worth. The sigmoid technique is used to process data from the previous covered state as well as data from the

present. The outcome of attributes was around between 0 and 1. The value closer to zero means that it should be ignored, while the value closer to 1 indicates that it should be kept.

**b) Input gate:**

The input gate is liable for the addition of data to the cell state. This addition of data is fundamentally three-step measure.

- Controlling what esteems should be added to the cell state by including a sigmoid function. This is essentially basically the same as the forget gate and goes about as a channel for all the data from  $h_{t-1}$  and  $x_t$ .
- Making a vector containing all potential qualities that can be added (as seen from  $h_{t-1}$  and  $x_t$ ) to the cell state. This is finished utilizing the tanh method, which yields esteems from - 1 to +1.
- Multiplying the worth of the regulatory channel (the sigmoid gate) to the made vector (the tanh capacity) and afterward adding this valuable data to the cell state through addition activity.

**c) Output gate:**

Lastly, we have the output gate. The output gate determines what the next concealed state should really be. Remember that the hidden unit contains information about previous inputs. Predictions are also made using the hidden state. To commence, we feed the previous concealed state and the current input into a sigmoid algorithm. The freshly modified cell state is then sent to the tanh method. To determine what data the disguised state should transmit, we multiply the sigmoid production by the tanh yield. The yield is the unnoticed condition. The specific cell condition and concealed are then carried over to the next time step.

Here, the equations of the input, forget, and output gates are represented by  $i_t$ ,  $f_t$ , and  $O_t$  respectively,  $w$  represents the weights, and  $\sigma$  is the sigmoid function:

$$i_t = \sigma(w_i [h_{t-1}, x_t] + b_i)$$

$$f_t = \sigma(w_f [h_{t-1}, x_t] + b_f)$$

$$O_t = \sigma(w_o [h_{t-1}, x_t] + b_o)$$

## LSTM model:

```
# defining the LSTM model
model = Sequential()
model.add(LSTM(300, input_shape=(X_modified.shape[1], X_modified.shape[2]), return_sequences=True))
model.add(Dropout(0.2))
model.add(LSTM(300))
model.add(Dropout(0.2))
model.add(Dense(Y_modified.shape[1], activation='softmax'))

model.compile(loss='categorical_crossentropy', optimizer='adam')
```

## 5.4 Language Used:

- **Python:**

Python is a general-purpose programming language with a high level of abstraction and an interpreter. It was originated by Guido van Rossum and was first delivered in 1991, Python incorporates philosophy of design that increases the code interpretability, notably utilizing significant whitespace. It works with constructors that empower clear programming on both the little and immense scopes.

Python people group has created numerous modules to assist software engineers with carrying out in the AI. In this, we will utilize numpy, scipy and scikit-learn modules. Python can be introduced by utilizing following command:

**pip install model\_name**

## 5.5 Tools & Libraries used:

## Tools

- **Jupyter Notebook:**

Jupyter Notebook is an open-source web application that permits clients to make and share codes and records.

It's anything but surroundings where developers can do coding, can run it, can take a gander at the result, picture information and see the outcomes without leaving the that environment. This makes it an exceptionally valuable tool for performing start to finish information science work processes – measurable demonstrating, information cleaning, building and preparing AI models, imagining information, and some more.

- **Anaconda command prompt:**

This is quite similar to that same command prompt that it allows you to use anaconda as well as conda instructions out from prompt without switching folders or your path.

At a point when we start Anaconda command prompt, it can be noticed that it prepends a lot of locations to the PATH. These locations contain commands and scripts that you can run. So insofar as we are in the Anaconda command prompt, we can utilize these commands.

At the time when we will install Anaconda, we will be given a choice to add these to the PATH as default, and whenever checked these commands can be used on the regular command prompt. However, the anaconda prompt will consistently work.

To the extent refreshing conda, on the off chance that it doesn't work in command prompt, following command can be executed:

```
conda update conda
```

## **Some of common libraries used:**

- **Numpy:**

NumPy is a notable broadly useful array handling bundle. A broad assortment of high intricacy numerical functions makes NumPy amazing to deal with enormous multi-dimensional arrays and matrices. This library is extremely helpful for taking care of linear polynomial math, Fourier changes, and arbitrary numbers. Various libraries such as TensorFlow utilizes NumPy at the backend for controlling tensors.

### **Installation:**

Execute this command on terminal: -

```
pip install numpy
```

- **Scipy:**

The SciPy library offers modules for linear variable based math, picture improvement, incorporation introduction, extraordinary functions, Fast Fourier change, sign and picture preparing, and other computational undertakings in science and analytics.

The hidden information structure utilized by SciPy is a multi-dimensional array provided by the NumPy module. It relies upon NumPy for the cluster control subroutines. This library works with NumPy arrays alongside giving easy to understand and productive mathematical methods.

### **Installation:**

Execute this command on terminal: -

```
pip install scipy
```

- **Matplotlib:**

Matplotlib is an information representation library which is utilized for 2D plotting to deliver distribution quality picture plots and figures in an assortment of configurations. It assists with creating scattered plots, histograms, error charts, bar diagrams with only a couple lines of code.

**Installation:**

```
pip install matplotlib
```

- **Scikit-learn:**

Scikit-learn was based on these two Python libraries – NumPy and SciPy and has now turned to be the most mainstream Python AI library for creating AI calculations.

This library has a wide scope of directed and unaided learning calculations that handles a predictable interface in Python. The library can likewise be utilized for information mining and information scrutinizing.

**Installation:**

```
pip install scikit-learn
```

- **Keras:**

This library includes neural-network elementary units such as layers, objectives, activation techniques, and analyzers. There are several highlights or features in Keras that aid when writing Deep Neural Network code when dealing with images

and text images. Apart from the basic neural network, it supports convolutional and repeating neural networks.

### **Installation:**

```
pip install keras
```

- **Pandas:**

This library is going up to be the well known library of python that is utilized for information examination with help for quick, adaptable, and expressive information structures intended to chip away at both "social" or "marked" information.

### **Installation:**

```
pip install pandas
```

- **Seaborn:**

It is a information representation library of python that is dependent on matplotlib. It's anything but an undeniable level interface for drawing alluring and useful factual illustrations.

### **Installation:**

```
pip install seaborn
```

- **NLP Libraries:**

- a. **NLTK(Natural Language Toolkit) :**

This is the main and extraordinary compared to other libraries in NLP for Python. It has about 100 corpus and related lexical materials that become regularly updated, such as WordNet, online Text Corpus, and others. It's also not a one of described models, which helps us assess tasks efficiently.

### **Installation:**

```
pip install nltk
```

### **b. Gensim:**

Gensim is an open-source vector space and subject displaying toolbox. It utilizes numpy and scipy and intended for information stemming, handle enormous content assortments and proficient gradual calculations.

### **Installation:**

```
pip install gensim
```



## **CHAPTER 6: IMPLEMENTATION**

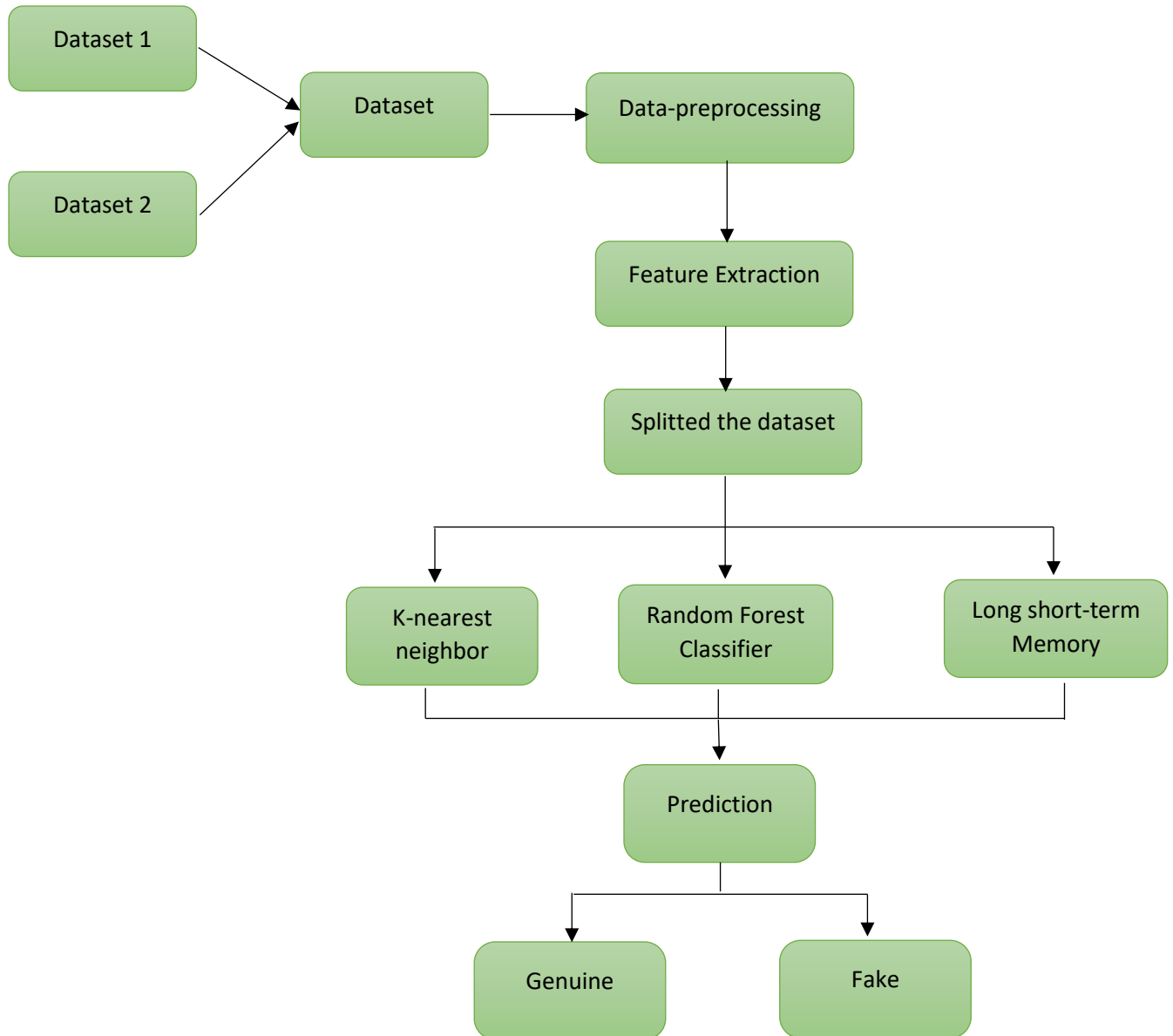


Fig13: Flow of implementation

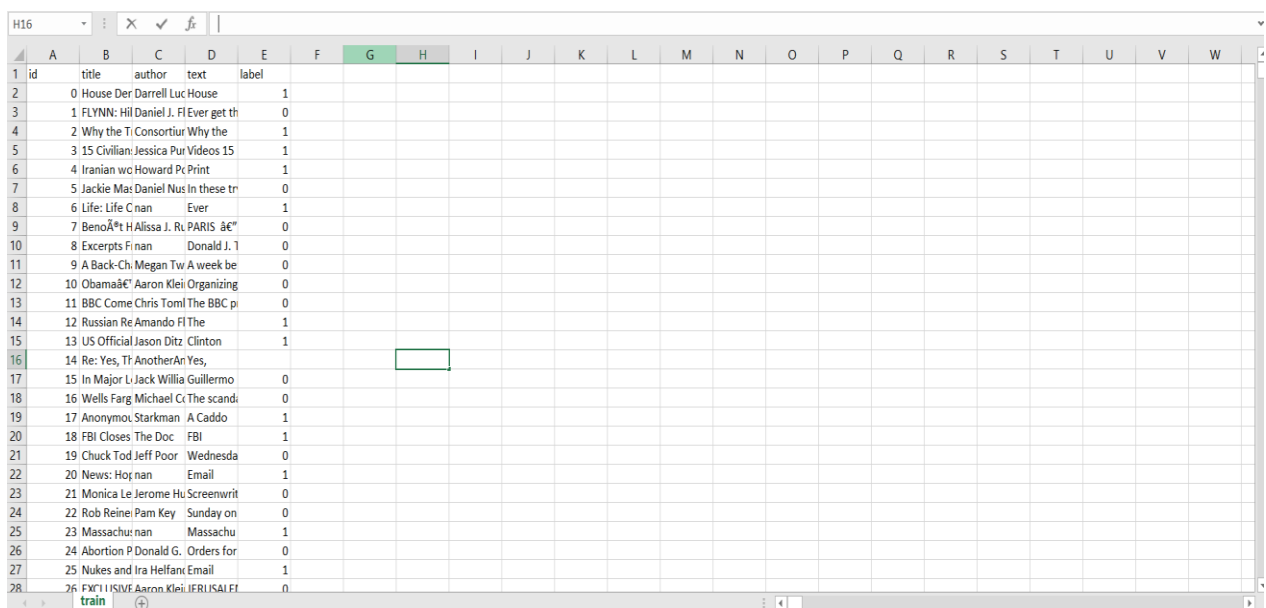
## 6.1 Dataset Description:

In this work, we have aggregated two different datasets to generate one larger dataset which have been used further in this project. The datasets utilized in this project were taken from KAGGLE and their descriptions are given below:

### I. Dataset I:

train.csv- This training dataset has around 20,800 columns of information from different articles on the web. A complete training dataset has the accompanying features:

1. id: This is a distinctive identification no. assigned to each article
2. title: This field contains title for a article
3. author: This field represents the author of the article
4. text: This field contain the full content of the article; fragmented at times
5. label: a tag name provided to each article to categorize it in two groups
  - 1: counterfeit
  - 0: genuine



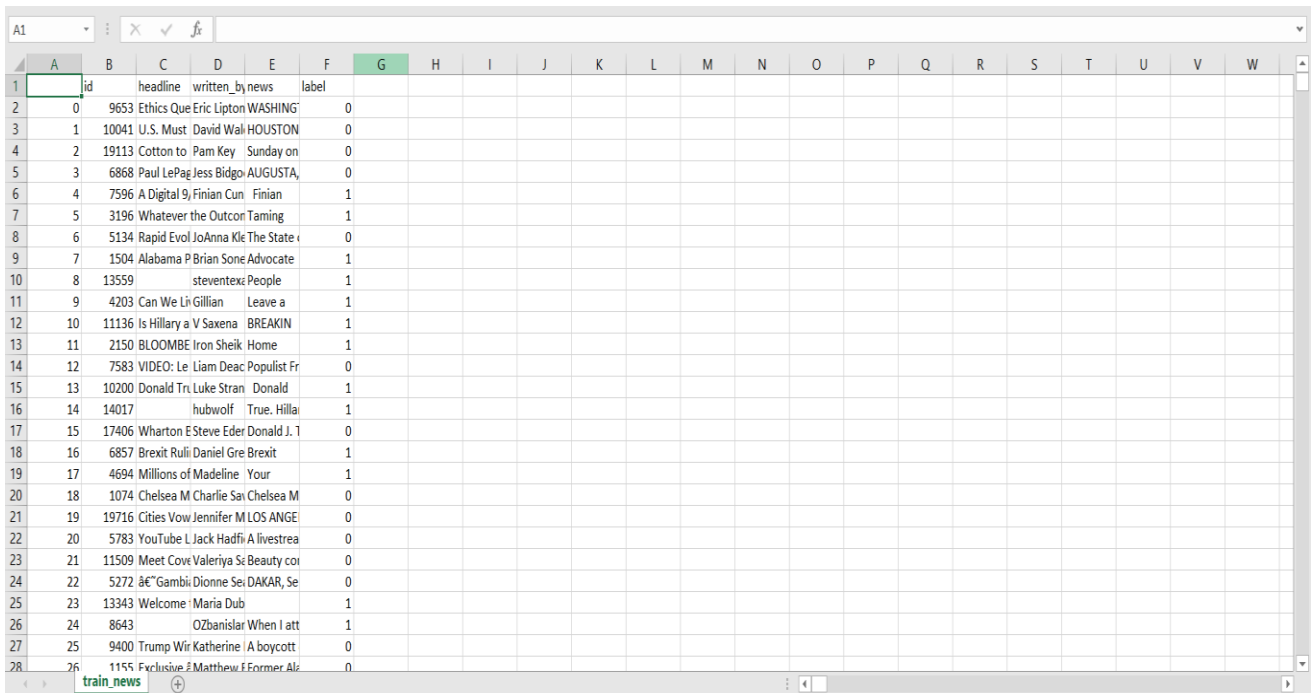
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	id	title	author	text	label																		
2	0	House Der Darrell Luc House			1																		
3	1	FLYNN: Hil Daniel J. Fl Ever get th			0																		
4	2	Why the Ti Consortiur Why the			1																		
5	3	15 Civilian: Jessica Pu Videos 15			1																		
6	4	Iranian wc Howard Pc Print			1																		
7	5	Jackie Mac Daniel Nus In these tr			0																		
8	6	Life: Life Cnan	Ever		1																		
9	7	BenoÂt H Alissa J. Ru PARIS â€			0																		
10	8	Excerpts Finan	Donald J. T		0																		
11	9	A Back-Ch Megan Tw A week be			0																		
12	10	Obamaâ€ Aaron Klei Organizing			0																		
13	11	BBC Come Chris Toml The BBC pi			0																		
14	12	Russian Re Amando Fl The			1																		
15	13	US Official Jason Ditz Clinton			1																		
16	14	Re: Yes, TF AnotherAn Yes,																					
17	15	In Major L Jack Willia Guillermo			0																		
18	16	Wells Farg Michael C: The scandi			0																		
19	17	Anonymoi Starkman A Caddo			1																		
20	18	FBI Closes The Doc	FBI		1																		
21	19	Chuck Tod Jeff Poor	Wednesda		0																		
22	20	News: Hoq nan	Email		1																		
23	21	Monica Le Jerome Hu Screenwrit			0																		
24	22	Rob Reine Pam Key	Sunday on		0																		
25	23	Massachu: nan	Massachu		1																		
26	24	Abortion P Donald G. Orders for			0																		
27	25	Nukes and Ira Helfant Email			1																		
28	26	FW: T LUSIVE Aaron Klei: JERUSALET			0																		

Fig14: train.csv

## II. Dataset II:

train\_news.csv- This dataset has around 20800 rows and 6 columns. The description of each of the column is given below:

1. id: This is a distinctive identification no. assigned to each news article
2. headline: It is the title of the news article.
3. news: It contains the full content of the news article
4. Unnamed:0: It is a serial number
5. written\_by: It represents the author of the news article
6. label: a tag name provided to each article to categorize it in two groups
  - 1:counterfeit
  - 0: genuine



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		id	headline	written_by	news	label																	
2	0	9653	Ethics Que	Eric Lipton	WASHINGTON	0																	
3	1	10041	U.S. Must	David Wal	HOUSTON	0																	
4	2	19113	Cotton to	Pam Key	Sunday on	0																	
5	3	6868	Paul LePag	Jess Bidgo	AUGUSTA,	0																	
6	4	7596	A Digital 9	Finian Cun	Finian	1																	
7	5	3196	Whatever the	Outcon	Taming	1																	
8	6	5134	Rapid Evol	JoAnna Kle	The State	0																	
9	7	1504	Alabama P	Brian Sone	Advocate	1																	
10	8	13559		steventex	People	1																	
11	9	4203	Can We Li	Gillian	Leave a	1																	
12	10	11136	Is Hillary a	V Saxena	BREAKIN	1																	
13	11	2150	BLOOMBE	Iron Sheik	Home	1																	
14	12	7583	VIDEO: Le	Liam Deac	Populist Fr	0																	
15	13	10200	Donald Tr	Luke Stran	Donald	1																	
16	14	14017		hubwolf	True. Hillai	1																	
17	15	17406	Wharton E	Steve Eder	Donald J. 1	0																	
18	16	6857	Brexit Ruli	Daniel Gre	Brexit	1																	
19	17	4694	Millions of	Madeline	Your	1																	
20	18	1074	Chelsea M	Charlie Sai	Chelsea M	0																	
21	19	19716	Cities Vow	Jennifer M	LOS ANGE	0																	
22	20	5783	YouTube L	Jack Hadfi	A livestrea	0																	
23	21	11509	Meet Cove	Valeriya S	Beauty cor	0																	
24	22	5272	â€œGambii	Dionne Sei	DAKAR, Se	0																	
25	23	13343	Welcome i	Maria Dub		1																	
26	24	8643		OZbanislar	When I att	1																	
27	25	9400	Trump Wir	Katherine	A boycott	0																	
28	26	1155	Exclusive	Matthew F	Former Al	0																	

Fig15: train\_news.csv

There was a lot of pre-processing needed to be done before the data could be utilized.

## 6.2 Data Preprocessing:

Reading datasets:

```
In [2]: data1 = pd.read_csv('train_news.csv')
        data2 = pd.read_csv('train.csv')
```

Merging all the field of dataset 1 that contains content of articles like 'headline', 'written\_by' & 'news' into one column with name 'text' which will be same as name of column in dataset 2, and then deleting these columns i.e; 'headline', 'written\_by', 'news' from dataset:

```
In [7]: data1['text'] = data1['headline'] + data1['written_by'] + data1['news']
        del data1['headline']
        del data1['written_by']
        del data1['news']
        data1.head()
```

```
Out[7]:
```

	id	label	text
0	9653	0	Ethics Questions Dogged Agriculture Nominee as...
1	10041	0	U.S. Must Dig Deep to Stop Argentina's Lionel ...
2	19113	0	Cotton to House: 'Do Not Walk the Plank and Vo...
3	6868	0	Paul LePage, Besieged Maine Governor, Sends Co...
4	7596	1	A Digital 9/11 If Trump WinsFinian Cunningham ...

Merging all the fields of dataset 2 that contains content of articles like column 'author', 'title' & 'text' into one column with name 'text' which will be same as name of column in dataset 2, and then deleting these columns i.e; 'author', 'title', 'text' from dataset :

```
In [9]: data2['text']=data2['author']+data2['title']+data2['text']
del data2['author']
del data2['title']
del data2['id']
```

```
In [10]: data2.head()
```

```
Out[10]:
```

	text	label
0	Darrell LucasHouse Dem Aide: We Didn't Even Se...	1
1	Daniel J. FlynnFLYNN: Hillary Clinton, Big Wom...	0
2	Consortiumnews.comWhy the Truth Might Get You ...	1
3	Jessica Purkiss15 Civilians Killed In Single U...	1
4	Howard PortnoyIranian woman jailed for fiction...	1

Concatenating both data frames i.e., ‘data1’ & ‘data2’ into one larger data frame i.e., ‘data’:

```
In [11]: data = pd.concat([data1, data2], ignore_index=True, sort=False)
data.head()
```

```
Out[11]:
```

	label	text
0	0	Ethics Questions Dogged Agriculture Nominee as...
1	0	U.S. Must Dig Deep to Stop Argentina's Lionel ...
2	0	Cotton to House: 'Do Not Walk the Plank and Vo...
3	0	Paul LePage, Besieged Maine Governor, Sends Co...
4	1	A Digital 9/11 If Trump WinsFinian Cunningham ...

Dropping null values from data frame:

```
In [12]: data.isnull().sum()
```

```
Out[12]: label      0
text      5030
dtype: int64
```

```
In [13]: data.dropna()
data.shape
```

```
Out[13]: (41600, 2)
```

The resulting data frame has 41600 rows and 2 columns i.e., ‘text’ & ‘label’.

Data preprocessing has been done that includes removing stop words, punctuation, and missing rows:

```
def textClean(text):
    text = re.sub(r"[^A-Za-z0-9^,!.\/'+-=]", " ", text)
    text = text.lower().split()
    stops = set(stopwords.words("english"))
    text = [w for w in text if not w in stops]
    text = " ".join(text)
    return (text)

def cleanup(text):
    text = textClean(text)
    text = text.translate(str.maketrans("", "", string.punctuation))
    return text
```

```
missing_rows = []
for i in range(len(data)):
    if data.loc[i, 'text'] != data.loc[i, 'text']:
        missing_rows.append(i)
data = data.drop(missing_rows).reset_index().drop(['index'], axis=1)

for i in range(len(data)):
    data.loc[i, 'text'] = cleanup(data.loc[i, 'text'])
```

Doc2Vec is utilizing two things when preparing your model, labels and the real data. The labels can be anything, yet to make it simpler each report document name will be its label.

Labelling sentences:

```
def constructLabeledSentences(data):
    sentences = []
    for index, row in data.iteritems():
        sentences.append(LabeledSentence(utils.to_unicode(row).split(), ['Text' + '_%s' % str(index)]))
    return sentences
```

```
x = constructLabeledSentences(data['text'])
y = data['label'].values
```

Doc2Vec model has been applied and data frame is splitted into train & test set.

- train set: 80%
- test set: 20%

Train set is fitted with model to train it and then test dataset is used to evaluate the fitted model:

```
text_model = Doc2Vec(min_count=1, window=5, vector_size=vector_dimension, sample=1e-4, negative=5, workers=7, epochs=10,
                    seed=1)
text_model.build_vocab(x)
```

```
train_size = int(0.8 * len(x))
test_size = len(x) - train_size

text_train_arrays = np.zeros((train_size, vector_dimension))
text_test_arrays = np.zeros((test_size, vector_dimension))
train_labels = np.zeros(train_size)
test_labels = np.zeros(test_size)

for i in range(train_size):
    text_train_arrays[i] = text_model.docvecs['Text_' + str(i)]
    train_labels[i] = y[i]

j = 0
for i in range(train_size, train_size + test_size):
    text_test_arrays[j] = text_model.docvecs['Text_' + str(i)]
    test_labels[j] = y[i]
    j = j + 1
```

## 6.3 Algorithms:

Now, first models are imported as per our requirements. We are using K-nearest neighbor, Random Forest algorithm and Long short-term memory in this project, so these all 3 models were imported:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from keras.layers import LSTM
```

Now, after importing the models, these models are fitted and trained with train set data and then using the test set data their accuracy score is calculated.

Training the KNN model:

```

model=KNeighborsClassifier()
model.fit(xtr1, ytr1)
ypred=model.predict(xte1)

```

Accuracy score is calculated for KNN:

```

accuracy1 = accuracy_score(yte1, ypred)
print('Accuracy: %.3f' % accuracy1)

```

Training the Random Forest Classifier with train set:

```

model=RandomForestClassifier()
model.fit(xtr1, ytr1)
ypred=model.predict(xte1)

```

Accuracy score is calculated for Random Forest Classifier:

```

accuracy2 = accuracy_score(yte1, ypred)
print('Accuracy: %.3f' % accuracy2)

```

LSTM model is fitted over 5 epoch and batch size of 64.

```

# Create the model
embedding_vector_length = 32
model = Sequential()
model.add(Embedding(top_words+2, embedding_vector_length, input_length=max_review_length))
model.add(LSTM(100))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
print(model.summary())
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=epoch_num, batch_size=batch_size)

```

Accuracy score is calculated for LSTM:

```

scores = model.evaluate(X_test, y_test, verbose=0)
print("Accuracy= %.2f%%" % (scores[1]*100))

```



## CHAPTER 7: RESULTS AND PERFORMANCE

### EVALUATION

After training the model, its performance is evaluated by calculating its accuracy.

Accuracy achieved by the models are:

- KNN:

```
from sklearn.metrics import precision_score, recall_score, plot
performance=[]
accuracy1 = accuracy_score(yte1, ypred)
performance.append(accuracy1*100)
print('Accuracy: %.2f%%' % (accuracy1*100))
```

Accuracy: 87.50%

- Random Forest:

```
accuracy2 = accuracy_score(yte1, ypred)
performance.append(accuracy2*100)
print('Accuracy: %.2f%%' % (accuracy2*100))
```

Accuracy: 93.05%

- Long Short-Term Memory (LSTM):

```
performance.append(scores[1]*100)
# Draw the confusion matrix
y_pred = model.predict_classes(X_test)
plot_cm(y_test, y_pred)
```

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 500, 32)	160064
lstm (LSTM)	(None, 100)	53200
dense (Dense)	(None, 1)	101

Total params: 213,365  
Trainable params: 213,365  
Non-trainable params: 0

```
None
Epoch 1/5
452/452 [=====] - 469s 1s/step - loss: 0.2294 - accuracy: 0.9094 - val_loss: 0.1099 - val_accuracy: 0.9634
Epoch 2/5
452/452 [=====] - 496s 1s/step - loss: 0.1327 - accuracy: 0.9553 - val_loss: 0.0982 - val_accuracy: 0.9661
Epoch 3/5
452/452 [=====] - 475s 1s/step - loss: 0.1156 - accuracy: 0.9602 - val_loss: 0.2223 - val_accuracy: 0.9262
Epoch 4/5
452/452 [=====] - 482s 1s/step - loss: 0.1583 - accuracy: 0.9421 - val_loss: 0.5394 - val_accuracy: 0.7490
Epoch 5/5
452/452 [=====] - 490s 1s/step - loss: 0.1609 - accuracy: 0.9388 - val_loss: 0.1429 - val_accuracy: 0.9568
Accuracy= 95.68%
```

The results showed up by computing the accuracies of the different models referenced previously.

The data represented below are the average values over progressive trials.

Model	Accuracy
K-NN	87.50%
Random Forest Classifier	93.04%
LSTM	97.20%

On the basis of accuracies achieved by model as shown in above Table, the chart in Figure below is built by putting the algorithms and accuracies on X and Y axis respectively. It is construed that LSTM achieves the most noteworthy accuracies accompanied by Random Forest Classifier and K-NN.

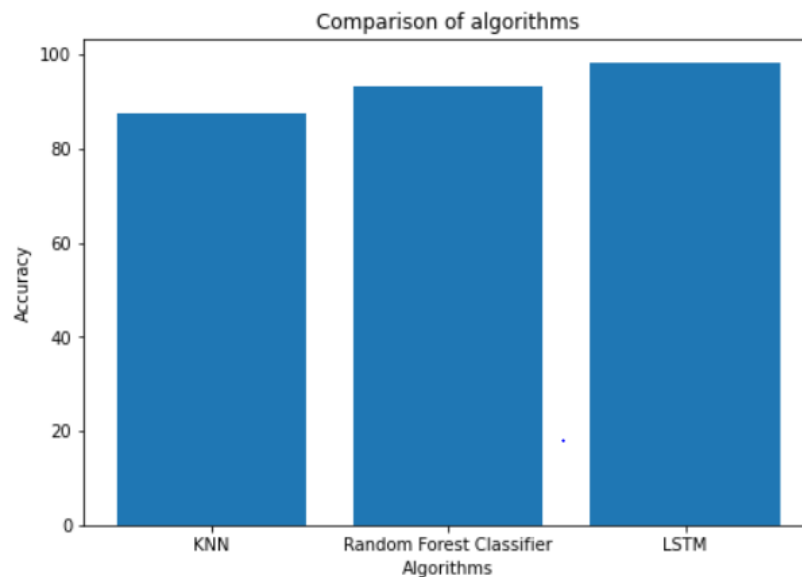


Fig 16: Comparison of Algorithm Results

## **CHAPTER 8: CONCLUSION**

The main purpose of this dissertation work is to evaluate the performance of some machine learning algorithms on extended dataset for detecting legitimacy of news articles. For obtaining the required result, we have trained some existing models on a larger dataset and also lot of data preprocessing has been done to improve its performance.

The models have been executed on larger dataset which is generated by merging two different datasets & then word embedding techniques such as Doc2Vec is applied on it followed by some existing machine learning model (including ensembled techniques) and natural language processing techniques such as K-nearest neighbors, Random Forest Classifier and Long short-term memory.

The most noteworthy accuracy is achieved by using LSTM (Long-short memory) technique i.e., 97.20%.

## **CHAPTER 9: REFERENCES**

- [1] Dataset 1:  
<https://www.kaggle.com/c/fake-news/data?select=train.csv>
- [2] Dataset2:  
[https://www.kaggle.com/surekharamireddy/fake-news-detection?select=train\\_news.csv](https://www.kaggle.com/surekharamireddy/fake-news-detection?select=train_news.csv)
- [3] <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>
- [4] de Oliveira, Nicollas R., Dianne SV Medeiros, and Diogo MF Mattos. "A Sensitive Stylistic Approach to Identify Fake News on Social Networking." *IEEE Signal Processing Letters*, 27, 1250-1254,2020.
- [5] Fathima Nada, Bariya Firdous Khan, Aroofa Maryam, Nooruz-Zuha, Zameer Ahmed, "Fake news Detection using Logistic Regression", *International Research Journal of Engineering and Technology (IRJET)*,2019.
- [6] Shu K., Sliva A., Wang S., Tang J., Liu H., Fake News Detection on Social Media: A Data Mining Perspective, *ACM SIGKDD Explorations Newsletter*,19(1), 2017.
- [7] Mykhailo Granik, Volodymyr Mesyura, Fake News Detection Using Naive Bayes Classifier, *IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017.
- [8] Akshay Jain, Fake News Detection, *IEEE International Students' Conference on Electrical, Electronics and Computer Sciences*, 2018.
- [9] Reis, Julio CS, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. "Supervised learning for fake news detection." *IEEE Intelligent Systems* 34, no. 2: 76-81,2019.
- [10] Zhou, Xinyi, Reza Zafarani, Kai Shu, and Huan Liu. "Fake news: Fundamental theories, detection strategies and challenges." In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 836-837. 2019.
- [11] <https://www.javatpoint.com/>
- [12] <https://www.tutorialspoint.com/index.htm>

- [13] LSTM model: <https://www.analyticsvidhya.com/blog/2021/06/lstm-for-text-classification/>
- [14] Python libraries theory: <https://www.upgrad.com/blog/top-python-libraries-for-machine-learning/>
- [15] Random Forest Classifier: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- [16] Text preprocessing: <https://www.geeksforgeeks.org/text-preprocessing-in-python-set-1/>