AN INDUSTRIAL TRAINING REPORT

ON

Data Analytics, Machine Learning using Python

(Project Title: FAKE NEWS DETECTION)

Submitted by

Name of the Student: Nidhi Gupta
Roll no.: 181500422

Department of Computer Engineering and Applications
Institute of Engineering and Technology

GLA University
Mathura-281406, India
2020

Department of computer Engineering and Applications
GLA University, Mathura
17 km. Stone NH#2, Mathura-Delhi Road, P.O. – Chaumuha,
Mathura – 281406

Declaration

I hereby declare that the work which is being presented in the Industrial Training **"Fake News Detection",** in partial fulfillment of the requirements for Industrial Training viva voce, is an authentic record of my own work carried under the supervision of "Diginique Techlabs".

Signature of Candidate: Nidhi Gupta

Name of Candidate: Nidhi Gupta

Roll. No. : 181500422

Course: Computer Engineering and Application

Year: 2020-2021

Semester: V semester

CERTIFICATES

- **Certificate of Training**

- **Internship Letter**

Internship Letter

To Whomsoever It May Concern

Date: July 02, 2020

This is to certify that **Nidhi Gupta** has successfully completed her internship in **Data Analytics, Machine Learning and AI using Python** conducted by Diginique TechLabs in association with **Cognizance IIT Roorkee**. She worked on several projects during this internship from May 18, 2020 to July 01, 2020 under the guidance of **Mr. Bipul Shahi** (CTO, Diginique TechLabs).

During the period, we found her punctual, hard-working and inquisitive. We wish her all the best for her future endeavors.

For Diginique TechLabs.

Amit Singh Tomar
HR Manager

- **Project Completion Letter**

Diginique TechLabs
Ideal for Geeks

Project Completion Letter

To Whomsoever It May Concern

Date: July 02, 2020

This is to certify that **Nidhi Gupta** has successfully completed her one month industry oriented project on **"Fake News Detection"** from June 02, 2020 to July 01, 2020 under the guidance of **Mr. Bipul Shahi** (CTO, Diginique TechLabs) and submitted the assigned project on time. With sincere efforts & excellent performance she has learned a lot of skills related to the same field and technology.

During this period, she demonstrated as a diligent and truthful person. Her interpersonal skills and learning methodology are outstanding. We wish her all the best for her future endeavors.

For Diginique TechLabs.

DIGINIQUE TECH LABS INDIA

Amit Singh Tomar
HR Manager

SYNOPSIS

Industrial Training Synopsis
B.Tech. (CSE)-Batch 2020-2021

Student Information:

Name: Nidhi Gupta	**University Roll. No.** 181500422
Mobile: 6396492523	**Email:** nidhi.gupta_cs18@gla.ac.in

Information about Industry/Organization:

Industry/Organization Name with full Address	Diginique Labs, A 1202,The Golden Palms sector -168, Noida(UP) 201305
Contact Person	**Name & Designation**: Amit Singh Tomar , HR Manager **Mobile/email**: 7409146082 / info@diginique.com

Project Information:

Title Of Project/Training/Task	Fake News Detector Using Machine Learning
Role & Responsibility	Completed the project individually
Technical Details	**Hardware Requirements**: Intel i5 processor(8th gen), Windows 10 ,4GB RAM **Software Requirements**: Anaconda(Jupyter Notebook) , Google Collab
Training Implementation Details	Fully Implemented
Training Period	**Start Date**: May 18, 2020 **End Date:** July 01, 2020 **Duration Of Training (In Weeks):** 6 weeks

Summary of the Training Work:

I completed my training in one of the most trending technology :-Machine Learning. Starting from the basic libraries of Python such as Numpy and Pandas ,I continued to learn some more of them like Matplot library and Open CV. And then took the next step to learn the two major faces of the coin – "Supervised Learning " and "Unsupervised Learning ". Machine Learning being a very vast field, I learnt a few of the algorithms and their implementation on the different types of datasets . Some of the algorithms were K-Neighbors Classifier, Linear Regression (Univariant and Multivariant variable), Polynomial Regression, Clustering Approach, Decision Tree Classifier and Regressor etc. Along with understanding the mathematics behind them, I learnt to implement it using sklearn library. And worked on a number of dataset given by the instructor during the training. And by the end of the training we just touched the hem of Deep Learning and hence giving me a road map to move further.

At the end of the training, I was told to submit A Final Project to evaluate what I learnt then. So I chose my project on the topic "FAKE NEWS DETECTOR" .where my task was to develop a Machine Learning program for text classification to identify when the news source may be producing fake news. My aim was to use the labeled data with two classes "*Real*" and "*Fake*" to develop a classifier that trains my model on the basis of the content from the dataset. In this project paper I used two datasets and both of the datasets were downloaded from Kaggle. And I applied a number of text classifying algorithms on both

the datasets along with *Feature extraction* at places. After applying each algorithm, I calculated the accuracy received on the dataset and plotted a confusion matrix for the

same to a view of how efficiently the algorithm worked.

Some of the algorithms I used were:

1- Naïve Bayes Algorithm: This algorithm works on the Bayes theorem of probability.

2- Support Vector Machine: This algorithm works by creating divisor lines between data points dividing it into two sub spaces (here as real and the fake one) and determine the best results between the vectors.

3- Passive Aggressive Algorithm: This family of algorithm is used to work upon large datasets. And since our dataset was large therefore we used this algorithm too for classification of the dataset.

In today's world, Machine Learning is an application of Artificial Intelligence.
Google says, "Machine Learning is the future" and it's really going to be very bright. As humans are getting more and more dependent and addicted to machines, we are to witness a new revolution which will take over the world and become its future…

Certifications:-

1. Internship Certificate
2. Project Letter

Acknowledgement

Presenting the ascribed project paper report in this very simple and official form, I would like to place my deep gratitude to Diginique TechLabs for providing us the instructor Mr. Bipul Shahi, our coordinate trainer and developer.

He has been teaching us the concept of Machine Learning in a very efficient way giving us real life examples and applications and more over teaching things graphically with all this experience and patience and has been constantly supervising our information regarding the project. Without his help, I wouldn't have been able to complete this project.

And at last but not the least I would like to thank my dear parents for helping me to grab this opportunity to get trained with Diginique TechLabs and also my colleagues who helped me find resources during the training.

Thanking You. **Nidhi Gupta**

ABSTRACT

Since the rise of social media, fake news has become a society problem, in some occasion spreading more and faster than the true information. Information preciseness on Internet is an increasingly important concern. Advances in technology and the spread of news through different types of media have increased the spread of fake news today. As such, the effects of fake news have increased exponentially in the recent past and something must be done to prevent this from continuing in the future as it causes chaos among people and sometimes create a wrong belief of a political issue or area as today's generation is getting much of their general awareness from social media itself and nobody bothers to look over real issues to form opinions. The credibility of social media network is also at stake when it comes to the spreading of fake news. Thus, it has become a research challenge to automatically check the information at the source, the content and publisher at verify whether the news is fake, fabricated or a real news. There are two methods by which machines could attempt to solve the fake news problem better than humans. The first is that machines are better at detecting and keeping track of statistics than humans but data powering was an issue. Secondly, machines may be more efficient in surveying a knowledge base to find all relevant articles and answering based on those many different sources but we decided to focus on supervised learning of machine.

Along with the fake news, today's technology has managed ways even to identify the between what's real and fake. Machine learning has played a very important role in the classification of information although with some limitations but to a great extent. Fake News is "fictitious article deliberately fabricated to deceive readers". So to classify the texts there has been a special field in Machine Learning called the Natural Language Processing which is an approach to work on the text data that I have been using in this project. Using text classification in Machine Learning automatically assigns tags and category to text. With data powering from various channels, it is hard to keep up with emails, chats, online reviews, etc. and is challenging for us to process all data so we use text classification (Natural Language Processing).

The purpose of the work is to come out with the solution that can be utilized by the users or the social media sites itself to detect and filter out sources containing misleading information.

CONTENT

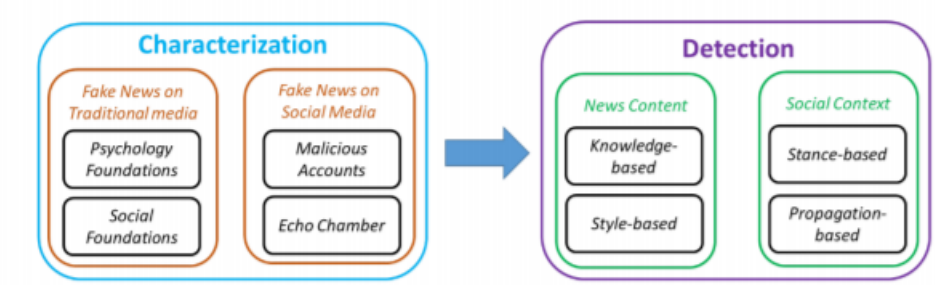
CHAPTER 1

INTRODUCTION TO THE PROJECT

These days' fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is "fake news" but lately blathering social media's discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints.

The importance of disinformation within American political discourse was the subject of weighty attention, particularly following the American president election. The term “fake news” became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views. In this paper, efforts have been made to produce a model that can accurately predict the likelihood that a given article is fake news. Facebook has been at the epicenter of much critique following media attention. They have already implemented a feature to flag fake news on the site when the user see it. They have also said publicly they are working on to distinguish these articles *in* an automated way. Certainly, it is not an easy task. However, in order to solve this problem, it is necessary to have an understanding on what Fake News is. Later, it is needed to look into how the techniques in the fields of machine learning, natural language processing help us to detect fake news. In order to build detection models, it is need to start by characterization to understand what is fake news before trying to detect them.

Fake news definition is made of two parts: authenticity and intent. Authenticity means that fake news content false information that can be verified as such, which means that conspiracy theory is not included in fake news as there are difficult to be proven true or false in most cases. The second part, intent, means that the false information has been written with the goal of misleading the reader.



Fake news on social media: characterization and detection

The goal of this project is to able to build a pseudo-model that is able to differentiate between the fake and the real news .For the completion of this task, we will be using the technology Machine Learning .Here we will be applying a number of algorithms and Natural Language Processing techniques for the fulfilment of our aim.

Machine Learning is an application of Artificial Intelligence that provides system the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it. The process of learning

begins with the observation of the data, such as example, direct experience or instructions, in order to look for pattern in data and make better decision in future based on examples. The prime aim is to allow the computers to learn automatically without human intervention. Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important. NLP is important because it helps resolve ambiguity in language and adds useful numeric structure to the data for many downstream applications, such as speech recognition or text analytics.

So we have used first the Natural Language Processing techniques to analyse our data and then the various machine learning algorithms to train the model to fulfil the purpose of the project.

CHAPTER 2

REQUIREMENT ANALYSIS

With the advancement of technology, digital news is more widely exposed to users globally and contributes to the increment of spreading hoaxes and disinformation online. Fake news can be found through popular platforms such as social media and the Internet. There have been multiple solutions and efforts in the detection of fake news where it even works with artificial intelligence tools. However, fake news intends to convince the reader to believe false information which deems these articles difficult to perceive. The rate of producing digital news is large and quick, running daily at every second, thus it is challenging for machine learning to effectively detect fake news.

Fake news paves the way for deceiving others and promoting ideologies. These people who produce the wrong information benefit by earning money with the number of interactions on their publications. The three most prevalent motivations for writing fake news and chosen only one as the target for this project as a means to narrow the search in a meaningful way. The first motivation for writing fake news, which dates back to the 19th century one-sided party newspapers, is to influence public opinion. The second, which requires more recent advances in technology, is the use of fake headlines as click baits to raise money. The third motivation for writing fake news, which is equally prominent yet arguably less dangerous, is satirical writing. [2] [3] While all three subsets of fake news, namely, (1) click baits (2), influential, and (3) satire, share the common thread of being fictitious, their widespread effects are vastly different.

- ## IMPACT OF FAKE NEWS

Spreading disinformation holds various intentions, in particular, to gain favour in political elections, for business and products, done out of spite or revenge. Humans can be gullible and fake news is challenging to differentiate from the normal news.

- **The goal of the media is altered**: Relations function of public relations is to share stories with reporters to get them to publish *accurate* information to influence others to care and act. However what fake news does is just the opposite. It creates a wrong belief in the

minds of the people getting them trapped into the false consipiracy framed by the click baits.

- **Political controversies**: Many a time most of the political leaders are being targeted with the help of fake news on social media. The anger and sense of hatred of a common individual against one can lead to mass anger and disbelief and chaos among citizens of a country.
- **Fake Reviews against a product**: Sometimes against a product ,the sales team of the company faces issues due to the unwanted ,illegible reviews of people hence the brand name gets affected .
- **Democratic Impacts**: Fake information played a major role once in American elections. That was the reason when media has spoken so much about the fake news phenomenon, because this was an important democratic issue.
- **Bullying and violence against innocent people**: False rumours that target specific individuals face. These individuals may be harassed on social media, and targeted by insults may face real life impacts. We must never rely on non-violated information circulating on social media to decide about the person's deed and should verify from the real, reliable sources.

In short impacts of fake news on social media and other things on internet are real and serious issues for sure and need to be tackled. As the technology is growing, we have solution for this problem. This project is the one among the solution and can be implemented.

- ## TYPES OF FAKE NEWS

Researchers have identified seven types of fake news, an advance that could help better spot misinformation, and create technology that can automatically detect misleading content.

Classifying fake news to seven basic agenda, includes false reviews, polarised content, satire, misreporting, commentary, persuasive information, and citizen journalism.

- ✓ **False Reviews:** Various Digital marketing websites advertise products of different brands and each product is followed by the reviews. Every

customer who buys an item comes to look for reviews and there fake reviews may spoil the brand image in the market.

- ✓ **Polarised content /Sloppy reporting** that fits an agenda – news that contains some grains of truth that are not fully verified, which are used to support a certain position or view.
- ✓ **Satire:** sites such as the Onion or Daily Mash publish fake news stories as humorous attempts to satirize the media, but have the potential to fool when shared out of context.
- ✓ **Misreporting:** Misleading news that's sort of true but used in the wrong context – selectively chosen real facts that are reported to gain headlines, but tend to be a misinterpretation of scientific research
- ✓ **Commentary** . Misleading news that's not based on facts, but supports an on-going narrative – news where there is no established baseline for truth, often where ideologies or opinions clash and unconscious biases come into play. Conspiracy theories tend to fall here!
- ✓ **Persuasive Information/Intentionally deceptive** – news that has been fabricated deliberately to either make money through number of clicks, or to cause confusion or discontent or as sensationalist propaganda. These stories tend to be distributed through imposter news sites designed to look like 'real' news brands, or through fake news sites. They often employ videos and graphic images that have been manipulated in some way.
- ✓ **Citizen Journalism:** Some citizens of the country sometimes possess hatred or grudge towards the government due to some political action or act in the name of religion, caste, creed, colour or sex and hence post antisocial reviews or news to mislead the information or the image of the current government.

Further adding to this, there are three more terms that are misinterpreted .They are distinguished between three types of problems: 'mis-information', 'dis-information', and 'mal-information':

1. Mis-information: false information disseminated without harmful intent.
2. Dis-information: created and shared by people with harmful intent.
3. Mal-information: the sharing of "genuine" information with the intent to cause harm.

- **DETECTING FAKE NEWS ONLINE**

Fake news has become increasingly prevalent over the last few years, with over 100 incorrect articles and rumours spread incessantly just with regard to the 2016 United States presidential election. These fake news articles tend to come from satirical news websites or individual websites with an incentive to propagate false information, either as click bait or to serve a purpose. Since they typically hope to intentionally promote incorrect information, such articles are quite difficult to detect. When identifying a source of information, one must look at many attributes, including but not limited to the content of the email and social media engagements. Specifically, the language is typically more inflammatory in fake news than real articles, in part because the purpose is to confuse and generate clicks. Furthermore, modelling techniques such as n-gram encodings and bag of words have served as other linguistic techniques to determine the legitimacy of a news source. On top of that, researchers have determined that visual-based cues also play a factor in categorizing an article, specifically some features can be designed to assess if a picture was legitimate and provides more clarity on the news. There are also many social context features that can play a role, as well as the model of spreading the news. Websites such as "Snopes" try to detect this information manually, while certain universities are trying to build mathematical models to do this task themselves.

There exists a large body of research on the topic of machine learning methods for deception detection online, most of it has been focusing on classifying online reviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining 'fake news' has also been the subject of particular attention within the literature.

Researchers outlines several approaches that seem promising towards the aim of perfectly classify the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech (POS) tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars (PCFG) have been shown to be particularly valuable in combination with n-gram methods and are able to achieve 85%-91% accuracy in deception related classification tasks using online review corpora.

CHAPTER 3

TECHNOLOGY USED

However there are various number of ways to detect the fake news online but according to the research I have gone through and have known by far, there are two methods by which machines could attempt to solve the fake news problem better than humans. The first is that machines are better at detecting and keeping track of statistics than humans, for example it is easier for a machine to detect that the majority of verbs used are “suggests” and “implies” versus, “states” and “proves.” Additionally, machines may be more efficient in surveying a knowledge base to find all relevant articles and answering based on those many different sources. Either of these methods could prove useful in detecting fake news, but we decided to focus on how a machine can solve the fake news problem using supervised learning that extracts features of the language and content only within the source in question, without utilizing any fact checker or knowledge base.

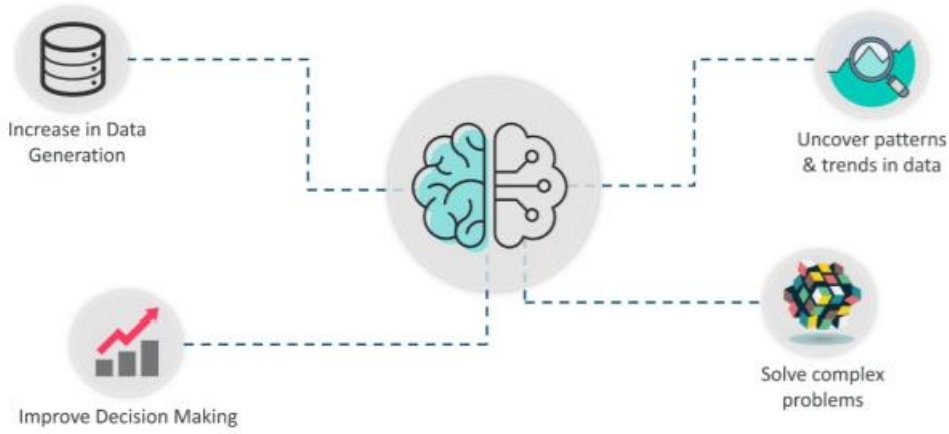
In this project to fulfil my aim I will be using one of the most trending and promising technology Machine Learning.

MACHINE LEARNING

Machine Learning is a sub-area of artificial intelligence, whereby the term refers to the ability of IT systems to independently find solutions to problems by recognizing patterns in databases. In other words: Machine Learning enables IT systems to recognize patterns on the basis of existing algorithms and data sets and to develop adequate solution concepts. Therefore, in Machine Learning, artificial knowledge is generated on the basis of experience.
In order to enable the software to independently generate solutions, the prior action of people is necessary. For example, the required algorithms and data must be fed into the systems in advance and the respective analysis rules for the recognition of patterns in the data stock must be defined. Once these two steps have been completed, the system can perform the following tasks by Machine Learning:

- Finding, extracting and summarizing relevant data
- Making predictions based on the analysis data
- Calculating probabilities for specific results
- Adapting to certain developments autonomously
- Optimizing processes based on recognized patterns

- **Need for Machine Learning**
  Ever since the technical revolution, we've been generating an immeasurable amount of data. As per research, we generate around 2.5 quintillion bytes of data every single day! It is estimated that by 2020, 1.7MB of data will be created every second for every person on earth. With the availability of so much data, it is finally possible to build predictive models that can study and analyse complex data to find useful insights and deliver more accurate results. Top Tier companies such as Netflix and Amazon build such Machine Learning models by using tons of data in order to identify profitable opportunities and avoid unwanted risks. Here's a list of reasons why Machine Learning is so important:



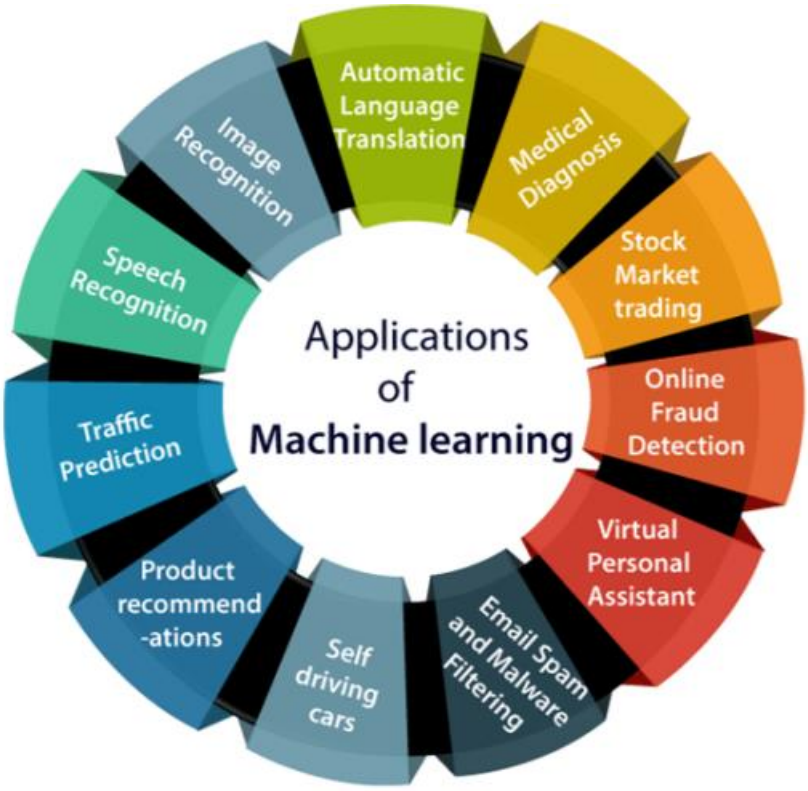
Increase in Data Generation: Due to excessive production of data, we need a method that can be used to structure, analyze and draw useful insights from data. This is where Machine Learning comes in. It uses data to solve problems and find solutions to the most complex tasks faced by organizations.

Improve Decision Making: By making use of various algorithms, Machine Learning can be used to make better business decisions. For example, Machine Learning is used to forecast sales, predict downfalls in the stock market, identify risks and anomalies, etc.

Uncover patterns & trends in data: Finding hidden patterns and extracting key insights from data is the most essential part of Machine Learning. By building predictive models and using statistical techniques, Machine Learning allows you to dig beneath the surface and explore the data at a minute scale. Understanding data and extracting patterns manually will take days, whereas Machine Learning algorithms can perform such computations in less than a second.

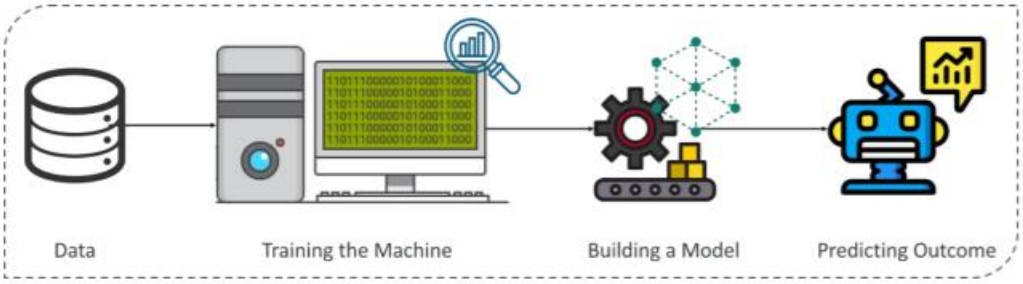
Solve complex problems: From detecting the genes linked to the deadly ALS disease to building self-driving cars, Machine Learning can be used to solve the most complex problems.

- **Machine Learning Applications**



- **Machine Learning Definitions**

The term Machine Learning was first coined by Arthur Samuel in the year 1959. According to Samuel - "Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed." The very first formal definition was given by Tom M. Mitchell (1997) - “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.” So, we can say that Machine Learning is a subset of Artificial Intelligence (AI) which provides machines the ability to lean automatically & improve from experience without being explicitly programmed.
A Machine Learning process begins by feeding the machine lots of data, by using this data the machine is trained to detect hidden insights and trends. These insights are then used to build a Machine Learning Model by using an algorithm in order to solve a problem.

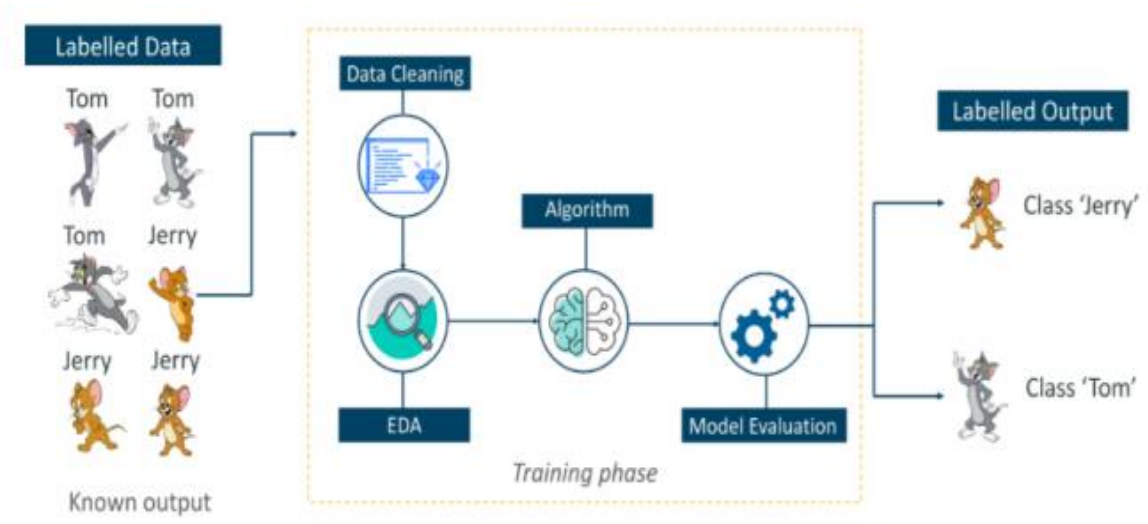


- **Machine Learning Types**

A machine can learn to solve a problem by following any one of the following three approaches. These are the ways in which a machine can learn:
1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Supervised Learning

"Supervised learning is a technique in which we teach or train the machine using data which is well labelled." To understand Supervised Learning let's consider an analogy. As kids we all needed guidance to solve math problems. Our teachers helped us understand what addition is and how it is done. Similarly, you can think of supervised learning as a type of Machine Learning that involves a guide. The labelled data set is the teacher that will train you to understand patterns in the data. The labelled data set is nothing but the training data set.



Consider the above figure. Here we're feeding the machine images of Tom and Jerry and the goal is for the machine to identify and classify the images into two groups (Tom images and Jerry images). The training data set that is fed to the model is labelled, as in, we're telling the machine, 'this is how Tom looks and this is Jerry'. By doing so you're training the machine by using labelled data.

In Supervised Learning, there is a well-defined training phase done with the help of labelled data. There are two kinds of supervise learning:

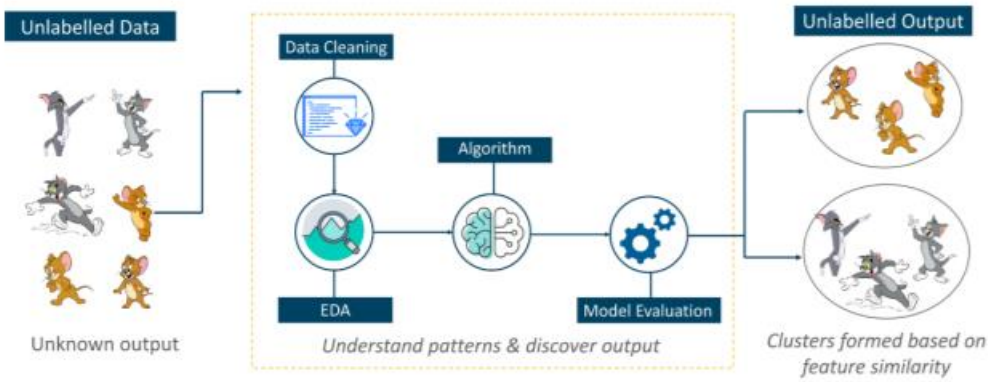
- Classification
- Regression

Difference between Classification and Regression

Classification	Regression
• Classification is the task of predicting a discrete class label	• Regression is the task of predicting a continuous quantity
• In a classification problem data is labelled into one of two or more classes	• A regression problem requires the prediction of a quantity
• A classification problem with two classes is called binary, more than two classes is called a multi-class classification	• A regression problem with multiple input variables is called a multivariate regression problem
• Classifying an email as spam or non-spam is an example of a classification problem	• Predicting the price of a stock over a period of time is a regression problem

Unsupervised Learning

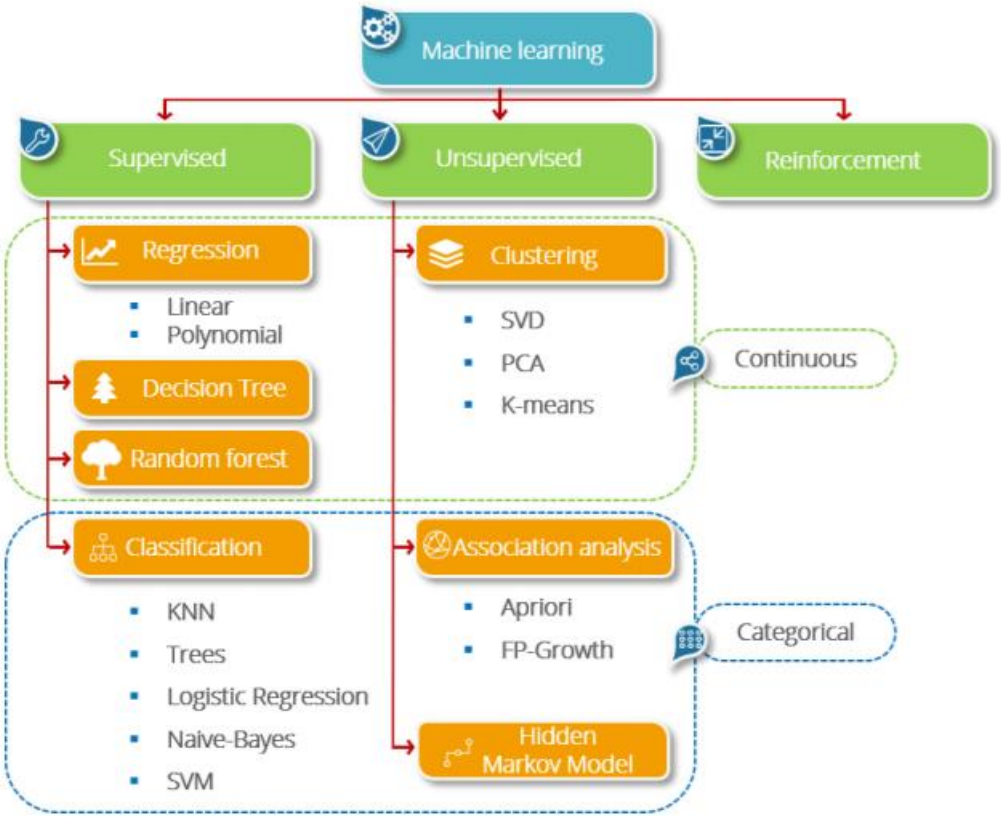
"Unsupervised learning involves training by using unlabelled data and allowing the model to act on that information without guidance." Think of unsupervised learning as a smart kid that learns without any guidance. In this type of Machine Learning, the model is not fed with labelled data, as in the model has no clue that 'this image is Tom and this is Jerry', it figures out patterns and the differences between Tom and Jerry on its own by taking in tons of data.



For example, it identifies prominent features of Tom such as pointy ears, bigger size, etc, to understand that this image is of type 1. Similarly, it finds such features in Jerry and knows that this image is of type 2. Therefore, it classifies the images into two different classes without knowing who Tom is or Jerry is. Google news is another example of unsupervised learning. Google news creates clusters of all the URLs of different e- news portal sharing the same news are mentioned at one place.

Reinforcement Learning

"Reinforcement Learning is a part of Machine learning where an agent is put in an environment and he learns to behave in this environment by performing certain actions and observing the rewards which it gets from those actions." Reinforcement Learning is mainly used in advanced Machine Learning areas such as self-driving cars, AplhaGo, etc.



NATURAL LANGUAGE PROCESSING

Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.

The study of natural language processing has been around for more than 50 years and grew out of the field of linguistics with the rise of computers. **Natural language processing** (**NLP**) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Natural Language Processing has various applications like sentimental analysis, text classification, spam detection and a many more. In my project we will focus on text classification.

TEXT CLASSIFICATION

Text classification also known as *text tagging* or *text categorization* is the process of categorizing text into organized groups. By using Natural Language Processing (NLP), text classifiers can automatically analyse text and then assign a set of pre-defined tags or categories based on its content.

Unstructured text is everywhere, such as emails, chat conversations, websites, and social media but it's hard to extract value from this data unless it's organized in a certain way. Doing so used to be a difficult and expensive process since it required spending time and resources to manually sort the data or creating handcrafted rules that are difficult to maintain. Text classifiers with NLP have proven to be a great alternative to structure textual data in a fast, cost-effective, and scalable way.

Text Classification is an example of supervised machine learning task since a labelled dataset containing text documents and their labels is used for train a classifier. An end-to-end text classification pipeline is composed of three main components:

1. Dataset Preparation: The first step is the Dataset Preparation step which includes the process of loading a dataset and performing basic pre-processing. The dataset is then splited into train and validation sets.
2. Feature Engineering: The next step is the Feature Engineering in which the raw dataset is transformed into flat features which can be used in a machine learning model. This step also includes the process of creating new features from the existing data.
3. Model Training: The final step is the Model Building step in which a machine learning model is trained on a labelled dataset.

CHAPTER 4

PROJECT DESIGN

There are two methods by which machines could attempt to solve the fake news problem better than humans. The first is that machines are better at detecting and keeping track of statistics than humans, for example it is easier for a machine to detect that the majority of verbs used are “suggests” and “implies” versus, “states” and “proves.” Additionally, machines may be more efficient in surveying a knowledge base to find all relevant articles and answering based on those many different sources. Either of these methods could prove useful in detecting fake news, but we decided to focus on how a machine can solve the fake news problem using supervised learning that extracts features of the language and content only within the source in question, without utilizing any fact checker or knowledge base**.**

- **Data Extraction:** Now that fake news has been defined and the target has been set, it is needed to analyse what features can be used in order to classify fake news. On a general basis if we analyse the news then there are majorly four principal raw data parts:-

➢ Source: Where does the news come from, who wrote it, is this source reliable or not.
➢ Headline: Short summary of the news content that try to attract the reader.
➢ Body Text: The actual text content of the news.
➢ Image/Video: Usually, textual information is agreed with visual information such as images, videos or audio.

I have been using **two datasets** in my project although one is enough but to have diversity in utility of an application is essential. In **one of the datasets** there have been **four features** that have been worked upon that is **unique id** of the news, **the label** of the particular instance, the **Headline** (title) and the **Body text.** And the **another dataset,** there are 3.6 million amazon reviews that are being scanned for fake or real and I am using some of them. This dataset has **two features the label** and the **text**. This is also an application where we classify and realize the importance of fake reviews that can create or destroy a brand name.

Fake news is used to influence the consumer, and in order to do that, they often use a specific language in order to attract the readers. On the other hand, non-fake news will mostly stick to a different language register, being more formal. This is linguistic-based features, to which can be added lexical features such as the total number of words, frequency of large words or unique words. This language testing can be done by using the Natural Language Processing.

- **Loading Data and splitting Data :**
  The dataset has been formed after analysing the data features and will be loaded on the platform we are performing the model.

- **Feature Engineering** :
  Cleaning, preparing and manipulating data to apply to the algorithms for our use. Raw text data is to be converted into feature vector and new features will be created using existing dataset.

  Count Vectors as features
  TF-IDF Vectors as features

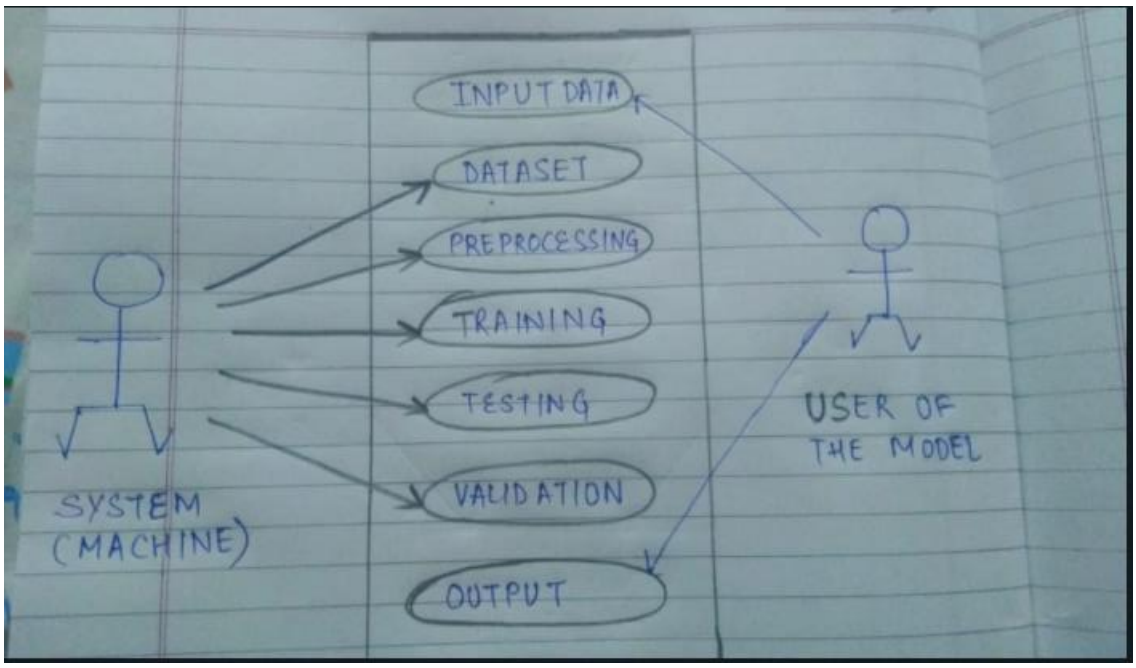
  - Word level
  - N-Gram level
  - Character level

  Word Embeddings as features
  Text / NLP based features

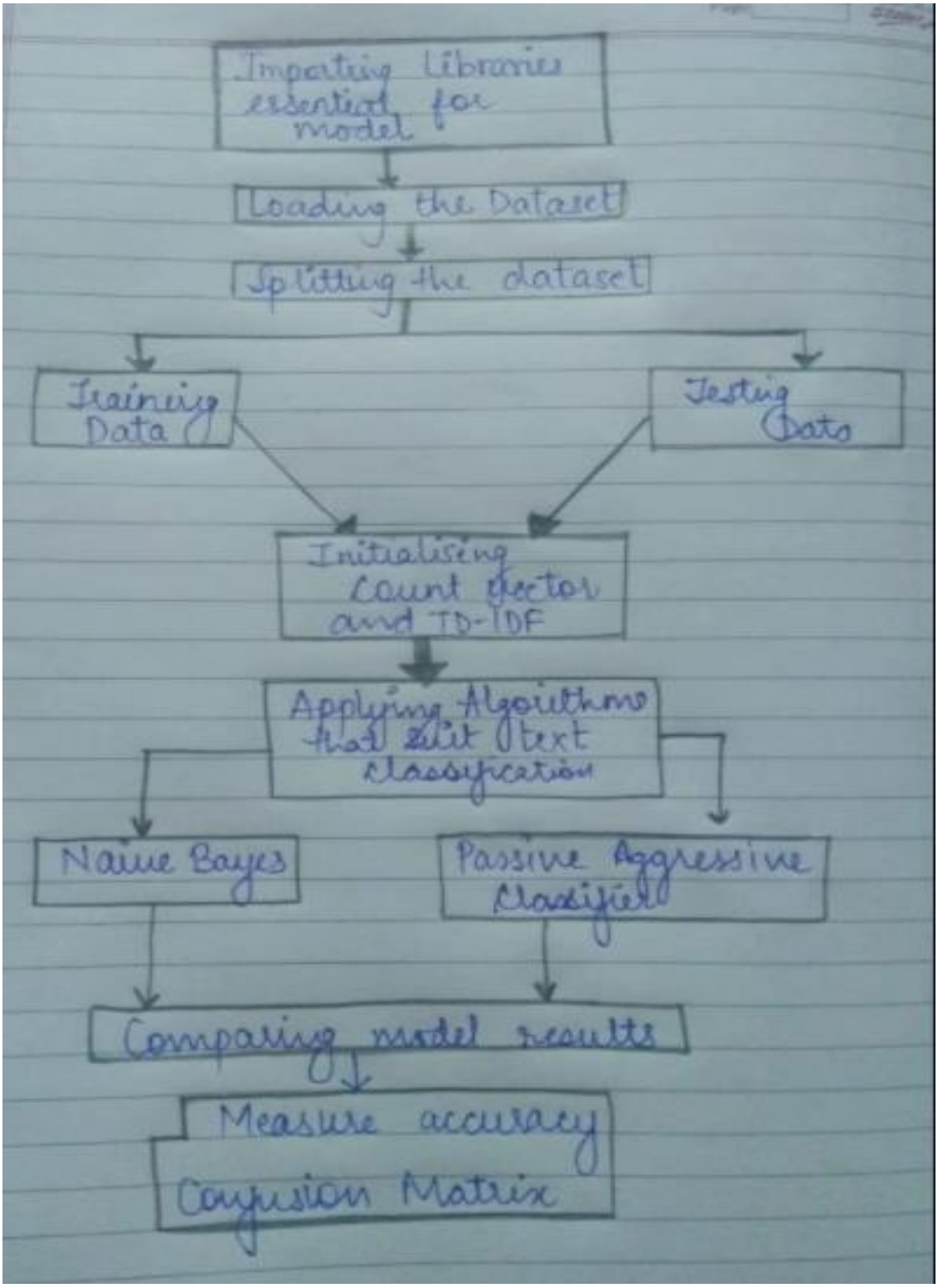
- **Train the model**
  The dataset connects to the algorithm and the algorithm leverages mathematical modelling to learn and develop predictions. This is what is called the experience to the model and this is where the model starts learning.

- **Testing the model**
  Originally the dataset was divided into two parts one for training the model and the other was to learn the accuracy and the efficiency of the pseudo model.

- **USE CASE OF THE MODEL**



- **FLOW CHART OF THE MODEL**


Importing Libraries essential for model
Loading the Dataset
Splitting the dataset
Training Data
Testing Data
Initialising Count Vector and TD-IDF
Applying Algorithms that suit Text Classification
Naive Bayes
Passive Aggressive Classifier
Comparing model results
Measure accuracy
Confusion Matrix

DATA PREPROCESSING

1. Load Data: The text is small and will load quickly and easily fit into memory. This will not always be the case and you may need to write code to memory map the file. Tools like NLTK will make working with large files much easier.

2. Split by Whitespace: Clean text often means a list of words or tokens that we can work with in our machine learning models. This means converting the raw text into a list of words and saving it again. We can also use select words to select the words (use the regex model (re) and split the document into words by selecting for strings of alphanumeric characters (a-z, A-Z, 0-9 and '_').

3. Remove Punctuation: We may want the words, but without the punctuation like commas and quotes. We also want to keep contractions together. One way would be to split the document into words by white space (as in "2. Split by Whitespace"), then use string translation to replace all punctuation with nothing (e.g. remove it).

4. Normalizing Case: It is common to convert all words to one case. This means that the vocabulary will shrink in size, but some distinctions are lost (e.g. "Apple" the company vs. "apple" the fruit is a commonly used example).

5. Filter out Stop Words (and Pipeline): Stop words are those words that do not contribute to the deeper meaning of the phrase. They are the most common words such as: "the", "a", and "is". For some applications like documentation classification, it may make sense to remove stop words. NLTK provides a list of commonly agreed upon stop words for a variety of languages, such as English, French.

6. Stemming: Stemming refers to the process of reducing each word to its root or base. For example "fishing," "fished," "fisher" all reduce to the stem "fish." Some applications, like document classification, may benefit from stemming in order to both reduce the vocabulary and to focus on the sense or sentiment of a document rather than deeper meaning. There are many stemming algorithms, although a popular and longstanding method

is the Porter Stemming algorithm. This method is available in NLTK via the Porter Stemmer class.

FEATURE EXTRACTION

- **Bag-of-Words Model**

  Bag of Words model is used to pre-process the text by converting it into a *bag of words*, which keeps a count of the total occurrences of most frequently used words. In practice, the Bag-of-words model is mainly used as a tool of feature generation. After transforming the text into a "bag of words", we can calculate various measures to characterize the text. The most common type of characteristics, or features calculated from the Bag-of-words model is term frequency, namely, the number of times a term appears in the text.

In Bayesian spam filtering, an e-mail message is modelled as an unordered collection of words selected from one of two probability distributions: one representing spam and one representing legitimate e-mail ("ham"). Imagine there are two literal bags full of words. One bag is filled with words found in spam messages, and the other with words found in legitimate e-mail. While any given word is likely to be somewhere in both bags, the "spam" bag will contain spam-related words such as "stock", "Viagra", and "buy" significantly more frequently, while the "ham" bag will contain more words related to the user's friends or workplace.

To classify an e-mail message, the Bayesian spam filter assumes that the message is a pile of words that has been poured out randomly from one of the two bags, and uses Bayesian probability to determine which bag it is more likely to be in.

- **Count Vectorizer**

  Count Vector is a matrix notation of the dataset in which every row represents a document from the corpus, every column represents a term from the corpus, and every cell represents the frequency count of a particular term in a particular document.

- **TD-IDF**
  - TF-IDF score represents the relative importance of a term in the document and the entire corpus. TF-IDF score is composed by two terms: the first computes the normalized Term Frequency (TF), the

second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

- TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
  IDF(t) = log_e(Total number of documents / Number of documents with term t in it)
- TF-IDF Vectors can be generated at different levels of input tokens (words, characters, n-grams)
- **a. Word Level TF-IDF :** Matrix representing tf-idf scores of every term in different documents
  b. N-gram Level TF-IDF : N-grams are the combination of N terms together. This Matrix representing tf-idf scores of N-grams
  c. Character Level TF-IDF : Matrix representing tf-idf scores of character level n-grams in the corpus

ALGORITHM USED

Naive Bayes

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature

Passive Aggressive Classifier

Passive Aggressive Classifier: The Passive Aggressive Algorithm is an online algorithm; ideal for classifying massive streams of data (e.g. twitter). It is easy to implement and very fast. It works by taking an example, learning from it and then throwing it away.

Passive Aggressive Algorithms are a family of online learning algorithms (for both classification and regression) proposed by Crammer at al. The idea is very simple and their performance has been proofed to be superior to many other alternative methods like Online Perception Classification

CHAPTER 5

TESTING

Evaluating your machine learning algorithm is an essential part of any project. Your model may give you satisfying results when evaluated using a metric ***say accuracy_score*** but may give poor results when evaluated against other metrics such as ***logarithmic_loss*** or any other such metric. Most of the times we use classification accuracy to measure the performance of our model. Since we are using text classification in our problem statement we can use the following performance evaluation measures.

- Accuracy
- Confusion Matrix
- F1-score
- Recall
- Precision

ACCURACY

It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

It works well only if there are equal number of samples belonging to each class.

For example, consider that there are 98% samples of class A and 2% samples of class B in our training set. Then our model can easily get **98% training accuracy** by simply predicting every training sample belonging to class A.

When the same model is tested on a test set with 60% samples of class A and 40% samples of class B, then the **test accuracy would drop down to**

60%. Classification Accuracy is great, but gives us the false sense of achieving high accuracy.

#CONFUSION MATRIX

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

There are 4 important terms :

True Positives : The cases in which we predicted YES and the actual output was also YES.

True Negatives : The cases in which we predicted NO and the actual output was NO.

False Positives : The cases in which we predicted YES and the actual output was NO.

False Negatives : The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the **"main diagonal"** i.e

$$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample}$$

#F1-SCORE

The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'. The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall. The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing. The formula for the standard F1-score is the harmonic mean of the precision and recall. A perfect model has an F-score of 1.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$
$$= \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

#PECISION AND RECALL

Precision : It is the number of correct positive results divided by the number of positive results predicted by the classifier.

Recall : It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$
$$= \frac{\text{retrieved and relevant documents}}{\text{all retrieved documents}}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$
$$= \frac{\text{retrieved and relevant documents}}{\text{all relevant documents}}$$

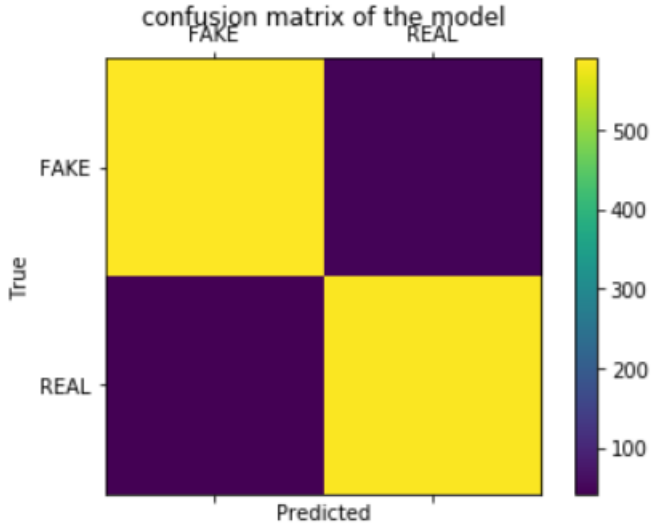
CHAPTER 6

EXPERIMENT ANALYSIS AND RESULTS

After applying the data pre-processing techniques and building a model that is created over the dataset, the algorithm Passive-Aggressive Classifier was found to give the highest accuracy of 92.9% .

The confusion matrix of the model is:-

```
[[590  48]
 [ 42 587]]
```



$$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample}$$

$$= \frac{590+587}{1267} = 92.89\%$$

CHAPTER 7

CONCLUSION

Result analysis

Some hypotheses can be made on why same models works very well on one dataset and does not work well on the other one. The first thing we can think of is that the original hypothesis on different styles of writing between fake and reliable news is only verified in one dataset, as texts come from online newspapers (or pretending to be), and thus capitalize on advertisements for making money. On the other hand this model can be used be used for social media fake news detection too (as used) and we see that on one of the dataset passive aggressive classifier gives excellent results but on the other dataset that I have tried and experienced writing this project paper it hasn't been working so well. So on further basis we can look forward to apply deep learning for this problem statement.

Project analysis

As per the goal of the project with the help of various algorithms applied on the data set after the application of Natural Language Processing, we have been able to classify the news as Fake and the Real News.

REFRENCES

- https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/

- https://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf

- ***https://www.geeksforgeeks.org/passive-aggressive-classifiers/***

- https://www.kaggle.com/priyeshsingh057/detecting-fake-news-with-python?scriptVersionId=24149425

Research Papers

- Machine Learning for Detection of Fake News by Nicole O'Brien
- Fake News Detection Using Machine Learning Author: Simon Lorent Supervisor: Ashwin Itoo