NEW HORIZON COLLEGE OF ENGINEERING
Autonomous College Permanently Affiliated to VTU, Approved by AICTE & UGC
Accredited by NAAC with 'A' Grade. Accredited by NBA

nirf TOP ENGINEERING COLLEGE OF INDIA as per NATIONAL INSTITUTIONAL RANKING FRAMEWORK, 2018
Ministry of Human Resource Development, Government of India

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

A

PROJECT REPORT

ON

FAKE NEWS DETECTION

Submitted in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

BY

PESALA SANATH(1NH16CS722)
SUMAN N(1NH16CS734)
SOMPALLI DINESH(1NH17CS731)

Under the guidance of

Ms. UMA
Senior Assistant Professor
Dept. of CSE, NHCE

NEW HORIZON COLLEGE OF ENGINEERING
Autonomous College Permanently Affiliated to VTU, Approved by AICTE & UGC
Accredited by NAAC with 'A' Grade. Accredited by NBA

nirf TOP ENGINEERING COLLEGE OF INDIA as per NATIONAL INSTITUTIONAL RANKING FRAMEWORK, 2018
Ministry of Human Resource Development, Government of India

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

It is hereby certified that the project work entitled **"Fake News Detection"** is a bonafide work carried out by **PESALA SANATH(1NH16CS722), SUMAN N(1NH16CS734), SOMPALLI DINESH(1NH17CS731)** in partial fulfilment for the award of **Bachelor of Engineering** in **COMPUTER SCIENCE AND ENGINEERING** of the New Horizon College of Engineering during the year **2020-2021.** It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said Degree.

.....	.....	.....
Signature of Guide	Signature of HOD	Signature of Principal
(Ms.Uma)	**(Dr. B. Rajalakshmi)**	**(Dr. Manjunatha)**

External Viva

Name of Examiner	**Signature with date**
1.	**.....**
2.	**.....**

Fake News Detection

ORIGINALITY REPORT

15%	%	15%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	S

PRIMARY SOURCES

1	"ICICCT 2019 – System Reliability, Quality Control, Safety, Maintenance and Management", Springer Science and Busine Media LLC, 2020 Publication
2	Shlok Gilda. "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection", 2017 IEEE 15th Student Conference on Research and Development (SCOReD), 201 Publication
3	Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu. "Fake News Detection on Social Media", ACM SIGKDD Explorations Newsletter, 2017 Publication
4	Gerardo Ernesto Rolong Agudelo, Octavio J Salcedo Parra, Julio Barón Velandia. "Chap 52 Raising a Model for Fake News Detectior Using Machine Learning in Python", Springe Science and Business Media LLC, 2018

ABSTRACT

This Project comes up with the applications of NLP (Natural Language Processing) techniques for detecting the fake news. As demonstrated by the widespread effects of the large onset of fake news, humans are inconsistent if not poor detectors of fake news. With this, efforts have been made to automate the process of fake news detection. While these tools are useful, to create a more complete end to end solution, we need to account for more difficult cases where reliable sources release fake news. As such, the goal of this project was to create a tool for detecting the language patterns that characterize fake and real news using machine learning and natural language processing techniques. The results of this project demonstrate the ability for machine learning to be useful in this task. We have built a model that catches many intuitive indications of real and fake news.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be impossible without the mention of the people who made it possible, whose constant guidance and encouragement crowned our efforts with success.

I have great pleasure in expressing my deep sense of gratitude to **Dr. Mohan Manghnani**, Chairman of New Horizon Educational Institutions for providing necessary infrastructure and creating good environment.

I take this opportunity to express my profound gratitude to **Dr. Manjunatha** , Principal NHCE, for this constant support and encouragement.

I am grateful to **Dr. Prashanth C.S.R**, Dean Academics, for his unfailing encouragement and suggestions, given to me in the course of my project work.

I would also like to thank **Dr. B. Rajalakshmi**, Professor and Head, Department of Computer Science and Engineering, for her constant support.

I express my gratitude to **Ms. Uma,** Senior Assistant Professor, my project guide, for constantly monitoring the development of the project and setting up precise deadlines. Her valuable suggestions were the motivating factors in completing the work.

Finally, a note of thanks to the teaching and non-teaching staff of Dept of Computer Science and Engineering, for their cooperation extended to me, and my friends, who helped me directly or indirectly in the course of the project work.

PESALA SANATH (1NH16CS722)

SUMAN N(1NH16CS734)

SOMPALLI DINESH (1NH17CS731)

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

There was a time once if anyone required any news, he or she would sit up for the next-day newspaper. With the expansion of on-line newspapers UN agency update news nearly instantly, individuals have found a more robust and quicker thanks to learn of the matter of his/her interest. Today social-networking systems, on-line news portals, and alternative on-line media became the most sources of reports through that fascinating and breaking news are shared at a fast pace news are shared at a fast pace.

Several news portals serve interest by feeding with distorted, part correct, and typically fanciful news that is probably to draw in the eye of a target cluster of individuals. Faux news has become a significant concern for being harmful typically spreading confusion and deliberate misinformation among the individuals.

The term faux news has become a buzz word lately. A united definition of the term faux news remains to be found. It may be outlined as a sort of info that consists of deliberate information or hoaxes unfold via ancient print and broadcast print media or on-line social media. These are revealed sometimes with the intent to mislead to wreck a community or person, produce chaos, and gain financially or politically.

Since individuals are usually unable to pay enough time to see reference and take care of the credibleness of reports, machine-driven detection of pretend news is indispensable. Therefore, it's receiving nice attention from the analysis community.

The previous works on faux news have applied many ancient machine learning ways and neural networks to detect faux news. They need targeted on police investigation news of specific variety.

Accordingly, they developed their models and designed options for specific datasets that match their topic of interest. It's probably that these approaches would suffer from dataset bias and are probably to perform poorly on news of another topic. Number of the present studies have

additionally created comparisons among totally different ways of pretend news detection. Prevaricator and experimented some existing models on the dataset. The comparison result hints U.S. Totally different models will perform on a structured dataset like prevaricator.

The length of this dataset isn't ample for neural network analysis and a few models were found tosuffer from overfitting. Several advanced machine learning models, e.g., neural network primarily based ones don't seem to be applied that are established best in several text classification issues.

1.2 PROBLEM-DEFINITION

Objective of Rumor detection is to classify a bit of knowledge as rumor or real. Four steps are concerned model Detection, Tracking, Stance & truthfulness that may facilitate to discover the rumors. These posts thought-about the vital sensors for crucial the believability of rumor. Rumor detection will more classes in four subtasks stance classification, truthfulness classification, rumor chase, rumor classification.

Still few points that need a lot of details to grasp the matter and additionally we are able to learn from the results that's it really rumor or not and if its rumor then what quantity for these queries we tend to believe that combination information of information and knowledge facet is needed to explore those areas that also inexplicable.

1.3 PROJECT PURPOSE

Learning from data and engineered knowledge to overcome fake news issue on social media. To achieve the goal a new combination algorithm approach shall be developed which will classify the text as soon as the news will publish online. In developing such a new classification approach as a starting point for the investigation of fake news we first applied available data set for our learning. The first step in fake news detection is classifying the text immediately once the news published online. Classification of text is one of the important research issues in the field of text mining. As we knew that dramatic increase in the content available online gives raise problem to manage this online textual data. So, it is important to classify the news into the specific classes i.e., Fake, Non fake, unclear.

1.4 PROJECT FEATURES

The main feature of this system is to propose a general and effective approach to predict the fake news or real news using data mining techniques. The main goal of the proposed system is to analyze and study the hidden patterns and relationships between the data present in the fake news dataset. The solution to problem can provide information to prevent fake or real news from taking place, and consequently generate great societal and technical impacts. Most of the existing work solves these problems separately by different models. Fake news detection is one of the vital things that is very important for the society, so dealing with this becomes more important. The analysis and prediction play an important role in the problem definition.

The social network collected in our study manifests noticeable polarized. Each user in this plot is assigned a credibility score in the range [−1, +1] computed as the difference between the proportion of retweeted true and fake news negative values representing fake are depicted in red and credible users are represented in blue. The node positions of the graph are determined by topological embedding computed via Latent Dirichlet algorithm, grouping together nodes of the graph that are more strongly connected and mapping apart nodes that have weak connections. We observe that credible and non-credible users tend to form two distinct communities, suggesting these two categories of tweeters to have mostly homophilic interactions. While a deeper study of this phenomenon is ahead of the scope , we note that comparable polarization has been seen before in social networks, e.g. In the context of political discourse, and might be related to echo chamber theories that attempt to explain the reasons for the difference in fake and true news propagation patterns.

1.5 MODULE DESCRIPTION

1.5.1 DATA GATHERING:

The first step during this project or in any data processing project is that the assortment of information to be studied or examined to search out the hidden relationships between the

information members. The necessary concern whereas selecting a dataset is that the information that we have a tendency to square measure gathering ought to be relevant to the matter statement and it should be massive enough in order that the logical thinking derived from the information is helpful to extract some necessary patterns between the information specified they will be wont to predict the longer term events or will be studied for additional analysis. The results of the method of gathering and making a group of information results into what we have tendency to decision as a Dataset. The dataset contains massive volume information of information which will be analyzed to induce some knowledge from the databases. This is often be a very important step within the method as a result of selecting the inappropriate dataset can lead USA to incorrect results.

1.5.2 DATA PREPROCESS:

The primary information collected from the web sources remains within the raw variety of statements, digits and qualitative terms. The data contains error, omissions and inconsistencies. It needs corrections once careful scrutinizing the finished questionnaires. The subsequent steps square measure concerned within the process of primary information. Large volume of data collected through field survey must be classified for similar details of individual responses.

Data Preprocessing may be a technique that's won't convert the raw information data information into a clean data set. In alternative words, whenever the information is gathered from completely different sources it's collected in raw format that isn't possible for the analysis.

Therefore, sure steps square measure dead to convert the information the info he information into low clean data set. This system is performed before the execution of unvaried Analysis. The set of steps is understood as information preprocessing the method comprises:

- Data cleanup
- Data Integration
- Data Reduction

Data Preprocessing is important owing to the presence of unformatted globe information principally globe information consists of:

• Inaccurate information - There square measure several reasons for missing information like data is not unendingly collected, a slip in information entry, technical issues with bioscience and far additional.

• The presence of clanging information - The explanations for the existence of clanging information might be a technological drawback of device that gathers information, a person's mistake throughout information entry and far additional.

• Inconsistent information - The presence of inconsistencies square measure because of the explanations specified existence of duplication at intervals information, human information entry, containing mistakes in codes or names i.e., violation of information constraints and far additional.

1.5.3 CLASSIFICATION

This technique is used to divide various data into different classes. This process is also similar to clustering. It segments data records into various segments which are known as classes. Unlike clustering, here we have knowledge of different clusters. Ex: Outlook email, they have an algorithm to categorize an email as legitimate or spam.

CHAPTER 2

LITERATURE SURVEY

2.1 DATA MINING

Literature survey is that the most vital step in code development method. Before developing the tool it's necessary to see the time issue, economy and company strength. Once these things are satisfied, then next steps is to determine which operating system and language can be used for developing the tool Once the programmers begin building the tool the programmers would like heap of external support. This support is obtained from senior programmers, from book or from websites Before building the system the on top of thought area unit taken under consideration for developing the projected system.. We have to analyze the Data mining Outline Survey:

2.1.1 Data Mining Survey

Data mining is a data analysis technique which allows us to study and identify different patterns and relationships between the data. In other words, data mining is a technique which can be employed to extract information from large and extensive datasets and convert the information into a prominent structure so that it can be used further for gaining inference and knowledge on the data so as to prevent the crimes.

Data mining contains techniques for analysis which involve various domains. For instance, some of the domains involved in data mining are Statistics, Machine Learning and Database systems. Data mining is additionally spoken as “Knowledge discovery in databases (KDD)”.

The real assignment of data mining systems is the semi-automatic or computerized analysis of huge volumes of data to extract earlier unknown relationships such as groups of data members(clustering analysis),unusual records(outlier or anomaly detection),and dependencies. Normally, this contains database techniques like spatial indices.

These relationships that are discovered can be used as input data or may also be used in depth analysis for example, in machine learning or predictive analysis.

Data mining can identify multiple groups in the data, that can be put to further than use for accurate projections by a decision support system.

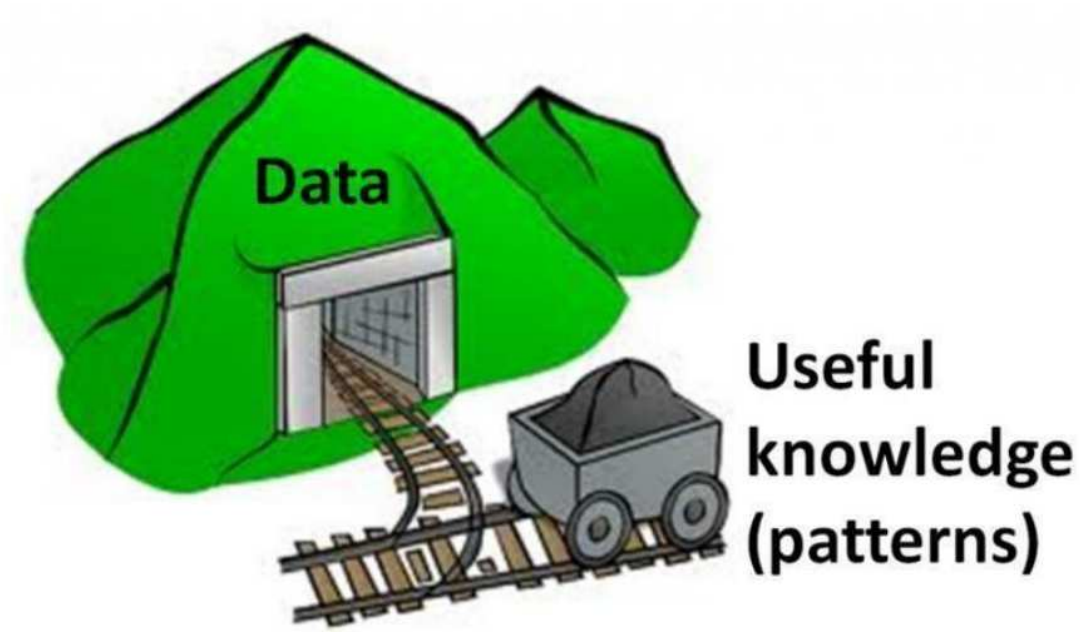


Fig 2.1: Data Mining

2.1.1 Stages in Data Mining

There are 4 major steps in data mining which are described as follows:

- **Data Sources:** This stage includes gathering the data or making a dataset on which the analysis or the study has performed. The datasets can be of many forms for instance, they can be new letters, databases, excel sheets or various other sources like websites, blogs and social media. An appropriate dataset must be chosen in order to perform an efficient study or analysis. The dataset must be chosen which is appropriate and well suited with respect to the problem definition.

- **Data Exploration:** This step includes preparing the data properly for analysis and study. This step is mainly focused on cleaning the data and removing the anomalies from the data. As there is a large amount of data there is always a great chance that some of the data might be missing or some data might be wrong. Thus, for efficient analysis we require the data to be maintained properly. This process includes removing the incorrect data and replacing the data which is missing with either mean or median of the whole data. This step is also generally known as data pre-processing.

- **Data Modeling:** In this step the relationships and patterns that were hidden in the data are examined and extracted from the datasets. The data can be modeled based on the technique that is being used. Some of the different techniques that can be used for modeling data are clustering, classification and association and decision trees.
- **Deploying Models:** Once the relationships and patterns present in the data are discovered we need to put that knowledge to use. These patterns can be used to predict events in the future and they can be used for further analysis. The discovered patterns can be used as inputs for machine learning and predictive analysis for the datasets.

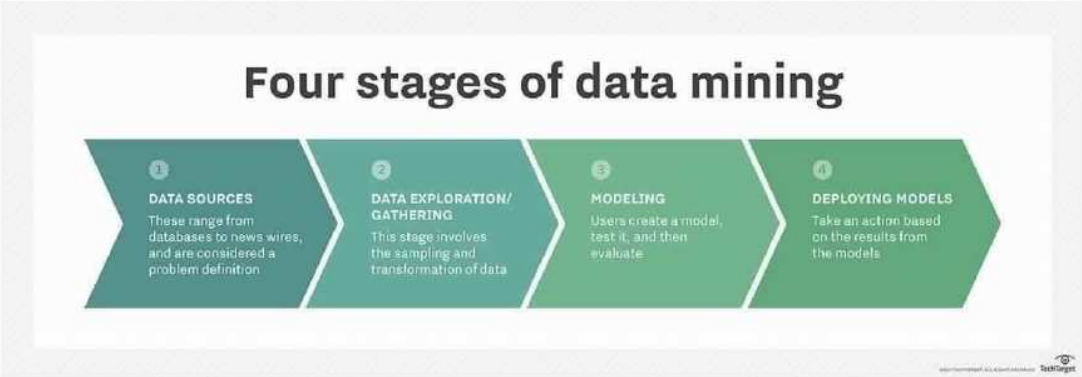


Fig 2.2: Stages in Data Mining

2.1.2 Techniques in Data Mining:

- **Classification:** This technique is used to divide various data into different classes. This process is also similar as clustering. It segments data records into various segments which are known as classes. Unlike clustering, here we have knowledge of different clusters. Ex: Outlook email, they have an algorithm to categorize an email as legitimate or spam.
- **Association:** This technique is used to discover hidden patterns in the data and identifying interesting relations between the variables in a database. Ex: It is used in retail industry.
- **Prediction:** This technique is used only for uses. It is used extract relationships between independent and dependent variables in the dataset. Ex: We use this technique to predict profit obtained from sales for the future.

- **Clustering:** A cluster is referred to as a group of data objects. The data objects that are similar in properties are kept in the same cluster. In other words we can tell that clustering is a process of discovering groups or clusters. Here we do not have prior knowledge of the clusters. Ex: It can be used in consumer profiling.

- **Sequential Patterns:** This is an essential aspect of data mining techniques its main aim is to discover similar patterns in the dataset. Ex: E-commerce websites suggestions are based on what we have bought previously.

- **Decision Trees:** This technique is a vital role in data mining because it is easier to understand for the users. The decision tree begins with a root which is a simple question. As they can have multiple answers we get our nodes of the decision tree also the questions in the root node might lead to another set of questions. Thus, the nodes keep adding in the decision tree. At last, we are allowed for making a final decision on it.

Apart from these techniques there are certain other techniques which allow us to remove noisy data and clean the dataset. This helps us to get accurate analysis and prediction results.

Fig 2.3: Data Mining Techniques

2.1.3 Benefits of Data Mining:

Data mining has various uses in various sectors of the society:

- In finance sector, it can be used for modeling risks accurately regarding loans and other facilities.
- In marketing, it can be used for predicting profits and can be used for creating targeted advertisements for various customers.
- In retail sector, it is used for improving consumer experience and increasing the amount of profits.
- Tax governing organizations use it to determine frauds in transactions.

2.2 EXISTING SYSTEM

There occurs a large body of study on the topic of machine learning methods for news detection, most of it has been concentrating on classifying online reviews and openly 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 attention within the literature.

Outlines many approaches that seem promising towards the aim of completely classify the false articles. They note that easy content related n-grams and shallow parts-of-speech (POS) tagging have demonstrated insufficient for the classification task, often failing to account for important context information. Instead, these methods have been shown valuable only in tandem with more complex techniques of analysis.

2.3 PROPOSED SYSTEM

Proposed method Due to the complexity of fake news detection in social media, it is evident that a feasible method must contain several aspects to accurately tackle the issue. This is why the proposed method is a combination of semantic analysis. The proposed method is entirely composed of Artificial Intelligence approaches, which is critical to accurately classify between the real and the fake, instead of using algorithms that are unable to cognitive functions. The three-part method is a combination between Machine Learning algorithms that divide into natural language

processing methods. Although each of these approaches can be solely used to classify and detect fake news, in order to increase the accuracy and be applicable to the social media domain, they have been combined into an integrated algorithm as a method for fake news detection.

It is important that we have some mechanism for detecting fake news, or at the very least, an awareness that not everything we read on social media may be true, so we always need to be thinking critically. This way we can help people make more informed decisions and they will not be fooled into thinking what others want to manipulate them into believing.

The procedure to be followed for the proposed system is given as follows:

- We collect the data and frame the dataset according to the problem definition to get the analysis correct and produce results which are efficient to meet the goals of the system. Then we must trim the dataset as per the needs of the problem definition and create a new dataset which contains the required fields, attributes and properties that are suitable for the analysis.

- Then we perform the data pre-processing procedure to replace any missing values with either mean value or median value of the given data. This is done to reduce the noise and inconsistency in the data. Then we perform the normalization operation on the dataset to remove any outliers in the dataset which can lead to inaccurate results in the analysis of the dataset.

- The results of the classification algorithms are stored in a data frame.

- After classifying the data we can import the data frame into an excel sheet so that it obtain fake news or real news.

2.4 SOFTWARE DESCRIPTION

2.4.1 JUPYTER NOTEBOOK:

The Jupyter Notebook App is a server-customer application that permits altering and running note pad records by means of an internet browser. The Jupyter Notebook App can be executed on a nearby work area requiring no web access as portrayed in this report or can be introduced on a remote server and got to through the web. A scratch pad part is a computational motor that executes the code contained in a Notebook record.

When you open a Notebook report, the related part is consequently propelled. At the point when the scratch pad is executed either cell-by-cell, the portion plays out the calculation and produces the outcomes. Contingent upon the sort of calculations, the piece may expend critical CPU and RAM. Note that the RAM isn't discharged until the part is closed down, he Notebook Dashboard is the part which is indicated first when you dispatch Jupyter Notebook App. The Notebook Dashboard is essentially used to open note pad archives, and to deal with the running portions. The Notebook Dashboard has different highlights like a record director, in particular exploring organizers, renaming and erasing documents**.**

2.4.2 MATPLOTLIB:

People are exceptionally visual animals, we comprehend things better when we see things envisioned. The progression to showing investigations, results or bits of knowledge can be a bottleneck, we probably won't realize where to begin or you may have as of now a correct configuration as a top priority, however then inquiries will have unquestionably gone over your brain.

When we are working with the Python plotting library Matplotlib, the initial step to responding to the above inquiries is by structure up information on themes.

Plot creation, which could bring up issues about what module we precisely need to import pylab, how we precisely ought to approach instating the figure and the Axes of our plot, how to utilize matplotlib in Jupyter note pads.

Plotting schedules, from straightforward approaches to plot your information to further developed

methods for picturing your information. Essential plot customizations, with an emphasis on plot legends and content, titles, tomahawks marks and plot format.

Since all is set for us to begin plotting your information, it's an ideal opportunity to investigate some plotting schedules. We'll regularly go over capacities like plot() and disperse(), which either draw focuses with lines or markers interfacing them, or draw detached focuses, which are scaled or shaded. In any case, as you have just found in the case of the primary area, we shouldn't neglect to pass the information that you need these capacities to utilize.

In conclusion, we will quickly cover two manners by which we can alter Matplotlib, with templates and the settings.

ax.bar()	Vertical rectangles
ax.barh()	Horizontal rectangles
ax.axhline()	Horizontal line across axes
ax.vline()	Vertical line across axes
ax.fill()	Filled polygons
ax.fill_between()	Fill between y-values and 0
ax.stackplot()	Stack plot

2.4.3 NUMPY

NumPy is one of the bundles that we can't miss when we are learning information science, principally in light of the fact that this library gives us a cluster information structure that holds a few advantages over Python records, for example, being increasingly reduced, quicker access in perusing and composing things, being progressively advantageous and increasingly productive.

NumPy is a Python library that is the center library for logical registering in Python. It contains an accumulation of apparatuses and strategies that can be utilized to settle on a PC numerical models of issues in Science and Engineering. One of these apparatuses is an elite multidimensional cluster object that is an incredible information structure for effective calculation of exhibits and lattices.

To work with these clusters, there's a tremendous measure of abnormal state scientific capacities work on these grids and exhibits. Since you have set up your condition, it's the ideal opportunity for the genuine work. In fact, you have officially gone for some stuff with exhibits in the above Data camp Light pieces. We haven't generally gotten any genuine hands-on training with them, since we originally expected to introduce NumPy all alone pc. Since we have done this current, it's a great opportunity to perceive what you have to do so as to run the above code pieces without anyone else. A few activities have been incorporated underneath with the goal that you would already be able to rehearse how it's done before we begin our own. To make a numpy exhibit, we can simply utilize the np.array () work. There's no compelling reason to proceed to retain these NumPy information types in case we are another client, but we do need to know and mind what information we are managing. The information types are there when we need more power over how our information is put away in memory and on plate. Particularly in situations where we are working with broad information, it's great that we know to control the capacity type.

2.4.4 PANDAS

Pandas is an open-source, BSD-authorized Python library giving elite, and simple to-utilize information structures and information examination instruments for the Python programming language. Python with Pandas is utilized in a wide scope of fields including scholastic and business areas including money, financial matters, Statistics, examination, and so on. In this instructional exercise, we will get familiar with the different highlights of Python Pandas and how to utilize them practically speaking.

This instructional exercise has been set up for the individuals who try to become familiar with the essentials and different elements of Pandas. It will be explicitly valuable for individuals working with information purging and examination. In the wake of finishing this instructional exercise, we will wind up at a moderate dimension of ability from where you can take yourself to more elevated amounts of skill. We ought to have a fundamental comprehension of Computer Programming phrasing. Library utilizes vast majority of the functionalities of NumPy. It is recommended that we experience our instructional exercise on NumPy before continuing with this instructional exercise.

2.4.5 ANACONDA

Anaconda constrictor is bundle director. Jupyter is an introduction layer. Boa constrictor endeavors to explain the reliance damnation in python where distinctive tasks have diverse reliance variants, in order to not influence distinctive venture conditions to require diverse adaptations, which may meddle with one another. Jupyter endeavors to fathom the issue of reproducibility in investigation by empowering an iterative and hands-on way to deal with clarifying and imagining code by utilizing rich content documentations joined with visual portrayals, in a solitary arrangement.

Boa constrictor is like pyenv, venv and minconda, it's intended to accomplish a python situation that is 100% reproducible on another condition, autonomous of whatever different forms of a task's conditions are accessible. It's somewhat like Docker, however limited to the Python biological system.

Jupyter is an astounding introduction device for expository work, where we can display code in squares, joins with rich content depictions among squares, and the consideration of organized yield from the squares, and charts created in an all around planned issue by method for another square's code. Jupyter is extraordinarily great in expository work to guarantee reproducibility in somebody's exploration, so anybody can return numerous months after the fact and outwardly comprehend what somebody attempted to clarify and see precisely which code drove which representation and end. Regularly in diagnostic work we will finish up with huge amounts of half-completed note pads clarifying Proof-of-Concept thoughts, of which most won't lead anyplace at first.

2.4.6 PYTHON

Python is a translated, object-arranged, unusual state programming language with dynamic semantics. Its unusual state worked in information structures, joined with dynamic composing and dynamic authoritative, make it attractive for Rapid Application Development, just as for use as a scripting or paste language to interface existing segments together. Python's basic, simple to learn language structure underlines intelligibility and hence decreases the expense of program support. Python underpins modules and bundles, which empowers program seclusion and code

reuse. The Python translator and the broad customary library are accessible in source or parallel structure without charge for every single significant stages.

Frequently, code engineers begin to look all starry eyed at Python on account of the expanded efficiency it provides. Since there is no aggregation step, the alter test-troubleshoot cycle is staggeringly quick. Troubleshooting Python programs is simple: a bug or awful information will never cause a division blame. Rather, when the mediator finds a blunder, it raises a special case. At the point when the program does not get the special case, the translator prints a stack follow. A source level debugger permits assessment of nearby and worldwide factors, assessment of discretionary articulations, setting breakpoints, venturing through the code a line at any given moment, etc. The debugger is written in Python itself, vouching for Python's contemplative power. Then again, frequently the speediest methodology to troubleshoot a program is to add a couple of print proclamations to the source: the quick alter test-investigate cycle makes this straightforward methodology successful.

Python is an item situated, abnormal state programming language with incorporated unique semantics essentially for net and application improvement. It is incredibly alluring in the field of Rapid Application Growth since it offers dynamic composing and dynamic limiting alternatives.

Python is generally basic, so it's anything but difficult to learn since it requires a one of a kind language structure that centers on coherence. Designers can peruse and interpret Python code a lot simpler than different dialects. Thusly, this decreases the expense of program upkeep and improvement since it enables groups to work cooperatively without huge language and experience obstructions.

Moreover, Python underpins the utilization of modules and bundles, which means that projects can be planned in a secluded style and code can be reused over an assortment of tasks.

A standout amongst the foremost encouraging advantages of Python is that both the standard library and the mediator are accessible for nothing out of pocket, in both parallel and source structure. There is no restrictiveness either, as Python and all the necessary instruments are accessible on every single real stage. In this way, it is a tempting alternative for designers who would prefer not to stress paying high improvement costs.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 FUNCTIONAL REQUIREMENTS

The functions of software systems are defined in functional requirements and the behavior of the system is evaluated when presented with specific inputs or conditions which may include calculations, data manipulation and processing and other specific functionality.

- Our system should be able to read the data and preprocess data.
- It should be able to analyze the fake data.
- It should be able to group data based on hidden patterns.
- It should be able to assign a label based on its data groups.
- It should be able to split data into train set and test set.
- It should be able to train model using train set.
- It must validate trained model using test set.
- It should be able to classify the fake and real data.

3.2 NON-FUNCTIONAL REQUIREMENTS

Nonfunctional requirements illustrate how a system must behave and create constraints of its functionality. This type of constraints is also known as the system's quality features. Attributes such as performance, security, usability, compatibility are not the feature of the system, they are a required characteristic. They are "developing" properties that emerge from the whole arrangement and hence we can't compose a particular line of code to execute them. Any attributes required by the user are described by the specification. We must contain only those needs that are appropriate for our design.

Some Non-Functional Requirements are as follows:

- Reliability
- Maintainability
- Performance
- Portability
- Scalability
- Flexibility

3.2.1 ACCESSIBILITY:

Availability is a general term used to depict how much an item, gadget, administration, or condition is open by however many individuals as would be prudent.

In our venture individuals who have enrolled with the cloud can get to the cloud to store and recover their information with the assistance of a mystery key sent to their email ids. UI is straightforward and productive and simple to utilize.

3.2.2 MAINTAINABILITY:

In programming designing, viability is the simplicity with which a product item can be altered as:

- Correct absconds

- Meet new necessities

New functionalities can be included in the task based the client necessities just by adding the proper documents to existing venture utilizing ASP. Net and C# programming dialects. Since the writing computer programs is extremely straightforward, it is simpler to discover and address the imperfections and to roll out the improvements in the undertaking.

3.2.3 SCALABILITY:

Framework is fit for taking care of increment all out throughput under an expanded burden when assets (commonly equipment) are included. Framework can work ordinarily under circumstances, for example, low data transfer capacity and substantial number of clients.

3.2.4 PORTABILITY:

Portability is one of the key ideas of abnormal state programming. Convenient is the product code base component to have the capacity to reuse the current code as opposed to making new code while moving programming from a domain to another. Venture can be executed under various activity conditions gave it meet its base setups. Just framework records congregations would need to be designed in such case.

3.3 HARDWARE REQUIREMENTS

- Processor : Any Processor above 500 MHz
- RAM : 4 GB
- Hard Disk : 500 GB
- System : Pentium IV 2.4 GHz

Any system with above or higher configuration is compatible for this project.

3.4 SOFTWARE REQUIREMENTS

- Operating system : Windows 7/8/9/10
- Programming language : Python
- IDE: Jupyter Notebook
- Tools: Anaconda

CHAPTER 4

DESIGN

4.1 DESIGN GOALS

Truth discovery plays a distinguished role in modern era as we need correct data currently over ever. Completely different application areas truth discovery is used particularly wherever we want to require crucial choice supported the reliable data extracted from different sources e.g. Healthcare, crowd sourcing and knowledge extraction.

Social media provides extra resources to the researchers to supplement and enhance news context models. Social models engagements within the analysis method and capturing the knowledge in numerous forms from a spread of views. After we check the present approaches we will class social modelling context in stance based mostly and propagation based. One necessary purpose that we want to focus on here that some existing social context models approaches used for pretend news detection. We are going to strive with the assistance of literature those social context models that used for rumor detection. Correct assessment of faux news stories shared on social media platforms and identification of faux contents mechanically with the assistance of knowledge sources and social judgment.

The main options of the planned system are:

- More economical.
- Better pretended news detector systems.
- It reduces the time quality of the system.
- System that contains easier design to grasp.

- Processing of enormous quantity of knowledge becomes easier.

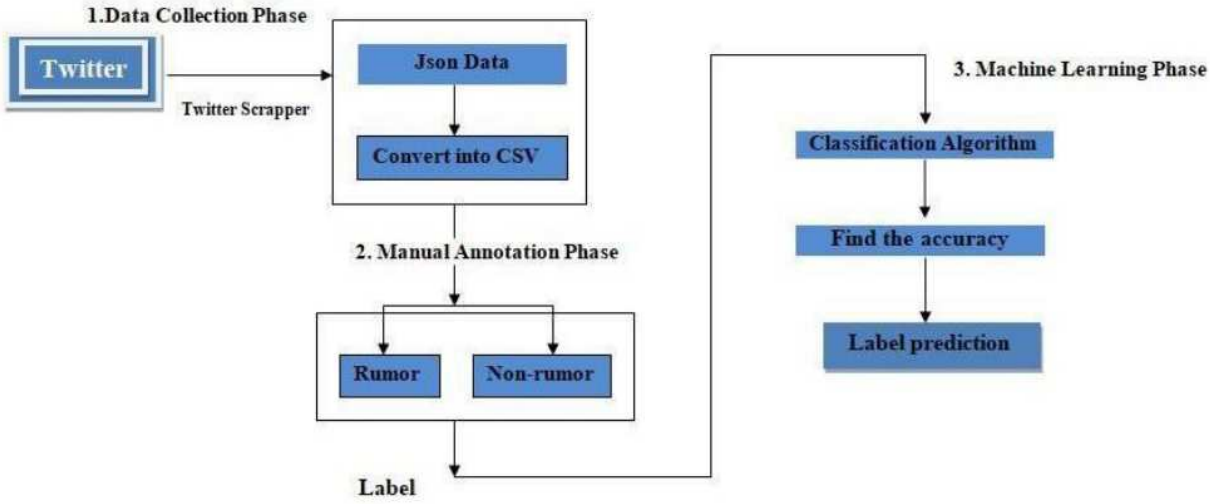


Fig 4.1 System design

4.2 USE CASE DIAGRAM

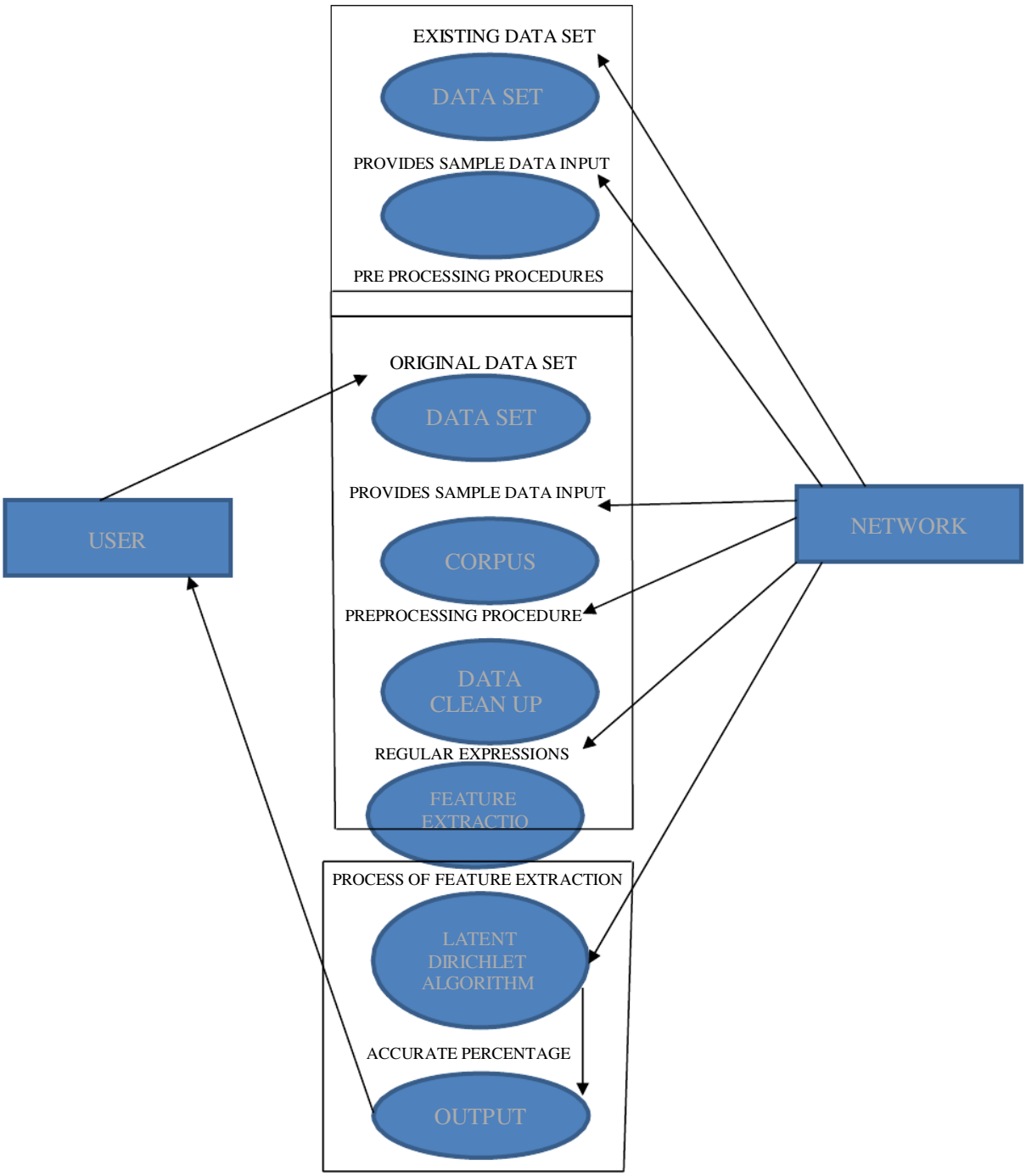


Fig 4.2: Use case diagram

CHAPTER 5

IMPLEMENTATION

5.1 DATASET

A data set could also be associate assortment of information. Most generally knowledge set corresponds to the contents of one data table, or one mathematics information matrix, wherever each column of the table represents a specific variable, and every row corresponds to a given member of the data set in question. The data set lists values for every of the variables, like height Associate in weight of associate object, for each member of the knowledge set. Each worth is known as knowledge purpose set would possibly comprise data for one or extra members, appreciate the quantity of rows.

The dataset consists of the following details regarding the faux incidents:

• Category - category of the faux news. This may be the target variable that goes to the expected.

• Descript - Description of the faux news incident.

• Day of week - the day of the week.

• Address - the approximate address of the news.

• X – meridian

• Y - Latitude

5.2 DATA PREPROCESSING

The primary data collected from net sources remains within the raw style of statements, digits and qualitative terms. The data contains error, omissions and inconsistencies. It needs corrections once careful scrutinizing the finished questionnaires. The subsequent steps area unit concerned within the process of primary knowledge. An enormous volume of data collected through field survey has to be sorted for similar details of individual responses.

Data Pre processing could be a technique that's accustomed convert the raw knowledge into a clean data set. In different knowledge, whenever the info is gathered from completely different sources it's collected in raw format that isn't possible for the analysis.

Therefore, sure steps area unit dead to convert the knowledge into a tiny low clean data set. This system is performed before the execution of unvarying Analysis. The set of steps is believed as data pre processing. The tactic comprises:

• Data improvement

• Data Integration

• Data Transformation

• Data Reduction

• Data Pre processing is very important attributable to the presence of unformatted planet data

Principle .

Inaccurate data - There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics and much more.

The presence of noisy data - The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.

Inconsistent data - The presence of inconsistencies due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more. The column Resolution is dropped because it does not provide any assistance and has no significance in helping to predict the target variable.

5.3 CLASSIFICATION

This technique is used to divide various data into different classes. This process is also similar of clustering. It segments data records into various segments which are known as classes. Unlike clustering, here we have knowledge of different clusters. Ex: Outlook email, they have an algorithm to categorize an email as legitimate or spam.

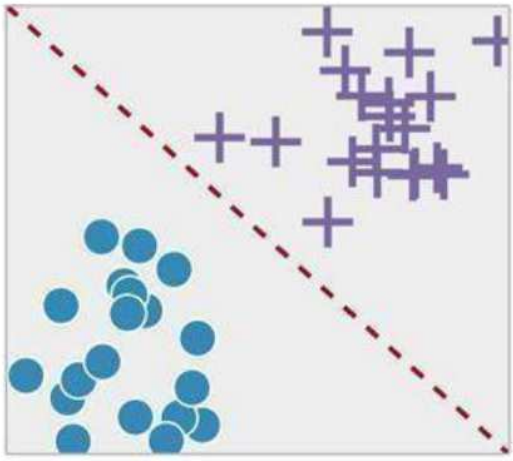


Fig 5.1: Classification

Some of the classification algorithms are:

- Linear Classifiers: Logistic Regression, Naive Bayes Classifier
- Support Vector Machines
- Decision Trees
- Boosted Trees
- Random Forest
- Neural Networks

5.3.1 LATENT DIRICHLET ALGORITHM

As additional data becomes offered, it becomes tougher to search out and find out what we want. We need tools to assist North American nation organize, search and perceive quantity of knowledge.

Topic modelling provides ways for mechanically organizing, understanding, searching, and summarizing giant electronic archives:

1. Discover the hidden themes within the assortment.
2. Annotate the documents consistent with these themes.
3. Use annotations to prepare, summarize, search, and kind predictions.

Some Assumptions:

• We have a set of documents D, D1, D2, D3 etc.

• Each document is just a collection of words or a “bag of words”. Thus, the order of the words and the grammatical role of the words (subject, object, verbs) are not considered in the model.

• Words like am/is/are/of/a/the/but/... can be eliminated from the documents as a preprocessing step since they don’t carry any information about the “topics”.

• In fact, we can eliminate words that occur in at least %80 ~ %90 of the documents!

• Each document is composed of NN “important” or “effective” words, and we want KK topics.

MODEL DEFINITION

- Each topic could be a distribution over words.

- Each document could be a mixture of corpus-wide topics.

- Each word is drawn from one in all these topics.

- We solely observe the words at intervals the documents and the alternative structure are hidden variables.

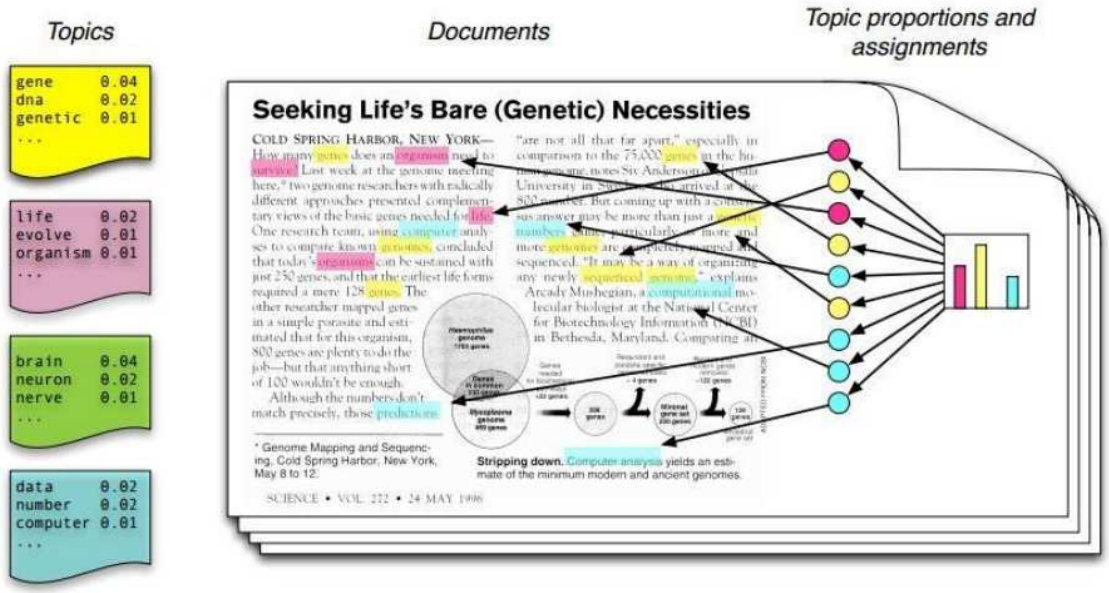


Fig 5.2: Process of algorithm

Our goal is to infer or estimate the hidden variables, i.e. computing their distribution conditioned on the documents p (Topic, Proportion, assignment)

• Nodes are RVs; edges indicate dependence.
• Shaded nodes are determined, and unshaded nodes are hidden.
• Plates indicate replicated variables.

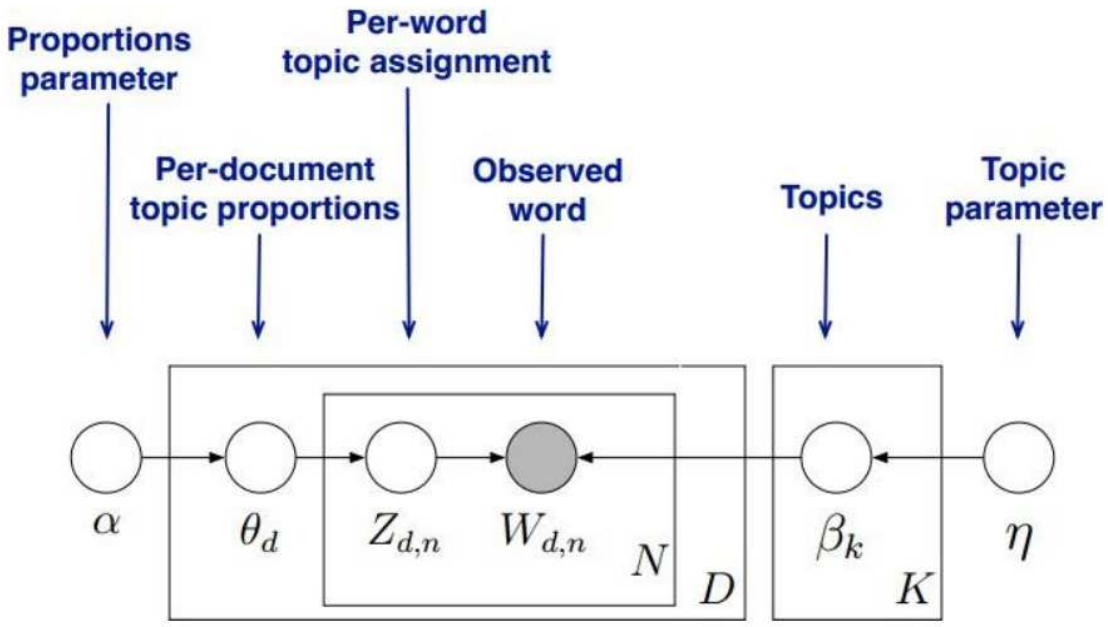


Fig 5.3. Sorting

5.3.2 Natural Language Processing

• **Natural language processing** (**NLP**) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

• Challenges in natural language processing frequently involve speech recognition, natural language understanding and natural language generation.

5.4 IMPLEMENTATION

```
import pandas as pd
import numpy as np


fake_df=pd.read_csv(r"C:\Users\sanath\Desktop\FakeNewProjectForSanath\FakeNewProjec
tForSanath\fake.csv")
real_df=pd.read_csv(r"C:\Users\sanath\Desktop\FakeNewProjectForSanath\FakeNewProject
ForSanath\real_news.csv")



print(fake_df.shape)
print(real_df.shape)


real_df2 = real_df[['title', 'content', 'publication']]
real_df2['label'] = 'real'
real_df2.head()


fake_df2 = fake_df[['title', 'text','site_url']]
fake_df2['label'] = 'fake'
fake_df2.head()


# let's obtain all the unique site_urls
site_urls = fake_df2['site_url']

# let's remove the domain extensions
site_urls2 = [x.split('.',1)[0] for x in site_urls]

# now let's replace the old site_url column
fake_df2['site_url'] = site_urls2
fake_df2.head()


# let's rename the features in our datasets to be the same
newlabels = ['title', 'content', 'publication', 'label']
real_df2.columns = newlabels
fake_df2.columns = newlabels

# let's concatenate the dataframes
```

```
frames = [fake_df2, real_df2]
news_dataset = pd.concat(frames)
news_dataset.head()



news_dataset.describe()


news_dataset.info()
!pip install nltk
import nltk
nltk.download('stopwords')
nltk.download('averaged_perceptron_tagger')


import re
from nltk.corpus import stopwords
import nltk

stop_words = set(stopwords.words('english'))

def cleanup(text):
    #print(text)
    text = re.sub('\d+', ' ', text)
    text = re.sub('<[A-Za-z /]+>', ' ', text)
    text = text.split()
    text = [w.strip('-') for w in text if not w.lower() in stop_words]
    text = ' '.join(text)
    text = re.sub(r"'[A-Za-z]", '', text)
    text = re.sub("[^A-Za-z -]+", '', text)

    temp = []
    res = nltk.pos_tag(text.split())
    for wordtag in res:
        if wordtag[1] == 'NNP':
            continue
        temp.append(wordtag[0].lower())
    text = temp

    return (text)

text = "This is a FABULOUS hotel James i would like to give 5 star. The front desk staff, the
doormen, the breakfast staff, EVERYONE is incredibly friendly and helpful and warm and
welcoming. The room was fabulous too."
cleanup (text)
```

```
# Remove punctuation
import string
news_dataset = news_dataset.dropna()
news_dataset["content"] = [text.translate(string.punctuation) for text in
news_dataset["content"]]


# White spaces removal
news_dataset["content"] = [text.strip() for text in news_dataset["content"]]

import nltk
nltk.download('punkt')

from nltk.tokenize import sent_tokenize, word_tokenize
news_dataset["Words"] = [word_tokenize(text) for text in news_dataset["content"]]
news_dataset.head()


!pip install matplotlib
import pandas as pd
from sklearn.model_selection import train_test_split
import sklearn
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import svm
from sklearn import metrics
from matplotlib import pyplot as plt
from sklearn.feature_extraction.text import HashingVectorizer
import itertools
import numpy as np


news_dataset=news_dataset.dropna()
y = news_dataset.label
news_dataset.drop("label", axis=1)
X_train, X_test, y_train, y_test = train_test_split(news_dataset['content'], y, test_size=0.33,
random_state=53)


# Initialize the `count_vectorizer`
count_vectorizer = CountVectorizer(stop_words='english')
# Fit and transform the training data
count_train = count_vectorizer.fit_transform(X_train)              # Learn the vocabulary
dictionary and return term-document matrix.
```

```
# Transform the test set
count_test = count_vectorizer.transform(X_test)
# Initialize the `tfidf_vectorizer`
tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.7)    # This removes
words which appear in more than 70% of the articles

# Fit and transform the training data
tfidf_train = tfidf_vectorizer.fit_transform(X_train)
# Transform the test set
tfidf_test = tfidf_vectorizer.transform(X_test)


# Support Vector Machine
# Training Performance
clf = svm.SVC()
clf.fit(count_train, y_train) # Model is trained here.
pred = clf.predict(count_test) # Predicting the output
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)


def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):


  plt.imshow(cm, interpolation='nearest', cmap=cmap)
   plt.title(title)
   plt.colorbar()
   tick_marks = np.arange(len(classes))
   plt.xticks(tick_marks, classes, rotation=45)
   plt.yticks(tick_marks, classes)

   if normalize:
      cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
      print("Normalized confusion matrix")

   else:
      print('Confusion matrix, without normalization')

   thresh = cm.max() / 2.
   for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
      plt.text(j, i, cm[i, j],
               horizontalalignment="center",
               color="white" if cm[i, j] > thresh else "black")
```

```
    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')


from sklearn import metrics
cm = metrics.confusion_matrix(y_test, pred, labels=['fake', 'real'])
plot_confusion_matrix(cm, classes=['fake', 'real'])


cm = metrics.confusion_matrix(y_test, pred, labels=['fake', 'real'])
plot_confusion_matrix(cm, classes=['fake', 'real'])


# Saving Model And Prediction on new data set

import pickle
pickle.dump(count_vectorizer, open(r'count_vectorizer.pickle', "wb"))
pickle.dump(tfidf_vectorizer, open(r'tfidf_vectorizer.pickle', "wb"))

filename = r'finalized_model_SVM.pkl'
file = open(filename, 'wb')
loaded_model = pickle.dump(clf,file)

file = open(filename, 'rb')
# load the unpickle object into a variable
model = pickle.load(file)


count_vectorizer1=pickle.load(open(r'count_vectorizer.pickle', "rb"))
tfidf_vectorizer2=pickle.load(open(r'tfidf_vectorizer.pickle', "rb"))



valid=count_vectorizer1.transform(pd.Series(""))

print("Given News Article Is: ",model.predict(valid)[0])
```

CHAPTER 6
TESTING

The reason for testing is to seek out blunders. Testing is that the manner toward endeavoring to seek out every doable blame or disadvantage in a very work item. It offers associate approach to visualize the quality of elements, sub gatherings, congregations similarly as a completed item it's the manner toward active programming with the goal of guaranteeing that the computer code. Framework lives up to its wants associated shopper needs and doesn't flop in an unsuitable manner. There area unit completely different types of check. every check kind tends to a selected testing requirement.

6.1 TYPES OF TESTS

6.1.1 UNIT TESTING

Unit testing includes the structure of experiments that approve that the inward program principle is functioning, which program inputs turn out substantial yields. All alternative branches and within code stream need to be approved. It's the attempting of individual programming units of the appliance. It's done when the finishing of a private unit before combination. This can be a basic testing, that depends on info of its development and is obtrusive. Unit checks perform elementary tests at half level and test a selected procedure, application, and to boot framework style. Unit tests guarantee that each extraordinary manner of a business procedure performs exactly to the recorded particular and contains clearly characterized info sources and anticipated outcomes.

6.1.2 INTEGRATION TESTING

Joining tests are intended to test incorporated programming segments to decide whether they really keep running as one program. Testing is occasion driven and is progressively worried about the fundamental result of screens or fields. Incorporation tests exhibit that despite the fact of the segments were separately fulfillment, as appeared by effectively unit testing, the mix of parts is right and reliable. Coordination testing is explicitly gone for uncovering the issues that emerge from the blend of segments.

6.1.3 VALIDATION TESTING

A building approval test (EVT) is performed on first building models, to guarantee that the essential unit performs to plan objectives and particulars. It is imperative in recognizing plan issue and fathoming them as right off the bat in the structure cycle as could reasonably be expected, is the way to keeping ventures on schedule and inside spending plan. Over and over again, item plan and execution issues are not identified until late in the item improvement cycle — when the item is prepared to be transported. The familiar saying remains constant. It costs a penny to roll out an improvement in building, a dime underway and a dollar after an item is in the field.

Check is a Quality control process that is utilized to assess whether an item, administration, or framework conforms to guidelines, details, or conditions forced toward the beginning of an improvement stage. Check can be being developed, scale-up, or creation. This is regularly an inside procedure.

Approval is a Quality affirmation procedure of setting up proof that gives a high level of confirmation that an item, administration, or framework achieves its planned prerequisites. This regularly includes acknowledgment of qualification for reason with end clients and other item partners.

The testing process overview is as follows:

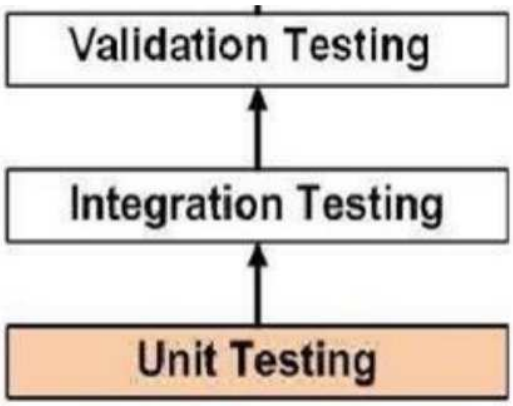


Figure 6.1: The testing process

6.1.4 SYSTEM TESTING

Framework testing of programming or equipment is trying led on a total, coordinated framework to assess the framework's consistence with its predetermined prerequisites. Framework testing falls inside the extent of discovery testing, and all things considered, ought to require no learning of the inward plan of the code or rationale.

When in doubt, framework testing takes, as its information, the majority of the incorporated programming segments that have effectively passed joining testing and furthermore the product framework itself coordinated with any relevant equipment systems.

Framework testing is a progressively constrained sort of testing, it looks to identify absconds both inside the between collections and furthermore inside the framework all in all.

Framework testing is performed on the whole framework with regards to a Functional Requirement Specifications as well as a System Requirement Specification.

Framework testing tests the structure, yet in addition the conduct and even the trusted desires for the client. It is likewise planned to test up to and past the limits characterized in the product necessities specifications.

CHAPTER 7

SNAPSHOTS

In [1]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
fake_df=pd.read_csv(r"C:\Users\sanath\Desktop\FakeNewProjectForSanath\FakeNewProjectForSanath\fake.csv")
real_df=pd.read_csv(r"C:\Users\sanath\Desktop\FakeNewProjectForSanath\FakeNewProjectForSanath\real_news.csv")



print(fake_df.shape)
print(real_df.shape)
```

```
(12999, 20)
(15712, 11)
```

In [4]:
```
real_df2 = real_df[['title', 'content', 'publication']]
real_df2['label'] = 'real'
real_df2.head()
```

```
c:\users\sanath\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

Out[4]:

	title	content	publication	label
0	House Republicans Fret About Winning Their Hea...	WASHINGTON — Congressional Republicans have...	New York Times	real
1	First, a Mixtape. Then a Romance. - The New Yo...	Just how is Hillary Kerr, the founder of ...	New York Times	real
2	Calling on Angels While Enduring the Trials of...	Angels are everywhere in the Muñiz family's ap...	New York Times	real
3	U.S. Plans to Step Up Military Campaign Agains...	ABU DHABI, United Arab Emirates — The Obama...	New York Times	real
4	272 Slaves Were Sold to Save Georgetown. What ...	WASHINGTON — The human cargo was loaded on ...	New York Times	real

Fig : 7.1

In [5]:
```
fake_df2 = fake_df[['title', 'text','site_url']]
fake_df2['label'] = 'fake'
fake_df2.head()
```

```
c:\users\sanath\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

Out[5]:

	title	text	site_url	label
0	Muslims BUSTED: They Stole Millions In Gov't B...	Print They should pay all the back all the mon...	100percentfedup.com	fake
1	Re: Why Did Attorney General Loretta Lynch Ple...	Why Did Attorney General Loretta Lynch Plead T...	100percentfedup.com	fake
2	BREAKING: Weiner Cooperating With FBI On Hilla...	Red State : \nFox News Sunday reported this mo...	100percentfedup.com	fake
3	PIN DROP SPEECH BY FATHER OF DAUGHTER Kidnappe...	Email Kayla Mueller was a prisoner and torture...	100percentfedup.com	fake
4	FANTASTIC! TRUMP'S 7 POINT PLAN To Reform Heal...	Email HEALTHCARE REFORM TO MAKE AMERICA GREAT ...	100percentfedup.com	fake

In [6]:
```
# Let's obtain all the unique site_urls
site_urls = fake_df2['site_url']

# Let's remove the domain extensions
site_urls2 = [x.split('.',1)[0] for x in site_urls]

# now let's replace the old site_url column
fake_df2['site_url'] = site_urls2
fake_df2.head()
```

```
c:\users\sanath\appdata\local\programs\python\python37\lib\site-packages\ipykernel_launcher.py:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

Fig : 7.2

Out[6]:

	title	text	site_url	label
0	Muslims BUSTED: They Stole Millions In Gov't B...	Print They should pay all the back all the mon...	100percentfedup	fake
1	Re: Why Did Attorney General Loretta Lynch Ple...	Why Did Attorney General Loretta Lynch Plead T...	100percentfedup	fake
2	BREAKING: Weiner Cooperating With FBI On Hilla...	Red State : \nFox News Sunday reported this mo...	100percentfedup	fake
3	PIN DROP SPEECH BY FATHER OF DAUGHTER Kidnappe...	Email Kayla Mueller was a prisoner and torture...	100percentfedup	fake
4	FANTASTIC! TRUMP'S 7 POINT PLAN To Reform Heal...	Email HEALTHCARE REFORM TO MAKE AMERICA GREAT ...	100percentfedup	fake

In [7]:
```
# Let's rename the features in our datasets to be the same
newlabels = ['title', 'content', 'publication', 'label']
real_df2.columns = newlabels
fake_df2.columns = newlabels

# Let's concatenate the dataframes
frames = [fake_df2, real_df2]
news_dataset = pd.concat(frames)
news_dataset.head()
```

Out[7]:

	title	content	publication	label
0	Muslims BUSTED: They Stole Millions In Gov't B...	Print They should pay all the back all the mon...	100percentfedup	fake
1	Re: Why Did Attorney General Loretta Lynch Ple...	Why Did Attorney General Loretta Lynch Plead T...	100percentfedup	fake
2	BREAKING: Weiner Cooperating With FBI On Hilla...	Red State : \nFox News Sunday reported this mo...	100percentfedup	fake
3	PIN DROP SPEECH BY FATHER OF DAUGHTER Kidnappe...	Email Kayla Mueller was a prisoner and torture...	100percentfedup	fake
4	FANTASTIC! TRUMP'S 7 POINT PLAN To Reform Heal...	Email HEALTHCARE REFORM TO MAKE AMERICA GREAT ...	100percentfedup	fake

In [8]:
```
news_dataset.describe()
```

Out[8]:

	title	content	publication	label
count	28031	28665	28711	28711
unique	27388	28134	253	2
top	Get Ready For Civil Unrest: Survey Finds That ...		Reuters	real

Fig : 7.3

```
In [10]: news_dataset.info()
         !pip install nltk
         import nltk
         nltk.download('stopwords')
         nltk.download('averaged_perceptron_tagger')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 28711 entries, 0 to 15711
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   title        28031 non-null  object
 1   content      28665 non-null  object
 2   publication  28711 non-null  object
 3   label        28711 non-null  object
dtypes: object(4)
memory usage: 1.1+ MB
Collecting nltk
  Downloading https://files.pythonhosted.org/packages/92/75/ce35194d8e3022203cca0d2f896dbb88689f9b3fce8e9f9cff942913519d/nltk-3.5.zip (1.4MB)
Collecting click (from nltk)
  Downloading https://files.pythonhosted.org/packages/d2/3d/fa76db83bf75c4f8d338c2fd15c8d33fdd7ad23a9b5e57eb6c5de26b430e/click-7.1.2-py2.py3-none-any.whl (82kB)
Requirement already satisfied: joblib in c:\users\sanath\appdata\local\programs\python\python37\lib\site-packages (from nltk) (0.14.1)
Collecting regex (from nltk)
  Downloading https://files.pythonhosted.org/packages/33/60/c9dbe875daa0f63e7fe5711493b77ef28e4e4e9fb0ac8941da2abad9c87a/regex-2020.5.7-cp37-cp37m-win_amd64.whl (271kB)
Collecting tqdm (from nltk)
  Downloading https://files.pythonhosted.org/packages/c9/40/058b12e8ba10e35f89c9b1fdfc2d4c7f8c05947df2d5eb3c7b258019fda0/tqdm-4.46.0-py2.py3-none-any.whl (63kB)
Installing collected packages: click, regex, tqdm, nltk
  Running setup.py install for nltk: started
    Running setup.py install for nltk: finished with status 'done'
Successfully installed click-7.1.2 nltk-3.5 regex-2020.5.7 tqdm-4.46.0
```

```
You are using pip version 19.0.3, however version 20.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\sanath\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\sanath\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping taggers\averaged_perceptron_tagger.zip.
```

```
Out[10]: True
```

Fig : 7.4

In [14]:
```
import re
from nltk.corpus import stopwords
import nltk

stop_words = set(stopwords.words('english'))

def cleanup(text):
    #print(text)
    text = re.sub('\d+', ' ', text)
    text = re.sub('<[A-Za-z /]+>', ' ', text)
    text = text.split()
    text = [w.strip('-') for w in text if not w.lower() in stop_words]
    text = ' '.join(text)
    text = re.sub(r"'[A-Za-z]", '', text)
    text = re.sub("[^A-Za-z -]+", '', text)

    temp = []
    res = nltk.pos_tag(text.split())
    for wordtag in res:
        if wordtag[1] == 'NNP':
            continue
        temp.append(wordtag[0].lower())
    text = temp

    return (text)

text = "This is a FABULOUS hotel James i would like to give 5 star. The front desk staff, the doormen, the breakfast staff, EVER
cleanup (text)
```

Out[14]:
```
['hotel',
 'would',
 'like',
 'give',
 'star',
 'front',
 'desk',
 'staff',
 'doormen',
 'breakfast',
 'staff',
 'incredibly',
 'friendly',
 'helpful',
 'warm',
 'welcoming',
 'room',
 'fabulous',
```

Fig : 7.5

In [15]:
```
# Remove punctuation
import string
news_dataset = news_dataset.dropna()
news_dataset["content"] = [text.translate(string.punctuation) for text in news_dataset["content"]]
```

In [16]:
```
# White spaces removal
news_dataset["content"] = [text.strip() for text in news_dataset["content"]]
```

In [17]:
```
import nltk
nltk.download('punkt')

from nltk.tokenize import sent_tokenize, word_tokenize
news_dataset["Words"] = [word_tokenize(text) for text in news_dataset["content"]]
news_dataset.head()
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\sanath\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

Out[17]:

	title	content	publication	label	Words
0	Muslims BUSTED: They Stole Millions In Gov't B...	Print They should pay all the back all the mon...	100percentfedup	fake	[Print, They, should, pay, all, the, back, all...
1	Re: Why Did Attorney General Loretta Lynch Ple...	Why Did Attorney General Loretta Lynch Plead T...	100percentfedup	fake	[Why, Did, Attorney, General, Loretta, Lynch, ...
2	BREAKING: Weiner Cooperating With FBI On Hilla...	Red State : +Fox News Sunday reported this mor...	100percentfedup	fake	[Red, State, :, +Fox, News, Sunday, reported, ...
3	PIN DROP SPEECH BY FATHER OF DAUGHTER Kidnappe...	Email Kayla Mueller was a prisoner and torture...	100percentfedup	fake	[Email, Kayla, Mueller, was, a, prisoner, and,...
4	FANTASTIC! TRUMP'S 7 POINT PLAN To Reform Heal...	Email HEALTHCARE REFORM TO MAKE AMERICA GREAT ...	100percentfedup	fake	[Email, HEALTHCARE, REFORM, TO, MAKE, AMERICA,...

In [19]:
```
!pip install matplotlib
import pandas as pd
from sklearn.model_selection import train_test_split
import sklearn
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import svm
from sklearn import metrics
from matplotlib import pyplot as plt
from sklearn.feature_extraction.text import HashingVectorizer
import itertools
import numpy as np
```

Fig : 7.6

In [19]:
```
!pip install matplotlib
import pandas as pd
from sklearn.model_selection import train_test_split
import sklearn
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import svm
from sklearn import metrics
from matplotlib import pyplot as plt
from sklearn.feature_extraction.text import HashingVectorizer
import itertools
import numpy as np


news_dataset=news_dataset.dropna()
y = news_dataset.label
news_dataset.drop("label", axis=1)
X_train, X_test, y_train, y_test = train_test_split(news_dataset['content'], y, test_size=0.33, random_state=53)
```

```
Collecting matplotlib
  Downloading https://files.pythonhosted.org/packages/b4/4d/8a2c06cb69935bb762738a8b9d5f8ce2a66be5a1410787839b71e146f000/matplotlib-3.2.1-cp37-cp37m-win_amd64.whl (9.2MB)
Collecting kiwisolver>=1.0.1 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/7e/e5/d8bd2d063da3b6761270f29038d2bb9785c88ff385009bf61589cde6e6ef/kiwisolver-1.2.0-cp37-none-win_amd64.whl (57kB)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\sanath\appdata\local\programs\python\python37\lib\site-packages (from matplotlib) (2.8.1)
Collecting cycler>=0.10 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/f7/d2/e07d3ebb2bd7af696440ce7e754c59dd546ffe1bbe732c8ab68b9c834e61/cycler-0.10.0-py2.py3-none-any.whl
Requirement already satisfied: numpy>=1.11 in c:\users\sanath\appdata\local\programs\python\python37\lib\site-packages (from matplotlib) (1.18.4)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in c:\users\sanath\appdata\local\programs\python\python37\lib\site-packages (from matplotlib) (2.4.7)
Requirement already satisfied: six>=1.5 in c:\users\sanath\appdata\local\programs\python\python37\lib\site-packages (from python-dateutil>=2.1->matplotlib) (1.14.0)
Installing collected packages: kiwisolver, cycler, matplotlib
Successfully installed cycler-0.10.0 kiwisolver-1.2.0 matplotlib-3.2.1

You are using pip version 19.0.3, however version 20.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.
```

In [20]:
```
# Initialize the `count_vectorizer`
count_vectorizer = CountVectorizer(stop_words='english')
# Fit and transform the training data
count_train = count_vectorizer.fit_transform(X_train)                # Learn the vocabulary dictionary and return term-document
# Transform the test set
count_test = count_vectorizer.transform(X_test)
```

Fig : 7.7

```
In [20]: # Initialize the `count_vectorizer`
         count_vectorizer = CountVectorizer(stop_words='english')
         # Fit and transform the training data
         count_train = count_vectorizer.fit_transform(X_train)                # Learn the vocabulary dictionary and return term-documen
         # Transform the test set
         count_test = count_vectorizer.transform(X_test)
         # Initialize the `tfidf_vectorizer`
         tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.7)    # This removes words which appear in more than 70% of th

         # Fit and transform the training data
         tfidf_train = tfidf_vectorizer.fit_transform(X_train)
         # Transform the test set
         tfidf_test = tfidf_vectorizer.transform(X_test)

In [21]: # Support Vector Machine
         # Training Performance
         clf = svm.SVC()
         clf.fit(count_train, y_train) # Model is trained here.
         pred = clf.predict(count_test) # Predicting the output
         score = metrics.accuracy_score(y_test, pred)
         print("accuracy:   %0.3f" % score)

         accuracy:   0.916
```

Fig : 7.8

In [22]:
```
def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):
    """
    See full source and example:
    http://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

    This function prints and plots the confusion matrix.
    Normalization can be applied by setting `normalize=True`.
    """
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")

    else:
        print('Confusion matrix, without normalization')

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')


from sklearn import metrics
cm = metrics.confusion_matrix(y_test, pred, labels=['fake', 'real'])
plot_confusion_matrix(cm, classes=['fake', 'real'])
```

```
Confusion matrix, without normalization
```

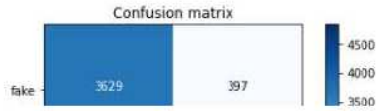
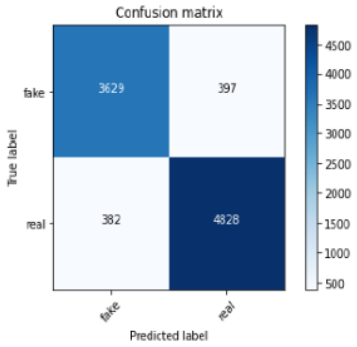


Fig :7.9

In [23]:
```
cm = metrics.confusion_matrix(y_test, pred, labels=['fake', 'real'])
plot_confusion_matrix(cm, classes=['fake', 'real'])
```

```
Confusion matrix, without normalization
```



In [24]:
```
# Saving Model And Prediction on new data set

import pickle
pickle.dump(count_vectorizer, open(r'count_vectorizer.pickle', "wb"))
pickle.dump(tfidf_vectorizer, open(r'tfidf_vectorizer.pickle', "wb"))

filename = r'finalized_model_SVM.pkl'
file = open(filename, 'wb')
loaded_model = pickle.dump(clf,file)

file = open(filename, 'rb')
# load the unpickle object into a variable
model = pickle.load(file)
```

In [25]:
```
count_vectorizer1=pickle.load(open(r'count_vectorizer.pickle', "rb"))
tfidf_vectorizer2=pickle.load(open(r'tfidf_vectorizer.pickle', "rb"))
```

Fig : 7.10

```
In [24]: # Saving Model And Prediction on new data set

         import pickle
         pickle.dump(count_vectorizer, open(r'count_vectorizer.pickle', "wb"))
         pickle.dump(tfidf_vectorizer, open(r'tfidf_vectorizer.pickle', "wb"))

         filename = r'finalized_model_SVM.pkl'
         file = open(filename, 'wb')
         loaded_model = pickle.dump(clf,file)

         file = open(filename, 'rb')
         # load the unpickle object into a variable
         model = pickle.load(file)
```

```
In [25]: count_vectorizer1=pickle.load(open(r'count_vectorizer.pickle', "rb"))
         tfidf_vectorizer2=pickle.load(open(r'tfidf_vectorizer.pickle', "rb"))
```

```
In [26]: valid=count_vectorizer1.transform(pd.Series("""Google Pinterest Digg Linkedin Reddit Stumbleupon Print Delicious Pocket Tumblr
         There are two fundamental truths in this world: Paul Ryan desperately wants to be president. And Paul Ryan will never be presider
         In a particularly staggering example of political cowardice, Paul Ryan re-re-re-reversed course and announced that he was back or
         â€" ABC News Politics (@ABCPolitics) November 5, 2016
         The Democratic Party couldnâ€™t have asked for a better moment of film. Ryanâ€™s chances of ever becoming president went down to
         The ringing endorsement of the man he clearly hates on a personal level speaks volumes about his own spinelessness. Ryan has post
         Whatâ€™s especially bizarre is how close Ryan came to making it through unscathed. For months the Speaker of the House refused to
         If 2016â€™s very ugly election has done any good itâ€™s by exposing the utter cowardice of the Republicans who once feigned moral
         Featured image via Twitter"
         """))

         print("Given News Article Is: ",model.predict(valid)[0])
```

```
Given News Article Is:  fake
```

Fig : 7.11

```
In [27]: valid=count_vectorizer1.transform(pd.Series("""With little fanfare this fall, the New York developer who had planned to build an
         Those who had rallied in opposition to the building because of its religious affiliation back in 2010 were exultant. â€œThe impor
         Itâ€™s all well and good that so many Republicans have condemned Donald Trumpâ€™s reprehensible call for â€œa total and complete
         When he was president, George W. Bush honorably put a lid on right-wing Islamophobia. He regularly praised American Muslims and s
         Thus, Trumpâ€™s embrace of a religious test for entry to our country did not come out of nowhere. On the contrary, it simply brou
         You donâ€™t have to reach far back in time to see why Trump figured he had the ideological space for his Muslim ban. Last month,
         Sen. Ted Cruz (Tex.) took a similar view, saying , â€œThere is no meaningful risk of Christians committing acts of terror.â€•
         Trump took limits on Muslim access to our country to their logical â€” if un-American and odious â€” conclusion. Vice President E
         The demagoguery began with the labeling of the controversy itself. As PolitiFact pointed out, â€œthe proposed mosque is not at or
         This didnâ€™t stop opponents from going over the top, and Newt Gingrich deserved some kind of award for the most incendiary comme
         When President Obama defended the right of developers to build the project, he was â€” surprise, surprise â€” accused of being ou
         â€œI think it does speak to the lack of connection between the administration and Washington and folks inside the Beltway and mai
         At the time, John Feehery, the veteran Republican strategist, put his finger on why Republicans were so eager to lambaste Obamaâ€
         The Republican establishment is now all upset with Trump, but he is simply the revenge of a Republican base that took its leaders
         You canâ€™t be â€œjust a littleâ€• intolerant of Muslims, any more than you can be â€œjust a littleâ€• prejudiced against Catholi
         Read more from E.J. Dionneâ€™s archive, follow him on Twitter or subscribe to his updates on Facebook.
         """))
         print("Given News Article Is: ",model.predict(valid)[0])

Given News Article Is:  real

In [28]: valid=count_vectorizer1.transform(pd.Series("""Real Disclosure! Secret Alien Base Found In Moon's Tycho Crater # Grey 52
         Real Disclosure is where you find something on the lunar surface that cannot possibly exist unless someone built it. NO WAY it's
         print("Given News Article Is: ",model.predict(valid)[0])

Given News Article Is:  fake
```

Fig : 7.12

CHAPTER 8

CONCLUSION

With the increasing quality of social media, additional individuals consume news from social media rather than ancient fourth estate. social media has conjointly been accustomed unfold pretend news, that has sturdy negative impacts on individual users and broader society. We have a tendency to explore the pretend news drawback by reviewing existing literature in two phases characterization and detection. Within the characterization part, we have a tendency to introduced the essential ideas and principles of faux news in each ancient media and social media. Within the detection part, we have tendency to reviewed existing pretend news detection approaches from a knowledge mining perspective, together with feature extraction and model construction. We have tendency to conjointly more mentioned the datasets, analysis metrics, and promising future directions in pretend news detection analysis and expand the sphere to alternative applications.