Insider Trade Monitoring System for Corporate Entities using NLP and Machine Learning

A

Project Report

submitted in partial fulfillment of the

requirements for the award of the degree of

MASTER OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND ARTIFICIAL NEURAL NETWROK

by

Name	Roll No.
Ravi Teja Panga	**R102214012 (SAP:500041708)**

Under the guidance

Mr. Vinod Nair and Ms. Jyoti Jhurani

L&T Infotech ,Mumbai.

and Dr. Venkatadri Marriboyina

Centre of Information Technology, UPES , Dehradun

Department of Computer Science & Engineering

Centre for Information Technology

University of Petroleum & Energy Studies

Bidholi, Via Prem Nagar, Dehradun, UK

April – 2016

UPES COLLEGE OF ENGINEERING STUDIES

The innovation driven E-School

CANDIDATE'S DECLARATION

I hereby certify that the project work entitled "**Insider Trade Monitoring System for Corporate Entities using NLP and Machine Learning** " in partial fulfilment of the requirements for the award of the Degree of MASTER OF TECHNOLOGY in COMPUTER SCIENCE ENGINEERING with specialization in ARTIFICIAL INTELLIGENCE AND ARTIFICIAL NEURAL NETWORK and submitted to the Department of Computer Science & Engineering at Center for Information Technology, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my work carried out during a period from **January, 2016** to **April, 2016** under the supervision of **Dr. Venkatadri Marriboyina, Assistant Professor(SS).**

The matter presented in this project has not been submitted by me for the award of any other degree of this or any other University.

Ravi Teja Panga
Roll No.: R102214012
Sap ID: 500041708

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: 12 04 2016

Dr. Venkatadri Marriboyina
(Project Guide)

Dr. Amit Agarwal
Program Head – M. Tech (AI & ANN)
Center for Information Technology
University of Petroleum & Energy Studies
Dehradun – 248 007 (Uttarakhand)

ACKNOWLEDGEMENT

I wish to express my deep gratitude to my guides **Dr.Venkatadri Marriboyina** *Centre of Information Technology, UPES, Dehradun*, for all advice, encouragement and constant support they have given to me throughout my project work. This work would not have been possible without their support and valuable suggestions.

I am heartily thankful to my course coordinator, **Mr. Vishal Kaushik**, for the precise evaluation of the milestone activities during the project timeline and the qualitative and timely feedback towards the improvement of the project.

I sincerely thank to our respected **Dr. Amit Agarwal**, Program Head of the Department, for his great support in doing our project in **M. Tech Artificial Intelligence and Artificial Neural Networks** at **CIT**.

I am also grateful to **Dr. Manish Prateek, Associate Dean** and **Dr. Kamal Bansal** Dean CoES, UPES for giving me the necessary facilities to carry out my project work successfully.

I would like to thank all my **friends** for their help and constructive criticism during my project work. Finally, I have no words to express my sincere gratitude to my **parents** who have shown me this world and for every support they have given me.

Name	**Ravi Teja Panga**
Roll No.	**R102214012**
Sap ID	**500041708**

ABSTRACT

As the business world continues to enlarge in global markets, trading of shares, bonds, derivatives and other instruments continues to surge. One form of trading that has received considerable attention in recent years is Insider Trading. Usually arises when an individual with potential access to non-public information about a Corporation buys or sells equities of that company. However, if trading is done in a manner that does not take benefit of non-public information, it is often permitted. Insider Trading is an expression which has increased great currency in the financial markets in the last two decades. Concisely, it refers to a false practice which is resorted to by most of the corporate entities which are registered in a recognized stock exchange. Insider trading is considered as a global phenomenon which needs desperate attention and if unrestricted it would lead to numerous economic problems like increase in gap between the rich and the poor, stock market failures and economic down turns. This is an approach to identify such illegal activities in Corporate entities by surveillance the chats, emails and Audio calls (Transcript texts) using some state of the art Artificial Intelligence techniques. Implementation includes use of Natural language processing techniques to convert the unstructured data to reasonable format, web scraping to get unpublished information from websites and Machine Learning for classification.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page No**

LIST OF TABLES

1. INTRODUCTION

1.1 History

The primary 'insider trader' case was listed in the US way back in 1792. William Duer, was the Assistant Secretary in US Department of Treasury utilized his official position to collect insider knowledge and was involved in speculative trading at the time He was accused and spent his rest of the days behind bars.

In the initial 1920s JP Morgan helped as an unofficial central bank of the U.S and reportedly utilized its high power with the Republican Party to make incomes. Then the inevitable thing happened. In the year 1929, there was an extraordinary boom in the New York Stock Exchange and in September the selloff started dropping the prices to triggering a panic among investors.

The Great Depression was set in. After this the chaotic economic scenario, which saw that saw financial excesses made in the 1920s also decade long Great Depression and the shift in public view and all these added to the introduction of harsh laws on insider trading done to manipulate on profits.

Insider trading was not considered illegal, even till a few decades of the 20th century. In fact, in the year 1960s the world followed the Massachusetts Supreme Court ruling in the Goodwin v Agassiz incident that insider information is a 'perk.' Although, in 1964 the Securities Act Amendments were laid down disciplinary panels for brokers and dealers.

In India, the first set of rules for insider trading was announced in the year 1992 by the Securities and Exchange Board of India or SEBI. The rules restricted insiders and companies in trading securities through times of possession of unpublished price-sensitive information and striped them from distributing it with any other person outside the company as well.

SEBI has also approved disclosure norms for any directors or officers, who is holding shares in the company apart from amending the Model Code of Conduct in 2008 that limits the directors or officers who ever have bought or sold shares from getting into an opposite transaction by the next six months to sell and buy shares.

The penalty is very high for insider trading which is almost three times the amount of profits made out by insider trading issue to a minimum of Rs.25 crores. Further, SEBI is authorized to examine into insider trading irregularities and also initiate criminal proceedings if found guilty.

One most famous insider trade case was registered in India in Satyam Computer Services Ltd. On December 21st, 2012 Capital market regulator SEBI imposed a penalty for alleged violations of Insider trading norms. Total penalty was around 65 lakhs on a senior official for allegedly involving in Insider trading, while he was in possession of UPSI (unpublished price sensitive information). This famous scandal was registered at the time of acquisition of Maytas Properties Ltd and Maytas Infra Ltd by Satyam Computer Services.

1.2 Requirement analysis

Prohibition of insider trading is tremendously huge and incredibly complex with the number of parameters involved such as communication could be verbal(Telephonic), email, documents and word of mouth. The primary thing, that can be done is to monitor trading activities in legal communication interfaces (Lync, Outlook, Phones), within a company.

The requirement of the company is to build a proof of concept for prohibition of insider trading. Which is end to end proof with a specific design and proof. I get input data in terms of a chat history of an employee of the company, transcripts of telephone calls within company telephone network and documents shared through email.

Chat data can be expected to be unstructured and could contain lot of irrelevant data. Dealing with unstructured data will be very hard than dealing with structured content, where pattern recognition is easier.

1.3 Main Objective

To identify the following stated rules as per Stock Exchange Board of India (SEBI, 2015)

Regulation 3.1 and 3.2 (Chapter 2), Unpublished Price Sensitive Information (UPSI)

- No Insider shall communicate, provide or allow access to any UPSI.
- No person shall procure from or cause the communication by any insider of UPSI

Regulation 4.1 (Chapter 2)

- No insider shall trade in securities that are listed or proposed to be listed on a stock exchange when in possession of UPSI. (SEBI, PROHIBITION OF INSIDER TRADING, 1992)

1.4 Sub Objectives

1.4.1 Unstructured data

Unstructured data doesn't follow certain format. It is said that only 20 percent of the data is structured in corporate entities, whereas remaining 80 percent data is Unstructured. Normally both Structured and Unstructured could be from Human or Machine. In this case most of the unstructured data can be expected from Humans. There is high chance that data generated from Machine could be structured. Human generated data. The most common form of human generated data is from Social media content, website content and all Company data such as survey results, confidential documents and e-mails.

Since input is unstructured and huge. I have to reduce the redundancy in order to improve the time complexity. Certain NLP techniques are useful in dealing with unstructured data such as Tokenization, Sentence segmentation, Stop words reduction, POS tagging, Chunking. based on the accuracy and time.

1.4.2 Insider Pattern

Insider's communication pattern is vital for classification. In order to understand the pattern, I met few Trade analysts, a Trade Broker and others. I found certain colloquial expressions that traders tend to use, which are incorporated in the model. Some key characters and words that makes sure that Unpublished price sensitive information has been shared.

Certain key words like Financial Year (FY), Percentage (%) and Increase or decrease by certain % like (+2.10, -3.00). Apart from these specific letters there is importance to certain verbs like increase, growth, raise, etc.

1.4.3 Trader pattern

Traders have a different pattern, such as no. of stocks, price. As both are not mutually exclusive, they share a band of common patterns. Apart from the Intersection part, I have found specific keywords, which are exclusive for the traders.

These expressions carry a special meaning in trading, which can be captured to identify the trade chat.

- It feels good
- Risk to reward ratio
- Price earnings ratio
- Company has good Prospects
- Net Outflow looks great

2 RELATED WORK

In most of the cases, courts have taken the position that employers have the right to monitor what employees do on the employer's computer systems and equipment. To start, that means boss can see any messages you send using your domain email. But that is not all, when you send an email from your work place, the company server will not know or even care whether that email is on your company email account or your personal Microsoft account it monitors almost everything, said Lewis Maltby, the president of National Work Rights Institute and In fact that's completely legal.

One gray area: A recent study in National Labor Relations Board ruling also found that employees have a presumptive right in order to use their employee email system for union organizing, although labor laws limit employers from observation of union organizing activities (N.L.R.B, 2014). That means the NLRB may ultimately conclude that employers would not be able to monitor emails related to union organizing, even if they were sent using the employee server or equipment.

In the year 2007, a survey by the American Management Association also found that 28% of employers were fired employees over "e-mail misuse." The most common kind of abuse: violation of company rules, inappropriate language, excessive personal use, or even breaking privacy. ("Internet misuse" was even more common; another 30% of employers said they had terminated employees for excessive personal use of the Internet at work, viewing inappropriate content at work, or other violations of the employer's electronic use policy.

When an insider trading activity is being reported, Compliance officer should take hold of everything and continue the process of verification. But unless someone is found suspicious, there is no chance of identifying an Insider trading activity. Even after finding someone suspicious, one has to go through enormous number of records in order to ensure illegal activity.

Worldwide email usage continues to grow at a healthy pace. In 2015, the number of worldwide email users will be nearly 2.6 billion (Radicati, 2015). While roughly 196.3 billion emails were being sent and received each day in the year 2014, the figure is to be expected to increase to 236.5 billion mails by the end of the decade. Although there are many software's for monitoring emails. They are not looking for a specific pattern and that are limited to only emails. The most widely used communication interfaces are office communicators such as Microsoft Lync.

Daily Email Traffic	2015	2016	2017	2018	2019
Total Worldwide Emails Sent/Received Per Day (B)	205.6	215.3	225.3	235.6	246.5
% Growth		*5%*	*5%*	*5%*	*5%*
Business Emails Sent/Received Per Day (B)	112.5	116.4	120.4	124.5	128.8
% Growth		*3%*	*3%*	*3%*	*3%*
Consumer Emails Sent/Received Per Day (B)	93.1	98.9	104.9	111.1	117.7
% Growth		*6%*	*6%*	*6%*	*6%*

Figure 2.1 Worldwide Daily Email traffic by The Radicati Group

A recent study from the Capital Markets Cooperative Research Centre (CMCRC) found that exchange potential to mitigate insider trading, but they also can boost the apparent profits of insider schemes. 1-standard-deviation improvement in trading rule specificity gives rise to a 23.43% reduction among the number of insider trading cases and a 53.17% surge in profits per case, on average. Likewise, it conservatively estimates that a one standard deviation development in surveillance will gives rise to a 67.0% decrease in the number of cases and 26.3% rise in profits per every case (Zhan, 2014).

India has put her efforts and has made a move towards the enactment of the new Insider Trading Regulations with a view to align its laws on Insider Trading with that of the developed countries. This would effectively help in combating Insider Trading to a very large extent. SEBI in order to modify the law on Insider Trading and ensure that it is in consonance with the global best practices, constituted a high level committee under the Chairmanship of Justice N.K.Sodhi which drafted the Prohibition (Lee Biggerstaff, 2012) of Insider Trading Regulations, 2015.

Text analytics spans across virtually all verticals. Frequent surge in text analytics use cases, such as in finance, insurance, media, and retail industries, but even oil and gas companies can derive value from text analytics. A typical text-analytics application in the finance industry focuses on compliance and fraud prevention. For example, Dodd-Frank states that all electronic communications at financial institutions—email, chats and instant messages—need to be monitored to reduce the risk of market manipulation, fraudulent account activities, anti-trust/collusion, outside business activities, illegal political contributions, and sharing sensitive customer information. The purpose of natural language processing in this use-case is to understand

the content of communication threads through semantic interpretation, and to identify relationships and entities across threads (e.g., Analyst Joe claimed the stock Enron is about to take off). Text analytics, however, is responsible for determining whether a given message, or set of messages, breaks compliance. Compliance departments benefit from combining structured data, like trades and transactions, alongside the information extracted from emails and instant messages. With both types of data assets, it is then possible to infer the intent behind a transaction.

Financial institutions face another fundamental compliance problem—anti-money laundering. Financial institutions are obligated to screen all transactions across the entirety of their business units for the purpose of preventing transactions between blacklisted parties. This task involves the analysis of free text contained within the transaction (e.g., the specified purpose of the transaction) and matching names and entities against watch lists from the Office of Foreign Assets Control (OFAC) and other governmental agencies. One important task is matching transliterated names to 'one' representation on a list. For example, the name Alexander can be transliterated to Aleksandr, Alex, or Alexandr, etc.). The match against multiple lists must be very precise as analysts can only manually review a small percentage of alerts.

In the insurance sector, insurance companies have large collections of unstructured call center, claim, billing, and adjuster notes text data. To get a better understanding of policyholders, these companies can utilize sentiment analysis to gauge if their customers are satisfied or dissatisfied with their products, services, and processes. Text analytics can identify problem areas with the products and procedures, and it can provide guidance for improving services or developing new products.

In the legal space, law firms collect millions of unstructured documents consisting of emails, case files, court documents and health records to name a few. Such collections of documents can be used to signal potential new class action suits through the identification of coherent subsets of documents relating to a particular subject or interest area. There is also growing interest in estimating juror voting propensities from their social media profiles. This is of particular value during jury selection to assemble the most favorable jury for a particular client or case.

As the recently proposed regulatory actions focused on communication flow between individuals, corporate entities are looking for less costly compliance risk identification techniques and more effective that surpass looking at trade data only. Some well-known regulators like the SEBI, SEC, ASIC, other policing organizations and internal compliance functions have historically focused

their analytics resources developing sophisticated ways to monitor and detect unusual trading patterns. This focus on trading activity is an important approach to locating suspicious transactions; however, the problem for the investigator often lies in identifying trades that may have been based on actual material non-public or sensitive information.

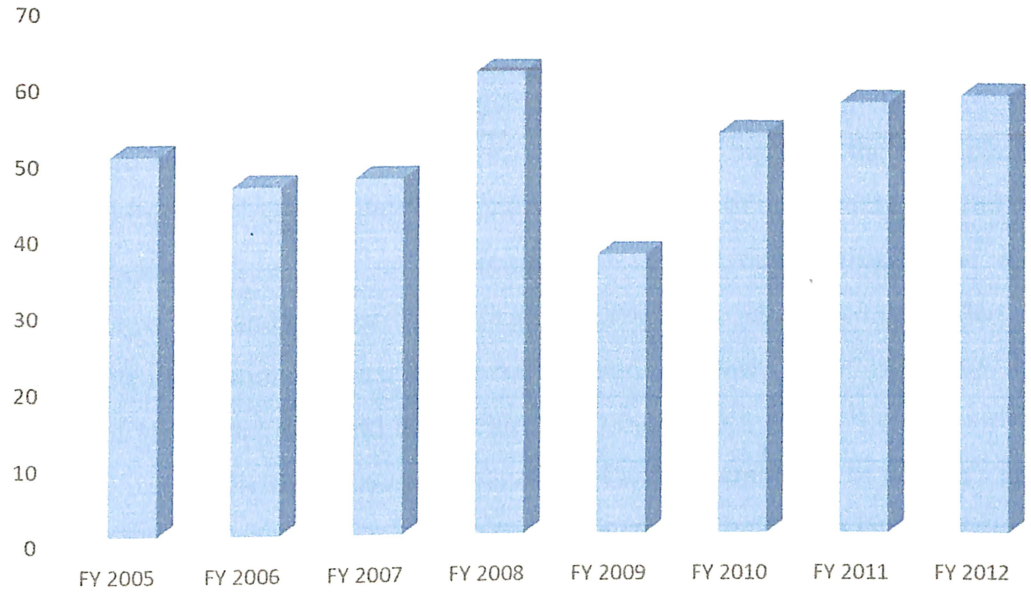

Figure 2.2 Insider trading Enforcement Actions

3. PROPOSED SYSTEM

Major growth in the volume and variety of data is because of the accumulation of unstructured text data—in fact, up to 9o% of the data is unstructured text data. Companies collect massive amounts of documents, emails, social media, and other text-based information to get to know about their customers better, to offer customized services, to comply with federal regulations and in my case insider trade monitoring. However, most of this data is unused and untouched.

Text analytics, with the use of natural language processing (NLP), holds the key to unlocking the business value within these vast data assets. In the period of big data, the right platform enables businesses to fully exploit their data lake and take advantage of the modern parallel text analytics and NLP algorithms. In such an environment, text analytics facilitates the integration of unstructured text data and structured data (e.g., customer transaction records) to derive deeper and more complete illustrations of business operations and customers (Acar Tamersoy, 2015).

Integration of structured data and unstructured text data into a single file and unified analytical environment, can facilitate the operationalization of a new generation of business improvements very quickly. Natural language processing and text analytics, outlined a large number of use cases where these are applied today. Some Studies pointed out scenarios where combining structured and unstructured analytics capabilities can deliver more powerful alerts, greater insight for business decisions, and also new types of process automation.

Text analytics refers to the extraction of useful information from text sources. It is a broad term that defines tasks from annotating text sources along with meta-information such as people and places mentioned in the text to a wide variety of models about the documents (e.g., text clustering, sentiment analysis, and categorization of text). To Illustrate, the term document is an abstract notion that can denote any coherent piece of text in a larger pool such as a single blog post in a collection of WordPress posts, a Times of India article, a page related to company investments, among others.

Proposed system uses NLP techniques to deal with the unstructured data. Then the processed data is given to Machine learning algorithm to classify. In order to help the classifier, we get the generally available information from stock websites.

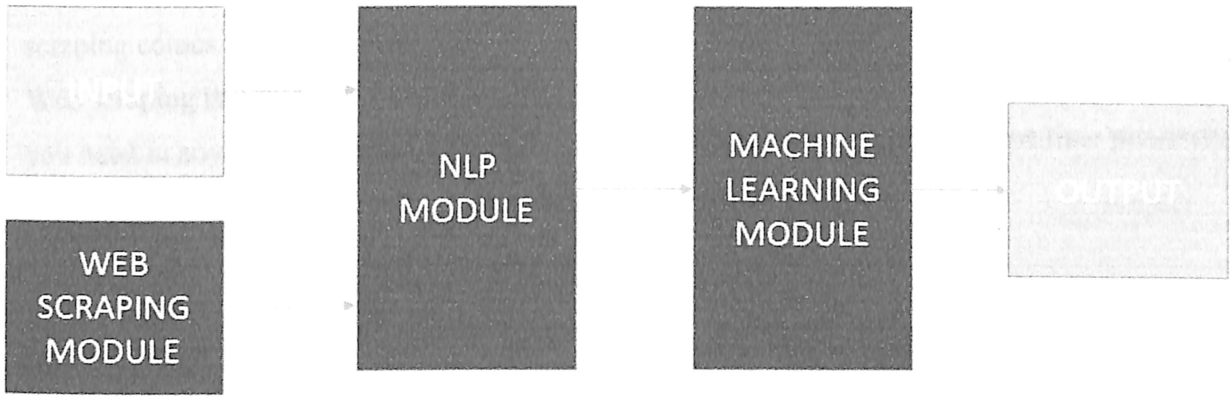

Figure 3.1 Overview of the proposed system.

There are three primary modules in the proposed system. Input is given to the Natural Language Processing Module, where we perform certain operations the unstructured input data. Later on one set of input is given from the web scraping module as well. Almost similar operations are performed for the scrapped data. Machine learning module is loaded with set of classifiers, which in turn helps in filtering the un published price sensitive information.

3.1 Modules

3.1.1 Natural Language Processing Module

Natural Language Processing is a subject of artificial intelligence concerned with making natural languages accessible to machines. Some tasks such as identifying sentence boundaries in text files, extracting relationships from document files, and searching and retrieving of documents, among others. NLP is most widely used to simplify text analytics by establishing structure in unstructured text to enable further analysis.

There are so many techniques like Tokenization, Sentence segmentation, Stemming/ Lemmatization, Part-of-Speech tagging, Parsing, Named entity recognition and Co-reference resolution that are widely used.

3.1.2 Web scraping module

Since web sites are written in HTML, we could also say each web page is a structured document. Which would be great to obtain some data from them and preserve the structure of the website.

Normally, web sites don't always provide their data in accessible formats such as csv or json. Web scraping comes into handy here.

Web scraping is the practice of using a program to go through a web page and extract the data that you need in any format that is most useful to you. In addition to that, at the same time preserving the structure of the data.

3.1.3 Machine Learning module

Machine learning is the science of getting programs mimic human behavior of learning ability. In the past decade or so, machine learning has given us some miraculous research projects and aided in capital of billion dollars. The most widely used areas include Image recognition, Speech recognition and Predictive analytics. Natural language processing benefited from machine learning vastly, getting us closer to one of the primary goal of artificial intelligence. In this project, I have used some machine learning classifier algorithms to make predictions on insider trading patterns.

4 DESIGN

Design of the proposed system contains three important modules. One is NLP module used for text analytics to handle the unstructured text data. Second module is web scraping module, used for scraping required text from the websites. The final module is Machine learning, used for classification of text and identifying whether published information is present or not. Following is the design.

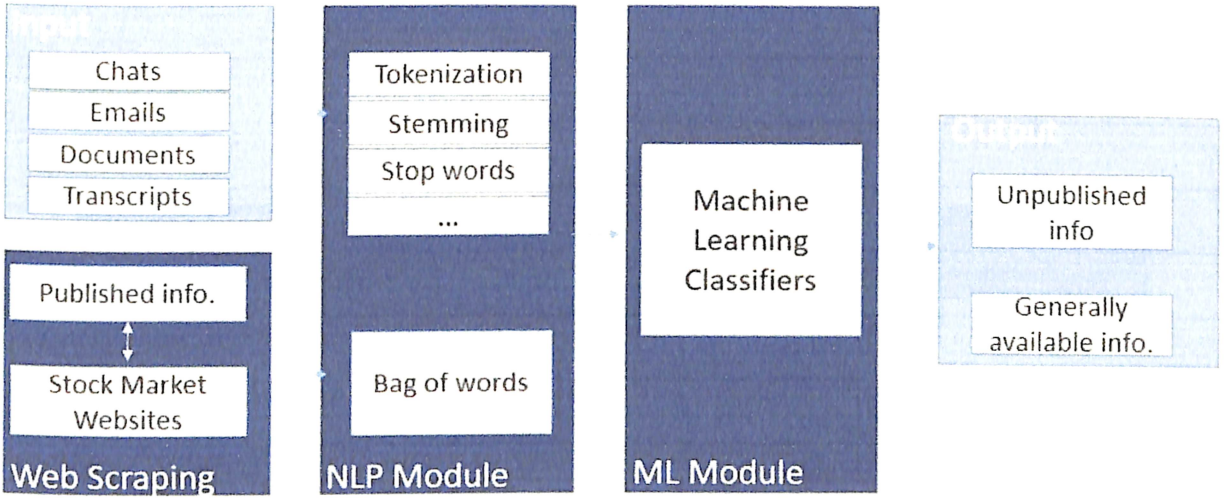

Figure 4.1 Design and Architecture

4.1 Input

Input could be of three forms (Chats, Emails & Transcripts) which are then converted to Excel or CSV. If Excel, Columns would be Date, Time, Name and One for text. NLP techniques are used on the text column. Data and time are very important in identifying insider trading. If the any information is present before the day and Until it gets published, it can't be considered as Unpublished info.

4.2 Webscraping

In order to extract generally available information from stock websites, I should scrape websites. I have used Scrapy- Open source framework in python. There is another popular way to use Beautiful Soup + Requests to scrape websites. Requests is a Python Module for HTTP library, written in Python.

Scrapy is an Open Source Web crawling framework developed in python. Scrapy tries to factor out the general things, that are required to write web scrapers. It also separates them from the extraction rules or XPath and CSS. This is where Scrapy comes handier than adhoc solutions like using Beautiful soup and Requests.

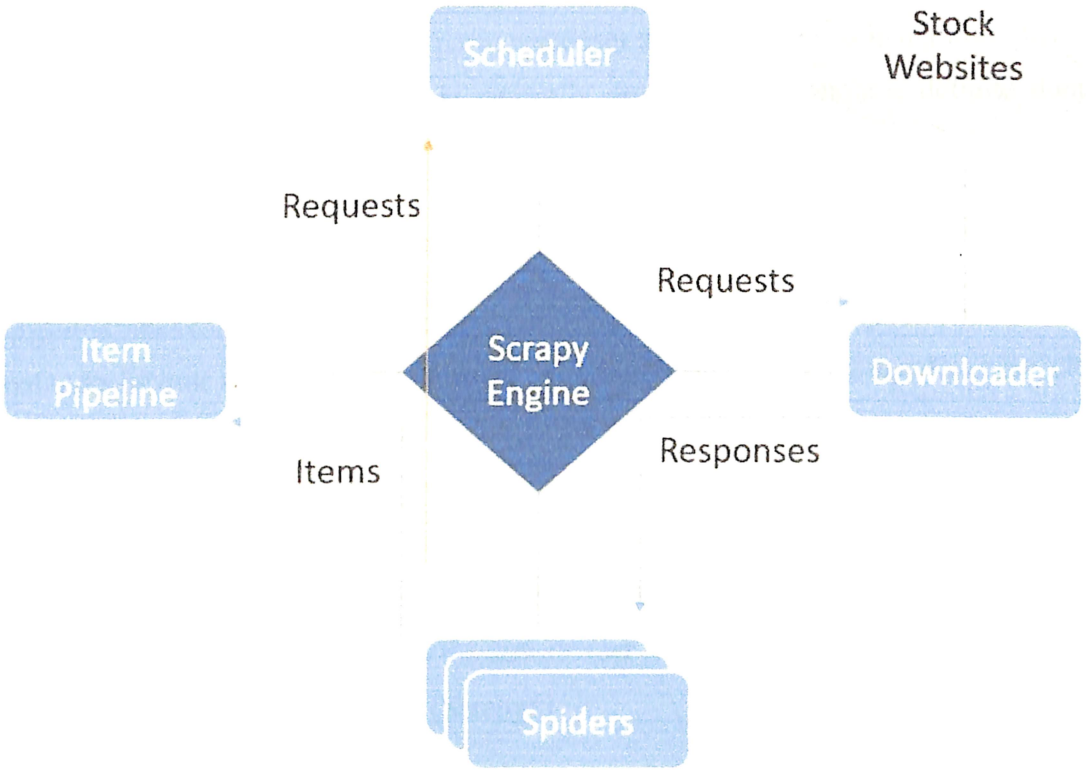

Figure 4.2 Overview of scraping module architecture

4.2.1 Scrapy engine

The core part is the engine, which controls the data flow among all the components of the system. It triggers all the events for specific actions.

4.2.2 Scheduler

Receiving requests from the engine and queuing them for feeding later is the responsibility of scheduler. This also includes feeding the engine also.

4.2.3 Downloader

The responsibility of downloader is to download the webpages. Once the requested websites are being downloaded, these websites are fed to the engine. Finally, engine in turn feeds those downloaded websites to the spiders.

4.2.4 Spiders

Spider is a custom class, developed to parse responses and mine scraped items from them. There could be multiple spiders. A spider is generally developed to handle a definite domain or sometimes group of domains.

4.2.5 Item Pipeline

Item pipeline takes care of processing all the items as soon as they are being scraped by the spiders. Normal tasks include checking for duplicates and dropping them, storing in a data base, cleansing HTML data and validating the scraped data.

4.2.6 Downloader middlewares

These lie between the downloader and the engine. The main task is to process the requests whenever requests are being passed to the downloader from the engine. Downloader middleware also provides plugging custom mode. (Scrapy)

4.2.7 Spider middlewares

These lie between the Spiders and the Engine. The main task is to process spider 's input responses and items, requests. Spider middleware also provides plugging custom mode.

4.3 NLP Techniques

4.3.1 Sentence Segmentation

Sentence segmentation is a process of identifying where one sentence ends and another begins. Punctuation often marks sentence boundaries, but as the example in the below shows, there are many exceptions in the usage of language. The construct would be: "Hi! What's up—Mr.

President?" can be viewed as a single sentence, although full stop is given after Mr.

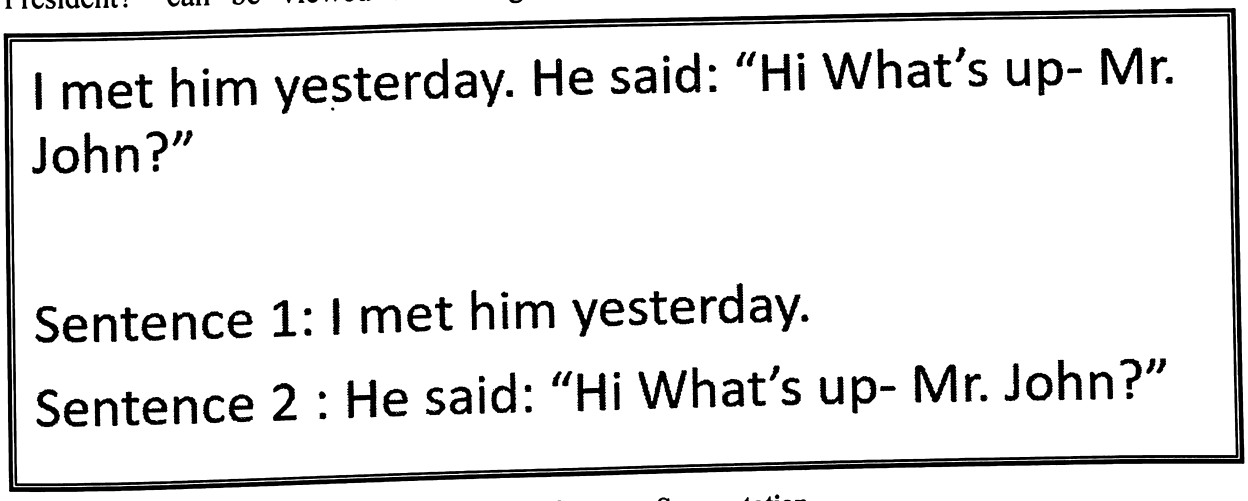

Figure 4.3 Sentence Segmentation

4.3.2 Tokenization

Tokenization is the process of isolating individual numbers, words, and other single coherent constructs. For example, Hashtags in Twitter feeds of constructs containing of alphanumeric and special characters that should be treated as one individual token. In languages such as Japanese and Chinese do not precisely delimit individual words in sentences, complicating the task of tokenization. In personal chats, it is not easily possible to identify words, as it is not structured like in text books.

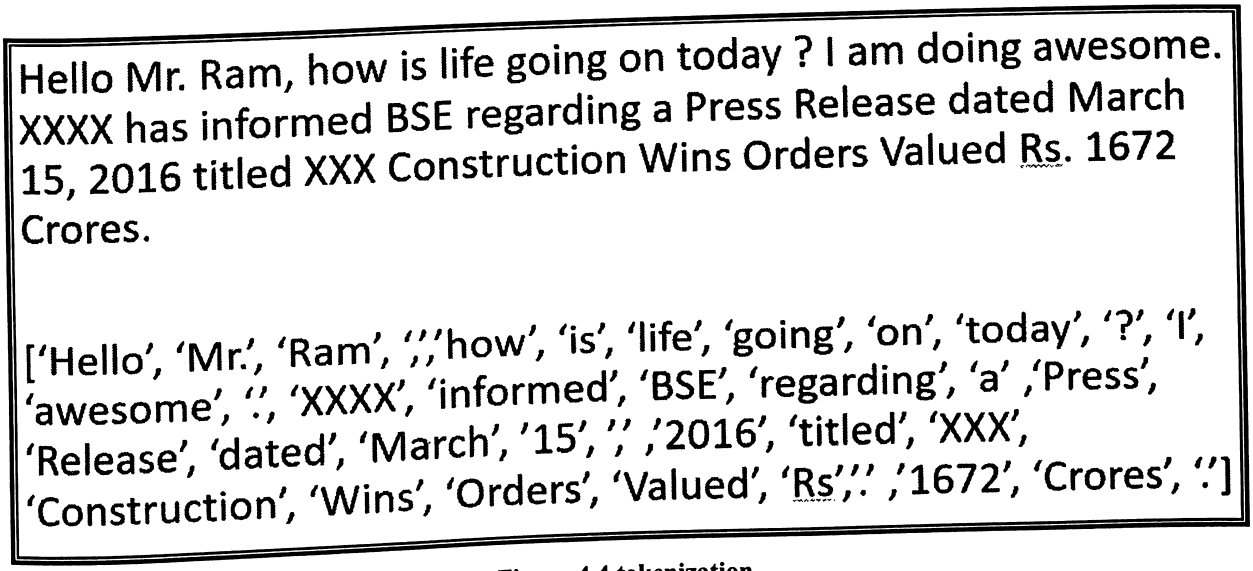

Figure 4.4 tokenization

4.3.3 Stop words

Stop words are the words that don't carry any meaning to the text. Removing stop words from a text reduces all the irrelevant, which in turn helps n concentrating the data that is important and impacting.

Stop words are generally used in sentiment analysis, spam email classification.

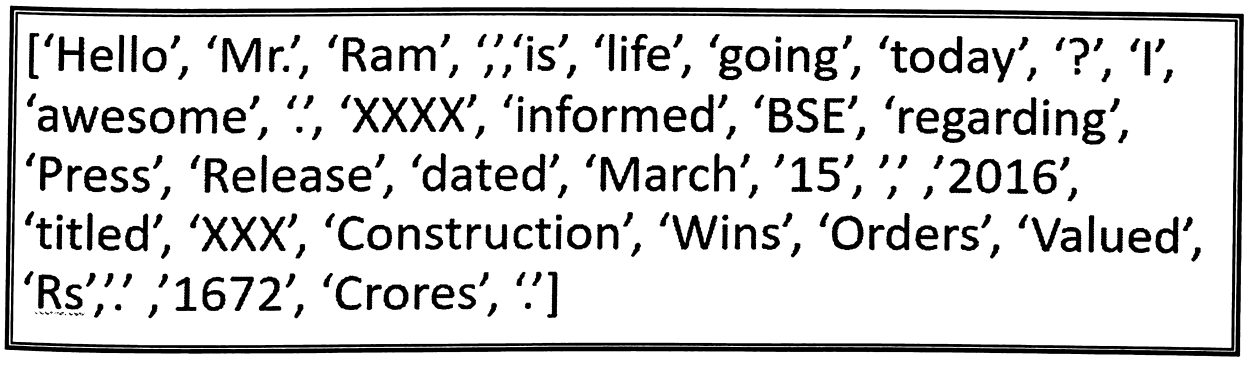

Figure 4.5 Stop words elimination on the same text

4.3.4 Stemming

Stemming stems out the 'ending' of words. Stemming is used to reduce the certain repetitive words like increasing, increase, increased, etc. Instead taking everything as different words, we can stem out to the least possible word. The stem of the word still conveys the meaning.

Stemming is widely use when data has multiple words with tenses and you are not worried about whether the sentence is in past tense or present tense and is also considered as a special case of Normalization.

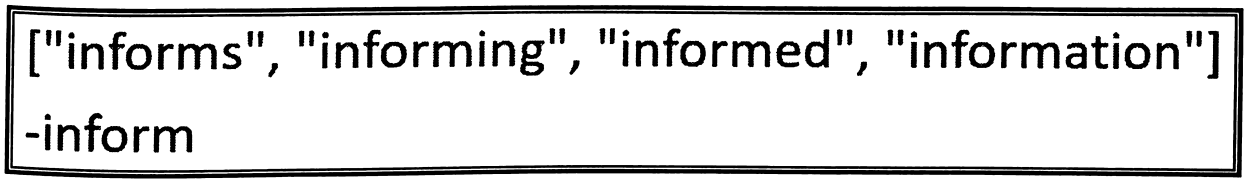

Figure 4.6 Stemming

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Figure 4.7 Comparison of three Stemming algorithms

4.3.5 Lemmatization

Sometimes, stemming alone won't work. By just chopping off the end part, we can never expect the root word. Lemmatization usually refers doing things accurately with usage of vocabulary and also morphological analysis of words. Lemmatization often gives the dictionary root word, which is known as lemma.

4.3.6 POS Tagging

POS tagging is known as Part of speech tagging, used to tag part of speech of every word. POS tagging has remarkable use when we are searching for a particular structure in a file. (Kristina Toutanova)

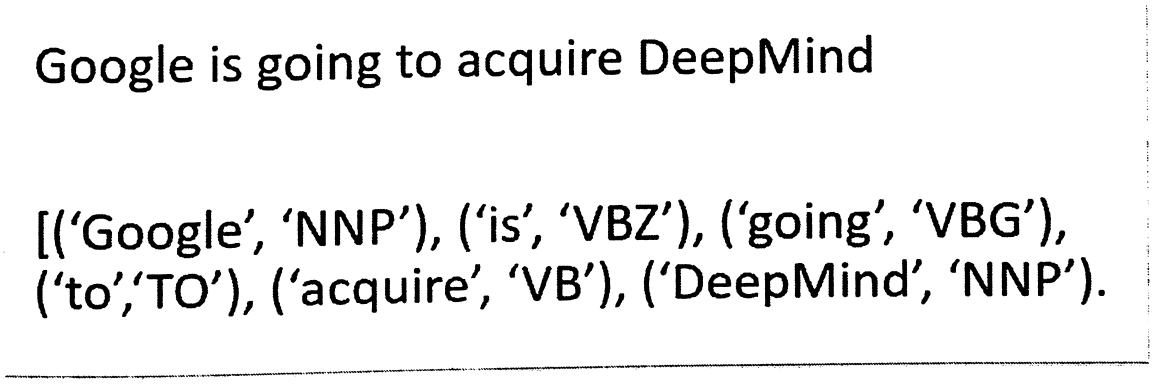

Figure 4.8 POS tagging

Tag	Description	Tag	Description	Tag	Description
CC	Coordinating conjunction	JJS	Adjective, superlative	POS	Possessive ending
CD	Cardinal number	LS	List item marker	PRP	Personal pronoun
DT	Determiner	MD	Modal	RB	Adverb
EX	Existential there	NN	Noun, singular or mass	RBR	Adverb, comparative
FW	Foreign word	NNS	Noun, plural	RBS	Adverb, superlative
IN	Preposition	NNP	Proper noun, singular	VB	Verb, base form
JJ	Adjective	NNPS	Proper noun, plural	VBD	Verb, past tense
VBP	Verb, Non – 3rd person singular	VBZ	Verb, Non – 3rd person singular present	WP	Wh- pronoun
JJR	Adjective, comparative	PDT	Pre determiner	VBN	Verb, past particle

Table 4.1 POS tags and their Descriptions

4.3.7 Chunking

Chunking the text is, chunking certain pattern in terms of Part of speech tags. Chunking is also considered as grouping of words. Chunking is helps when we are looking for a group of fixed patterns. For example, Noun phrase followed by Verb Phrase and ending again with Adjective. This is given as a regular expression in the code to look for.

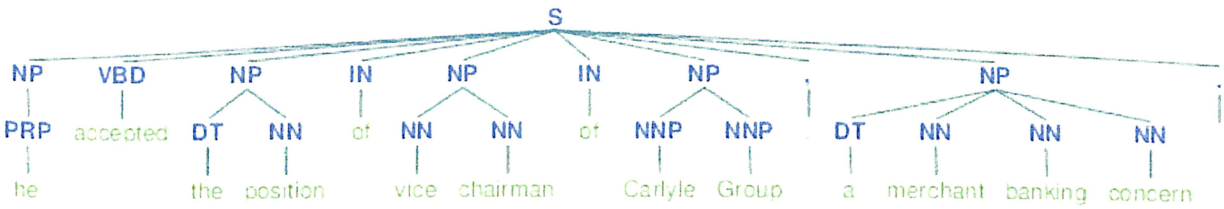

Figure 4.9 Chunking

4.3.8 WordNet

WordNet is considered as global lexical database for English language. WordNet helps in providing syntactic meaning of words. Synonyms and Antonyms of a word could be a great help, when someone tries to trade using different words and transfer messages.

Similarity measures of words would be of great help in order to identify the semantic similarity of a word. WordNet offers 3 measures of relatedness and 6 measures of similarity, all these measures are calculated with the help of lexical database of WordNet. All these measures takes in two notions as input and gives a numeric value as output. The generated numeric value represents the degree to how similar or related those two concepts. (Ted Pedersen, 2004)

In WordNet version 2.0, there were nearly 600 verb hierarchies and even nine noun hierarchies which included 80,000 concepts together were added up to 13,500 concepts. (Banerjee, 2003)

In the paper published by Jay J. Jiang David W. Conrath in the year 1991. Proposed a new approach for calculation of similarity between the concepts. The proposed system combines a corpus statistics information with lexical taxonomy structure, the idea is that to better quantify the computation derived using distributional analysis of corpus data of the semantic distance between the nodes in a semantic space, which is constructed by the taxonomy.

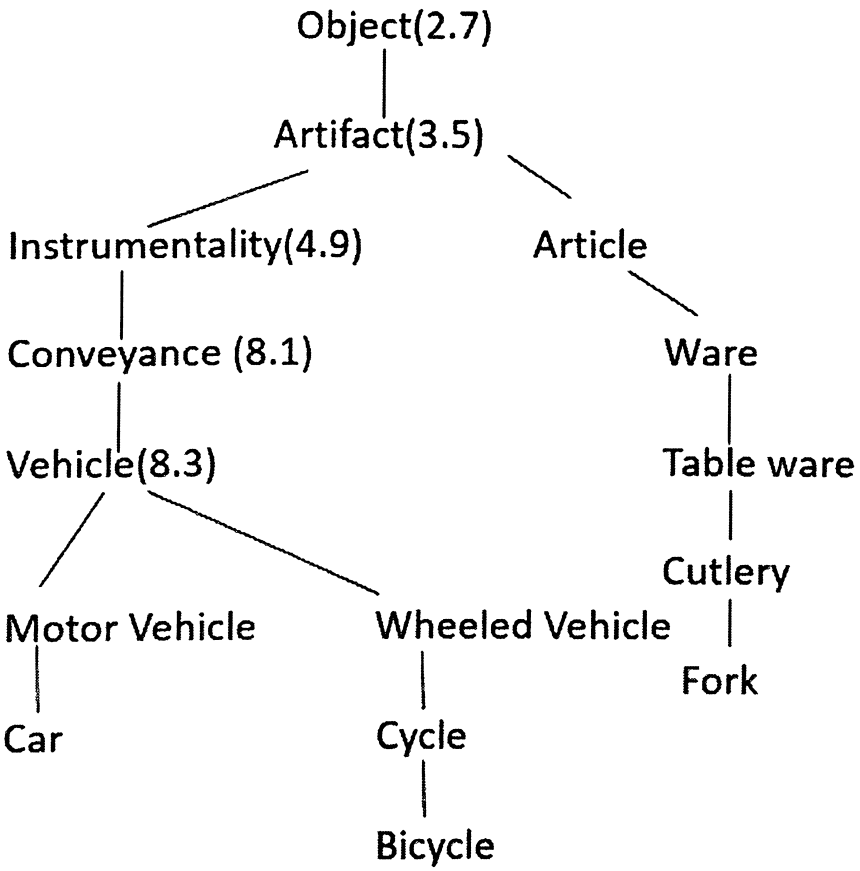

Figure 4.10 Fragments of WordNet Noun Taxonomy

4.4 Machine Learning Module

Machine Learning is often said to be evolved from computational learning theory and pattern recognition in the field of Artificial Intelligence. Machine learning's basic goal is to automate analytical model building, which is achieved by algorithms that iteratively learns from input data without being explicitly programmed. (Leacock, 1998)

4.4.1 Text Classification

The idea of classification is very broad and has many applications within and past the information retrieval (IR). For example, in the field of computer vision, a classifier may be used to divide images into classes like human presence or absence in an image and whether image contains

vegetation or not, if so how much? Is it very high, normal or very high? My focus was here on information retrieval such as classifying documents, chats, etc.

Apart from manual classification from the Human employees and hand-crafted rules, there is a third methodology to text classification, like machine learning-based text classifications. By incorporating machine learning module, both set of rules and the decision principle of the text classifier, can be taught automatically from training data. Normally, this kind approach is also known as *statistical text classification* since the learning method is statistical. In statistical text classification, all we need are well classified example documents (aka training documents) for every class. The need for manual classification cannot be eliminated, as all the training documents has to come from a human expert, who has labeled them – where by *labeling,* it refers to the process of marking each document with its class. However, labeling task is an easier job than writing rules.

There are many popular machine learning text classification algorithms that can be used to classify the text. Since my data is unstructured and more unexpected patterns. I am selecting few algorithms and check the accuracy to make sure which would do the best.

I chose the trained classifier whichever proved to be the best at that point of time for that particular data. Then I choose best classifier for classification process. Some popular text classification algorithms are Naïve Bayes classification, Support Vector Machines, Logistic Regression etc.,

The accuracy method is mentioned as the confidence measure of an algorithm, on one particular data set.

The below figure is the accuracy rate calculated with the increase in number of words. The below stated result is the part of research work by Brill and Banko on spelling correction, Microsoft. Naïve bayes is proved to the best for the kind of problem that am focusing on. (Brill, 2001)

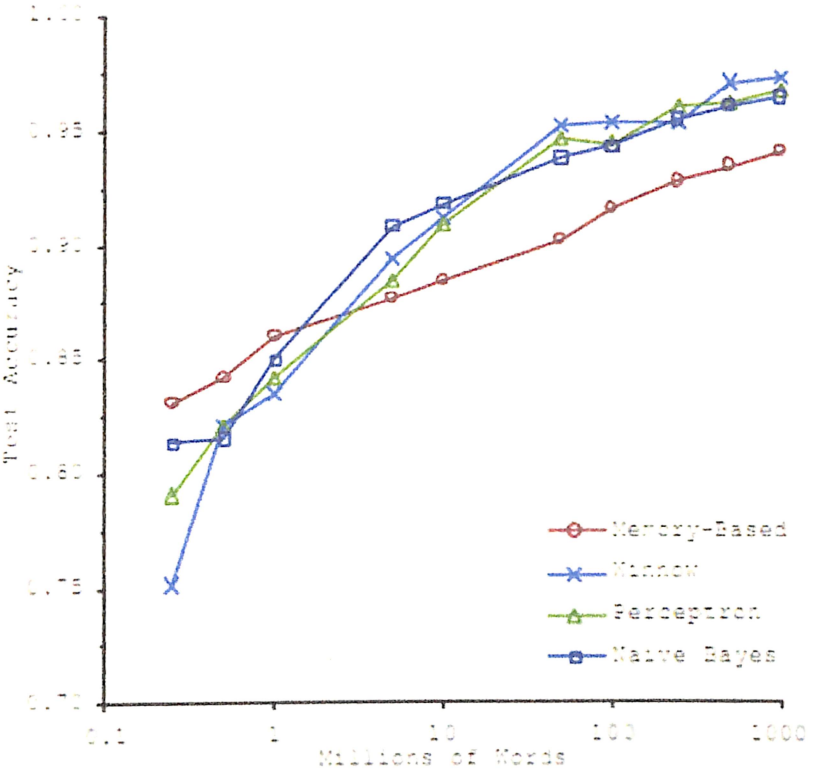

Figure 4.11 Brill and Banko on spelling correction

4.4.2 Text corpus

Text corpora is collection of a wide variety of data from various domains. Collection of a corpus from a wide variety of data helps in diversity of data and increases the learning ability of a machine learning algorithm.

Text corpora is generally used to as training and testing data to check how good an algorithm works. The accuracy percentage on a corpus labeled data would tell about the goodness measure of the algorithm we choose.

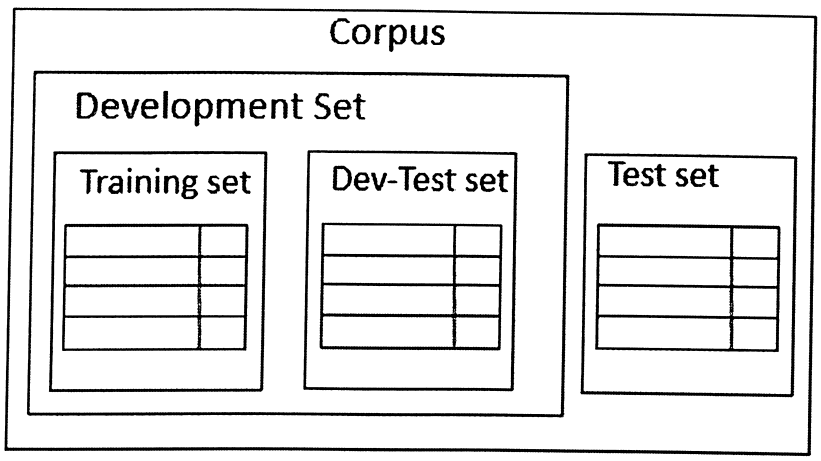

Figure 4.12 Organization of corpus data for training supervised classifier

4.4.3 Bag of words

Bag of words representation is a kind of simplifying way of representation used Classification task in Natural Language Processing and Information retrieval. Bag of words can also be treated as Labelled data. Bag words is given in a text file and based on the that file, Classifier algorithm classifies the text. In some cases, Bag of words is given with score alongside of each word.

4.5 Output

Output of the proposed system is

- Identifying the people who shared insider trade information from the given input file
- Identifying whether trade related talks was discussed or not.

Output will be given in the form of chat or file -Id or most of the cases the employee name with the suspicious text highlighted.

5 Modeling

5.1 Generally available Information

Generally available information is scraped from the stock market website would look like

ID	DATE	TIME	L&T Publsihed Information
			Larsen & Toubro Ltd has informed BSE regarding a Press Release dated March 15, 2016 titled "L&T Construction Wins Orders
	15/03/16	10:17	Valued Rs 1672 Crores"
	04/03/16	18.27	Larsen & Toubro Ltd has informed BSE regarding "Sale of Casting Manufacturing Unit Located At Coimbatore To Bradken Limited"
			Larsen & Toubro Ltd has informed BSE that Ms Naina Lal Kidwai has been appointed as an Independent Non-Executive Director
	01/03/16	10.26	of the Company, w.e.f. March 01, 2016.
			Larsen & Toubro Ltd has informed BSE that Ms. Naina Lal Kidwai has been appointed as an Independent Non-Executive Director
	01/03/16	9:35	of the Company, w e.f. March 01, 2016.
			Larsen & Toubro Ltd has informed BSE regarding a Press Release dated February 25, 2016 titled "L&T Hydrocarbon enters into
	25/02/16	11 38	Long-Term Agreement with McDermott for Emerging Deepwater Market in India".
			Larsen & Toubro Ltd has informed BSE regarding a Press Release dated February 19, 2016 titled "L&T Construction Wins Orders
	19/02/16	11.19	Valued Rs.1404 Crores".
			With reference to news reported - in The Economic Times on February 15, 2016 : "L&T eyes Rs 1 5 L cr Worth of orders from Navy
	15/02/16	14.56	in 4 yrs", Larsen & Toubro Ltd has submitted to BSE a copy of Clarification is enclosed.
			The Exchange has sought clarification from Larsen & Toubro Ltd February 15, 2016 with reference to news reported - in The
	15/02/16	11:52	Economic Times on February 15, 2016 . "L&T eyes Rs.1.5 L cr Worth of orders from Navy in 4 yrs"
			Larsen & Toubro Ltd has informed BSE about :
	06/02/16	13:46	Result Presentation for the period ended December 31, 2015
			Larsen & Toubro Ltd has informed BSE that the listed, Secured Non-Convertible Debentures of the Company aggregating to Rs
			400 crore issued on January 2009 and outstanding as on December 31, 2015 are secured by way of first mortgage/charge on the
	6/2/2016	13:01	Company's various properties and the asset cover thereof exceeds hundred percent of the principal amount of the said debentures
	03/02/16	16:02	Larsen & Toubro Ltd has informed BSE regarding "Interest and Redemption amount payment".
	02/02/16	10:00	Larsen & Toubro Ltd has informed BSE regarding "Schedule of Analyst / Institutional Investor Meet"
			Larsen & Toubro Ltd has informed BSE that Mr. M. V. Satish has been appointed as a Whole-time Director of the Company w.e.f.
	29/01/16	17:45	January 29, 2016 for a period of 5 years.

Figure 5.1 General available information scraped from a stock website

5.2 Algorithms

Text Classification algorithms:

Supervised Machine Learning algorithms are chosen as it is possible to train with the labeled information. Among many classification algorithms that falls under supervised category. For now, I chose to test on 3 algorithms as stated below.

Input:

document $\boldsymbol{d}$

set of classes $\boldsymbol{C} = \{ c_1, c_2, \ldots, c_J \}$

training set $\boldsymbol{m}$ *labeled documents* $(d_1,c_1),\ldots,(d_m, c_m)$

Output:

a learned classifier $\gamma: d \rightarrow c$

5.2.1 Bayes Classification

Thomas Bayes stated Bayes theorem in probability theory in the 18th century. Since then, we have been finding it incredible useful not only in the field of Computer Science but also in various diverse fields. In Philosophy, it is known for clarifying the relationship among Evidence and Theory. From Machine learning perspective Bayes Theorem was widely used in classification such spam/Non Spam email filtering systems. Bayes rule for a document and a class is introduced below.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Where A and B are events.

P(A) and P(B) are the

Where 'd' is a document, 'c' is a class

5.2.1.1 Naïve Bayes classifier

MAP is 'maximum posteriori is the most likely class. Argmax function is being used instead of max function. Where max of f(x) gives the maximum of x, Argmax function gives in terms of f. Argmax(x) tells what is the input what is the maximum of a function. Argmax gives which input gives us that maximum.

$$C_{MAP} = \text{argmax}_{c \epsilon C}\ P(c|d) \quad (2)$$

As per Bayes rule,

$$C_{MAP} = \text{argmax}_{c \epsilon C}\ \frac{P(d|c)P(c)}{P(d)}$$

Dropping the denominator as it is constant,

$$C_{MAP} = \text{argmax}_{c \epsilon C}\ P(d|c)P(c) \quad (3)$$

Document d represented as features $x_1, x_2, \ldots\ldots, x_n$

$$C_{MAP} = \text{argmax}_{c \in C}\ P(x_1, x_2, \ldots\ldots x_n|c)P(c) \quad (4)$$

5.2.1.2 Multi nominal Naïve Bayes

Some assumptions are made as following

- Bag of words assumption: Assuming the position doesn't matter.
- Conditional Independence: Assuming the feature probabilities $P(x_i|c_j)$ are independent given the class c.

$$P(x_1, x_2, \ldots\ldots x_n|c) = P(x_1|c).P(x_2|c).P(x_3|c) \ldots\ldots P(x_n|c)$$

$$C_{NB} = \text{argmax}_{c \in C}\ P(c_j) \prod_{x \in X} P(x|c) \quad (5)$$

Laplace add-1 smoothing for Naïve Bayes,

$$P\hat{}(w_i|c) = \frac{count(w_i,c)}{\sum_{w \in V}(count(w,c))} \quad (6)$$

5.3 Libraries

All the software libraries used are Open Source and most of modules are edited according to the requirement of the project. The following are the libraries that are used in the project.

5.3.1 SciPy

SciPy is an Open Source Library for Scientific Tools in Python language. Mostly used by scientists and analysts for scientific computing. SciPy was built on the NumPy array object and is part of the NumPy stack.

SciPy also contains modules for Optimization, integration, linear algebra, FFT, signal processing and Image processing. SciPy library is currently distributed under the BSD license and development is sponsored and also supported by an open community of developers.

Scikit learn has all the required machine learning algorithms and its variations. Scikit learn gives the flexibility to change the parameters of various machine learning algorithms, which in turn helps in achieving better efficiency.

5.3.2 NumPy

NumPy is an extension to the python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on multi-dimensional arrays.

It is the most basic package for scientific computing. At the crux of the NumPy package there is ndarray object, which encapsulates n-dimensional arrays of similar data types, with most of the operations being performed in compiled code for performance.

5.3.3 NLTK

The Natural Language Toolkit is a well-known platform, basically used for building Python programs that work on human language data in order to apply statistical natural language processing.

It also contains text processing libraries for stemming, tagging, tokenization, parsing, classification, and semantic reasoning. NLTK also includes graphical representations and sample data sets as well. NLTK supports corpus data and models as well.

6 Limitations and Future Enhancements

Limitations of this system is word of mouth. Communication of the trade information could be word of mouth, which can't be monitored. But all other communication interfaces can be monitored in a corporate entity such as Emails, Shared documents, Chat and transcripts of audio calls. Limitations of the system include word similarity of certain words.

This Monitoring system can be able to find the most probable insider trading activities. Which is not 100% accurate yet.It might not be as good as a human expert in monitoring in all the situations. These enhancements will be made in next releases, with more capturing patterns. All future enhancements will be reached with the expertise of a human. Future enhancements also include improvement in time complexity as well. In search for synonyms and Antonyms the time complexity should be improved.

7 Output Screens

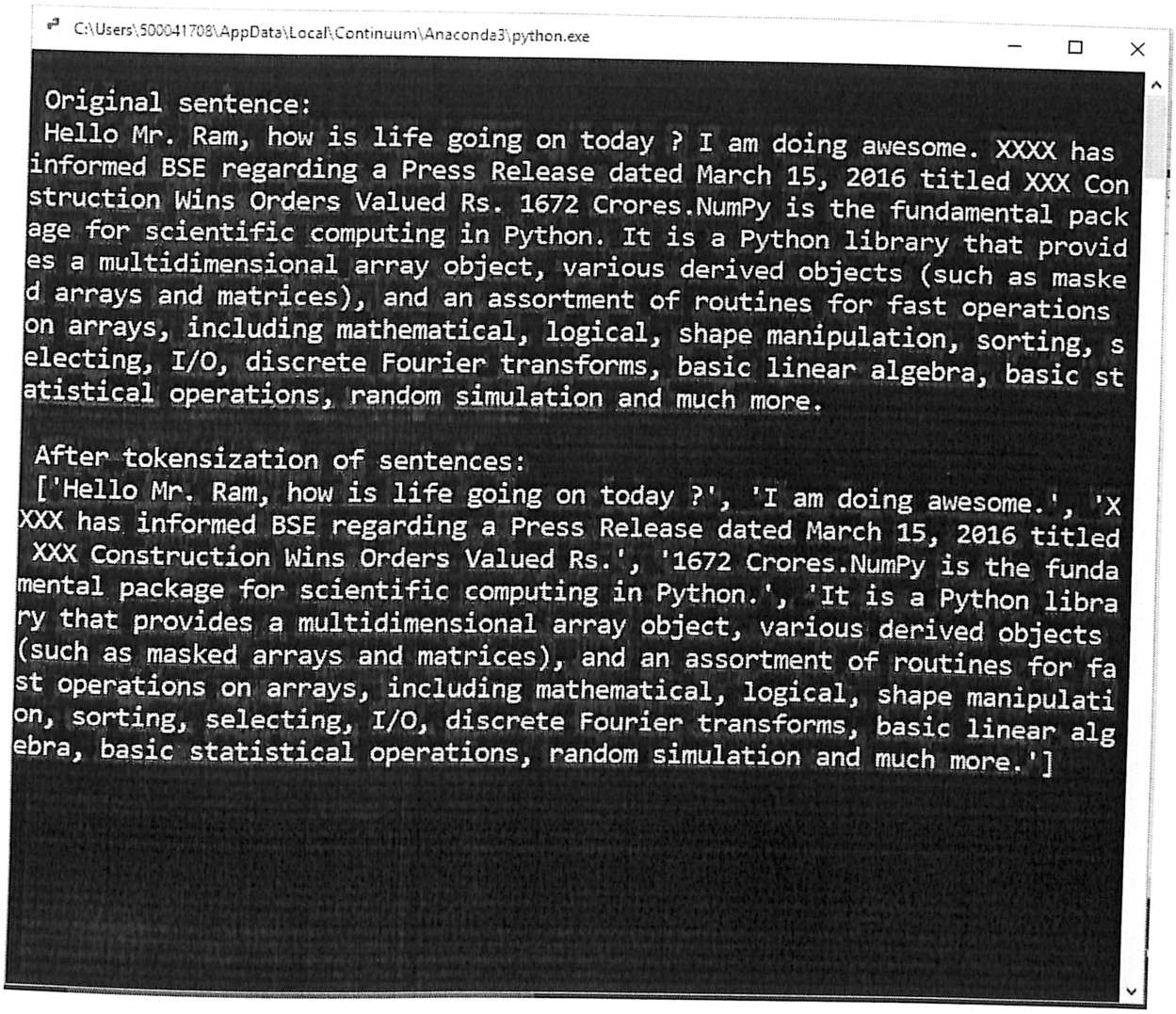

Figure 7.1 Tokenization of a paragraph

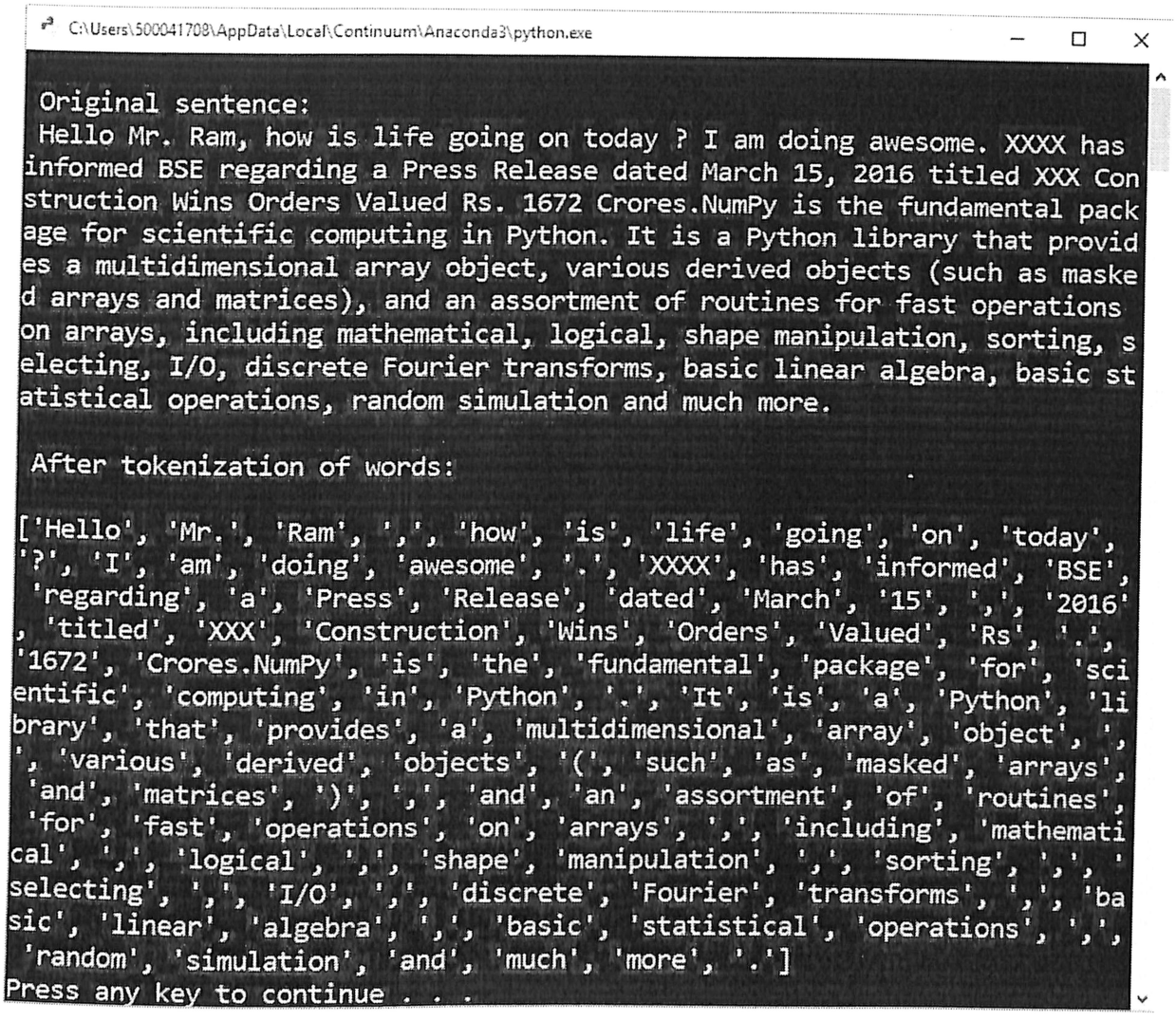

Figure 7.2 Tokenization of paragraph into words

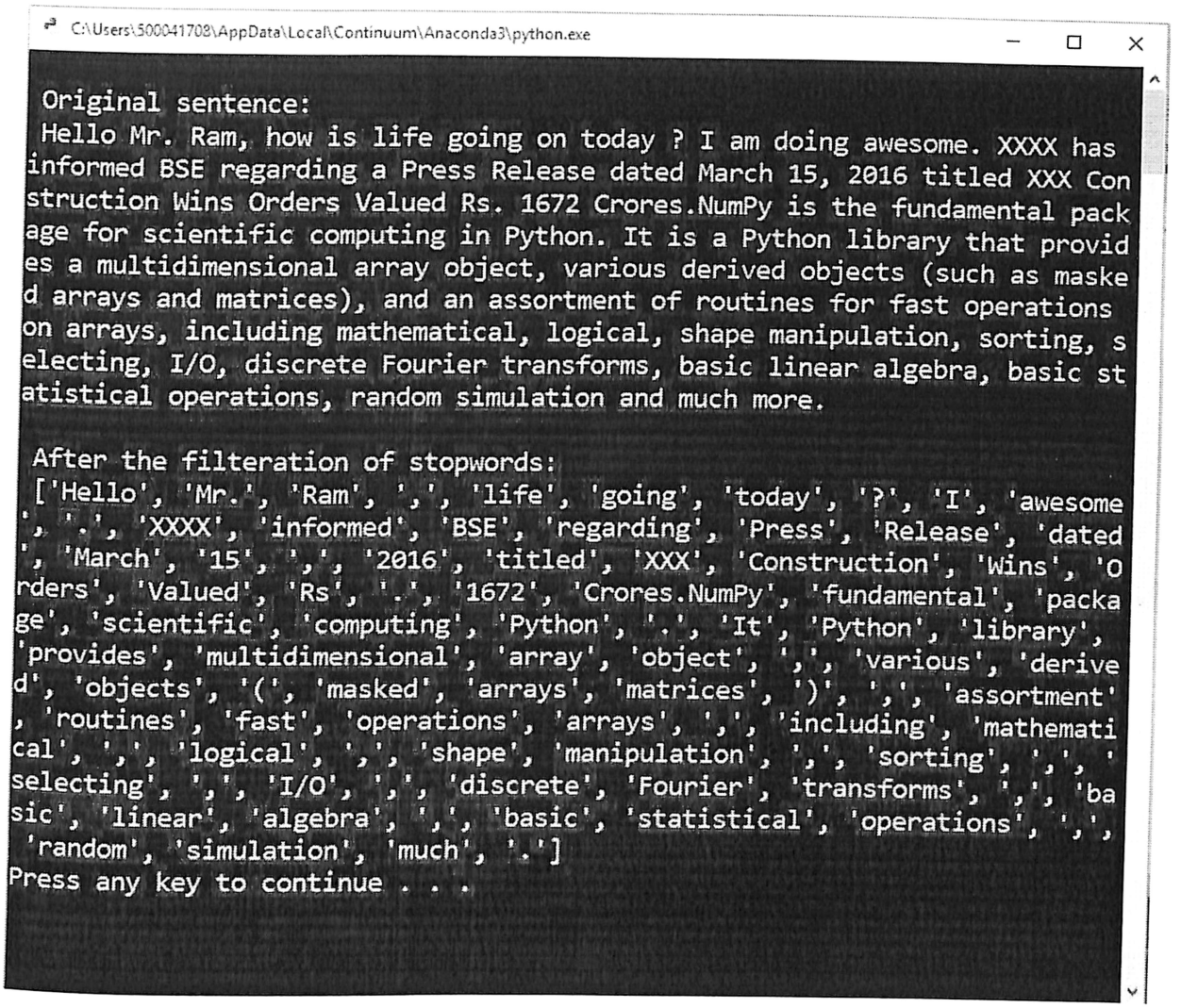

Figure 7.3 Stop words elimination

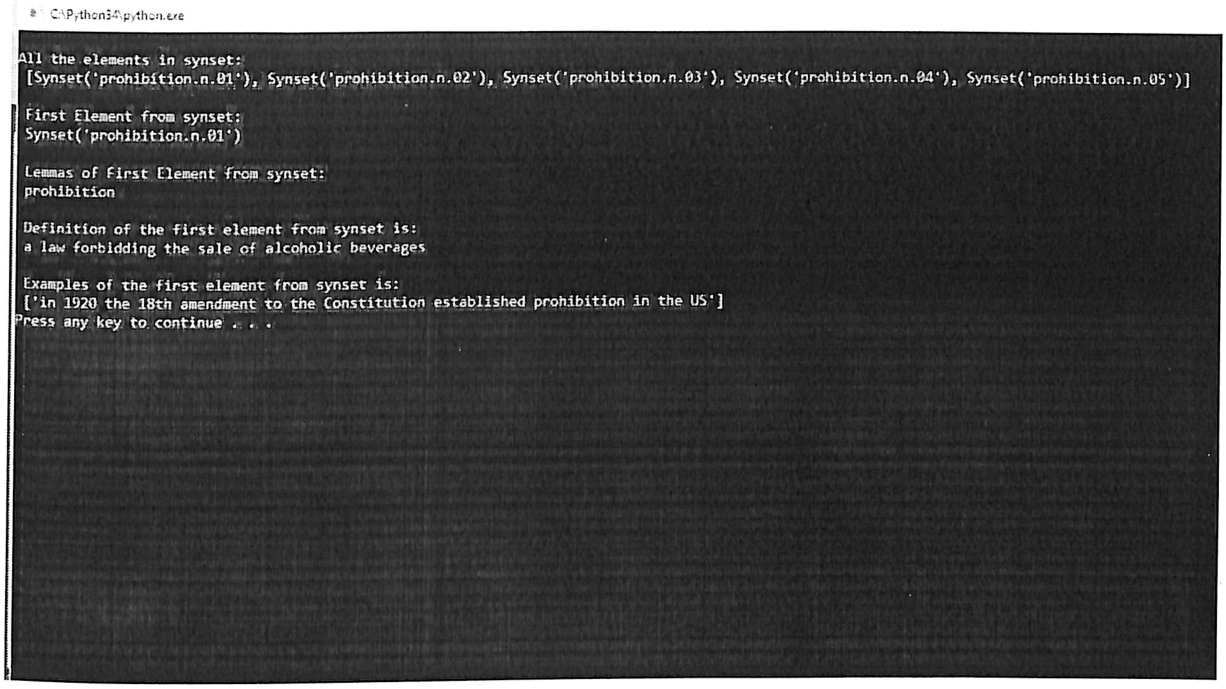

Figure 7.4 Definition and synonyms

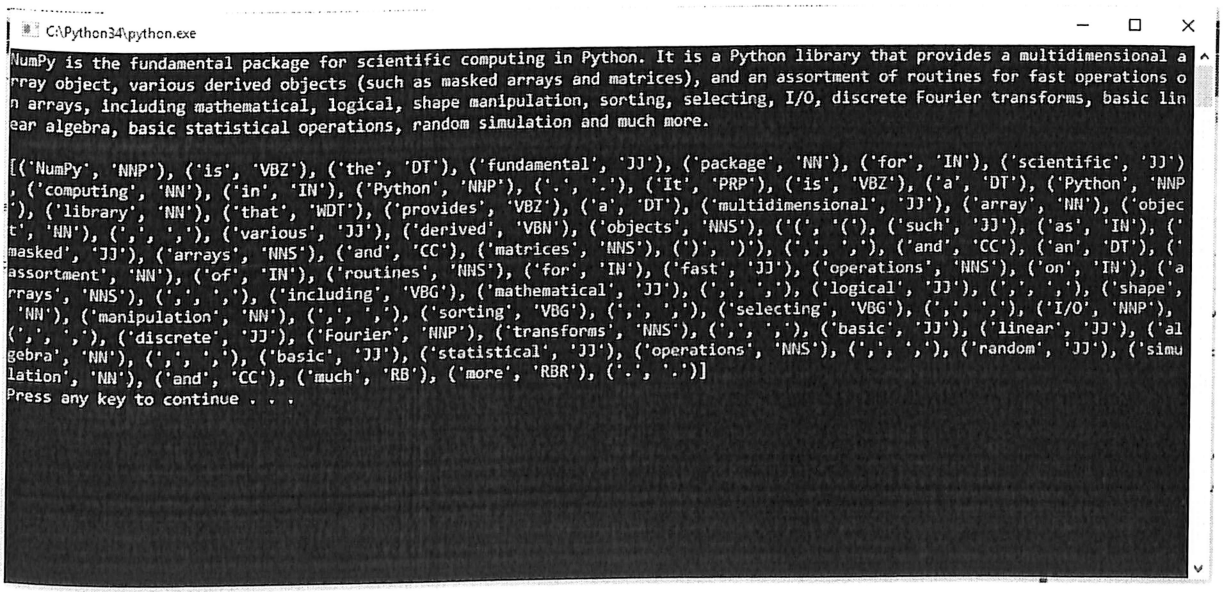

Figure 7.5 POS Tagging

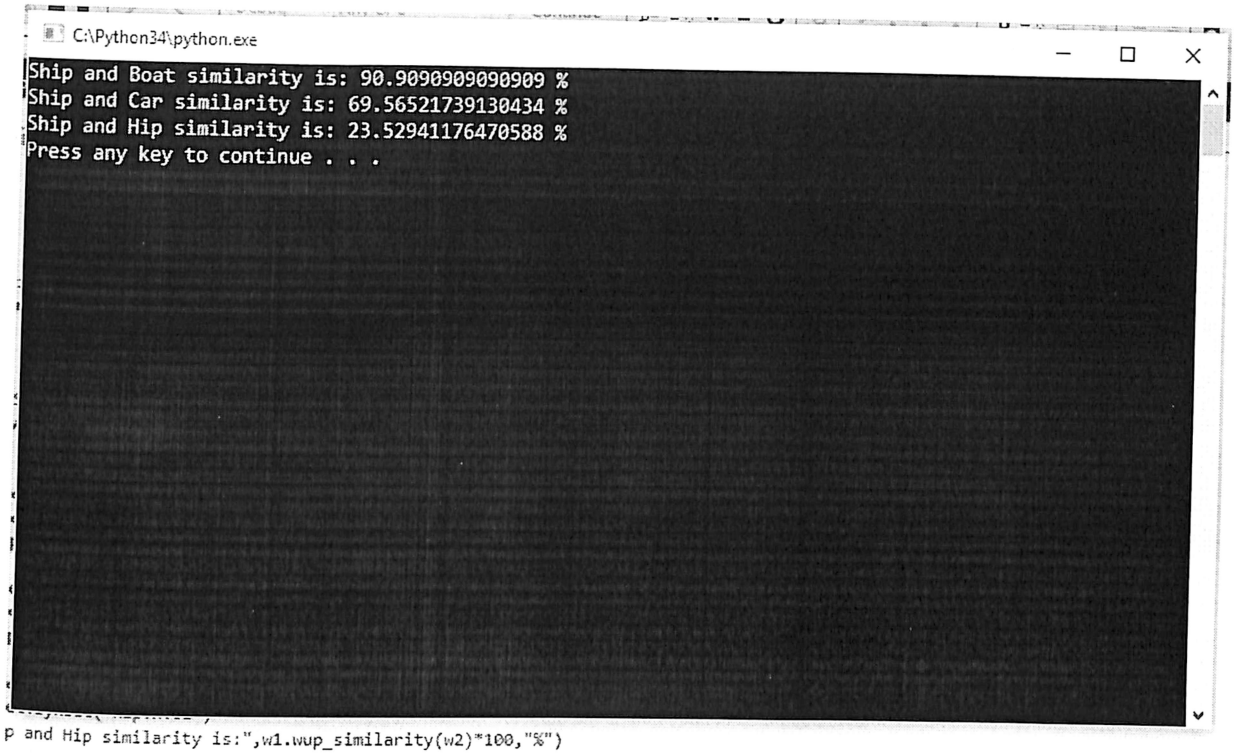

Figure 7.6 Similarity measure

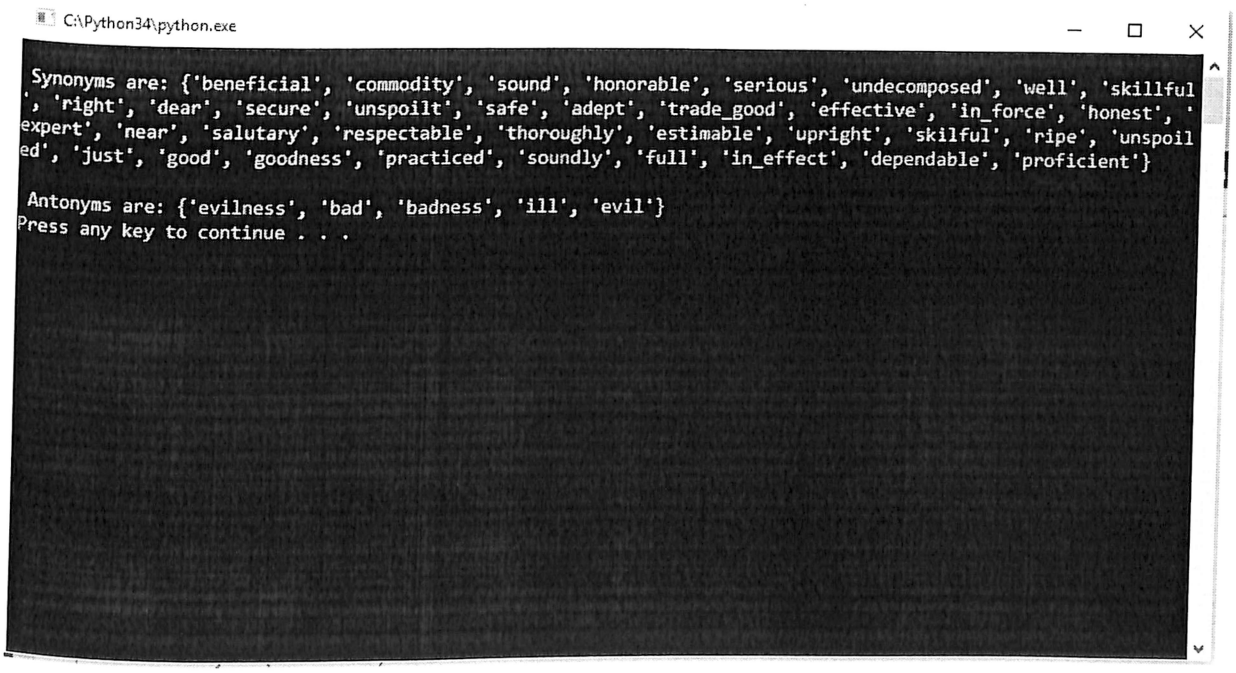

Figure 7.7 Synonyms and Antonyms

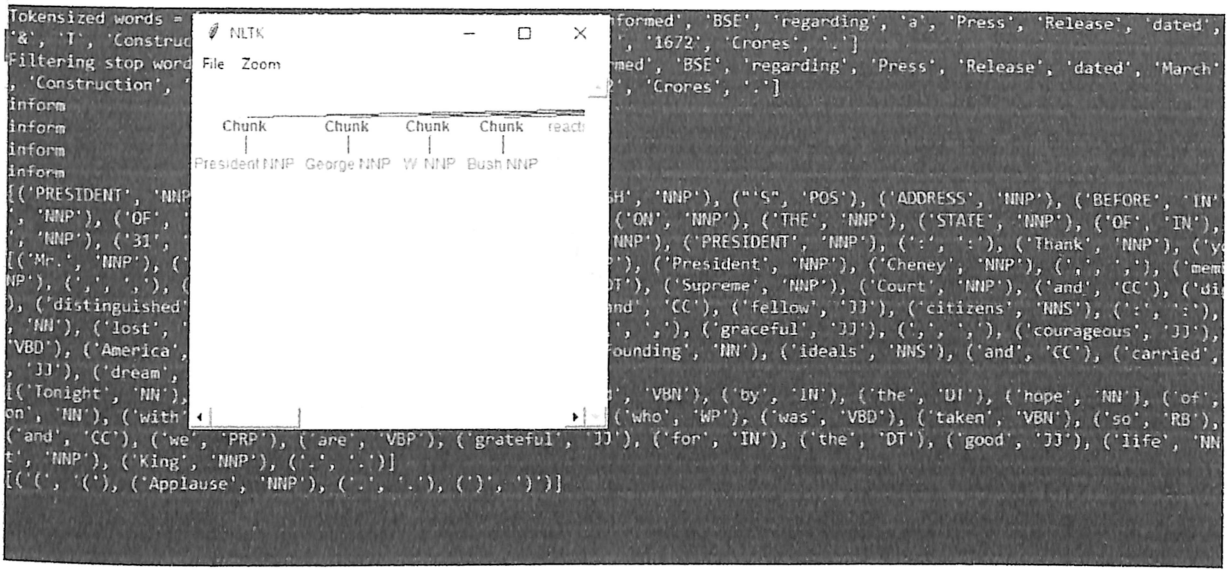

Figure 7.8 Chunking

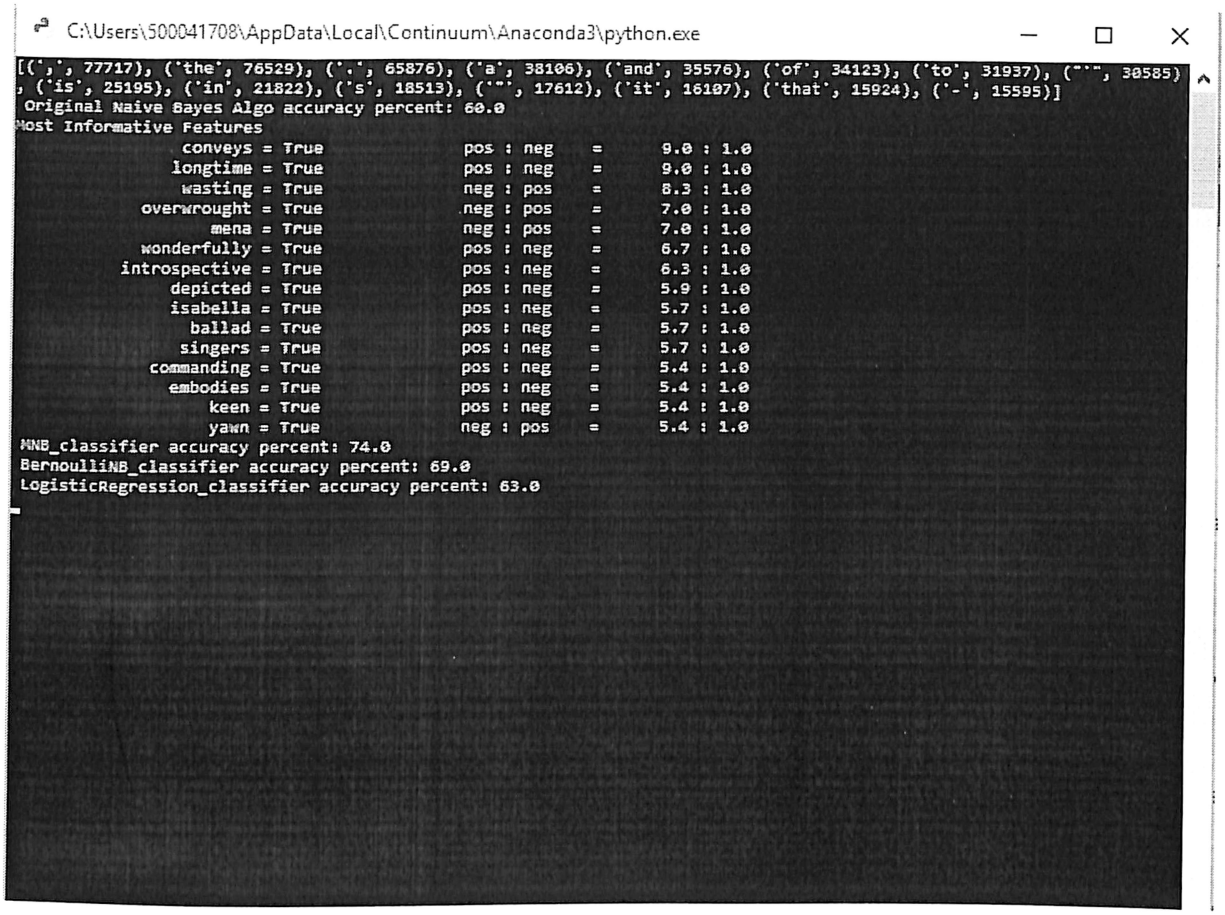

Figure 7.9 Classifiers accuracy

8 Conclusion

Considering the recent regulatory actions taken by the authorities, the main focus was on the flow of information between individuals in the corporate entities. As the ultimate goal of every corporate entity is to look for the effective and less expensive compliance risk detection techniques.

The proposed system will help in identifying the following

- If any Unpublished price sensitive information in corporate entities has been shared with in the intranet of the corporate entities.
- If trade related files, messages are being discussed.

The proposed system aids the compliance team in multiple ways and reduces the work load on the back offices and prevents the market crashes.

9 References

1. Acar Tamersoy, B. X. (2015). Inside Insider Trading: Patterns & Discoveries from a Large Scale Exploratory Analysis . gatech.

2. Andrew Y. Ng, M. I. (2002). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes.

3. Banerjee, S. a. (2003). Extended gloss over-laps as a measure of semantic relatedness. (pp. 805-810). In Proceedings of Eighteenth International Joint Conference on Artificial Intelligence.

4. Brill, M. B. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. Microsoft Research.

5. Rish, J. H. (2001). An analysis of data characteristics that affect naive Bayes performance. . IBM T.J Watson Research Center.

6. I.Rish. (n.d.). An empirical study of the naive Bayes classifier.

7. Kristina Toutanova, D. K. (n.d.). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. nlp.Standford.

8. Leacock, C. a. (1998). Combining local context and WrdNet similarity for word sense identification. IN (pp. 265-283). In Fellbaum, C., ed.,: WordNet: An electronic lexical database. MIT press.

9. Lee Biggerstaff, D. C. (2012). Insider Trading Patterns. Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is it time for some Linguistics? Departments of Linguistics and Computer Science Standford University.

10. McCallum, A. (n.d.). A Comparison of Event Models for Naive Bayes Text Classification. N.L.R.B. (2014).

11. Purple Communications, Inc. and Communications. Radicati, T. (2015). Email Statistics Report. The Radicati Group.

12. Scrapy. (n.d.). doc.scrapy.org. Retrieved from doc.scrapy.org.

13. SEBI, S. A. (1992). PROHIBITION OF INSIDER TRADING. REGULATIONS 1992 STOCK EXCHANGE BOARD OF INDIA.

14. SEBI, S. A. (2015). (PROHIBITION OF INSIDER TRADING) REGULATIONS,THE GAZETTE OF INDIA. PUBLISHED BY AUTHORITY, NEW DELHI, Part- 3, Section-4.,New Delhi.

15. Ted Pedersen, S. P. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts.

16. Zhan, F. (2014). Insider Trading: How Effective Are Exchange Rules and Surveillance? Boler School of Business,John Carroll University,CMCRC.