**

**

***Stock Market Analysis***

***Using NLP***

*Project Members*

*Gaurav Harariya- (PRN 180840125012)*

*Nikhil Bhatia - (PRN 180840125024)*

*SagarDubey - (PRN 180840125036)*

*YaswanthVudumula-(PRN180840125058)*

***Project Guide - Mr. Shridhar Page***

*Jan, 2019*

**DECLARATION**

We hereby declare that the project work entitled “Stock Market Analysis Using NLP” submitted to ACTS, CDAC, Pune; as part of the curriculum for Diploma in Big Data Analytics course is a record of an original work done by us under the guidance of

Mr. Shridhar Page,Head of IT Department, National Insurance Academy and this project work has not formed the basis for the award of any other degree / diploma/ associate-ship /fellowship and similar project if any.

*Gaurav Harariya - (PRN 180840125012)*

*Nikhil Bhatia - (PRN 180840125024)*

*SagarDubey - (PRN 180840125036)*

*YaswanthVudumula - (PRN180840125058)*

## **Acknowledgements**

## This project is using the news headlines from

## www.moneycontrol.com for most active companies

## in the past 18 years and by analyzing the past

## performance of the companies, We attempt to

## demonstrate the viability of using natural

## language processing to predict stock price volatility

## after a company experiences a major event.

## We thank the [www.moneycontrol.com](http://www.moneycontrol.com)and

## **www.bseindia.com** for making the data available publically.

## https://www.moneycontrol.com/

## **https://www.bseindia.com/markets/equity/EQReports/StockPrcHistori.aspx?scripcode=512289&flag=sp&Submit=G**

## We thank the **staff of ACTS, CDAC** and Course Coordinator **Mrs. ShilpiShalini**, for guiding and providing support for this project.

Certificate of Successful Completion of the Project

Contents

[Acknowledgements 3](#_Toc520892200)

Introduction……………………………………………………………………………………………………………………………………………6

[Domain / Business Problem Framing 6](#_Toc520892205)

[Conceptual Background of the Domain Problem 6](#_Toc520892206)

[Review of Literature 6](#_Toc520892207)

[Motivation for the Problem Undertaken 7](#_Toc520892208)

[Mathematical / Analytical Modeling of the Problem 8](#_Toc520892209)

[News Data : 8](#_Toc520892210)

[Stock Data: 8](#_Toc520892211)

Data Preprocessing [9](#_Toc520892213)

[Data Inputs – Logic – Output Relationships 9](#_Toc520892214)

[State the set of assumptions 10](#_Toc520892215)

[Hardware and Software Requirements & Tools Used 10](#_Toc520892216)

[Key metrics for success in solving problem under consideration 10](#_Toc520892217)

[Web Scrapping 11](#_Toc520892220)

[Reading the Data 13](#_Toc520892221)

[Cleaning the Data 14](#_Toc520892222)

[Tagging Part of Speech 14](#_Toc520892225)

[Dependency Parsing 15](#_Toc520892226)

[Function call to various Algorithms 16](#_Toc520892227)

[TF-IDF 17](#_Toc520892228)

[Light GBM 18](#_Toc520892229)

[Sentiment Analysis Code 19](#_Toc520892230)

[Visualizations 21](#_Toc520892232)

[Interpretation of Results 26](#_Toc520892233)

[Key findings & conclusions of the Study 27](#_Toc520892234)

Introduction

# Domain / Business Problem Framing

# A vast amount of new information related to companies

# listed on the stock market appears constantly, with immediate

# impact on stock prices. Monitoring such information

# in real time is important for big trading institutions but out

# of reach of the individual investor. We present a news monitoring

# and stock prediction system, designed from the position

# of the individual investor without access to real-time

# trading tools.

# Conceptual Background of the Domain Problem

# Making decision in stock market is not really easy because a lot of factors are involved with every choice we make. Therefore, a lot of analysis is required to make an optimal move on stock market which may involve price trend, market's nature, company's stability, different news and rumors about stocks etc. The objective of this study is to extract fundamental information from relevant news sources and use them to analyze or sometimes forecast the stock market from the common investor's viewpoint. We surveyed the existing business text mining researches and proposed a framework that uses our text parser and analyzer algorithm with an open source natural language processing tool to analyze (machine learning and text mining), retrieve (natural language processing), forecast (compare with historic data) investment decisions from any text data source on stock market.

# Review of Literature

https://www.moneycontrol.com

https://towardsdatascience.com

https://www.bseindia.com/markets/equity/EQReports/StockPrcHistori.aspx?scripcode=512289&flag=sp&Submit=G

# 

# Motivation for the Problem Undertaken

* The news headlines of various companies published by www.moneycontrol.com is reliable, comprehensive and well suited for applying data analytics methods.
* Using Descriptive Analysis it will be possible to visualize the past state of stocks.
* Using predictive techniques and machine learning algorithms, models can be developed which can help to predict the behavior of the stock market on the basis of news,which will be very beneficial for the investors.

Analytic Problem Framing

# Mathematical / Analytical Modeling of the Problem

The news accumulated will be processed through various descriptive analysis techniques and visualizations will be prepared to help create static inferences about the stocks.

The Pre-processed news will be used to create a Machine Learning based model which can predict the trend of stock market.

1. **Data Sources & their Formats**

## News Data :

The data is in a csv file format and includes following details of the various companies. There are about 11k records in the dataset from year 2000 to 2019.

* Date
* News
* Company Name
* Loss/Profit
* Percentage Change

There are about 11k records in the dataset from year 2000 to 2019.

**StockData :**

The data is in a csv file format and includes following details of the various companies. There are about 36k records in the dataset from year 2000 to 2019.

* Date
* Open Price
* High Price
* Low Price
* Close Price
* WAP
* No. of Shares
* No. of Trades
* Total Turnover(Rs)
* Percentage Deliverable Quantity to Traded Quantity
* SpreadHigh-Low
* SpreadClose-Open
* Company Name
* Percentage Change
* Loss/Profit

**News data:**

The information displayed represents a listing of news headlines from Jan 1, 2000 to Dec 31, 2018, that is given by www.moneycontrol.com, we did web-scraping with the usage of requests library and beautiful soup package .

The Data which is Scrapped is stored in the Excel(.xlsx) file.

**Stock Data:**

The Stock Data Comprises of various Columns but the Open price and Close price column are responsible for providing us with Loss/profit value which can

Be either 0 or 1.

Here 1 stands for profit and 0 stands for Loss.We have merged the data of various companies into one single file which comprises of approximately 36k records and than used that file for Data Visualiztaion using Tableau.

1. **Data Preprocessing Done (wherever applicable)**
   * **News Data:**

After we got the Scraped Data initially we removed all the stop words from

the dataand than we are left with Bag Of Words we further used various

techniques such as lemmatization,TF-IDF(Term Frequency Inverse

Document Frequency),Tokenizing to get the clean data.

* + **Stock Data:**

We removed the duplicate date column which we were getting after

Applying VLOOKUP function in excel for aggregating the News Data

excel Sheet with Stock Data sheet.

1. **Data Inputs – Logic – Output Relationships**

**Prediction Model**

**Machine Learning Algorithms**

**News Data**

**Visualizations**

**Aggregations**

**Stock Data**

**Loss/Profit**

**VLOOKUP In Excel**

**Stock Data**

1. **State the set of assumptions (if any) related to the problem under consideration**
   * The News data captured by the www.moneycontrol.com is accurate and

Is no under- reporting or bias.

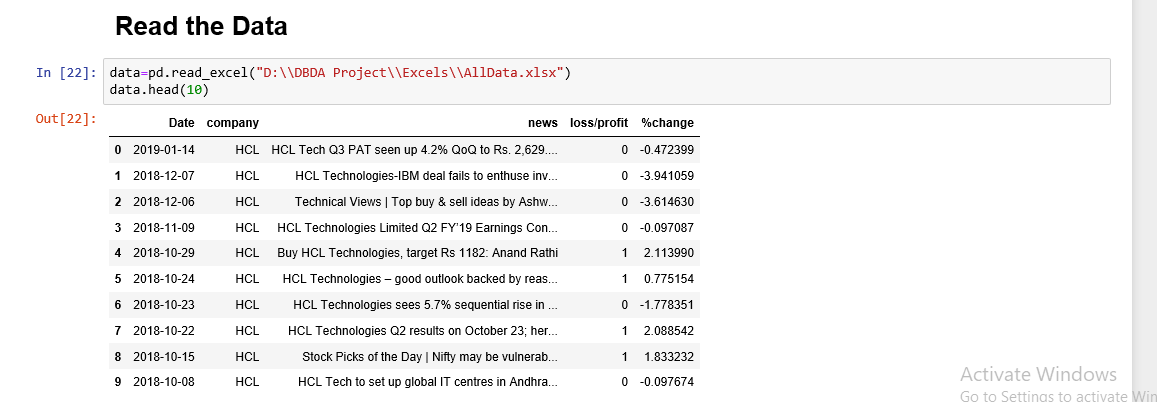
* + The Stock daily perfomances historic data from [www.bseindia.com](http://www.bseindia.com) was

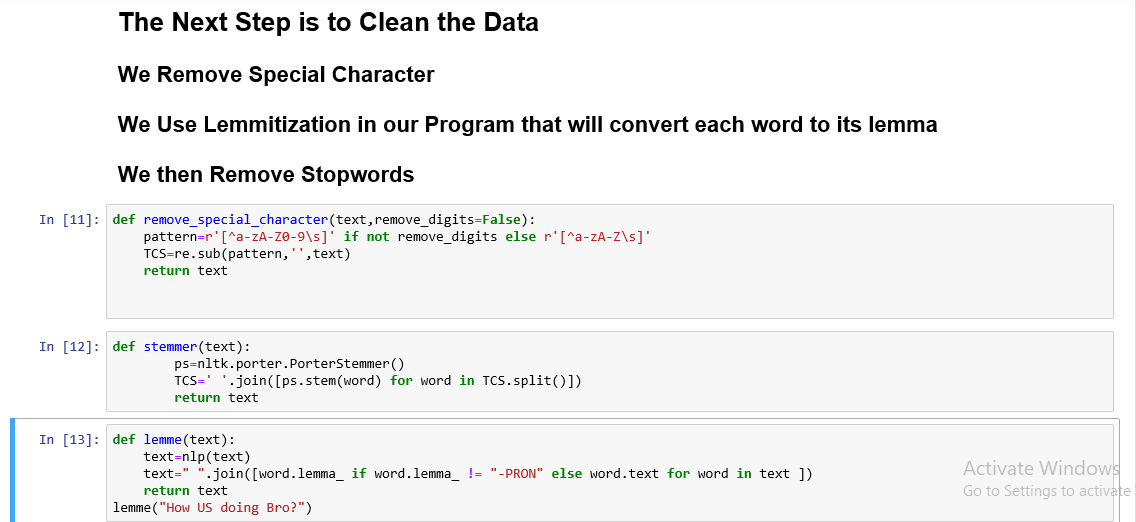
Precise and reliable.

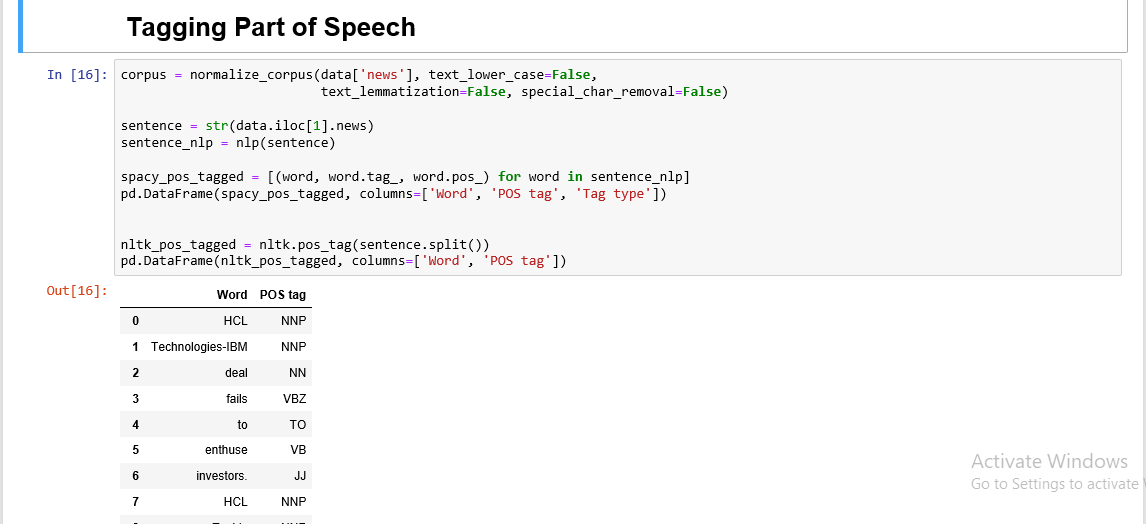
1. **Hardware and Software Requirements & Tools Used**
   * Python - Panda, Scikit learn
   * Tableau - Windows
   * Web Scrapinng - requests,Beautiful Soup package
   * Excel – VLOOKUP
   * Machine Learning Algorithms –Xgboost,Support Vector Machine,LGB,LogisticRegression
2. **Key metrics for success in solving problem under consideration**
   * High Accuracy Score of the Stock Prediction Model
   * Actionable insights from Descriptive Analysis

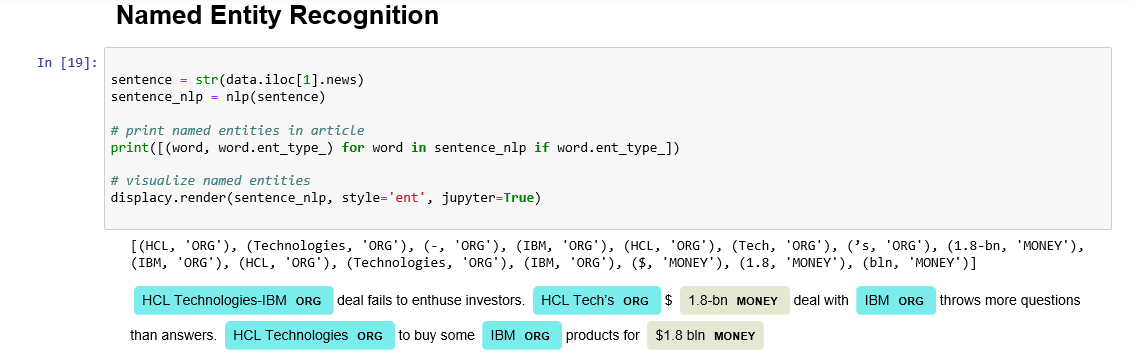
**Web Scraping**

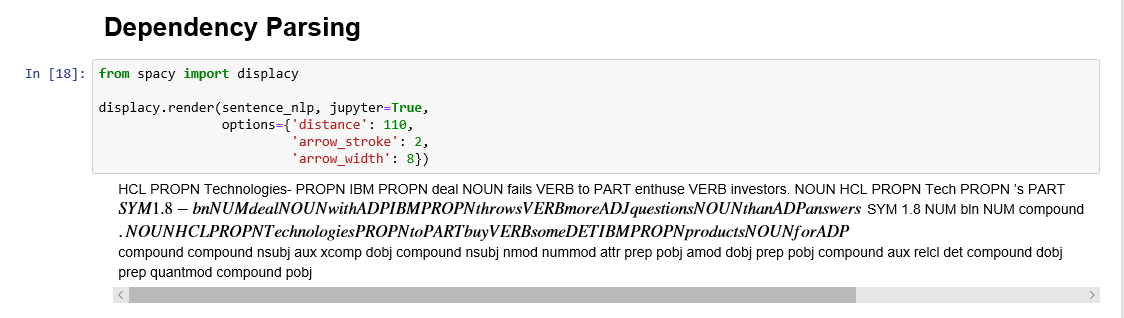


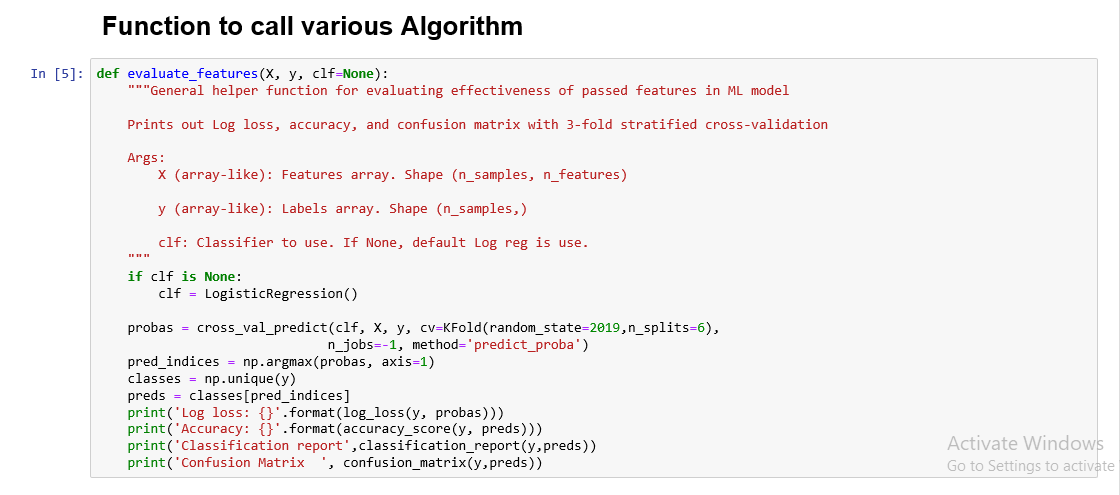


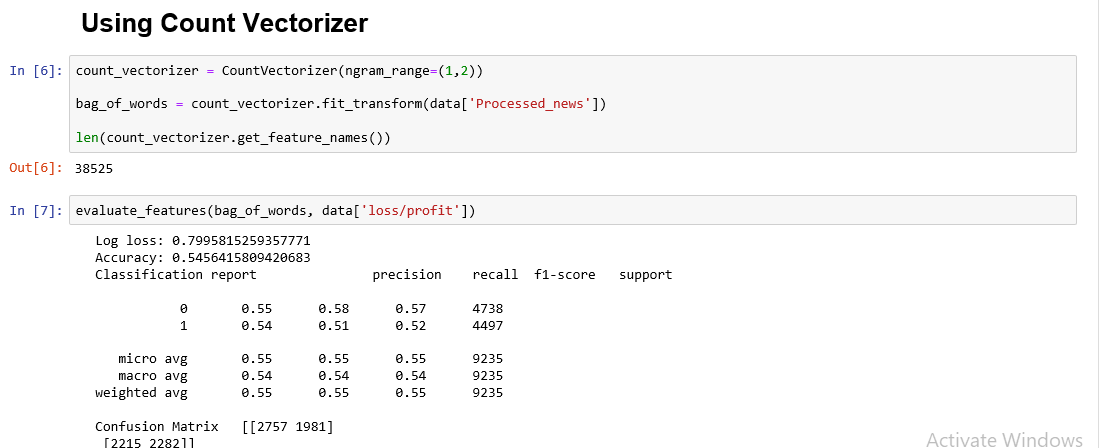










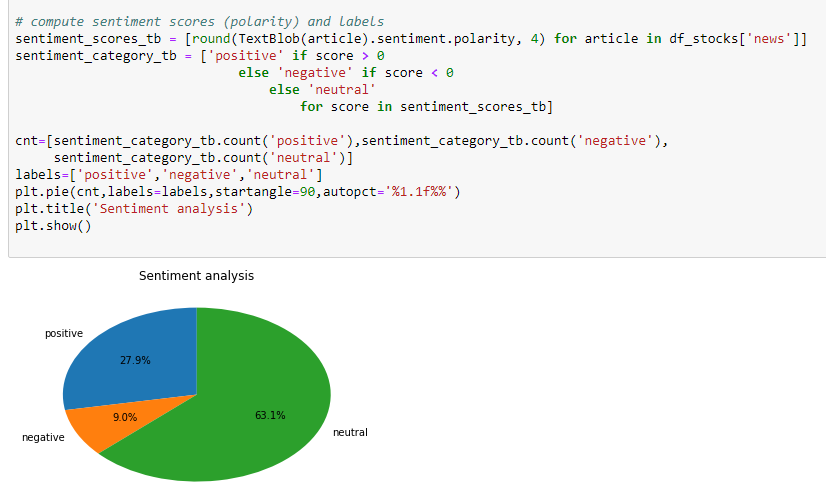


## 

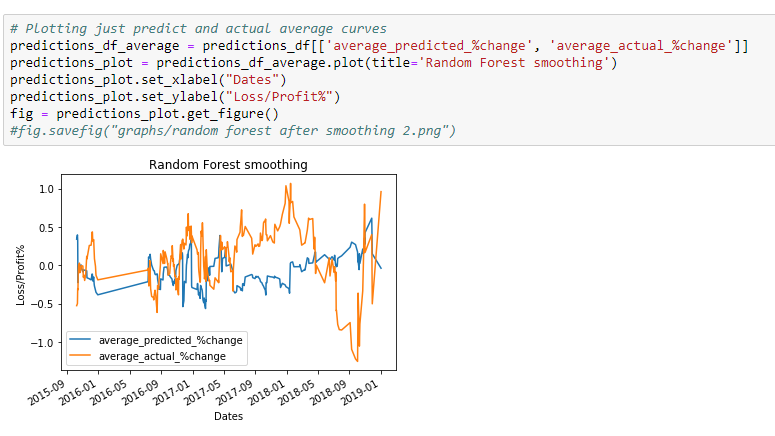
## 

**Sentiment Analysis code**



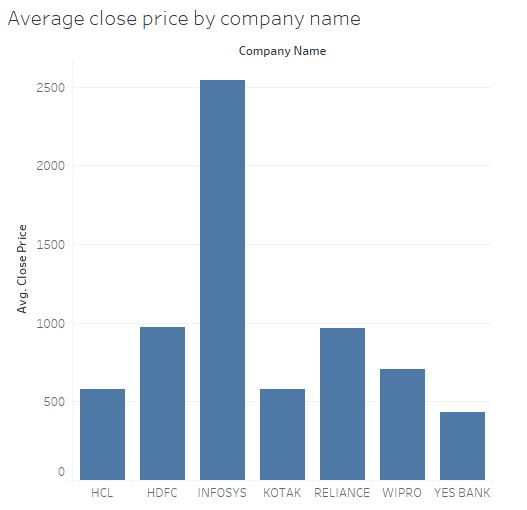


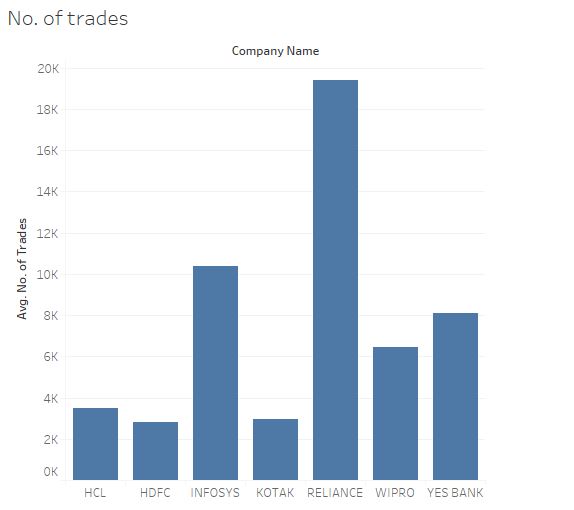
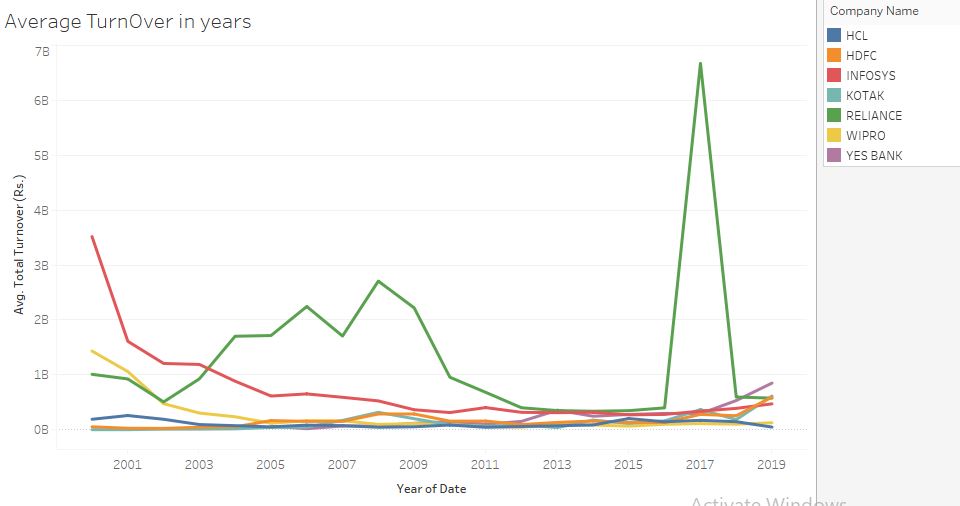


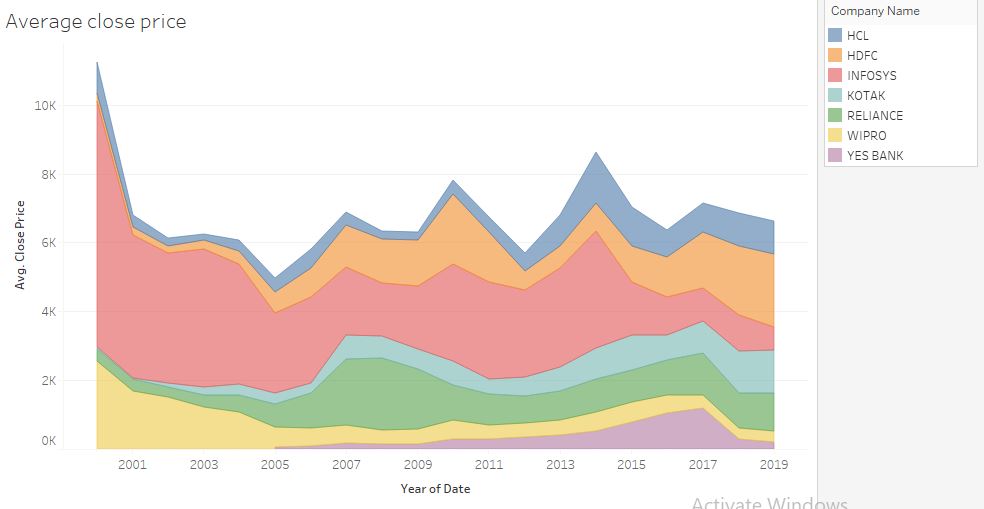


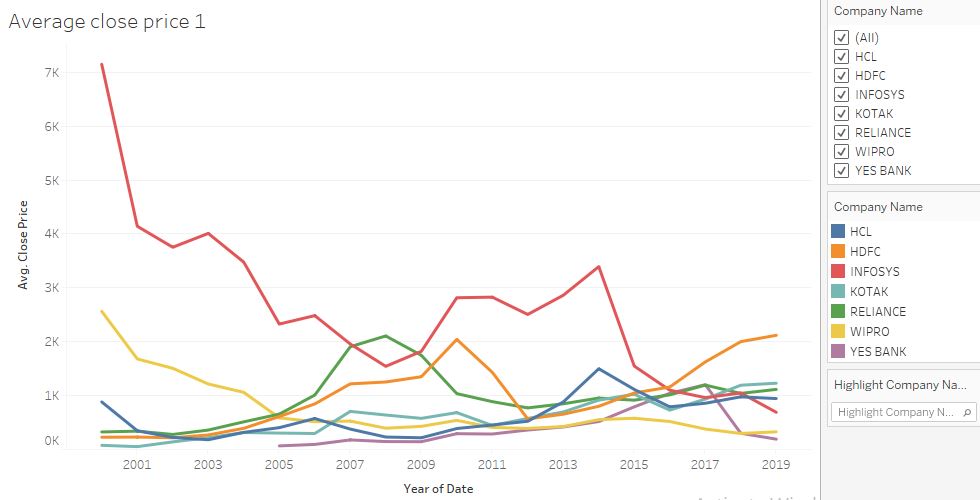
# Visualizations

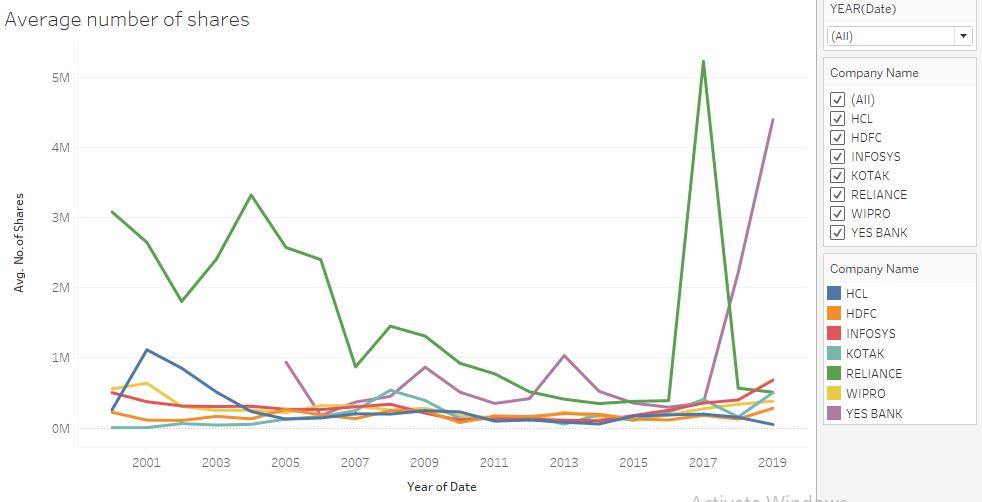


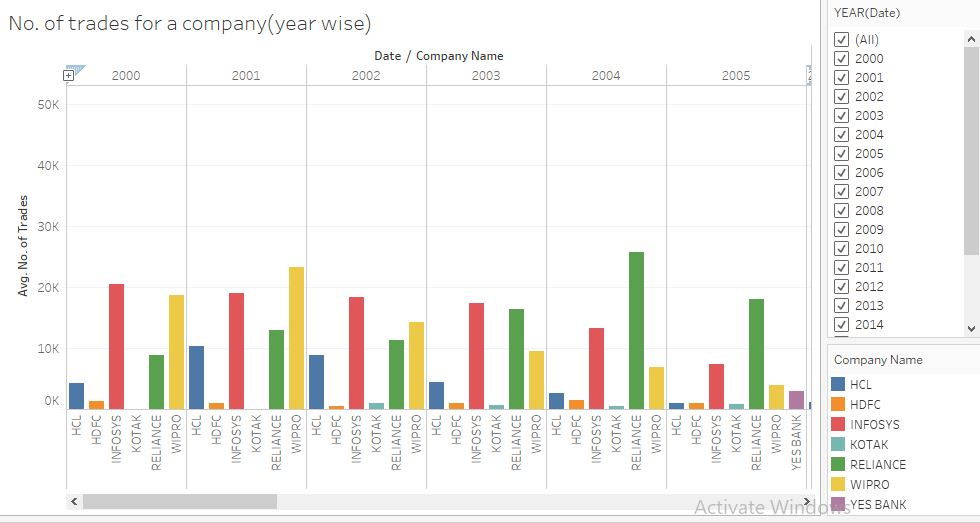










# Interpretation of Results

* **Descriptive Analysis :**
  + The top 3 words in our corpus are buy, target, say except stock names.
  + Positive news came more times than negative news.
  + Reliance is having highest turnover among all stocks.
  + The average turn over of reliance increased in 2017 after Jio profits.
* **Prediction Model:**
  + Support vector machines based model fits well to the news data available.
  + The model shows % accuracy in predicting the profit or loss based on given news.
  + Above model can be used to predict the stock behavior.
  + This can be used for various scenarios like selling, buying or holding stocks etc.

Conclusions

# Key findings & conclusions of the Study

* The Stock prices are increasing over a period.
* Positive news came more times than negative news.
* The average turn over of reliance increased in 2017 after Jio profits.
* The sentiment analysis graph is having similarity with stock behavior.
* The Text classification model can be used to predict stock behavior for next few days.

# Learning outcomes of this study

* The data validation, cleaning and preparation is most important and takes most of the time.
* Support vector machine method is effective in modeling Text classification model.

# Limitations of this Work & Scope for Future Work

* Volatility of market depend on various factors, so exact precision is not feasible to achieve.
* This is for short term prediction only based on news of stocks.
* We can analyze sector wise stock news also.
* We can enhance this model for trading the stocks in real time based on news without need of human interaction.
* We can send notification alerts to users when any negative or positive news comes for user interested stocks.

## 