Introduction to Big Data Phase 4:
Project Report

Abhinandan Desai ad2724@rit.edu
Jaideep Bhide jb3273@rit.edu
Sarthak Gupte sg7179@rit.edu
Swapnil Shah sss4174@rit.edu
Mihir Shah ms8830@rit.edu

Abstract:
The dataset 'News Popularity in Multiple Social Media Platforms' contains a large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn. This project will dwell on exploring, cleaning, analyzing and then visualizing the acquired knowledge. The results will be used to draw out conclusions and use it for further discussions.

Introduction - Dataset Description
The dataset contains news items and their respective social feedback on multiple platforms like Facebook, Google+ and LinkedIn. The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about 100,000 news items on four different topics:

- Economy.
- Microsoft.
- Obama.
- Palestine.

This data set will be tailored for evaluative comparisons in predictive analytics tasks, although tasks in other research areas such as topic detection and tracking, sentiment analysis in short text, first story detection or news recommendation can also be carried out to produce different type of results from the data.

R programming language will be used to upload, read and carry out the data mining tasks on the dataset.

The dataset contains news headlines and their respective information. They include

- Facebook news
- Google-Plus news
- LinkedIn news

The attributes for each of the tables are :

- IDLink (numeric): Unique identifier of news items
- Title (string): Title of the news item according to the official media sources
- Headline (string): Headline of the news item according to the official media sources
- Source (string): Original news outlet that published the news item
- Topic (string): Query topic used to obtain the items in the official media sources
- Publish-Date (timestamp): Date and time of the news items' publication
- Sentiment-Title (numeric): Sentiment score of the text in the news items' title
- Sentiment-Headline (numeric): Sentiment score of the text in the news items' headline

- Facebook (numeric): Final value of the news items' popularity according to the social media source Facebook
- Google-Plus (numeric): Final value of the news items' popularity according to the social media source Google+
- LinkedIn (numeric): Final value of the news items' popularity according to the social media source LinkedIn

The second portion of the dataset contains the Social Feedback data.

Social Feedback Data contains scores which display the popularity of a certain article during different time slices and the final popularity score after 2 days of the article's publication.
IDLink is the unique identifier for the news articles.

Data Exploration and Summarization

#Command used to load file into a variable in R
LinkedIn_Microsoft <- read.csv("LinkedIn_Microsoft.csv", header = T)

# This command displays the type of the class of the object inside the parentheses.
class(LinkedIn_Microsoft)
Output - [1] "data.frame"

#This command displays the dimensions of the table that is number of rows and number of columns
dim(LinkedIn_Microsoft)
Output - [1] 20702 145

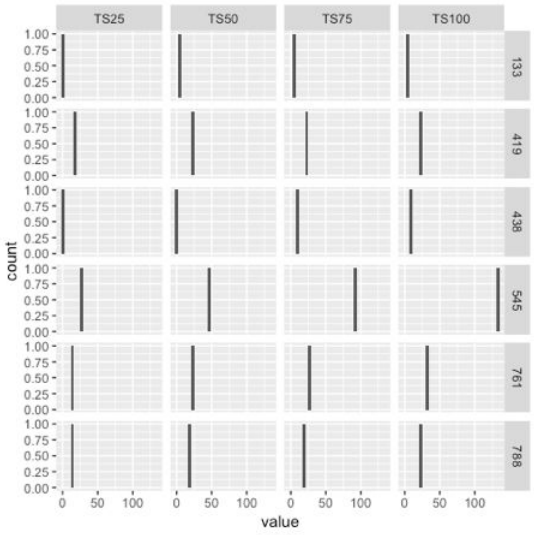
#This command displays the name of the column or variable of the table
names(LinkedIn_Microsoft)
Output - [1] "IDLink" "TS1" "TS2" …… "TS144"

#This command gives the brief description of the table data type of the rows name of the column.
str(LinkedIn_Microsoft)
Output -
$ IDLink: int 101 102 103 104 105 106 107 108 109 110 ...
$ TS1 : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
$ TS2 : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
$ TS3 : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 …
.
.
.
$ TS97 : int 18 8 10 0 186 0 5 36 95 0 ...
$ TS98 : int 20 8 10 0 186 0 5 36 95 0 ...
[list output truncated]

The tables and the csv file for the next phase of the project are of two types first the news type which stores IDLink of the headlines source and the likes hit(by the users), while another type of table consists of IDlinks and the total number of hits obtained in the respective time slot recorded within the span of two days with the time difference of 20 mins.
Using R we could explore the data like as shown, this graph shows IDlink on Y axis and how they perform or gets like in over different time slots.

Analyzing the data we could observe that by the end of 100th time slot amongst the 6 headlines IDLinks most hits were obtained by the headline with IDLink = "545", i.e., 133, while least hits were obtained by the headline with IDLink = "438", I.e., 7 . We

can plot more for every IDLink and analyze with time how many hits a headline can get



Code used to create and print the graph, is given below. The graph was created using the library package ggplot which a powerful graphic language tool used to plot complex graphs. It was not possible to fit the graph of the whole table with more than 20k rows here is the snippet of the part of the table.

```
> dat <- data.frame(Library = c("133","419","438","545","761","788"),
+                   TS25 = c(0,18,0,28,16,16),
+                   TS50 = c(6,21,0,48,23,19),
+                   TS75 = c(6,21,7,93,29,20),
+                   TS100 = c(6,21,7,133,30,22)
+                  )
> dat <- data.table::melt(dat,id.vars="Library")
> library(ggplot2)
> ggplot(dat,aes(x = value)) + geom_histogram() + facet_grid(Library~variable)
```

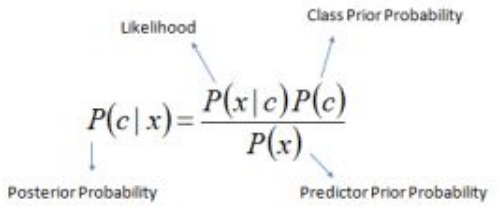
Proposed Data Mining Technique

Naive Bayes Classification

The technique of classification can be used to segregate the news articles into four groups namely - Obama, Palestine, Economy, and Microsoft. We can perform supervised learning on the given data using the Topic attribute which provides the group identity of a news article.

Classification is an appropriate choice for this dataset as the class assignments are known and the given data can be trained such that the classes of news articles can be correctly predicted and the articles can be assigned to their target classes or groups on the basis of their titles.

The Naive Bayes classification algorithm can be used here. Naive Bayes will work by finding a posterior probability of each class for a given headline and the class with the highest probability will be the target class for the news article. This algorithm will work well here because the classes are majorly disconnected from each other and so the assumption of independence among the predictors will work. Secondly, the dataset is very large and Naive Bayes has been known to outperform other sophisticated algorithms when it comes to large datasets.



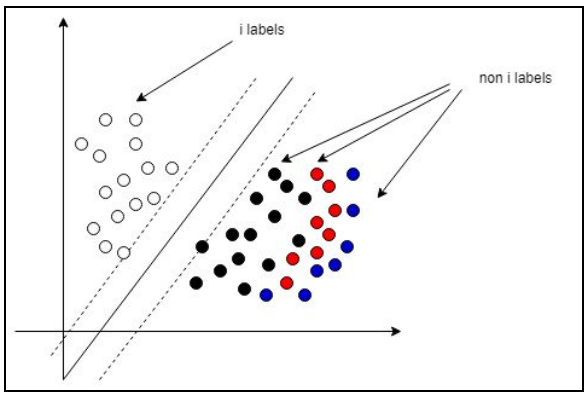
$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

```
classifier <- naiveBayes(mat4, train$Topic)
inspredicted <- predict(classifier,mat5)
table(as.character(test$Topic), as.character(inspredicted))

          economy microsoft obama palestine
economy     20840        66   108      3443
microsoft     246     12527    31      2933
obama         228        40 19322       948
palestine     149         1    61      6189

compare <- table(test$Topic, inspredicted)
PercentageAccuracy = sum(diag(compare)/sum(compare))*100

print(PercentageAccuracy)
1] 87.70482
```

SVM Classifier

The objective of using a Support Vector Machine is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

For multiclass classification using svm the most common technique called as One versus All is been used. For eg. Let's assume we have N different classes. One versus All will train one classifier per class in total N classifiers. For class "i" it will assume i-labels as positive and all the other labels as negative. Hence, the hyperplane will be between the "i" labels and all the other labels as shown in the figure below.



For the given dataset classification has to be done for 4 different classes i.e. Microsoft, Palestine, Obama and Economy. Hence it is viable to use SVM for this type of multiclass classification.Also SVM has a feature space which can be determined by a Kernel Function. This function helps us to map the classified data in a higher dimensional space.

```
>
> mat1 <- DocumentTermMatrix(corpus1)
>
> mat5 <- weightTfIdf(mat1)
> mat5 <- as.matrix(mat5)
> matdf1 <- as.data.frame(mat5)
> svm_model<-
+   svm(news_title_topic$Topic ~ ., data = matdf, type = "C-classification",
+       kernel = "linear", scale = FALSE)
Error in model.frame.default(formula = news_title_topic$Topic ~ ., data = matdf,  :
  variable lengths differ (found for 'bernie')
> pred_train <- predict(svm_model, matdf)
Error in eval(predvars, data, env) : object 'additions' not found
> mean(pred_train == news_title_topic$Topic)
[1] 0.2972103
```

As seen in the image given the probability is 0.29 for the given class for the given new topic

Linear Regression

Also, Regression can be used for using predictive analysis of news headlines based on the type of the headline and the date when the news article was published.
Regression would be most favored as ,in statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

The proposed technique involves using linear regression to predict the feedback of a particular news topic on a social media platform.

The dataset contains various news headlines at various time slices. The dataset contains multiple scattered values for a particular time slice and so a linear regression curve wouldn't be possible to create and so to derive a standardised result, a mean value of the popularity of all the news articles in a particular time slice was used to map it on a plot.

Using this information 75% of the dataset will be used to train the model. 25% of the dataset will be used to test the dataset.

Testing of the dataset will involve taking a news article and predicting the popularity of the news article at a specific given time. The results of linear regression are given below: