ASSIGNMENT – WSMA

Post-graduation in Data Science & Business Analytics

**By**

**Jithesh Janardhanan**

Contents

[1. Problem statement and Objective 2](#_Toc83091658)

[Topic Selection 2](#_Toc83091659)

[Scrapping of Tweets 2](#_Toc83091661)

[2. Exploratory Data Analysis 2](#_Toc83091662)

[Data Preprocessing and Cleaning 3](#_Toc83091663)

[Analyzing Text Frequency 3](#_Toc83091664)

[Creating word cloud 6](#_Toc83091665)

[Word Association and Correlation Charts 8](#_Toc83091666)

[10](#_Toc83091667)

[Topic Modelling using LDA 11](#_Toc83091668)

[Twitter Sentiment analysis 12](#_Toc83091669)

[Polarity Analysis 13](#_Toc83091671)

# Problem statement and Objective

Objective: To do a brand perception analysis (Social media analytics - Text mining)

* Identify a brand - any global or Indian product, celebrity, company, etc.
* For the identified brand download minimum 1000 twitter messages for the most recent period.
* Perform EDA and Data Cleaning.
* Conduct text mining on the data – Correlation, Frequency, Topic Modelling using Word Association, Sentiment Analysis.

## Topic Selection

## The topic selected in this analysis was on a trending subject which hit the news headlines recently which is the resignation of Punjab chief minister Mr. Amarinder Sigh from the rule as several MLAs from the cabinet opposed his leadership as the chief minister. He had to step down from the post immediately after the congress high command asked him to do so. This was one of the trending topics on 19th September Sunday, as the twitteraties erupted the internet with tweets and messages all over the social media. The subject got mixed response from the public as we analyze the data

The purpose of this analysis is to find out the response and reaction of the public in response to the above-mentioned event. Many of them were in favor of the resignation while many supported Mr. Amarinder Singh. This analysis and report try to do a text mining and sentiment analysis of the recent tweets on this event.

## Scrapping of Tweets

The tweets were scraped using the rtweet package with the below lines of code in R. Around 5000 tweets in English language were scrapped for the analysis and we will now go detail into the text mining exercise

A picture containing text

Description automatically generated

# Exploratory Data Analysis

From the data we scrapped, we only consider the text column out of it and the other columns we disregard with respect of the interest of this analysis and save it as a data frame.

The data contains about 4944 tweets and 90 columns out of which we consider only the tweet text column in the interest of the analysis as mentioned above

## Data Preprocessing and Cleaning

The first step towards preprocessing of the data was to create a corpus using the text column in the data. Once the corpus of data is created, we a re ready for data preprocessing

Following are the steps done towards pre-processing and cleaning

* **Remove Punctuation:** The basic approach to deal with this is to remove everything that isn’t a standard number or letter. In our case, we will remove all punctuation
* **To lower case:** Now we change all the text to lower case letters in English. This is done because R is case sensitive, and R considers same word in upper case and lower case as different entities
* Another pre-processing task we must do is to remove meaningless terms to improve our ability to understand sentiments. Transformations in text done via the tm\_map() function. Basically, all transformations work on single text documents and tm\_map() just applies them to all documents in a corpus.
* Removing the stop words which does not add any value to the analysis
* Removing of all numbers from the texts
* Strip off white spaces
* Remove all URLs and @usernames
* Remove anything except the English language (including nos.) and space
* Remove single letter words

We looked at the tweets individually to some extent to see if there are any gaps in the cleaning stage of the data

## Analyzing Text Frequency

Once we have pre-processed our data, we’re now ready to extract the word frequencies used in our twitter data. The tm package provides a function called Term Document Matrix that generates a matrix where the rows correspond to documents, in our case tweets, and the columns correspond to words in those tweets.

The values in the matrix are the counts of how many times that word appeared in each document. Document matrix is a table containing the frequency of the words. Column names are words and row names are documents.

Below is the snapshot of the Document term matrix with the frequencies of mostly appeared terms from the tweets

Table

Description automatically generated

* Words having frequency of at least 100 in the corpus are

A picture containing text

Description automatically generated

* Words having frequency of at least 150 in the corpus are

Text

Description automatically generated

* Words having frequency of at least 200 in the corpus are

Text

Description automatically generated

Now let’s plot the graphs of frequent words at 100, 150 and 200.

Chart, table

Description automatically generated with medium confidence

Chart

Description automatically generated

Chart

Description automatically generated

Based on the “TermDocumentMatrix()” output tried to sort the keywords based on their frequency. The word with high frequency is “captain” as tweeted by the users.

Chart, bar chart, histogram

Description automatically generated

## 

## Creating word cloud

We create an overall word cloud first followed by Positive word cloud and negative word cloud. Mostly appeared words in the overall word cloud are “captain”, “replaced”, “command”, “good”, “exit” etc.

Text

Description automatically generated

Below is the word cloud for **positive sentiments**. Most appeared words are “graceful”, “good”, “won”, “support” etc.

Text

Description automatically generated with low confidence

Below is the word cloud of only **negative sentiments**. Most frequent negative words are “sinking”, “fallen”, “lying”, “destroy”, “oppose” etc.

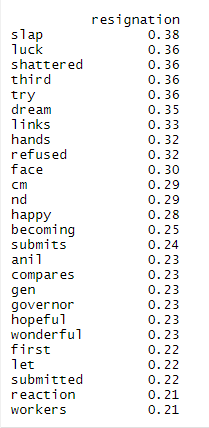
Text, timeline

Description automatically generated

## Word Association and Correlation Charts

We find out the association between main key words that came out through the analysis. We have selected 4 key words such as **“resignation”, “fallen”, “command” and “sinking”** to find out their association with other words in the analysis

***1.Resignation***



Chart, bar chart

Description automatically generated

The words “slap”, “luck” and “shattered” etc. are showing high correlation with the word “**resignation**”

***2.Fallen***

“Lying”, “sword”, and “aide” etc. shows high correlation with the word **“fallen”**

Chart, bar chart

Description automatically generatedTable

Description automatically generated

***3.Command***

**Command** has high correlation “replaced”, “voters”, “voted”, “won” etc.

# Chart, bar chart Description automatically generatedA picture containing table Description automatically generated

***Sinking***

The word “sinking” is highly correlated with “Voted”, “replaced”, “Voters” and “won”

Table

Description automatically generated

# Chart, bar chart Description automatically generated

## Topic Modelling using LDA

We do topic modelling to identify latent/hidden topics using LDA technique. Using this we have found out 7 hidden topics and first 7 terms for these topics.

Text, letter

Description automatically generated

From the topic modelling done above, we can make out the following meanings

Topic1 – The fight has ended by Rahul in time

Topic2 – About the humiliation faced by the CM

Topic3 – About India watching the explanations about the change in reign

Topic4 – About the command name getting replaced

Topic5 – About the situation of next chief minister

Topi6 – About the CM as an army and becoming a weapon against congress

Topic7 – About the hurried and graceful exit of the CM

The below plot shows the topic discussion in twitter mainly on Sept 19th, when the decision and resignation came out and people talked about topic 5 and 6 mainly on 19th of September

Chart

Description automatically generated

## Twitter Sentiment analysis

### Sentiment analysis is done to study the people’s opinions, sentiments, evaluations, attitudes, and emotions from written language. The sentiments are categorized as Positive, Anticipation, Fear, Joy, Surprise and Negative. This is done to understand the emotional variance in the tweets and social medias

Chart, bar chart

Description automatically generated

The score for positive tweets is much higher than any other emotion. This means people in general have reacted positively towards Amarinder Singh’s resignation from the post of chief minister. This is followed by Trust which can be interpreted as people have trust in the decision made

People have got high negative sentiments also about the decision made by congress party and that had come out as mixed feelings of Anger, Anticipation, and disgust amongst the public

Surprise and sadness are low amongst the public as they were expecting the decision any moment and is being reflected in their reactions in Twitter

## Polarity Analysis

Chart

Description automatically generated

This is about the polarity of tweets on different days and whether it is increasing or decreasing.

We could see that; the tweets came high in numbers between Sunday and Monday after the decision was announced to the public