Approach

Sentiment Analysis on Twitter data for BRExit

Neha Pai

Apurva Mokal

19 August, 2016

Introduction

In the Internet Age, the anonymity the platform provides makes Social media the most forthcoming and impactful in terms of information, opinions, and sentiments. This implies increasing need for understanding public voice, and thus demands ways for Analysing Sentiments. This documentation presents one of these ways where Tweets are used for analysing what the people have to say about 'BRExit', and present their views as Statistical data for further Data Analytics.

The approach is explained as: -

1. Extracting Tweets
   - Creating an App on twitter
   - Using '#brexit' to search tweets (Corpus Creation)
   - Other tweet processing methods
2. Sentiment Analysis
   - Importance of Opinion Lexicon
   - Using Opinion Lexicon for Scoring tweets
   - Classifying processed tweets into Sentiments
   - Preparing the data for further Data Analysis
3. Data Analysis
   - Introduction to Tableau
   - Representing the Data in charts
   - Creating Dashboard

Extracting Tweets

Creating an App on twitter: For extracting tweets from twitter, Authentication must be done. For this we must create an app in twitter by logging into dev.twitter.com and apps.twitter.com

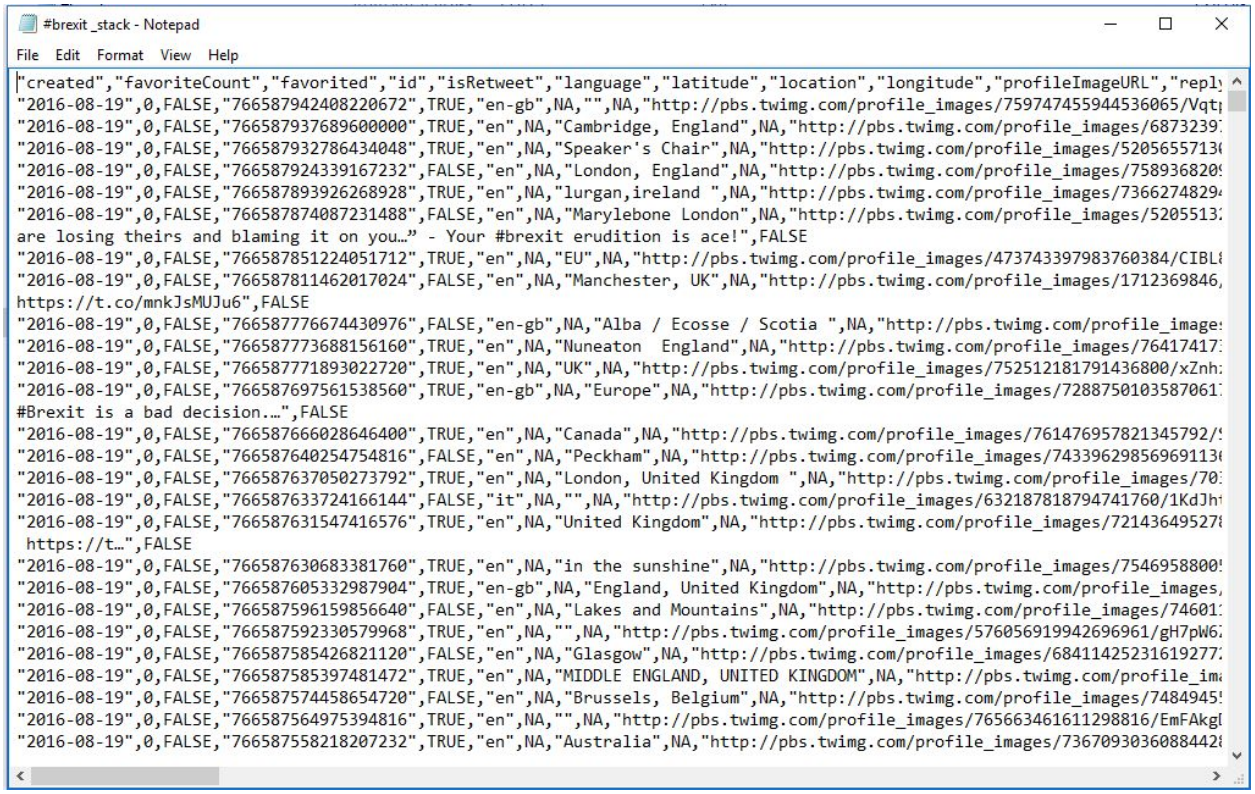
Following is the app we created and used in our project.

On creation of the app we are provided with Consumer Keys and Access Tokens. These keys will duly authenticate our R script, enabling us to use 'twitteR' package API for extracting tweets by using 'searchTwitter()' function.

Using '#brexit' to search tweets (Corpus Creation): A corpus can be defined as a collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular natural language or language variety. Corpora provide a material basis and a test bed for building NLP systems. By using #brexit tag we have retrieved the data about brexit from twitter. This data is transformed into corpus.

```
s <- searchTwitter('#brexit', n = 15000, since = '2016-07-23',  locale = 'en', geocode = '51.507351,-0.127758,100mi')
```

This function will create a Comma Separated Value File (.csv) file which will be used for further processing.

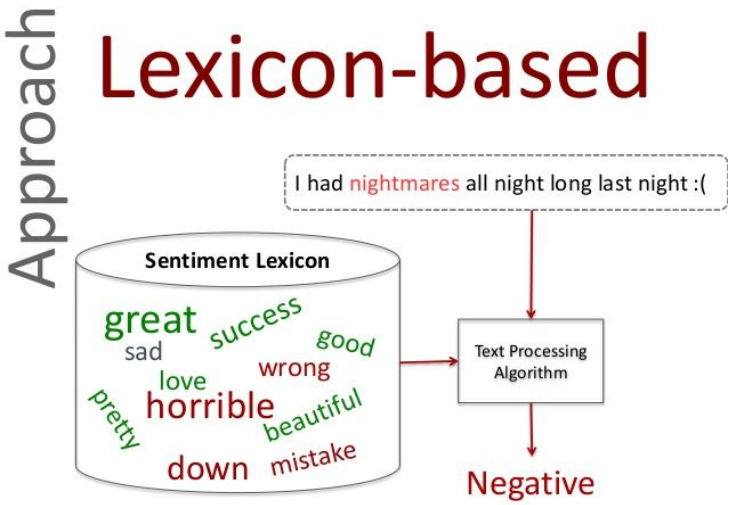


Other tweet processing methods: The Corpus is then treated with certain Cleaning Functions, i.e. Removing 'at(@) mentions', deleting duplicate tweets, removing Punctuations, followed by converting text into Lowercase, and Tokenization of the entire corpus.

Sentiment Analysis

Importance of Opinion Lexicon: Opinion Lexicon is a collection of words which have been determined as either positive or negative, with scores of each and every word in terms of polarity. The general idea is to calculate a sentiment score for each tweet so we can know how positive or negative is the posted message.

negative-words	3/12/2011 11:46 AM	Text Document	46 KB
positive-words	3/12/2011 11:46 AM	Text Document	21 KB

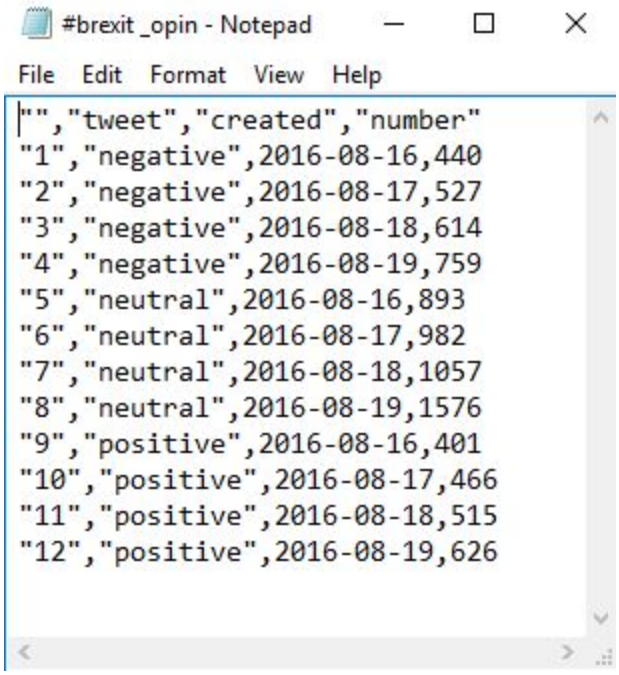


Using Opinion Lexicon for Scoring tweets:

Tokens, previously generated, will all be judged against the lexicon and each tweet will be scored on this basis. This will allow better classification in terms of statistics since tweets with negative impact on the sentiment will have lower scores compared to positively oriented tweets. These scores are then saved into a .csv file as well.

```
write.csv(scores, file = paste(searchterm, '_scores.csv')
```

Classifying processed tweets into Sentiments: Once all the scores are obtained, the corpus is classified with respect to Negative tweets, Positive tweets and Neutral tweets, each with individual scores, which is stored in opin.csv.



Data Analysis

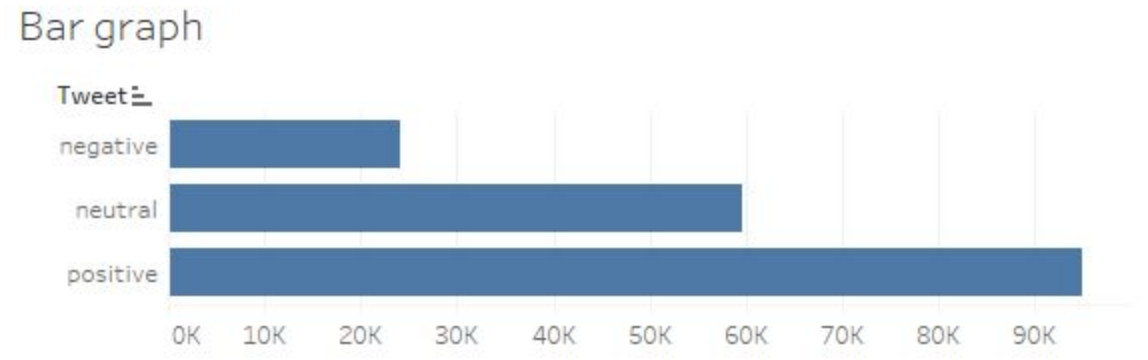
Introduction to Tableau: Tableau is a tool that can be described as an instrument that combines interactive visual analytics, sophisticated visualization methods, an easy to use graphical user-interface and a lot of connectors for dozens of different data sources ranging from standard Excel to more “Big Data” like things such as different Hadoop distributions and NoSQL databases. What is really impressing is that you can create interactive visualizations in just no time.

Representing the Data in charts: A picture is worth more than a thousand words. So it is better to represent data in forms of charts, as it makes data more expressive and efficient to read.

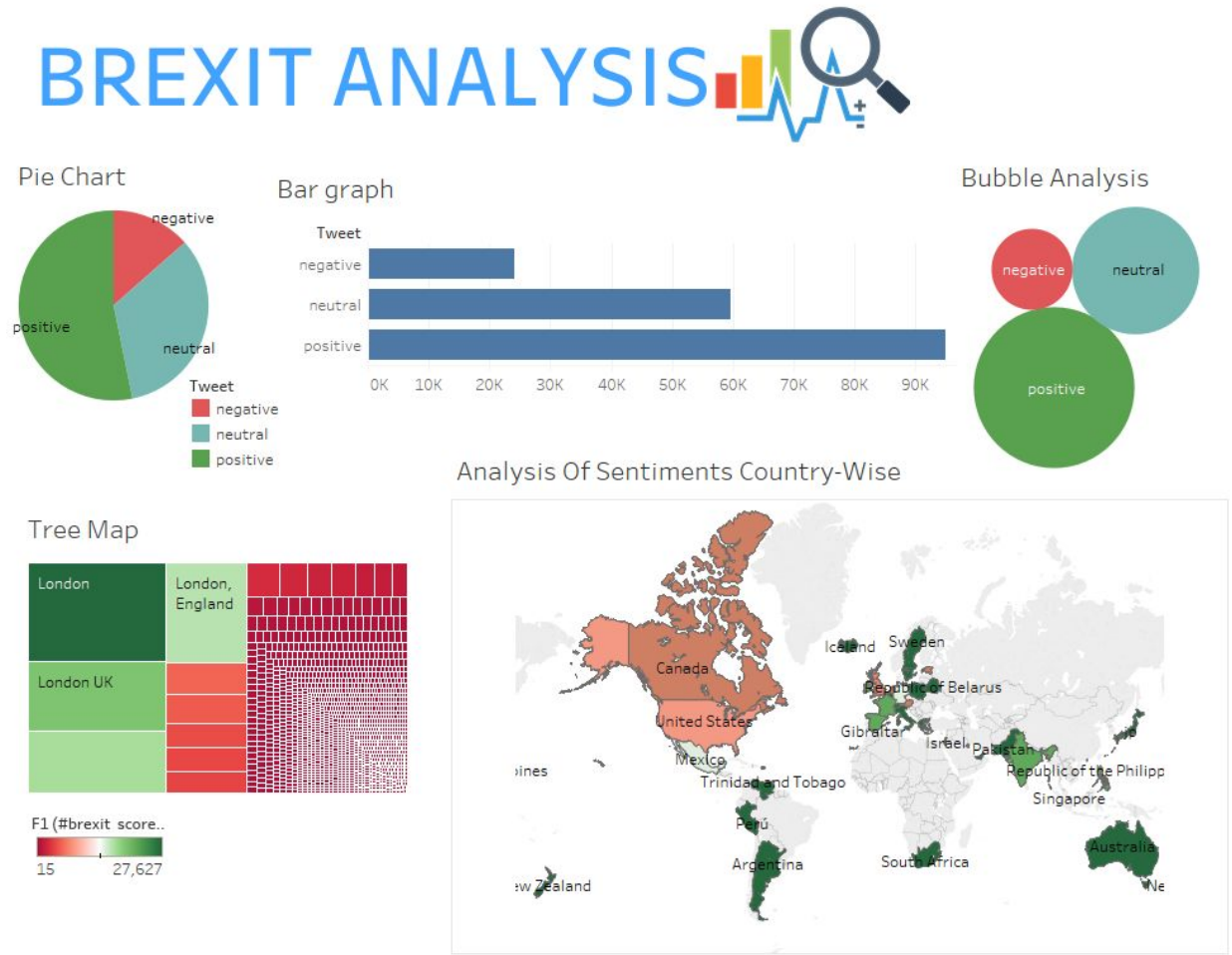
We have used tableau, an excellent visual analytics tool.

We have depicted sentiments using bar graph, pie chart and bubble chart as shown by creating individual sheets.

Further to analyse tweets geographically by location, we used a tree map and map analysis



Creating Dashboard:



Conclusion

- 53.17% and 13.49% of tweets exhibit a positive and negative response respectively about BRExit and its impact
- 33.33% tweets are analyzed to be neutral
- Countries like USA, Canada, and parts of Europe including UK itself and Netherlands etc scale more negative responses over positive