International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 379-383
e-ISSN:2278-621X

TEXT MINING AND SENTIMENT ANALYSIS USING CLOUD PLATFORMS A NOVEL APPROACH FOR TWITTER AND TEXT ANALYTICS

Arun Jose[1]

Abstract- To obtain patterns from unstructured and semi-structured data using R and API's for Text Analytics. In this paper, we are focus on Text mining and sentiment Analysis on data. We use specific Text Analytics API and packages for the analysis of data. Our main aim is to mine text from text documents or text file and check whether a given text is positive or negative, also charts and plots are used to visualize the data obtained from the analysis.
Keywords – API, Sentiment Analysis, Opinion mining, Text analytics, Text mining

1. INTRODUCTION

Text Mining, also called text data mining is used to find meaningful information from the larger database or dataset. The Text mining deals with unstructured data and semi-structured data. Text mining is a process in which larger quantities of natural language text is analyzed and to detects lexical patterns from the text to extract meaningful information. Sentiment analysis or opinion mining is the process in which identifying and detecting subjective information using natural language processing, text analysis, and computational linguistics. In short, the aim of sentiment analysis is to extract information on the attitude of the writer or speaker towards a specific topic or the total polarity of a document.

2. PROPOSED SYSTEM

2.1 Using Microsoft Azure and Cognitive service

2.1.1 Created Text Analytics API

We use Text Analytics API to analyze sentiment, extract key phrases, and detect language for any kind of text. In order to use Text Analytics API, we need to create Cognitive service in Azure.

2.1.2 Created a Power BI Streaming Dataset

The Streaming Dataset allow the user to get data from tweeter and result of sentiment analysis. It would be used for Data Visualization.

2.1.3 Created a connection using Microsoft Flow

The Microsoft Flow extract twitter feeds from Twitter and is connected to Cognitive service in order to search text from the twitter, we need to input a keyword for the Twitter data extraction. Finally, the output is inputted into Power BI Streaming Dataset using Microsoft Flow.

2.1.4 Report generation of Twitter and Sentiment analysis of each tweet

A detailed report is generated using Power BI from streaming dataset. I gave a sentiment scale 0 to 1 value. The output of the data would be either Positive, Negative or Neutral.

2.2 Using IBM Bluemix, Node-Red and DSX

2.2.1 Created Node-Red Service and Cloudant NoSQL in Bluemix

Node-Red and Cloudant Database service are created from Boilerplates container. Here we use Twitter API, Sentiment, function to get values from message payload and stored to CloudantDB.

2.2.2 Created user credential from Cloudant NoSQL Dashboard

Access credential is created in order to access from Text Analytics Tools. In this case, we use Rstudio in DSX.

2.2.3 Configuration in DSX RStudio

Installed R package 'sofa' to get access to CloudantDB and fetch relevant data for the analysis.

[1] MCA V semester, St. Aloysius College, AIMIT, Mangalore, Karnataka, India

2.2.4 Data Extraction of specified field from NoSQL

Tweeter data is extracted by using CloudantDB R query function and is stored in data frame.

2.2.5 Text mining and Sentiment analysis from the tweet

The Tweeter data is obtained with the help of Node-Red. Sentiment node will extract tweeter data to tokens, then we will take words from it. Again, we will categorize the words into positive and negative. The sentiment node module uses the AFINN-165 wordlist and Sentiment Emoji Ranking to perform a task.

3. EXPERIMENT AND RESULT

3.1 implementation in IBM BlueMix and DSX for Twitter Analysis

	Tweets....tweets	Score....sentimentscore	Sentiment
1	RT @ACLU: President Donald Trump today provided t...	1	Positive
2	RT @crusher614: Would you be in favor of a Donald T...	2	Positive
3	RT @jasonsfolly: Dizzy with joy. Like finding out Sant...	3	Positive
4	RT @Pappiness: Dear Donald Trump, Please don't kill ...	-1	Negative
5	RT @PalmerReport: Donald Trump makes panic move...	-4	Negative

Fig 1

In the above diagram, there are five tweets and their sentiment score.The sentiment score scale is in between -5 to 5 if the value of sentiment score is Zero it means the tweet is neutral.

3.2 Token words

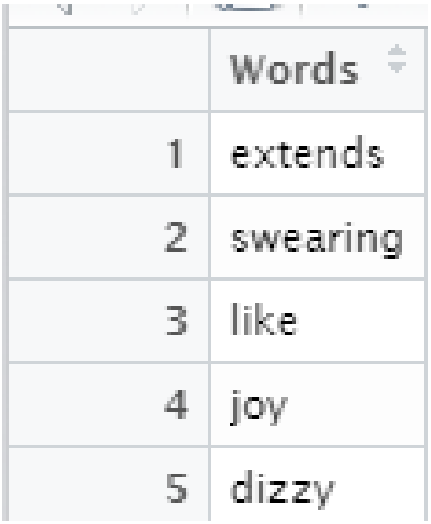
	Token
1	rt
2	jasonsfolly
3	dizzy
4	with
5	joy
6	like
7	finding
8	out
9	santas
10	real
11	bob
12	marley
13	has
14	a
15	new
16	album
17	and
18	swearing

Showing 1 to 19 of 21 entries

Fig 2

These are some of the token words that are derived from a selected tweet (fig. 2).it gives words that are used in the text.The token words eliminate symbols, number and other special characters from the text.

3.3 English Words

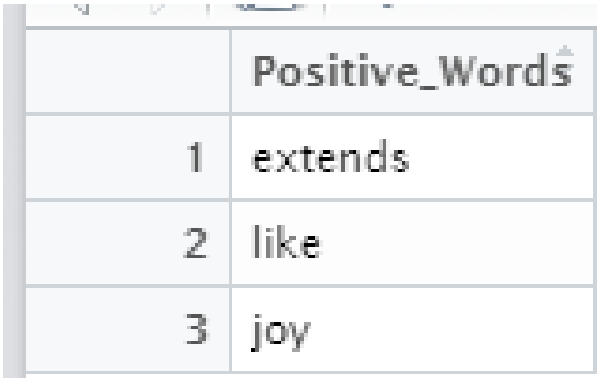


	Words
1	extends
2	swearing
3	like
4	joy
5	dizzy

Fig 3

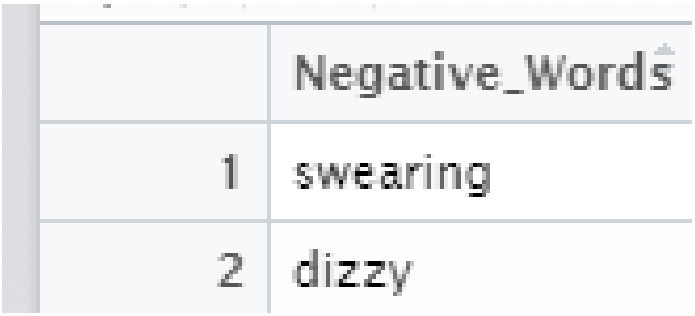
These are some of the English words that are taken from the word token.For every tweet, it will take English words and eliminate all other words that are not much related to it.

3.4 Positive and Negative Words



	Positive_Words
1	extends
2	like
3	joy

Fig 4



	Negative_Words
1	swearing
2	dizzy

Fig 5

These are some of the positive and negative words that are derived from the word list.According to sentiment module, it will rank for positive and negative words in order to get the sentiment score.

3.5 Sentimental Analysis in Microsoft Power BI

TweetBy	Tweet	Average of Retweetcount	Average of Score	sentiment
___RocioMartin	RT @DavidPapp: Donald Trump Wants to Spend Big on Anti-Drug Ads. Here's Why They Don't Work https://t.co/O4r4mA6lZO	5.00	0.14	Negative
__aireanna	RT @RobMRosenberg: Nov. 8th, 2016: We can now project Donald J. Trump will be the 45th President of the United States. Statue of Liberty: h...	8.00	0.50	Neutral
__Howard__1	RT @wesley_jordan: Donald J Trump is a repeat sex offender who calls himself the victim & his victims liars. It's time he's held accountabl...	228.00	0.00	Negative
__J_A_E__	RT @PapiiQuann_: B spent 10 million for this ad to get Donald Trump impeached Yall better Re fucking tweet for the respect htt...	58,736.00	0.50	Neutral
__JoeKool__	RT @DavidPapp: Donald Trump Wants to Spend Big on Anti-Drug Ads. Here's Why They Don't Work https://t.co/O4r4mA6lZO	11.00	0.14	Negative
_AmyanimalLover	RT @MichaelSkolnik: Donald J. Trump is a co-founder of the #WhiteLivesMatter movement.	299.00	0.50	Neutral
_Andrewadams11	RT @JacobAWohl: @realDonaldTrump No administration in history has been more transparent than the Donald J. Trump Administration!	629.00	0.96	Positive
Blessed07	RT @DavidJHarrisJr: Donald J. Trump everyone!!! https://t.co/hJvGDyMpac	2.00	0.50	Neutral
_cbackus	RT @wesley_jordan: Donald J Trump is a repeat sex offender who calls himself the victim & his victims liars. It's time he's held accountabl...	351.00	0.00	Negative
_CLewandowski	RT @Scavino45: Statement by President Donald J. Trump on the Apprehension of Mustafa al-Imam for His Alleged Role in the 9/11/12 Attacks in...	3,299.00	0.50	Neutral
_comatose1	RT @realDonaldTrump: "Statement by President Trump on the Apprehension of Mustafa al-Imam for His Alleged Role in Benghazi Attacks" https:/...	7,960.00	0.50	Neutral
_ecwagner	RT @joncoopertweets: It's been a long time comin' And the table's turned around 'Cause one of us is goin' One of us is goin' down	223.00	0.50	Neutral
Total		**2,244.64**	**0.51**	**Neutral**

Fig 6

The above given is the Sentiment analysis computed using Power BI, It gives a clear picture of the tweet, sentiment score obtained by each tweet.The sentiment score is given within a scale of 0 to 1. In this case, 0.50 is neutral, Zero means negative and one gives positive.Here the average of a score is 0.51 and so it means the search text "Donald Trump" is neutral in the twitter.

3.6 Word Frequency table in DXS from a Text document

	word	freq
will	will	56
american	american	30
america	america	28
must	must	20
new	new	19
country	country	17
world	world	17
and	and	16
americans	americans	15
people	people	15
but	but	14

Fig 7

In this given diagram it describes the word frequency, the number of times words occurred in the text file is counted in the word frequency table.Here "will" word is mostly used in the text file.

3.7 Word cloud using Word frequency table

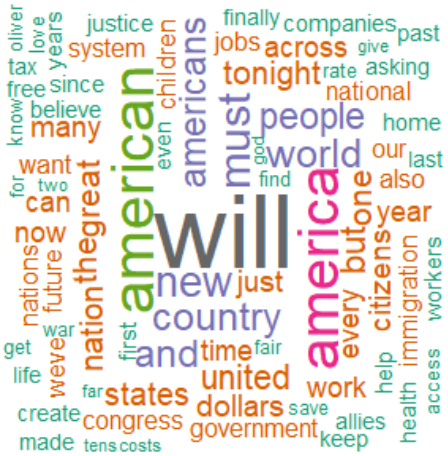


Fig 8

The word cloud is a visualization of word frequency table, Here the word "will" is highlighted from the frequency table and secondly word "american".The detailed coloring is given according to the frequency of the words.

4. CONCLUSION

In this research paper, we discuss how we can use cloud-based for text mining and sentiment analysis. With the use of Cognitive service and advance analytical tools like Microsoft Power BI and RStudio for the data visualization, analysis. We also use back-end connectors like Node-Red and Microsoft flow online flow based UI to connect cloud-based services to Bluemix and Azure. We found that impact of sentiment analysis will predict whether a text document is positive or negative based on the word score that is obtained from the sentiment module. We believe that studies on Text mining and Sentiment analysis give a deep root in the area of research and emotional attitude of the person towards a written document or similar data.

We assume that Text mining and Sentiment analysis will become a part of future application and services. The adaptation of Text mining and Sentiment analysis techniques will standardize the future products for more accurate and meaningful information, also we can infer that research in the area of Text mining and sentiment analysis can provide considerable advancement to natural language processing, computational linguistics, and AI.

5. REFERENCES

[1] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 246–253

[2] M. Q. Patton, Qualitative evaluation and research methods. Sage Publishers, 1990

[3] Feinerer I, Hornik K (2015). tm: Text Mining Package. R package version 0.6-2, URL https: //CRAN.R-project.org/package=tm

[4] Williams GJ (2011). Data Mining with Rattle and R: The art of excavating data for knowledge discovery. Use R! Springer, New York.

[5] Hansen KD, Gentry J, Long L, Gentleman R, Falcon S, Hahne F, Sarkar D (2016). Rgraphviz: Provides plotting capabilities for R graph objects. R package version 2.12.0.