**Twitter Sentiment Analysis**

SENTIMENT ANALYSIS

Sentiment Analysis is the process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral. It’s also known as **opinion mining**, deriving the opinion or attitude of a speaker. It is the process of finding out the sentiment of a particular sentence/paragraph/text.

Examples

* I do not dislike cabin cruisers. ([Negation](https://en.wikipedia.org/wiki/Negation) handling)
* Disliking watercraft is not really my thing. (Negation, inverted [word order](https://en.wikipedia.org/wiki/Word_order))
* Sometimes I really hate [RIBs](https://en.wikipedia.org/wiki/Rigid-hulled_inflatable_boat). ([Adverbial](https://en.wikipedia.org/wiki/Adverbial) modifies the sentiment)
* I'd really truly love going out in this weather! (Possibly [sarcastic](https://en.wikipedia.org/wiki/Sarcastic))

The data sources for sentiment analysis includes the following:

* Twitter tweets
* Facebooks comments
* Online reviews
* Sms/emails/messages
* Novels, books
* Any other text based sources like books/newspapers/magazines/etc

What is Sentiment Analysis?

Sentiment essentially relates to feelings; attitudes, emotions and opinions. Sentiment Analysis refers to the practice of applying Natural Language Processing and Text Analysis techniques to identify and extract subjective information from a piece of text. A person’s opinion or feelings are for the most part subjective and not facts. Which means to accurately analyze an individual’s opinion or mood from a piece of text can be extremely difficult. With Sentiment Analysis from a text analytics point of view, we are essentially looking to get an understanding of the attitude of a writer with respect to a topic in a piece of text and its polarity; whether it’s positive, negative or neutral.

Why Text Processing using R?

With the increasing importance of computational text analysis in research , many researchers face the challenge of learning how to use advanced software that enables this text analysis. Currently, one of the most popular environments for computational methods and the emerging field of “data science” is the R statistical software. However, for researchers that are not well-versed in programming, learning how to use R can be a challenge, and performing text analysis in particular can seem daunting. In this step by step guide one can learn that performing text analysis in R is not as hard as some might fear.

This is an introduction into the use of common techniques, with the aim of helping researchers get acquainted with computational text analysis in general, as well as getting a start at performing advanced text analysis studies in R.

Why R?

R was specifically designed for statistical analysis, which makes it highly suitable for data science applications. Although the learning curve for programming with R can be steep, especially for people without prior programming experience, the tools now available for carrying out text analysis in R make it easy to perform powerful, cutting-edge text analytics using only a few simple commands. One of the keys to R’s explosive growth has been its densely populated collection of extension software libraries, known in R terminology as packages, supplied and maintained by R’s extensive user community. Each package extends the functionality of the base R language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package. The best-known package repository, the Comprehensive R Archive Network (CRAN), currently has over 10,000 packages that are published.

Why Twitter Data?

Twitter is an online microblogging tool that disseminates more than 400 million messages per day, including vast amounts of information about almost all industries from entertainment to sports, health to business etc. One of the best things about Twitter — indeed, perhaps its greatest appeal — is in its accessibility. It’s easy to use both for sharing information and for collecting it. Twitter provides unprecedented access to our lawmakers and to our celebrities, as well as to news as it’s happening. Twitter represents an important data source for the business models of huge companies as well.

WHAT IS TWITTER SENTIMENT ANALYSIS?

Twitter is an online news and social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but those who are unregistered can only read them. Twitter Sentiment Analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention.

WHY TWITTER SENTIMENT ANALYSIS?

The applications for sentiment analysis are endless. ​It is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics However, it is also practical for use in business analytics and situations in which text needs to be analysed.

Sentiment analysis is in demand because of its efficiency. Thousands of text documents can be processed for sentiment in seconds, compared to the hours it would take a team of people to manually complete. Because it is so efficient (and accurate – Semantic has 80% accuracy for English content) many businesses are adopting text and sentiment analysis and incorporating it into their processes. Applications: The applications of sentiment analysis are broad and powerful. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

For example, the Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages a head of 2012 presidential election.

The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television adverts

Tweets are imported using R and the data is cleaned by removing emoticons and URLs. Lexical Analysis is used to predict the sentiment of tweets and subsequently express the opinion graphically through ggplots, histogram, pie chart and tables.

SYSTEM REQUIREMENTS

The following are the major requirements for performing sentiment analysis is mentioned below:

Installation of R Studio.

Twitter Authentication to access API.

FEATURES

1. Extraction of Tweets

(i) Create twitter application

(ii) twitteR - Provides an interface to the Twitter web API

(iii) ROAuth - ​R Interface For OAuth

(iv) ​Create twitter authenticated credential object, It is done using consumer key, consumer secret, access token, access secret.

(v) During authentication, we are redirected to a URL automatically where we click on Authorize app as shown in the image below and enter the unique 7-digit number to get linked to the account from which feeds are being taken.

​2. Cleaning Tweets

The tweets are cleaned in R by removing:

● Extra punctuation

● Stop words (Most commonly used words in a language like​ the, ​is, ​at, which, and, ​on,.)

● Redundant Blank spaces

● Emoticons

● URLS

3. Loading Word Database

A database, created by Hui Lui containing positive and negative words, is loaded into R. This is used for Lexical Analysis, where the words in the tweets are compared with the words in the database and the sentiment is predicted.

For movie tweets, Naive Bayes Machine Learning Algorithm is used. AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated. The version used is: AFINN-111: Newest version with 2477 words and phrases.

4. Algorithms used

● Lexical Analysis:​ ​By comparing uni-grams to the pre-loaded word database, the tweet is assigned sentiment score - positive, negative or neutral and overall score is calculated.

5. Calculating percentage

​I have presented the scores, the tweets as well as the percentage of positive/negative emotion in the text. This calculated using simple arithmetic to understand the overall sentiment in a more better manner

6. Histogram tab : histogram plot

Histograms of positive, negative and overall score are found under the Histogram tab for graphically analyzing the intensity of emotion in the tweeters.

PACKAGES USED

● twitteR​: Provides an interface to the Twitter web API

● stringr​: String operations in R

● ROAuth​: Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice.

● RCurl​: Provides functions to allow one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server.

● ggplot2​: An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources.

● reshape​: Flexibly restructure and aggregate data using just two functions: melt and cast

● tm ​: A framework for text mining applications within R.

● RJSONIO​: This is a package that allows conversion to and from data in JavaScript object notation (JSON) format. This allows R objects to be inserted into JavaScript code and allows R programmers to read and convert JSON content to R objects

● plyr​: Tools for Splitting, Applying and Combining Data

LIMITATIONS

1. The Twitter Search API can get tweets up to a maximum of 7 days old.

2. Cannot get 100% efficiency in analysing sentiment of tweets.

FUTURE WORK

* Detect sarcasm in tweets
* Analyse images for emotions.
* Add Telugu words to dataset.
* Find no of mentions of n particular organizations.
* Parallelizing code.
* Apply better Machine Learning Algorithms (Like Support Vector Machine).

REFERENCES

1. Gottschalk, Louis August, and Goldine C. Gleser. The measurement of psychological states through the content analysis of verbal behavior. Univ of California Press, 1969.
2. [USA Issued 7,136,877](http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=2&p=1&f=G&l=50&d=PTXT&S1=(fogel.INNM.+AND+volcani.INNM.)&OS=in/fogel+and+in/volcani&RS=(IN/fogel+AND+IN/volcani)), Volcani, Yanon; & Fogel, David B., "System and method for determining and controlling the impact of text", published June 28, 2001.
3. Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics. pp. 417–424. [arXiv](https://en.wikipedia.org/wiki/ArXiv):[cs.LG/0212032](https://arxiv.org/abs/cs.LG/0212032).
4. Pang, Bo; [Lee, Lillian](https://en.wikipedia.org/wiki/Lillian_Lee_(computer_scientist)); Vaithyanathan, Shivakumar (2002). ["Thumbs up? Sentiment Classification using Machine Learning Techniques"](http://www.cs.cornell.edu/home/llee/papers/sentiment.home.html). Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.
5. Pang, Bo; Lee, Lillian (2005). ["Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales"](http://www.cs.cornell.edu/home/llee/papers/pang-lee-stars.home.html). Proceedings of the Association for Computational Linguistics (ACL). pp. 115–124.
6. Snyder, Benjamin; Barzilay, Regina (2007). ["Multiple Aspect Ranking using the Good Grief Algorithm"](http://people.csail.mit.edu/regina/my_papers/ggranker.ps). Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). pp. 300–307.
7. Qu, Yan, James Shanahan, and Janyce Wiebe. "Exploring attitude and affect in text: Theories and applications." In AAAI Spring Symposium) Technical report SS-04-07. AAAI Press, Menlo Park, CA. 2004.
8. Vryniotis, Vasilis (2013). [The importance of Neutral Class in Sentiment Analysis](http://blog.datumbox.com/the-importance-of-neutral-class-in-sentiment-analysis/).
9. Koppel, Moshe; Schler, Jonathan (2006). "The Importance of Neutral Examples for Learning Sentiment". Computational Intelligence 22. pp. 100–109. [CiteSeerX](https://en.wikipedia.org/wiki/CiteSeerX" \o "CiteSeerX) [10.1.1.84.9735](https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.9735).

R Code

#Used Libraries

library(RColorBrewer)

library(wordcloud)

library(tm)

library(twitteR)

library(ROAuth)

library(plyr)

library(stringr)

library(base64enc)

library(SnowballC)

library(ggplot2)

library(maps)

#Inserting Access Keys

consumerKey <- "eOtcK9qJpjcObGUPv6SfTFZFE"

consumerSecret <- "TedGta8dDigiQcpKJ68AleIlfaHL7nqdVdCH2qppgD7wLYE0Tm"

accessToken <- "896758586243653632-6RCrl6Yia8uR0YbjBor38RxbjCWCqFm"

accessTokenSecret <- "QYLajXrqdbMYFKwno9itPNo4LVPsVuccOBSyxzfMfdCUF"

requestURL<- "https://api.twitter.com/oauth/request\_token"

accessURL<- "https://api.twitter.com/oauth/access\_token"

authURL<- <https://api.twitter.com/oauth/authorize>

#Setting Up Authentication

setup\_twitter\_oauth(consumerKey,consumerSecret,accessToken,accessTokenSecret)

#Search Twitter

users<- searchTwitteR("@ravithejamuthu", resultType="recent",n=3000, lang="en")

#Converting into DataFrame

tweet.df = do.call("rbind",lapply(users,as.data.frame))

#Plotting data on map

map('world')

points(tweets.df$longitude,tweets.df$latitude, pch=20, cex=1, col="red")

#Viewing the data

View(tweet.df)

#Reading sentiment analysis data from Txt document

pos.words = scan('./positive-words.txt', what='character', comment.char=';')

neg.words = scan('./negative-words.txt', what='character', comment.char=';')

#Appending some more words to actual words

pos.words = c(pos.words, 'new','nice' ,'good', 'horizon')

neg.words = c(neg.words, 'wtf', 'behind','feels', 'ugly', 'back','worse' , 'shitty', 'bad', 'no','freaking','sucks','horrible')

#Converting Into DataFrame

test <-ldply(users,function(t)t$toDataFrame())

#Calculating Result

result <- score.sentiment(test$text,pos.words,neg.words)

#Summarizing Data

summary(result$score)

#Histogram

hist(result$score,col="yellow", main="Score of tweets",ylab=" Count of tweets")

#Count No of Tweets

count(result$score)

#Plotting the tweets on qplot

qplot(result$score,xlab = "Score of tweets")

#Score Sentiment Function

#Used to remove all unwanted data

score.sentiment = function(sentences, pos.words, neg.words, .progress='none')

{

require(plyr)

require(stringr)

scores = laply(sentences, function(sentence, pos.words, neg.words) {

sentence = gsub('[[:punct:]]', '', sentence)

sentence = gsub('[[:cntrl:]]', '', sentence)

sentence = gsub('\\d+', '', sentence)

sentence = tolower(sentence)

word.list = str\_split(sentence, '\\s+')

words = unlist(word.list)

pos.matches = match(words, pos.words)

neg.matches = match(words, neg.words)

pos.matches = !is.na(pos.matches)

neg.matches = !is.na(neg.matches)

score <- sum(pos.matches) - sum(neg.matches)

return(score)

}, pos.words, neg.words, .progress=.progress )

scores.df = data.frame(score=scores, text=sentences)

return(scores.df)

}