Human Sentiment Analysis On Social Media

Bachelor Thesis

By
Shruti Sharma(CSE/16037)
Shikha Sinha (CSE/16035)

Project thesis submitted to

Indian Institute of Information Technology, Kalyani

For the partial fulfillment of the degree of

Bachelors of Technology in Computer Science & Engineering

November, 2018

Certificate

This is to certify that the thesis entitled “Human Sentiment Analysis on Social Media” being submitted by Shruti Sharma and Shikha Sinha, undergraduate students (Regis No. 0000192 and 0000190 respectively) in the Department of Computer Science, Indian Institute of Information Technology, Kalyani, India, for the award of Bachelors of Technology in Computer Science, is an original research work carried by him under my supervision and guidance. The thesis has fulfilled all the requirements as par the regulation of IIIT Kalyani and in my opinion, has reached the standards needed for submission. The works, techniques and the results presented have not been submitted to any other university or Institute for the award of any other degree or diploma.

(Sanjoy Pratihar, Ph.D)
Assistant Professor
Department of Computer Science
Indian Institute of Information Technology
Kalyani, Nadia, India.

Declaration

I hereby declare that the work being presented in this project report entitled, "Human Sentiment Analysis", submitted to Indian Institute of Information Technology Kalyani in partial fulfilment for the award of the degree of Bachelor of Technology in Computer Science and Engineering during the period from July, 2018 to November, 2018 under the supervision of Prof. Sanjoy Pratihar, Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, West Bengal 741235, India, does not contain any classified information.

Shikha Sinha (CSE/16035/190)
Shruti Sharma(CSE/16037/192)
Department of Computer Science
and Engineering
Indian Institute Of
Information Technology,
Kalyani,Webel IT Park,Bengal,India
Pin code-741235

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

.....
(Prof. Sanjoy Pratihar)
Assistant Professor
Departmental of Computer Science and Engineering
Indian Institute of Information Technology Kalyani,
Webel IT Park, West Bengal 741235

Acknowledgement

Firstly, we would like to thank our supervisor Prof. Sanjoy Pratihar, for his support and guidance to complete this project work without whom we would not be able to make out this far. We would also like to thank our friends who supported us greatly and were always willing to help us. We are very grateful to Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, West Bengal 741235, India, for providing us this wonderful opportunity. Lastly, we would like to thank our parents and God for their never ending grace.

Shikha Sinha (CSE/16035/190)
Shruti Sharma(CSE/16037/192)
Department of Computer Science
and Engineering
Indian Institute Of
Information Technology,
Kalyani,Webel IT Park,Bengal,India
Pin code-741235

Abstract

The goal of this project is to show how sentimental analysis can help improve the user experience over a social network. The algorithm will learn what our emotions are from the text and then help us make predictions about our mood. This knowledge will change our social activities accordingly in order to be rational with our feelings. As per our survey, people move towards social media when they are bored, in need of company so after learning from this algorithm , it can suggest appropriate options to them to lighten their mood. It can reduce the impact of those which the mood worse. The project aims to implement these in the social network community for making our lives better and our experience richer and efficient.

Introduction

Sentiment Analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer services to clinical medicine.

Sentiment Analysis builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.:

- *Polarity*: if the speaker express a *positive* or *negative* opinion,
- *Subject*: the thing that is being talked about,
- *Opinion holder*: the person, or entity that expresses the opinion.

Currently, sentiment analysis is a topic of great interest and development since it has many practical applications. Since publicly and privately available information over Internet is constantly growing, a large number of texts expressing opinions are available in review sites, forums, blogs, and social media.

With the help of sentiment analysis systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics, or any topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, product feedback, and customer service.

Social Sentiment Analysis is an algorithm that is tuned to analyze the sentiment of social media content, like tweets and status updates. The algorithm takes a string, and returns the sentiment rating for the “positive,” “negative,” and “neutral.” In addition, this algorithm provides a compound result, which is the general overall sentiment of the string.

Methodology

This project of analyzing sentiments of tweets comes under the domain of "Pattern Classification" and "Data Mining". To implement sentiment analysis in our project we took a step forward by not classifying the tweets simply positive or negative but also associating different emotions to it like happy, anger, anticipation, surprise etc. Every-time the algorithm runs, we create a new corpus in a .csv (comma separated value) file. This constitutes the data extraction step. Next, we make the collected data appropriate for our use by removing those data-fields which do not serve our purpose. The data provided comes with emoticons, usernames, URL's, references and hashtags which are required to be processed and converted into a standard form. The words are also a mixture of misspelled words / incorrect, extra punctuations, and words with many repeated letters. Therefore, tweets must be preprocessed to standardize the dataset. This completes text mining. We then extract the replies to a particular tweet from our corpus and use the sentiment of the replies to deduct the sentiment of that tweet. It is mainly a content-based classification problem used in Natural Language Processing and Machine Learning.

Tools used :

- R Sentiment Analysis tools
- R Studio
- Twitter Timeline

For Twitter Analysis, the proposed system has the following steps involved:

- Creating Twitter Application.
- Execute Twitter API code through R-Studio.
- Collecting Twitter data.
- Classifying the data with R-Tool commands.
- Running R commands for processing the tweets.
- Establishing R Plotter to view results.

The plan is to use R's rich Sentiment Analysis library to program our analyser. Our model will use Syuzhet which has built in four sentiment dictionaries and also provides tools for sentiment extraction tools developed under NLP. Sentiment analysis is performed as an intersection of a term-document (built from the mined text) and a lexicon of choice (provided by Syuzhet).

The get_nrc_sentiment() function from the Syuzhet library.This function takes in new_sentence and compares it with the nrc emotion lexicon to return the scores.

We collected data by using by making Twitter application for accessing Twitter API and getting the required OAuth permissions. We will use this data as the training data. R-Studio is the environment developed for statistical analysis and a Graphical view of the large data sets.
Using this data and the replies generated to it, we can determine the polarity of the text which further determines the polarity assigned to the person that data belongs to.

Implementation (in R)

Some used Function:
get_nrc_sentiment() : Get Emotions and Valence from NRC Dictionary
Description Calls the NRC sentiment dictionary to calculate the presence of eight different emotions and their corresponding valence in a text file. Usage get_nrc_sentiment(char_v, cl = NULL, language = "english") Arguments char_v A character vector cl Optional, for parallel analysis language A string Value A data frame where each row represents a sentence from the original file. The columns include one for each emotion type as well as a positive or negative valence. The ten columns are as follows: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative", "positive."

Code:

```
library(twitteR)        #extract tweets and followers from Twitter
library(syuzhet)        #sentiment analysis library
library(ggplot2)        #Visualization

#FOR SETTING UP CONNECTION TO TWITTER ACCOUNT  FOR GETTING INFO.
setup_twitter_oauth("consumer_key”,"consumer_secret
","access_token","access_secret")
tweets <- userTimeline("@elonmusk" , n=200, includeRts=FALSE)          #max limit is 3200


#FORMATTING OF TEXT DATA
tweets.df <- twListToDF(tweets)
tweets.df <- gsub("http.*","",tweets.df$text)
tweets.df <- gsub("https.*","",tweets.df)
```

```
tweets.df <- gsub("#.*","",tweets.df)
iconv(tweets$text, from="UTF-8", to="ASCII", sub="")
write.csv(tweets.df,"/Users/shruti/Documents/data/tweets.csv")

#FOR SORTING ON THE BASIS OF DATE AND TIME
y=vector()
for(i in 1:length(tweets.df))
{
   x <- as.POSIXct(tweets[[i]][["created"]], format='%Y-%m-%d %H:%M:%S')
   y[i] <- strsplit(format(x, "%Y-%m-%d %H:%M:%S"), ' ')
}

tweets <- twListToDF(tweets)  # converting tweets list to DataFrame
queryString = paste0("to:","@elonmusk")     # building queryString to fetch retweets
Id = tweets[15,"id"]                # retrieving tweet ID for which reply is to be fetched
rply = searchTwitter(queryString, sinceID = Id,lang = "en")   ## fetching all the reply  to
username
rply = twListToDF(rply)
rply <- gsub("http.*","",rply$text)
rply<- gsub("#.*","",rply)
word.df <- as.vector(rply)
emotion <- get_nrc_sentiment(word.df)
emotion1 <- cbind(rply, emotion)




##FOR BARGRAPH
barplot(
  sort(colSums(prop.table(emotion[, 1:10]))),
  horiz = TRUE,
  cex.names = 0.9,
  las = 1,
  main = "Emotions in text", xlab="Percentage"
)

##FOR DISTINGUISHING SENTIMENT
affect <- subset(affect_wordnet)
term_scores <- with(affect, unclass(table(term, emotion)))  #term_scores is a matrix
with #entry (i,j) indicating the number of times that term i appeared in the affect lexicon
with #emotion j.
ncat <- rowSums(term_scores > 0)
```

```
term_scores[ncat > 1, c("Positive", "Negative", "Ambiguous")] <- c(0, 0, 1)
#term appearing in two or more categories is ambiguous
term_scores[term_scores > 1] <- 1     #Replacing larger values by 1 for terms appearing in 2 #or more categories

chunks <- text_split(tweets, "tokens", 200)   #to segment texts into blocks, 2nd arg = unit, #3rd arg= max segment size
n <- text_ntoken(chunks)
x <- term_matrix(chunks, select = rownames(term_scores))
#For the count of each emotion category in each segment,we form a matrix of counts
#For the occurrence rates, we divide the counts by the segment sizes. We then multiply by #500 so that rates are given as occurrences per 500 tokens.
text_scores <- x %*% term_scores
unit <- 150
rate <- list(pos = text_scores[, "Positive"] / n * unit,
             neg = text_scores[, "Negative"] / n * unit,
             ambig = text_scores[, "Ambiguous"] / n * unit)
rate$total <- rate$pos + rate$neg + rate$ambig
se <- lapply(rate, function(r) sqrt(r * (unit - r) / n))   #binomial variance formula to get the standard errors

##GRAPH  ANALYSING  DIFFERENT SENTIMENTS
i <- seq_len(nrow(chunks))    # set up segment IDs


# set the plot margins, with extra space below the plot

par(mar = c(4, 4, 11, 9) + 0.1, las = 1)

# set up the plot coordinates; put labels but no axes

xlim <- range(i - 1, i + 1)
ylim <- range(0, rate$total + se$total, rate$total - se$total)
plot(xlim, ylim, type = "n", xlab = "Segment", ylab = "Rate \u00d7 500", axes = FALSE, xaxs = "i")
usr <- par("usr")

axis(1, at = i[i %% 5 == 0], labels = FALSE)
axis(1, at = i[i %% 10 == 0], labels = TRUE)

# defaults for the y axis
axis(2)

# put vertical lines at chapter boundaries
```

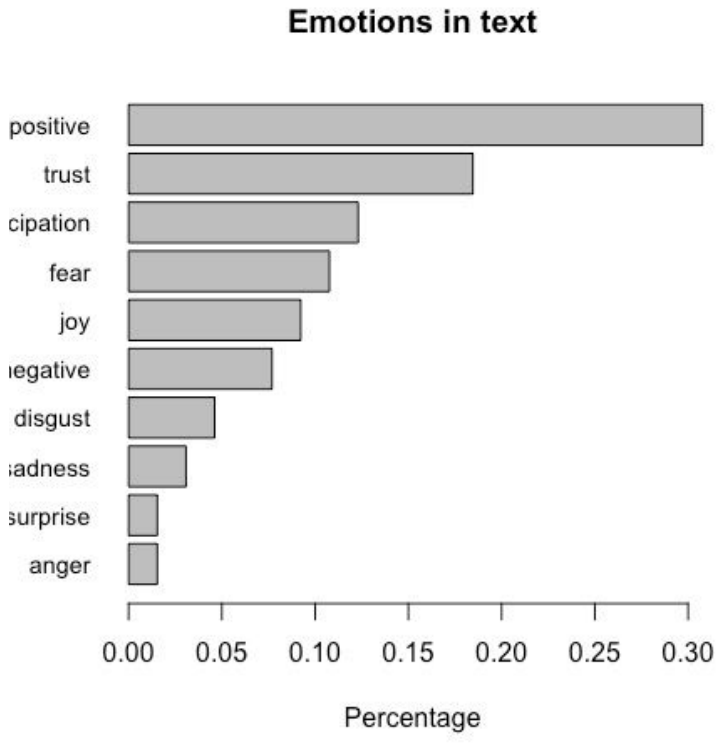
```
abline(v = tapply(i, chunks$parent, min) - 0.5, col = "gray")

box()   #frame the plot
col <- c( pos = "#FC8D62", neg = "#8DA0CB", ambig = "#66C2A5")       #colors     for
different emotions

#plot each rate type
for (t in c("ambig", "neg", "pos"))
{
  r <- rate[[t]]
  s <- se[[t]]
  cl <- col[[t]]

  # add lines and points
  lines(i,r, col = cl)
  points(i, r, col = cl, pch = 10, cex = 0.5)
}
```

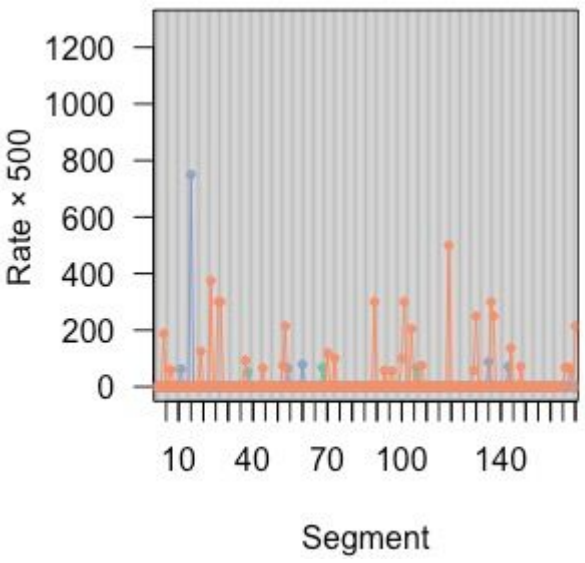
Result

The tweets are saved in a .csv file after running the code. It then gives following results. The tweets are classified as positive/negative sentiment and further into emotions like anger, disgust, anticipation, trust, joy, surprise etc. The accuracy of result may vary depending on the bias of dataset. A dataset having more of positive tweets will show more positive result.

Microsoft Excel - tweets

	A	B
1		x
2	1	@InsideEVs We will fight Big Tequila!
3	2	@garos82 @nichegamer That would be awesome
4	3	@jimmyzhong_iost @SpaceX Easily accessible energy
5	4	@nichegamer Send me a note!
6	5	@nichegamer Would be fun to add to Tesla
7	6	@MackenzieTulip @nextspaceflight @SpaceX @NASA Yes
8	7	@nextspaceflight @SpaceX @NASA Will try again next month
9	8	Congratulations @SpaceX team! Thanks @NASA, much appreciated.
10	9	Rest in peace, Stan Lee. The many worlds of imagination & delight you created for humanity will last forever.
11	10	@_sheateher @MalibuWine Is he ok?
12	11	It naturally follows that a hotter system will have more energetic events. For example, as a pot of water becomes c<U+2026>
13	12	@karaswisher Scary sign of times to come. It will get worse.
14	13	@kenlacovara Agreed, technically, mined hydrocarbons come primarily from ancient, decayed marine organisms & partia<U+2026>
15	14	Makes me so mad when smart, ethical scientists I know are accused of publishing climate papers for <U+201C>grant money<U+201D>. T<U+2026>
16	15	@SeanGVarney That is true & should be applauded. Right move is for oil companies truly to think of themselves as en<U+2026>
17	16	We know we<U+2019>ll run out of dead dinosaurs to mine for fuel & have to use sustainable energy eventually, so why not go<U+2026>
18	17	@zandywithaz Because people on the ground know more about what<U+2019>s actually needed than politicians do
19	18	@JackCydia Good, but not hospital grade. S & X were designed to be proof against an actual bioweapon attack. Requir<U+2026>
20	19	If Tesla can help people in California wildfire, please let us know. Model S & X have hospital grade HEPA filters.<U+2026>
21	20	@tizmagik @TeslaSupport Fair point, will reenable
22	21	@mihk101 Agreed, top priority for Autopilot team. Being cautious for max safety.
23	22	Please lmk what you<U+2019>d most like improved/fixed about your Tesla. Thanks!
24	23	@exzacklyright @MKBHD @austinnotduncan @Tesla Cool
25	24	@Tesla With similar track wheels/tires/brakes, Model S P100D is faster
26	25	@Erdayastronaut @Tesla Yes
27	26	@Tesla Would like to thank Robyn for joining the team. Great respect. Very much look forward to working together.
28	27	@Erdayastronaut @annerajb No, we<U+2019>re building a BFR dev ship to do supersonic through landing tests in Boca Chica, Texas
29	28	@Erdayastronaut @annerajb Won<U+2019>t land propulsively for those reasons. Ultra light heat shield & high Mach control su<U+2026>
30	29	@annerajb Aiming for orbital flight by June
31	30	Mod to SpaceX tech tree build: Falcon 9 second stage will be upgraded to be like a mini-BFR Ship
32	31	But comments are <U+0001F44C>
33	32	@Benioff @boringcompany LA/Hawthorne demonstration tunnel activates Dec 10. Runs parallel to 105 freeway from Crenshaw Blvd to 405.





Conclusion

Microblogging nowadays became one of the major types of the communication. A recent research has identified it as online word-of-mouth branding (Jansen et al., 2009). The large amount of information contained in microblogging web-sites makes them an attractive source of data for opinion mining and sentiment analysis. In our research, we have presented a method for an automatic collection of a corpus that can be used to train a sentiment classifier. We used sentiment analysis tools and observed the difference in distributions among positive, negative and neutral sets. From the observations we conclude that authors use syntactic structures to describe emotions or state facts. Some POS-tags may be strong indicators of emotional text. We used the collected corpus to train a sentiment classifier. Our classifier is able to determine positive, negative and neutral sentiments of documents. The classifier is based on the multinomial Na¨ıve Bayes classifier that uses N-gram and POS-tags as features.

Future Work

The project on sentiment analysis can be extended further to deeper levels. We can come up with a methodology to improve user experience like bringing in Unfriend option on Facebook for people which have negative impact on the user's frame of mind. The project is mainly designed for Twitter. We can have a similar application for Facebook wherein the intended features could be designed and incorporated.
Another future task is to have a hate-email classifier in addition to spam.

References

- https://begriffs.com/posts/2015-02-25-text-mining-in-r.html
- https://knightlab.northwestern.edu/2014/03/15/a-beginners-guide-to-collecting-twitter-https://joparga3.github.io/Udemy_text_analysis/#example-selecting-all-text-from-page---not-selecting-contents-images-etc
- Twitter Sentiment Analysis System - Shaunak Joshi and Deepali Deshpande , International Journal of Computer Applications (0975 – 8887) Volume 180 – No.47, June 2018
- P. Basile, V. Basile, M. Nissim, N. Novielli, V. Patti.“Sentiment Analysis of Microblogging Data”. To appearin Encyclopedia of Social Network Analysis and Mining, Springer. In press.data-and-a-bit-of-web-scraping/