Sentiment Analysis on News Channel at Twitter with R

Ali Akbar - 13514080
Informatics Major
School of Electrical Engineering and Informatics
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
13514080@std.stei.itb.ac.id

***Abstract* — Sentiment analysis is technique to acquire people opinion from a sentence. With sentiment analysis, company can acquire people sentiment about them. Social media like Twitter is a good resource to do sentiment analysis task. It's because so many people use Twitter to post their opinion about something. In this project we use simple lexicon-based to do sentiment scoring on Twitter data. We use tweets mentioning news channel account as the resource data to be analyzed. Writer use R language and R tools to compute the analysis.**

***Keywords* — R, Sentiment Analysis, Twitter**

I. Introduction

Social media can give a variety data for many purposes. It's because people post real-time message about their opinion in those platform. Company can use this as a tools to analyze what people think about their product. So the company can develop the product based on customer opinion and people requirement nowadays. One of the most used social media nowadays is Twitter.

Twitter has a lot of feature. But the one that really attract people is they can post something (called tweet) and directly mention the subject. And also user can label or highlight the topic. They also can see what is the hottest topic and the most spoken topic right now. Mostly tweet will mention the subject related to what happen at a certain time. The subject can be an artist, actor, musician, government and also news channel. It used to express their feeling upon them. With this subject related sentences, Twitter can give a summary about people sentiment on something.

Data provided by twitter is only a list of plain sentence. It's a challenge for a company to analyze those data and give an overall sentiment. The scientist has created a way to extract expression in a sentence. This method called sentiment analysis.

Sentiment analysis or sometimes called opinion mining is a method to determine subjective sentiment on a text. There are several approach to do sentiment analysis. It can be done with machine learning approach, lexicon-based approach, or rule-based approach.

In this paper, writer used twitter data to see people sentiment on a certain of news channel at America such as New York Times, ABC, Fox News, Washington Post and CNN. The data that being used is only the tweets that mention those news channel. Writer do lexicon-based approach to do sentiment scoring. Writer calculate how many positive words and negative words used from a lexicon on a tweets. The positive scoring result incline people give a positive sentiment and vice versa. Data retrieved from twitter using Twitter REST API.

II. Literature Study

A. Sentiment Analysis

Sentiment analysis is a method that use Natural Language Processing to extract or identify subjective information from a text. This task can be done at many levels of granularity such as at document level, sentence level, or phrase level [1].

There are several methods that can be done to do sentiment analysis such as,

- Machine learning approach, this method uses supervised learning on a training dataset to determine sentiment. To create model it can use several algorithms such as Naïve Bayes, Neural Network, SVM, and etc. [3]
- Lexicon-based approach, this method uses a bag of words that have been classified to score the sentiment. It uses the semantic orientation of words to calculate the result. [3]
- Rule-based approach, this method will find the opinion words and classify the sentiment with that. This method considers many aspects for classification such as dictionary polarity, negation words, idioms, emoticons, and etc. [3]

Twitter data is a good source for doing sentiment analysis task because it's contain people opinion at the real time. But there would be any bias on the source because twitter REST API data uses determined search keyword to search tweet. For a further experiment writer suggest to use STREAM API that retrieve tweet real time on a user timeline.

In this project writer use simple lexicon-based

sentiment scoring approach to do sentiment analysis on twitter data.

B. R Language

R is a language and environment for statistical computing and graphics. It is an GNU open source project that is similar to S language. R can do so many things. It includes

- Effective data handling
- Graphical facilities for data visualization.
- Simple syntax
- A well-developed library for machine learning, API consumer on social media, and etc.

In this project writer use R because R data frame can be easily recorded to a document, good data visualization graphic, simple syntax and easy library to handle some work such as Twitter REST API handler, problem solving handler, and many more.

III. Methods and Tools

A. R Library/Package Used and Data Resource

To run this project, we use the latest R and the latest package in R. Some package used are,

- twitteR
  This package is used to easily retrieve data using Twitter API. It's also easily convert the tweets retrieved into R data frame. But this package have dependencies on some package such as httr and base64enc
- httr
  Package to handle http request.
- base64enc
  Handle base64 decrypt and encrypt process.
- ggplot2
  Create a graph
- plyr
  A set of tools to solves common problem such as breaking some object and put all the pieces together again, summarize, and etc.
- stringr
  Wrapper package for string. It handles data like "NA" and zero length vector more consistent than other R package.

Those package must be installed/loaded before we run the data retrieval script and sentiment scoring script.

B. Data Retrieval

In this paper we use Twitter REST API to retrieve data source. Because we only want to see people sentiment about certain news channel, writer just use a tweet that mentioning the account related. The search key is *@abc*, *@washingtonpost*, *@cnn*, *@nytimes*, *@msnbc*. For each account writer retrieve 1500 tweets to be analyzed.

To use Twitter REST API. First we must register our application at apps.twitter.com. After that generate consumer key, consumer secret, access token and access secret. You need all those key to gain access on Twitter REST API. Those key used for OAuth authentication that used at Twitter REST API.

OAuth authentication has been handled by twitteR package on R. We can simply call setup_twitter_oauth method with those 4 keys as the input parameter. After that we can use searchTwitter method with filter option as input parameter to retrieve tweets with certain search key. For all those tweets retrieved, we can store those data on a file with .RData. Here is the script that writer use to retrieve tweets.

```
#Consumer key and secret
consumer_key <- "Es5KZjpnpT52nLqXzYOoawunZ"
consumer_secret <- "uvE8MUUYh5wtKxcFgT2RV81JVoAAsl2gNFC3fXKXp2yupWrx5W"
access_token <- "2574387926-gejKNOghqubr9JVu5OSsk19LYqVuPg5wClu38a6"
access_secret <- "AiFYCZ3mXoh5GpxBSzJ0gYkXO3OXleK6biCi3eulk1179"

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

#Fetch data
nytimes.tweets = searchTwitter('@nytimes', n=1500)
save(nytimes.tweets, file=file.path(dataDir, 'nytimes.tweets.RData' ), ascii=T)
print(nytimes.tweets)

abc.tweets = searchTwitter('@abc', n=1500)
save(abc.tweets, file=file.path(dataDir, 'abc.tweets.RData' ), ascii=T)

washingtonpost.tweets = searchTwitter('@washingtonpost', n=1500)
save(washingtonpost.tweets, file=file.path(dataDir, 'washingtonpost.tweets.RData'),
     ascii=T)

msnbc.tweets = searchTwitter('@msnbc', n=1500)
save(msnbc.tweets, file=file.path(dataDir, 'msnbc.tweets.RData' ), ascii=T)

foxnews.tweets = searchTwitter('@FoxNews', n=1500)
save(foxnews.tweets, file=file.path(dataDir, 'foxnews.tweets.RData' ), ascii=T)
```

Figure 1. Twittter Data Retrieval Script

C. Lexicon-based Sentiment Analysis for Positive and Negative Scoring

We use a lexicon of positive and negative words retrieved from Jeffrey Breen github project as a source for our scoring. With those source, we score a sentiment on sentence with the sum of positive words used minus sum of negative words used. Writer use Jeffrey Breen code to do this scoring [5]. Here is the script that Jeffrey Breen have made:

```
score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
{
  require(plyr)
  require(stringr)

  # we got a vector of sentences. plyr will handle a list or a vector as an "l" fo
  # we want a simple array of scores back, so we use "l" + "a" + "ply" = laply:
  scores = laply(sentences, function(sentence, pos.words, neg.words) {

    # clean up sentences with R's regex-driven global substitute, gsub():
    sentence = gsub('[[:punct:]]', '', sentence)
    sentence = gsub('[[:cntrl:]]', '', sentence)
    sentence = gsub('\\d+', '', sentence)
    # and convert to lower case:
    sentence = tolower(sentence)

    # split into words. str_split is in the stringr package
    word.list = str_split(sentence, '\\s+')
    # sometimes a list() is one level of hierarchy too much
    words = unlist(word.list)

    # compare our words to the dictionaries of positive & negative terms
    pos.matches = match(words, pos.words)
    neg.matches = match(words, neg.words)

    # match() returns the position of the matched term or NA
    # we just want a TRUE/FALSE:
    pos.matches = !is.na(pos.matches)
    neg.matches = !is.na(neg.matches)

    # and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum():
    score = sum(pos.matches) - sum(neg.matches)

    return(score)
  }, pos.words, neg.words, .progress=.progress )
```

Figure 2. Sentiment Scoring Script

Before you use Jeffrey Breen word list you must also see the tweets that you've retrieved. Some words are not

listed on that list such as “WTF”, “Nigga”, “LMAO” and other urban slang to express something. You can add it by yourself to make the analysis become better. On this project writer have add some negative word such as “WTF”, “LMAO”, “Fckin”, “FML” and many more urban slang.

D. The Main Program

Now we came to build the main program to analyze the tweets that we’ve retrieved. First we must get all the tweet text from tweets object. We can use laply function with our Twitter data object and getText() function as an input parameter. Here is some example script:

```
nytimes.text = laply(nytimes.tweets, function(t) t$getText())
```

Do those laply process on each tweets data that we’ve retrieved before. After that do sentiment scoring on each text with sentiment scoring function defined before:

```
nytimes.scores = score.sentiment(nytimes.text, pos.words, neg.words, .progress='text')
```

After that create a label variable to help us create the data visualization.

```
nytimes.scores$news = 'NY Times'
```

Then we visualize those data using ggplot2. First bind all score result to a single variable (at this script I named it all.scores) using rbind.

```
all.scores = rbind( score1, … , scoreN)
```

Use ggplot to create histogram. To create ggplot bind the all score result variable as the data of ggplot. Add geom_histogram to make the visualization as a histogram (you can also adjust the bin width with those function). Split those histogram based on their label using facet_grid. And then give a color with theme_bw() and scale_fill_brewer() function. Here is the script,

```
# ggplot function
g.hist = ggplot(data=all.scores, mapping=aes(x=score, fill=news) )

# add a bar graph layer.
g.hist = g.hist + geom_histogram( binwidth=1 )

# make a separate plot for each news
g.hist = g.hist + facet_grid(news~.)

# plain display, nice colors
g.hist = g.hist + theme_bw() + scale_fill_brewer()

print(g.hist)
```

Figure 3. Histogram with ggplot Script

IV. RESULTS

After running the program we get a histogram to see the distribution of sentiment scoring for each news channel account.

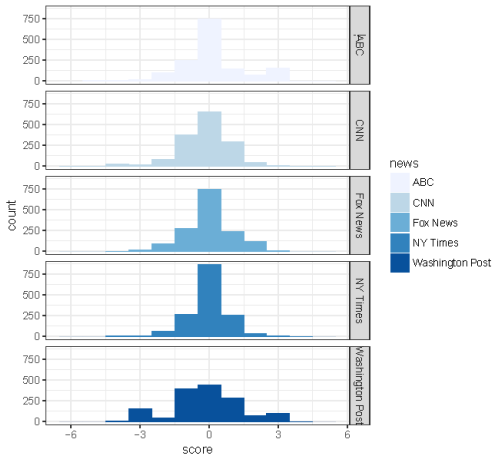

Figure 4. Sentiment Score on News Channel Histogram

As we can see on the result above ABC and CNN has a tendency to have a negative score. Fox News and NY Times has a tendency to have a neutral sentiment score. And Washington Post has a more distributed result. Some people give a more negative sentiment and the others also give a more positive sentiment. But overall Washington Post has a tendency to have a negative score.

After reading some article, a negative sentiment given by users on some channel is related to fake news rumor that spread among people of America.

V. DISCUSSION

Simple scoring using a defined lexicon of positive and negative words is enough to see people sentiment on something. But this method can’t see the context of negative sentiment. And also cannot handle the allegory words and very depend on the words that listed. This method also only can work with a language that used to be the lexicon language (in this project for example, the tweets must be in English so the scoring can work).

Beside the method that we use, the data that we’ve got is also not good enough. Using a defined search key can lead the result to be bias. This is the disadvantage of using REST API to retrieve data. And also we must filter again the tweets that outside our research context.

For a further experiment, we can use modified technic using machine learning to classify the sentence. In hope the result would be more contextual, precise and accurate. There are so many paper about sentiment analysis on twitter data using other technique that can be reproduced such as twitter sentiment analysis by Go et al. (2009) [4], Bermingham and Seaton (2010) [2] and Pak and Paroubek (2010) [6].

VI. CONCLUSION

Sentiment analysis is technique to acquire people opinion from a sentence. There are several approach that can be used such as machine learning approach, lexicon-based approach and rule-based approach. In this project we use simple lexicon-based approach to do sentiment

scoring on tweets mentioning news channel account as the resource data. Writer use R language and R tools to compute the analysis.

We use 5 news channel account to be analyzed. There are CNN, ABC, NY Times, Washington Post and Fox News. 3 of them (CNN, ABC, Washington Post) have a tendency to have a negative sentiment score by this analysis and the others (NY Times and Fox News) have a tendency to have neutral sentiment score.

For a further experiment, writer suggest to use Twitter STREAM API instead of defined search key using Twitter REST API to avoid bias data. And also writer suggest to use other technique such as machine learning approach to classify the positive and negative sentiment sentences. In hope the other technique can lead to a more precise and accurate result.

REFERENCES

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30-38). Association for Computational Linguistics.

[2] Bermingham, A., & Smeaton, A. F. (2010, October). Classifying sentiment in microblogs: is brevity an advantage?. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1833-1836). ACM.

[3] Collomb, A., Costea, C., Joyeux, D., Hasan, O., & Brunie, L. (2014). A study and comparison of sentiment analysis methods for reputation evaluation. *Rapport de recherche RR-LIRIS-2014-002*.

[4] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12).

[5] Jeffrey Breen. (2011). *Slides from my R tutorial on Twitter Text Mining*. Retrieved 5 May, 2017, from https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/

[6] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, No. 2010).

DECLARATION

I hereby declare that this paper is my own work, not a copy or translation of others' works, and not an act of plagiarism.

Bandung, 5 May 2017

Ali Akbar
13514080