Extracting and Analyzing Social Media Data

PANCHAMI RUDRAKSHI, The University of Texas at Dallas
LAHARI GANESHA, The University of Texas at Dallas
KAVYASHREE ANANTHA RAMAN, The University of Texas at Dallas

Social networking sites are websites designed for human interaction. Online social networks are now used by hundreds of millions of people and have become a major platform for communication and interaction between users. The use of social network websites is increasing day by day and this paves the way for useful information extraction and analysis.

The main goal of social network analysis is the study of structural properties of networks. Social networks exhibit properties that make them very suitable for Opinion Mining and Sentiment Analysis activities. Sentiment Analysis and Opinion Mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, Facebook, Instagram and social networks.

Additional Key Words and Phrases: Social networking sites, sentiment analysis, information extraction, opinion mining, APIs

ACM Reference Format:

Gang Zhou, Yafeng Wu, Ting Yan, Tian He, Chengdu Huang, John A. Stankovic, and Tarek F. Abdelzaher, 2010. A multi-frequency MAC specially designed for wireless sensor network applications. *ACM Trans. Embedd. Comput. Syst.* 9, 4, Article 39 (March 2010), 6 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Twitter, Facebook and other social media encourage frequent user expressions of their thoughts, opinions and random details of their lives. Tweets and status updates range from discussions of important political issues to inane comments. Several Twitter studies have demonstrated that aggregating millions of messages can provide valuable insights into a population based on several aspects like their geographic locations.

1.1 IDEA

The idea behind the project is to perform analysis on social media sites namely Twitter and Facebook to obtain useful information on various issues and learn from it in order to improve social networking. To be more specific in terms of social media sites:

- Twitter: To showcase two frequent analyses that rely on Twitter data- a map of geo-located tweets to examine the geographic distribution of human activity and an estimation of the favorability of a popular personality based on tweets that mention the person using sentiment analysis.
- Facebook: Collect general information about the users and different public pages to build a word cloud to analyze popularity of a person/page. Gender distribution analysis to find most popular names.

2. BACKGROUND

As people started making use of social media, many attempts were made to analyze social media like Twitter, Facebook and Instagram.

One of the approach to analyze Twitter Data: Here data is Twitter text data of @RDataMining used in the example of Text Mining, and it can be downloaded as file "termDocMatrix.rdata" at the Data webpage. Putting it in a general scenario of social networks, the terms can be taken as people and the tweets as groups on LinkedIn, and the term-document matrix can then be taken as the group membership of people. We will build a network of terms based on their co-occurrence in the same tweets, which is similar with a network of people based on their group memberships. At first, a term-document matrix, termDocMatrix, is loaded into R. After that, it is transformed into a term-term adjacency matrix, based on which a graph is built. Then we plot the graph to show the relationship between frequent terms, and also make the graph more readable by setting colors, font sizes and transparency of vertices and edges.

Another significant effort for sentiment classification on Twitter data is by Barbosa and Feng (2010). They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. The approach was extending by using real valued prior polarity, and by combining prior polarity with POS.

We see that most of the background work was constrained to twitter data analysis. In our project, we made efforts to analyze equally important social media like Facebook.

3. APPROACH/ANALYSIS

The extraction and analyses has been done with R programming language using a handful of libraries namely twitterR, streamR to collect data from Twitter's REST and Streaming API (for Twitter analysis) and Rfacebook packages to collect data from public pages on Facebook.

3.1 TWITTER

Firstly, we needed data for the analysis. Two different methods were used to collect data from Twitter:

1. REST API
    - Queries for specific information about users and tweets. Examples: user profile, list of followers and friends, tweets generated by a given user, user lists, follower lists, etc.
    - Library used: twitteR
2. Streaming API
    - To connect to the "stream" of tweets as they are being published (NOTE: tweets can only be downloaded in real time)
    - Three of them that were mainly used are – Filter stream (tweets filtered by keywords), Geo stream (tweets filtered by location and thereby examine the geographic distribution of human activity) and Sample stream (1% random sample of tweets)
    - Library used: streamR

Tweets collected from the said APIs are stored in JSON format.
Below is a sample tweet:

{"created_at":"Thu Nov 24 04:31:05 +0000 2016",
"id":8.01644284567028e+017,
"id_str":"801644284567027712",
"text":"President Obama lets bad puns fly at turkey pardoning tinged with sadness https://t.co/HYwpprIPxz https://t.co/JfVARnTmrw",
"source":"SocialFlow",
"user":{
"id":807095,
"id_str":"807095",
"name":"The New York Times",
"screen_name":"nytimes",
"location":"New York City",
"description":"Where the conversation begins. Follow for breaking news, special reports, RTs of our journalists and more from https://t.co/YapuoqX0HS.",
"url":http://t.co/ahvuWqicF9",
"followers_count":31739925,
"friends_count":969,

"listed_count":182750,
"created_at":"Fri Mar 02 20:41:42 +0000 2007",
"favourites_count":13411,
"time_zone":"Eastern Time (US & Canada)",
"statuses_count":257226,"lang":"en"}
"profile_image_url":"http://pbs.twimg.com/profile_images/758384037589348352/KB3RFwFm_normal.jpg",
"coordinates":null,
"place":null, "retweet_count":200",
"lang":"en"}

Extraction of data and analysis:

An OAuth token is created to authenticate. This access token is required for an application (in this case, our project) to make authorized calls to Twitter's APIs on behalf of a Twitter user. Basic user information is extracted and tweets that contain a keyword are searched using the REST API mentioned earlier. Also, tweets filtered by keywords and location are collected via the Streaming API.
First part involves analyzing geo-located tweets to study the map distribution of tweets in the US (giving an outline on the number of tweets from each state).
Second part is about measuring opinions on Twitter i.e., perform supervised sentiment analysis using a dictionary of positive and negative words and counting the number of times they appear (the Dictionary method). Also, we obtain training data by manually labelling random sample of posts as positive or negative. A classifier is used which learns from this manual labelling and predicts sentiment of unseen posts.

3.2 FACEBOOK

Facebook's more than one billion users make it a cultural, economic and social phenomenon. Because people use Facebook pages for the sake of marketing and more frequently as Fan's page, there is a lot of useful information that we can have by analyzing data on Facebook page.

Example, if a company posts different offers and sales on a Facebook page, by analyzing number of likes, comments and shares we can analyze "which" offer attracted customers the most. "What" are the needs of the customers and "what" is it that the company should invest on and the corrections to be done during the next offer season.

By analyzing likes and comments on a Fan's page, we see "which" is the most popular show that people are interested in. And "what" are the strengths of the celebrity. What is the "probability" of the celebrity winning an election, etc.

This is done as follows:

Packages Used:

1. Rfacebook: This package is used to collect the data that belongs to a particular page on Facebook. It extracts top posts, profiles of the people who liked and commented on those top posts.
2. Tm and wordcloud: These packages are used to print the most popular posts/words in the form of a cloud. Here, the most used word/post is printed in the highest font and in Bold and decreases thereafter. Least used post/word is printed in least font.

After installing the required packages, the access token is obtained as follows:

Procedure to generate Facebook token:

At the URL: https://developers.facebook.com/tools/explorer, click on "Get Access Token" to get the user access token for the application.
With the obtained token, the following data is collected:

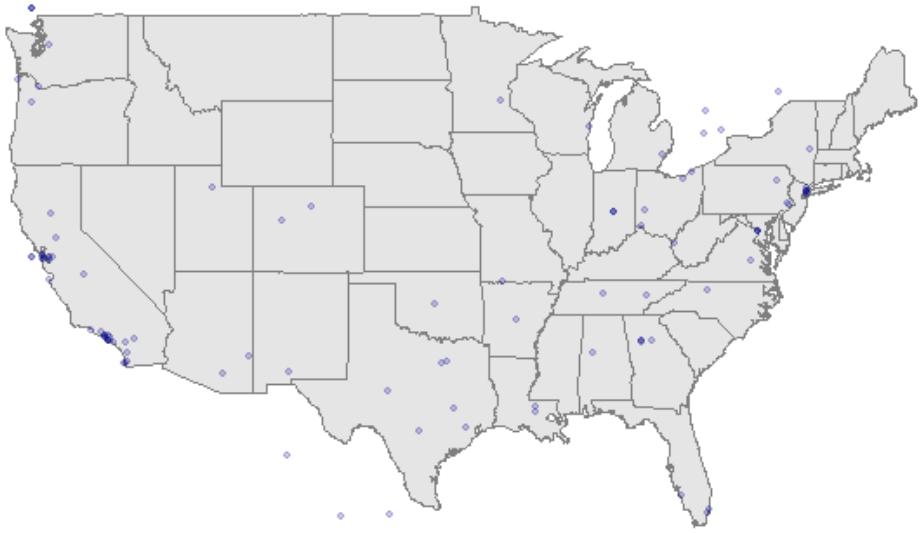
- Extract information about one or more Facebook users using getUsers().
- Collect top 5 list of posts from a public Facebook page with most likes and comments, not users with public profiles using getPage().
- Analyze the number of likes, average likes, number of comments, number of shares for a given public page.
- Print the posts in the form of a word cloud by printing the most popular post in the biggest font and least popular in the least font using tm and word cloud Package. This is done by creating a corpus of the text data, removing all the punctuation, stop words and converting it into a text document matrix to print the word cloud.
- Collect the list of the users who liked a specific post, so that we can analyze “what” kind of people are interested in a particular post using getPost() by collecting list of users associated with a particular page.
- Gender distribution analyses by getting top N most common names of the users.
- Collect the comments on a particular post and analyze the most popular post based on no of likes for that comment. This also tells us about the opinions of people on a particular post.

By analysis of this collected information, a lot can be predicted in the field of politics and Marketing. Summarizing this analysis can be priceless information to business man and artists.

4. RESULTS

4.1 Twitter:

Geolocation of tweets



The above image shows the distribution of clusters depicting the geographic distribution of human activity.

Further, extraction of number of tweets coming from each state gives the following result:

```
> states <- map.where(database="state", x=tweets$lon, y=tweets$lat)
> head(sort(table(states), decreasing=TRUE))
States
california              texas            Georgia                 ohio
        24                  6                  4                    4
district of columbia    indiana
         3                  3
```

Sentiment Analysis:

The sentiment analysis gives an outline about the sentiment ascertained from the tweets containing the word “Obama”.

Random sample of positive and negative words gives:

```
> sample(pos.words, 10)
 [1] "crusader"        "regard"          "accolade"        "lawfully"
 [5] "adoring"         "euphoric"        "enthusiastically" "indubitable"
 [9] "stable"          "historic"
> sample(neg.words, 10)
 [1] "cartoon"         "restlessness"    "demon"           "incautious"
 [5] "gaga"            "incompatibility" "oppression"      "self-serving"
```

[9] "slime" "disquietude"

Cleaning the text (i.e., tweets) gives the following result:

```
> text <- clean_tweets(tweets$text)
> text[[1]]
 [1] "reality"          "check"            "obama"
 [4] "will"             "go"               "down"
 [7] "in"               "history"          "as"
[10] "top"              "5"                "worst"
[13] "us"               "presidents"       "httpstco3dj22pga3v"
> text[[7]]
 [1] "rt"       "nia4trump"   "under"      "obama"      "dems"
 [6] "lost"     "900"         "state"      "legislature" "seats"
[11] "12"       "governors"   "69"         "house"      "seats"
[16] "13"       "senate"      "seats"      "wh"         "amp"
[21] "scotusâ€"    ""
```

Applying the function below to classify individual tweets

```
classify <- function(words, pos.words, neg.words){
  # count number of positive and negative word matches
  pos.matches <- sum(words %in% pos.words)
  #pos.matches
  neg.matches <- sum(words %in% neg.words)
  #neg.matches
  return(pos.matches - neg.matches)
}
```

And the function below to aggregate over many tweets,

```
classify <- function(words, pos.words, neg.words){
  # count number of positive and negative word matches
  pos.matches <- sum(words %in% pos.words)
  #pos.matches
  neg.matches <- sum(words %in% neg.words)
  #neg.matches
  return(pos.matches - neg.matches)
}
```

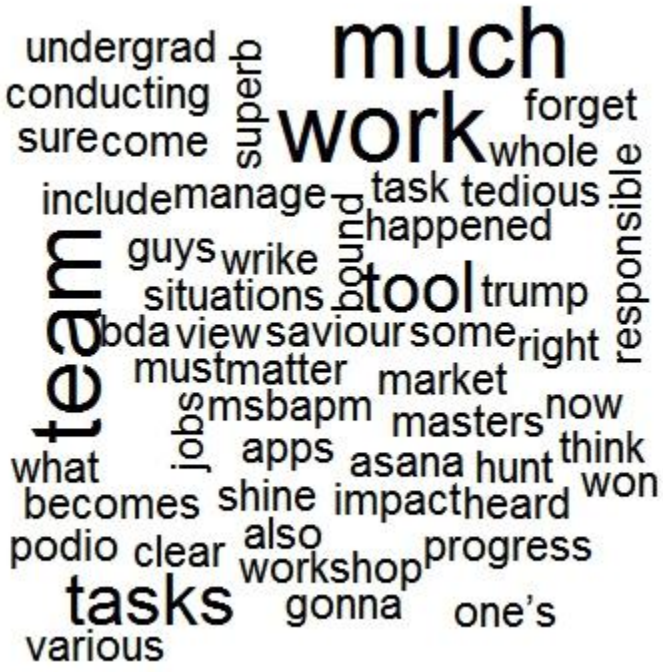
the sentiment classification of tweets, showcasing the favorability of a famous personality("Obama" in our case) gives:

```
> classifier(text, pos.words, neg.words)
```

234 tweets: 27 % positive, 18 % negative, 55 % neutral

Out of 234 tweets containing the word "Obama", 27% (~63 tweets) express positive opinion about "Obama", 18% (~42 tweets) express negative opinion and 55% (~129 tweets) express neutral opinion.

4.2 Facebook:




```
> token
[1]
"EAACEdEose0cBAJvMhlB5SF4tWi4g1Ov9bbVr01kSUaKT9pmBiofvry0y8cT9RNfZCWIAa
WZAuS47UIZCJZCZBWSLDXmtHFBV0IypwhHN7lS7hGpQqziNUhZAnbfWZBfsdKEOZCp
ypgFTldtE6velJkJe45ZBgx8QZBJEZCWxCxhsFMTSgZDZD"
> user_info
 id              name              username first_name middle_name last_name gender
1 305057203206522 Panchami Rudrakshi       NA   Panchami          NA  Rudrakshi female
locale category likes picture
en_US       NA     1      NA

 > page = getPage(page="msbapm",token=token,n=5,feed=T)
 5 posts
 > page
         from_id                          from_name
 1 861979060516514                    Mark Yusuf Ishmaeel
 2 108873275863636 MS in Business Analytics and Project Management
 3 108873275863636 MS in Business Analytics and Project Management
 4 108873275863636 MS in Business Analytics and Project Management
 5 161434714265280                       Amira Biswas

 message
 1
```

What now that trump has won, how do you guys think its gonna impact jobs there ?Especially for masters' students

2

Shine your light event...

3 <NA>

4 MSBAPM students conducting a R workshop for our BDA undergrad students

5 No matter how much responsible you are about your work, you are bound to forget tasks at times. It becomes tedious at times to manage tasks of a whole team; monitoring each one's task and also having a clear view as to how much progress has happened in work. In such situations, team management apps come as a saviour. In hunt for a perfect tool for project management, you sure must have heard of Basecamp as a project management tool, but there are various Basecamp alternatives outthere in the market. Some of the basecamp alternatives include Asana, Team Work,Wrike, Podio and so much more. Look at some superb basecamp alternatives right here:

	created_time	type
1	2016-11-09T08:11:18+0000	status
2	2016-10-30T00:09:02+0000	photo
3	2016-10-26T19:42:32+0000	link
4	2016-10-21T20:44:40+0000	photo
5	2016-10-17T16:43:24+0000	link

link

1 <NA>

2

https://www.facebook.com/msbapm/photos/a.835883766495913.1073741832.108873275863636/1175014709249482/?type=3

3

http://ow.ly/uTut305z5je

4

https://www.facebook.com/msbapm/photos/a.835883766495913.1073741832.108873275863636/1167178813366405/?type=3

5

http://bit.ly/basecamp-alternative

	id	likes_count	comments_count	shares_count
1	108873275863636_1185913984826221	0	0	0
2	108873275863636_1175014975916122	18	0	0
3	108873275863636_1171816802902606	12	0	0
4	108873275863636_1167178953366391	9	0	0
5	108873275863636_196091624132922	1	0	0

```
> posts
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 5
> tdm = TermDocumentMatrix(posts)
<<TermDocumentMatrix (terms: 65, documents: 5)>>
Non-/sparse entries: 66/259
Sparsity           : 80%
Maximal term length: 12
Weighting          : term frequency (tf)

> head(likes)
           from_name          from_id
1 Tr<U+1EA7>n Van Xuân   516061495171456
2    Nguyên b<U+1EE5>i  1564882647082020
3      Ismael Jabareen  1397022797274043
4 Basrani Rajesh Malhi   279030638940214
5       Henrik Koukku 10201885788894963
6  Sherry Purvis Abram   819194488122780
> head(sort(table(users$first_name), decreasing=TRUE), n=10)

   Maria Christina     Laura       Amy     David   Jessica    Ashley     Chris       Jay   Melissa
         7         6         6         5         5         5         4         4         4         4
 > head(comments)
           from_id      from_name
1   700980843295639        Simo As
2 10205744928247276  Alex Heminger
3 10203668676827926    Joe Walling
4 10201458129573423     John Brown
5 10155059170760398   Patrick Drew
6 10203578416082243 Jennifer Stacy

 message
 1 .
 \nlike                          plzzzzz\n(y)\nhttp://www.facebook.com/pages/Marrakech-The-Magic-
 City/329268487148931?ref=hl
 2
 Repeal Obamacare!  Repeal Obama!
 3                                                    FBO...
 4                                                    obama sucks!
 5                                                    GOD BLESS AMERICA
 6                                                    Not enough but a start
 created_time            likes_count                id
```

```
1 2012-10-29T19:50:54+0000     1     103164709846980_8190
2 2012-10-29T19:50:56+0000     0     103164709846980_8191
3 2012-10-29T19:51:02+0000     0     103164709846980_8192
4 2012-10-29T19:51:03+0000     0     103164709846980_8194
5 2012-10-29T19:51:06+0000     5     103164709846980_8195
6 2012-10-29T19:51:13+0000     6      103164709846980_8199

> comments[which.max(comments$likes_count),]
from_id                from_name
246 10203113125422564 David Alan-Hartmann Mingle
message
246
For all my Republican people waiting on Hurricane Sandy to arrive at your front door....Remember this....You are AGAINST government handouts!!! So....With that being said....Don't expect any government assistance from FEMA or the president when your house gets flooded/blown away and you run out of food and water. Be that STRONG patriotic American that you are!!! Remember, because YOU built that!!! lmfao

created_time likes_count                id
246 2012-10-29T19:58:11+0000         39 103164709846980_8502
```

5. FUTURE WORK

Our study of using Twitter and Facebook to mine public information focused on producing data that correlates with public interest metrics and knowledge. Further this application can be enhanced by analyzing sample Streaming over different phases of time, at different locations. Also, sentiment analysis done through geo-located twitter tweets can be used to predict if the person can be re-elected. Gender distribution analyses in Facebook can be extended to predict most commonly occurring names that will be kept for babies. Based on most popular Facebook post, sentiment analyses over a period of time can be done. Further, what new information can be learned by studying Twitter and Facebook, potentially supporting political and health informatics hypotheses can be analyzed. We plan to consider more specific applications with the goal of learning new things using web mining from Twitter and Facebook in future to get better accuracy. In addition, we will also try to understand more social data sources which are useful for including in the system and our future research.

6. CONCLUSION

Social media empowers customers to share their comments on products instantly and provides an open and accessible resource to allow enterprises to become closer to their customers. The data extracted from Facebook and Twitter is believed to have shed more light on the structure and activities of social network. We were successfully able to perform sentiment analysis on the tweets extracted based on a popular personality (“Obama” in our case) and find the favorability of the person considered. Cluster analysis showing distribution of human activity for geo-located tweets

extracted over a period of time, collect a random sample of tweets, analyze most common hashtags right now, find most re-tweeted tweets. We also captured tweets mentioning multiple keywords. General information about the users and page was collected to build a word cloud to analyze popularity. Gender distribution analysis was performed using user's information list. On a whole, an important outcome of this project is that we learnt how to harvest social media data for performing important data analysis at the industry level.

ACKNOWLEDGMENTS

We would like to thank Dr. Bhavani Thuraisingham for providing immense support in accomplishing this project.

REFERENCES

1) Michael J. Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health
2) Bing Liu, 2012. Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies.
3) Mariam Adedoyin-Olowe etal. A Survey of Data Mining Techniques for Social Network Analysis
4) Ting etal. Analysing Multi-Source social data for extraction and Mining Social Networks