International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 23
December 2018

Sentiment Analysis of Twitter Data

Devang Kumbhabhai Chavda
U & P U Patel Department of Computer Engineering,
Chandubhai S. Patel Institute of Technology
Charotar University of Science and Technology
chavdadevang23@gmail.com

***Abstract**:*

Sentiment analysis is very helpful to derive insights from raw data.Data is flowing in thousands of terabytes.Twitter is generating on an average 6000 tweets every second and 500 miliion tweets everyday. We have large amount of meaningful data and sentiment analysis is one of the use case of that massive data. Realtime tweets helps to understand the latest trends. Behavior of political leaders on social media make huge impact on public opinions so, it makes sense to analyze this massive data in order to know whether the election strategy is working good and making change or not.

Keywords

Sentiment analysis, opinion mining, twitter data, Natural language processing, gender detection.

1. Introduction

Microblogging websites is the platform where people are sharing their opinion or review on something it is very helpful when we need to know how people think of their product or in our case how people are reacting to the political parties initiatives or changes in laws.

Think of any product based company if they want to manufacture a mobile phone and they are not sure that it will make profit. They can do a quick analysis of market ,they can analyze latest trend in mobile phone or they can go for sentiment analysis for their demo version . They just need to manufacture demo version of product and receive sentiments on that it will provide powerful insights and company can take decisions on it.

This paper is divided in three parts follows.

In **Part 1:** Section 1: preparation of data set for analysis.

Part 2: Section 2: pre-processing tweets to make it analysis ready!!. Section 3 : Detect user's gender. In section 4 Sentiment score derivation.

Part 3: Section 5: visualization of dataset.

Part 4: Section 6 : challenges ,
section 7 :conclusion ,section 8 :future work in research.

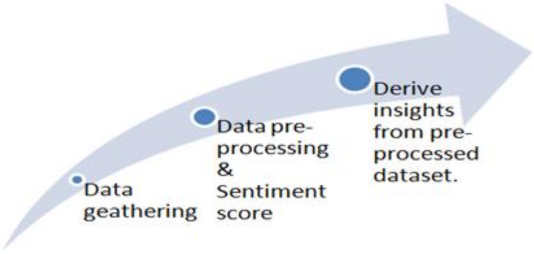

Data geathering
Data pre-processing & Sentiment score
Derive insights from pre-processed dataset.

Fig. 1: Workflow of Data

1) Data preparation

- Tools:
  * Twitter streaming api
  * Tweepy

Twitter streaming api:

Twitter streaming api is very simple and powerful tool to get real time tweets. Twitter API has certain criteria which is used when streaming live tweets and those criteria are actually works like an filter that what topic do you want to search .which regions do you want to get tweets from how many tweets do you want to stream in what language tweet should be and many more.

- Following queries are used while streaming tweets.
  - lang="en", #language filter
  - q=query + " -rt", #without ReTweet
  - count=number, #number of tweets
  - result_type="recent", #tweets version
  - geocode="22.20775,76.97021,200km", #origin of tweet

Tweepy:

It is a python library which is use to access twitter api.

- Dataset columns :

User_name	Tweet	Created_at	location

2) Data pre-Processing:

- **Remove Stop-words**: commonly occurring words are called stop-words like (a,an,the,etc)those should be removed because while analyzing the sentiment of the sentence those words have nothing to contribute.stop-words has 0 sentiment value so these are unnecessary for our purpose.
- **Remove of Punctuations**: Again same as stop-words punctuation are also unnecessary. Has 0 sentiment value so they should be removed.

3) Detect user’s gender:

Gender api: An api that detects the gender by the first name.

After applying gender api our dataset will appended with new colunmn called gender.

{first_name “Devang “ ;gender:”male”,accuracy:100}

- Updated Dataset:

User_name	Tweet	Created_at	location	gender

4) Polarity Scoring:

Text Blob :

It is an API that performs different Natural Language Processing (NLP) tasks like

- Part-of-Speech Tagging
- Noun Phrase Extraction
- Sentiment Analysis
- Classification like naïve bayes and decision trees. Language Translation and Detection
- Spelling Correction

TextBlob is built on nltk.

Sentiment Analysis:

Sentiment analysis means classifying text into three main category.

- Positive
- Negative
- Neutral

We can further divide the main categories like,

- Extremely positive
- Extremely negative

Polarity is float value which is between -1 to 1 where 1 means positive sentense and -1 means a negative sentence

- **Create a text blob:**

Code:

```
from textblob import TextBlob

statment= TextBlob("bjp and congress who is better don't know but aap is not deserving for sure")

tweet.tags

                    [('bjp', 'NN'),
                     ('and', 'CC'),
                    ('congress', 'NN'),
                     ('who', 'WP'),
                     ('is', 'VBZ'),
                     ('better', 'RB'),
                     ('do', 'VBP'),
```

("n't", 'RB'),
('know', 'VB'),
('but', 'CC'),
('aap', 'VB'),
('is', 'VBZ'),
('not', 'RB'),

- ○ **Feature Extractors**

the naïve bayes classifier uses a simple feature extractor that shows which words in the train set is in the document.

For example, the sentence *"bjp is best"* might have the

features contains(best): True or contains(worst): Fa lse.

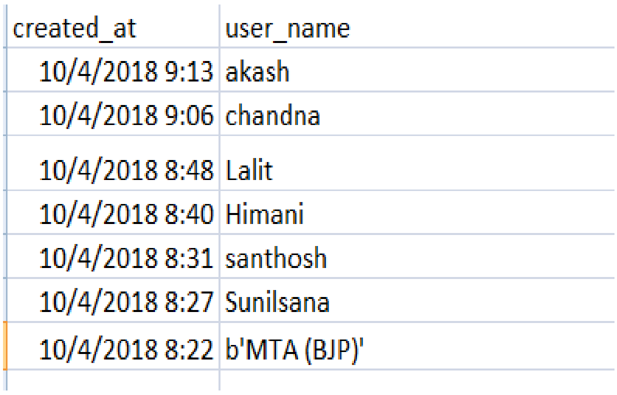
A feature extractor is function with document
neg : pos = 1.6 : 1.0
neg : pos = 1.4 : 1.0
pos : neg = 1.4 : 1.0
pos : neg = 1.3 : 1.0

- ○ Updated Dataset:

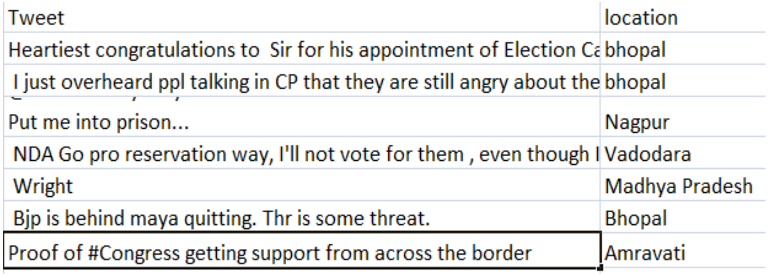
U_name	Tweet	Created_at	loc	gender	sentiment

- **Prepared Dataset:**

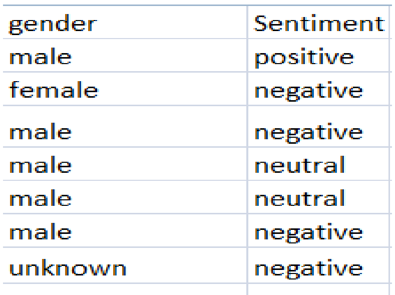
Here I have presented the small snippet of the prepared dataset.The whole dataset contains 3000 rows and 6 columns.1500 tweets are bjp related and rest are of congress.Here Data is cleaned enough but it is still difficult to derive insights from this large dataset. Visualization will help.



created_at	user_name
10/4/2018 9:13	akash
10/4/2018 9:06	chandna
10/4/2018 8:48	Lalit
10/4/2018 8:40	Himani
10/4/2018 8:31	santhosh
10/4/2018 8:27	Sunilsana
10/4/2018 8:22	b'MTA (BJP)'



Tweet	location
Heartiest congratulations to Sir for his appointment of Election Ca	bhopal
I just overheard ppl talking in CP that they are still angry about the	bhopal
Put me into prison...	Nagpur
NDA Go pro reservation way, I'll not vote for them , even though I	Vadodara
Wright	Madhya Pradesh
Bjp is behind maya quitting. Thr is some threat.	Bhopal
Proof of #Congress getting support from across the border	Amravati



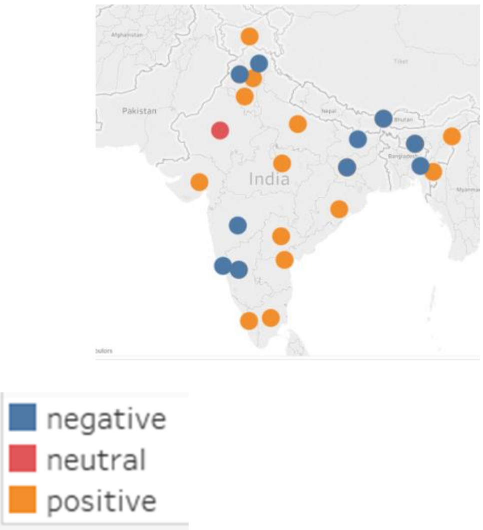
gender	Sentiment
male	positive
female	negative
male	negative
male	neutral
male	neutral
male	negative
unknown	negative

2)Data visualization

*Tool used: Maps in tableau

Following are the results

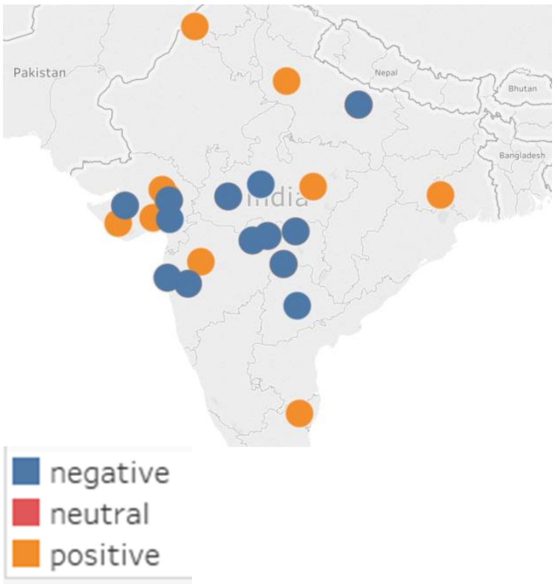
- BJP



➢ Circles are showing group of tweets from different cities in india. from the map representation insights can be derived easily. Following questions can be answered?

- Which state have most negative sentiment?
- Which state political party should work more?
- Which state is most dominating?

- **CONGRESS**



Results:

➢ Considering this analysis each party can improve their performance in respective cities where they are getting more negative opinions.

Challenges

- less accuracy of twitter streaming api:

➢ When working with twitter streaming api it will not ensure 100% accuracy of tweets . It often happens you search for any subject and get something else in return but again it's a free for use api and it has really good performance but for serious purpose analysis you need to upgrade to twitter firhose api it will provide 100% accuracy of tweets here you get exactly what you searched for of course it is not free.

- username

- Usernames are messy too like tweets.following are the examples of some usernames fetched during research.

- b'MTA (BJP)'
- b'TDPTrending\xc2\xae'
- b'Director Sports, MP'

there is no way to deal with this type of usernames so we cannot identify gender of user by its name.

another way of reorganization of gender is profile picture but again it is not 100% accurate.

- age of user:

Twitter streaming api is not providing age of user but for analysis of political parties it is very crucial factor. Age of user provide very helpful insights . **which age group is more sensitive in which region?** This question can be answered if we have age of user. We can use Microsoft azure face api or google face api to recognize age of a person by their profile picture but again it will not provide 100% accuracy .following are the cases where age recognition is not possible .

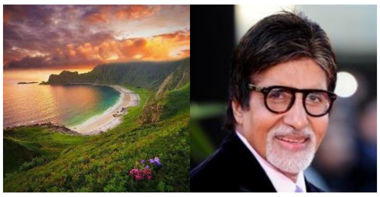


Fig. 1:Sample Profile pictures

- So many users are not uploading their own profile picture and that is a big challenge. above are two examples. In those cases solution of these problem only can solved by twitter it self. some restriction or some face recognition mechanism should be there to authenticate the user's profile picture.

Conclusion

I have presented the results of sentiment analysis of twitter on two Indian political parties BJP vs Congress.I have used twitter's streaming api to gather tweets and used python libraries for sentiment scoring and tableau for visualization. Discussed about various challenges and how to overcome those to some extent.

Sentiment analysis is very popular now a days because we have very large amount of data flowing everywhere in every field and making right use of that will lead to better future.

Future work in research

Future task will be detecting gender and age of user by using face api and improving accuracy of it. We can further add the machine learning algorithm and predict the result from gathered the data.

References

[1] Boyd, D., Golder, S., and Lotan, G. *Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In System Sciences* (HICSS), 2010 43rd Hawaii International Conference on (2010), IEEE, pp. 1-10.

[2] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. Predicting *elections with twitter: What 140 characters reveal about political sentiment.* ICWSM 10 (2010), 178{185.

[3]Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. *Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings* of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

[4]Luciano Barbosa and Junlan Feng. 2010. Robust sentiment *detection on twitter from biased and noisy data*. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

[5]Cohen, R., and Ruths, D. *Classifying political orientation on twitter: Its not easy!* In Proceedings of the 7th International Conference on Weblogs and Social Media (2013).