Sentiment Analysis of Twitter Data:Case Study on Digital India

Prerna Mishra
Research Scholar-IT
Amity University
Lucknow, India
prerna21.mishra@gmail.com

Dr. Ranjana Rajnish
Assistant. Professor
Amity University
Lucknow, India
ranjanavyas@rediffmail.com

Dr.Pankaj Kumar
Assistant Professor
SRMGPC
Lucknow, India
pk79jan@gmail.com

Abstract— **Nowadays Opinion Mining has become an emerging topic of research due to lot of opinionated data available on Blogs & social networking sites. Tracking different types of opinions & summarizing them can provide valuable insight to different types of opinions to users who use Social networking sites to get reviews about any product, service or any topic. Analysis of opinions & its classification on the basis of polarity (positive, negative, neutral) is a challenging task. Lot of work has been done on sentiment analysis of twitter data and lot needs to be done. In our work we are trying to perform sentiment analysis of the twitter data set that expresses opinion about Modi ji's Digital India Campaign. In my work, I have collected these sentiments and classified polarity of sentiments in these opinions w.r.t. Positive, Negative or Neutral. Twitter data is collected for analysis using Twitter API. Out of the two widely used approaches used for sentiment analysis, Machine Learning & Dictionary Based approach, we are using Dictionary Based approach to analyze data posted by different users. Then polarity classification of this data is done**

In this paper we discuss sentiment analysis of twitter data, existing tools available for sentiment analysis, related work, framework used, case study to demonstrate the work followed by the results section. Results clearly demonstrate that the 50% of the collected opinions are positive, 20% are Negative and rests 30% are neutral.

Keywords-Opinion Mining, Sentiment Analysis, Governance, Digital India, Natural Language Processing, Machine Learning, Dictionary Based approach.

I. INTRODUCTION

Huge volume of data is available on various websites where users are sharing & exchanging their ideas and opinion. With the increase in the use of facebook, twitter and other social networking sites to express views on topics of interest/ concern and having discussions on them is making such sites a pool of opinions. These opinions are also associated with sentiment of the user who is expressing the opinion.

Opinion Mining is a type of Natural Language processing technique that is used to mine the reviews/opinions about any particular topic, product, service or Prediction of Elections, Stock Market etc [1]. Nasukawa & YI first introduced the term Sentiment Analysis & Opinion Mining in the year 2003. SA analyzes the user's thoughts/sentiments by determining the polarity (Positive, Negative and Neutral) from huge amount of data availabilty on Internet. According to researchers Opinion Mining is classified at three different levels as "Document Level", "Sentence Level" and "Aspect level". In this paper we did analysis on Sentence Level.

Textual information can be broadly categorized into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are subjective expressions that describe an individual's sentiments, appraisals or feelings toward entities, events and their properties [12]. People express their opinions not only about products and services, but also about various topics and issues especially from social domains [8].

This paper comprises of VIII Sections. In Section II & III we have discussed about the popular Social Networking site Twitter and various tools available for the task of sentiment analysis. Section IV discusses about the related work done in this field. Section V discusses the framework used for the task of Sentiment analysis and different processing steps involved in doing the same. Section VI presents the Case study of Digital India with different tweets collected followed by the Section VII discusses the results of the work. It also, shows graphical representation of Polarity classifications done. Paper is concluded by discussing future scope & various challenges of Sentiment analysis in section VIII.

This work can be extended to provide summarization of opinions so that government can have participation of citizens, by knowing their views/opinions, in formulation and implementation various governance policy.

II. SENTIMENT ANALYSIS OF TWITTER DATA

Twitter is popular online social networking service launched in March 2006.It enables users to send and read tweets with about 140 characters length.Currently Twitter acts as opinionated Data Bank with large amount of data available used for sentiment anlysis. Twitter is very convinient for research because there are very large numbers of messages, Many of which are publicly available, and obtaining them is technically simple compared to scarping blogs from the web [9].

Twitter data is collected for analysis using Twitter API. Two widely used approaches used for the same are Machine Learning & Dictionary Based approach. We are using Dictionary Based approach for analyzing the sentiments of data posted by different users. Then polarity classification of this data is done i.e. Tweets collected after analysis are classified into three categories as Positive, Negative and Neutral. Result of this is depicted by using PIE Chart.Sentiment analysis is done by using NLTK toolkit.

III. TOOLS AVAILABLE FOR SENTIMENT ANALYSIS

S.No	Tool	Applications
1	NLTK	NLTK toolkit is widely used nowadays for sentiment analysis task. Main features of NLTK used in Sentiment analysis process are Tokenization, Stop Word removal, Stemming and tagging. This tool is written in Python language and can be downloaded from www.nltk.org.
2	GATE	General Architecture for Text Engineering (GATE) is information Extraction System consisting of modules like Tokenizer, Stemming and Part of speech tagger. This tool is written in Java language. https://gate.ac.uk/
3	Red Opal	This tool is widely used for users who want to buy any products based on different features. Users can search for any product depending upon the feature selected and can get reviews related to their search.
4	Opinion Finder	Opinion Finder is used for analysis of different Subjective sentences related to any topic & classification of sentences is done based on their polarity. It's written in Java and is platform Independent tool.

Table-1 Sentiment Analysis Tools

IV. RELATED WORK

[1] In this author aims to extend the machine learning approach for aggregating public sentiment.He sused case Study of UK for analysis and comapred by using two main approaches as "Dictionary Based approach " and "Machine Learning approach".Proposed an framework for analysis and visualization of public sentiment & the result obtained indicates that there is a reasonable correlation between scores produced by both the approaches.

[2] According to Kun-Lin Twitter has become popular online Micro blogging service so in this paper he presented a novel model called Emoticon Smoothed language Model (ESLAM) to handle noisy data. He used this model to deal with misspelled words, slang, acronyms which cannot be easily handled by fully supervised methods. He compared ESLAM model with fully supervised method with accuracy and F-score.

[3] In this paper author wants to accurately identify the semantic orientation of opinions expressed. Semantic orientation we mean whether the opinion is positive, negative or neutral. Author proposed a Holistic Lexicon –based approach by resolving two main problems with the existing methods 1-Opinion words whose semantic orientation are context dependent. 2-Aggregating multiple opinions words in same sentence.

[4] Author of this paper analyzed the people's opinions & review by using Case Study of Automotive Industry. He used 3 popular (BMW, Mercedes, Audi) automotive companies of Europe for analysis & polarity classification.

[5] In this paper author had presented a Sentiment Analysis System "TwiSent" & addressed the different problems related to Opinion Mining like Spams, Structural Anomalies in Text & pragmatics embedded in text.

V. FRAMEWORK USED

The framework used for this analysis is depicted in below figure. Different processing steps had their own important role. We discussed about all steps below.

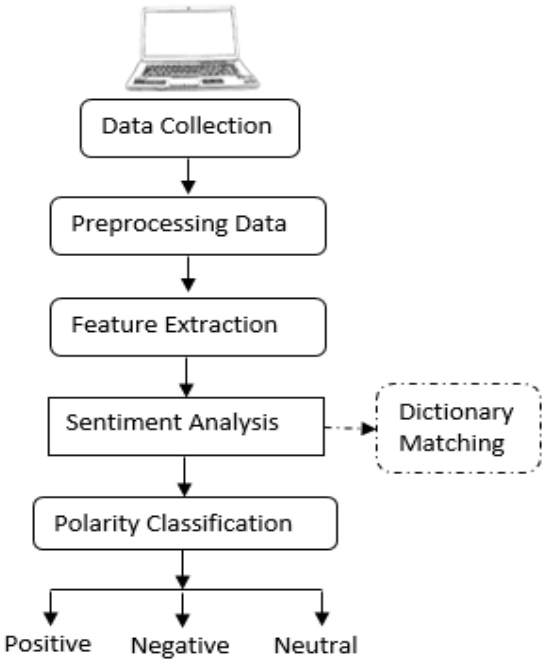

Figure-1 Opinion Mining Framework

A. Data Collection:

Collection of data is an important part of Sentiment Analysis. Various data Sources like Blogs, Review Sites, Online Posts & Micro Blogging like Twitter, Facebook are used for Data Collection. We used Twitter for Data Collection process.

B. Data Preprocessing:

Now before Sentiment Analysis we need to process the collected data using the following steps of data processing-

1) Stemming- In this process we remove the postfix from each words like "ing","tion" etc.

2) Tokenization- This process is very important for Data preprocessing as it includes several sub steps like "Removal of Extra spaces", "Emoticons (☺,☹) used replaced with their actual meaning like Happy, Sad by using Emoticon data set available on Internet", "Abbreviations like OMG, WTF are replaced by their actual meanings", "Pragmatics handling like hapyyyyyyy as happy, gudddddd as good etc."

3) Stop Word Removal- In this we remove stop words which are not of any use in analysis like Prepositions (a, an) and Conjunctions (and, between) used.

For all the above steps described we used NLTK 3.0 tool Kit with Python. Toolkit can be downloaded from (http:www.nltk.org)

C. Feature Extraction:

Feature extraction specifies the type of features used for opinion Mining [6]. There are different types of features used like-

1) Term Frequency- Frequency of any term in a document carries weightage. [6]
2) Term Co-occurrence- Repeatedly occurrence of a word like Unigram, Bigram or n-gram etc. [6]
3) Part of Speech-For each tweet we have features for counts of the number of Verbs, adjectives, nouns. [7]

D. Sentiment Analysis & Polarity Classification:

Emotions, opinions and sentiments play an important role in all human life. Mining such opinions termed as sentiment analysis [10]. Performing task of Sentiment analysis & polarity classification is a challenging task. We did sentiment analysis by using "Dictionary Based approach". This approach uses a predefined dictionary of positive and Negative words. SentiWord net is a standard dictionary used by most researchers today for sentiment analysis. Task of Polarity classification we mean the reviews collected are classified depending upon the emotions expressed as Positive, Negative and Neutral.

VI. CASE STUDY: DIGITAL INDIA

Now we have to do sentiment analysis & polarity classification of all the collected tweets which are now preprocessed by above steps. For this analysis we are taking a case study related to "DigitalIndia" mission of Government launched in year 2015 [http://www.digitalindia.gov.in/].This mission was envisioned with aim to digitally empower the people of country. Main factors of this mission are as-

1) High Speed Internet services to Citizens.
2) Business related Services.
3) Free Wi-Fi in Trains & Railway Stations.
4) Smart City Project.
5) Boast New Scheme-Digi Locker

Following are the steps with respect to the case study as discussed above in section V.

A. Data Collection:

Data collected from Twitter by using the Twitter API (twitter 4j) is shown below. Twitter has created its own API for tweets retrieval. We have used this Twitter API in our Python code for Twitter corpus Retrieval related to "#DigitalIndia". We were able to successfully retrieve 500 tweets from Twitter using our Python code.

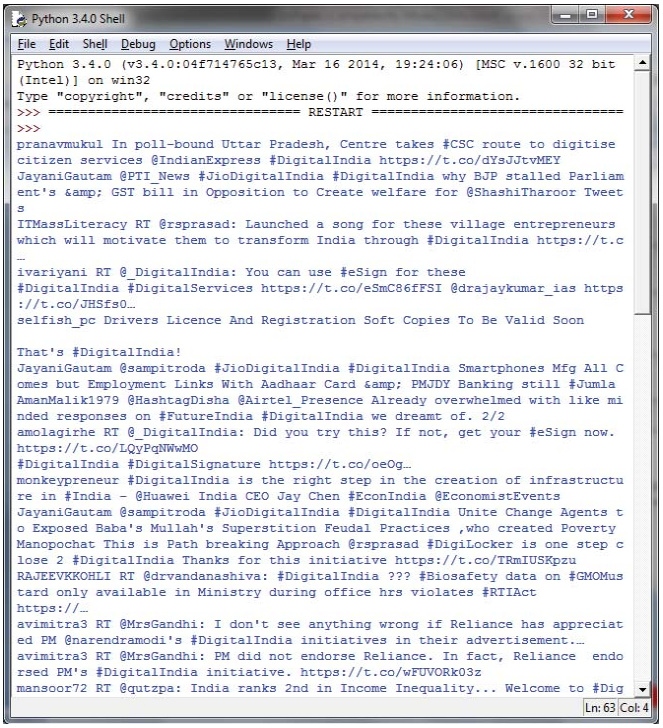

Figure-2 Data Collected From Twitter

B. Data Preprocessing & Feature Extraction:

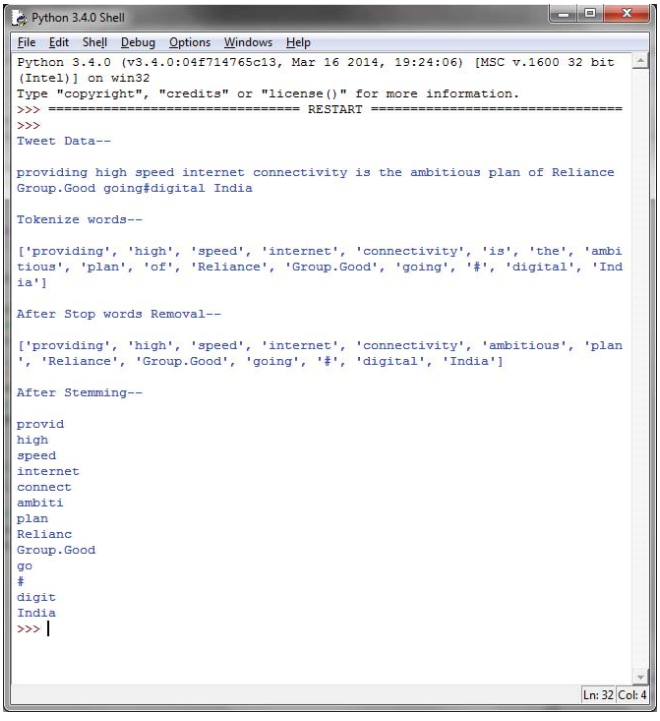

Figure-3 Data Preprocessing of Tweets

Data preprocessing is done using NLTK 3.0 modules integrated with Python code. Task includes StopWords Removal, Tokenization and Stemming.

C. Sentiment Analysis & Polarity Classification:

As discussed above for sentiment analysis we have used Dictionary based approach. In this approach collected tweets are matched against a dictionary which is collection of Positive & Negative words.

Tweets Collected	Dictionary	Polarity Classification
Providing high speed internet is the ambitious plan of Reliance Group.Good going # DigitalIndia	Positive Words High Ambitious Good Negative Words Letdown Dead no	Positive
@UIDAI plz fix d Aadhar android app SMS verification issue otherwise this will be alet down issue 4 @DigitalIndia @NarendraModi		Negative
@Airtel_Presence 48 hrs landline dead no Internet no action.Is this the #DigitalIndia		Negative

Table-2 Tweets Sentiment Analysis

As we can see in above table, tweets collected are matched against Positive & negative words used from Dictionary & then tweets are classified as positive & negative. Remaining tweets are classified as Neutral (Tweets which are neither positive & negative).

VII. RESULTS

Our goal for this study was two-fold. First, we wanted to extract the related tweets from the twitter data set. Then, we wanted to classify the tweets, retrieved using our python code, on the basis of polarity of sentiments.

Table 3 shows some of the related tweets that were retrieved (related to Digital India) from Twitter account. Classification results demonstrate that our code retrieved 250 positive, 150 neutral and 100 negative opinions.

Tweets	Polarity
Postal department now enjoying a glory as never before, all due to #DigitalIndia initiatives & e-comm business. Now indispensable.	Negative
48 hrs landline dead no Internet no action. Is this #DigitalIndia	Negative
Indian #ecommerce space may soon have a new giant, if #government has its way. #digitalIndia	Positive
Providing high speed internet connectivity is the ambitious plan of Reliance Group. Good going #Digital India	Positive
#DigitalIndia has new avenues in Future	Neutral

Table-3 Sample of tweets retrieved

Classification results are graphically represented using Pie Chart By this we can clearly see that 50% result was positive, 30% neutral and 20% negative opinions.

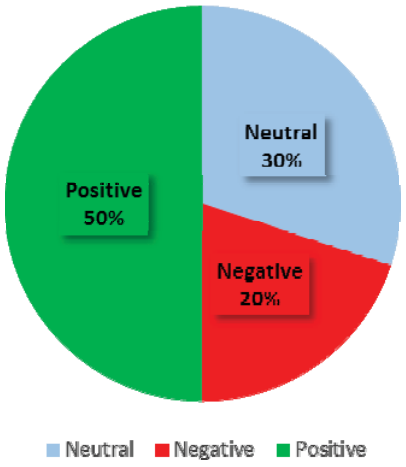

Figure-4 Polarity Classification Result in %

Based on the sentiment analysis of tweets posted by the users on micro blogging site Twitter, results of our study demonstrate that Digital India mission of Government of India is liked and is found useful by majority of Indian Citizens. 50% of the users have positive opinion about the campaign, 30% of them are neutral and only 20% of them have a negative opinion.

So were successful in achieving both the goals as mentioned in beginning of the results section.

VIII. CONCLUSION

In this paper we have seen various steps used to perform sentiment analysis. We also saw various tools available for sentiment analysis.

Our focus in this paper was to capture polarity of the sentiments captured from twitter data. We have used Case Study of Digital India mission to achieve our goals. We can see that results are encouraging as we are able to segregate sentiments as 250 positive, 150 neutral and 100 negative sentiments. We got these results on a small data set of 500 tweets, which is quite small for this case study but we would try to implement the same on larger data set of twitter corpus.

We will also try to overcome the different challenges related to task of sentiment analysis like Negation handling, handling of sarcasm sentences and sentences which use emoticons as way of expressing their opinions. Credibility of reviews is also an important challenging part of Sentiment analysis. We will try to improve these drawbacks and present an approach with better accuracy & efficiency.

IX. REFERENCES

[1] Vu Dung Nguyen, Blesson Vaghese, "Royal Birth of 2013:Analysing and Visualising Public Sentiment in the UK using Twitter," Research Gate, 2013

[2] Kun-Lin Liu, Wu-Jun Li, "Emoticon Smoothed Language Models for Twitter sentiment Analysis," AAAI, 2012

[3] Xiaowen Ding, Bing Liu- "A Holistic Lexicon-Based approach to Opinion Mining," Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, 2008

[4] Sarah E. Shukri, RawanI.Yaghi, "Twitter Sentiment Analysis: Case Study In Automotive Industry," IEEE-Jordan Conference on Applied Electrical Engineering and Computing Technologies, 2015.

[5] Subhabrata Mukherjee, Akshat Malu," TwiSent:A MultiStage System For analyzing Sentiment in Twitter," Proceedings of the 21st ACM international conference on Information and knowledge management, 2012

[6] Blessy Selvam, S. Abirami, "A Survey on OM Framework," International Journal of Advance Research in Computing & Communication engine- vol 2 Issue 9, Sept 2013

[7] Efthymios Kouloumpis, Theresa Wilson, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," Proceedings of the fifth International AAAI Conference on Weblogs and Social Media.

[8] Mostafa Karamibekr,Ali a.Ghorbani, ”A structure for opinions in Social Domains,” Social Computing (SocialCom), International Conference, 2013

[9] Brendan O’ Connor, Ramnath Balasubramanyan, ”From Tweets to polls:Linking Text Sentiment to Public opinion time Series,” Proceedings of the fourth International AAAI Conference on Weblogs and Social Media, 2010.

[10] M.Lovelin Ponn Felciah,R.Anbuselvi, ”A Study on Sentiment Analysis of Social Media Reviews,” IEEE-Second Conference on Innovations in Information Embedded & Communication Systems, 2015

[11] Prerna Awasthi, Ranjana Rajnish, ”Systematic Study on approach & tools used for opinion mining,”. https://www.scribd.com/doc/315237279/Opinion-Mining

[12] Anant Arora, Chinmay Patil, Stevina Correia, ”Opinion Mining: An Overview,” International Journal of Advance Research in Computer and Communication Engineering, 2015

[13] B.Pang and L.Lee, “Opinion Mining and Sentiment Analysis,” The Essence of Knowledge, 2008

[14] Erik Cambaria, Bjourn Schuller, Yunqing Xia, Catherine Havasi, ”New Avenues in Opinion Mining and Sentiment Analysis,” IEEE, 2013

[15] Asmita Dhokrat, Sunil Khillare, C. Namrata Mahender, “Review on Techniques and Tools used for Opinion Mining,” IJCAT, 2015