IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

A Mixed Approach for Data and Sentiment Analysis on Twitter

T. Jhansi Rani[1], K. Anuradha[2], P. Vijayapal Reddy[3]

Assistant Professor, IT, GITAM University, Hyderabad, India [1]

Professor, CSE, GRIET, Hyderabad, India [2, 3]

Abstract: Sentiment Analysis also termed as opinion mining is a process of obtaining knowledge on a specific subject from source content. Sentiment Analysis uses techniques from various disciplines such as machine learning, text mining and natural language processing. Data analysis is a process of extracting information from knowledge source for the purpose of supporting decision making process. The various phases in data analysis are data extraction, cleaning, transformation, integration and finally knowledge extraction. In this paper, a model has proposed to address the two issues such as sentiment analysis and data analysis on the data which is collected from Twitter. Various dimensions in the form of a new science behind the famous blog service known as Twitter are extracted. Twitter blog is used by the users to post the micro messages and their opinions know as tweets on a particular subject. The results are obtained for the tweets on the topic and classify them into positive, negative and neutral opinions. It is also analyzed the tweets arriving source, aggregate number of tweets generated by the users on a particular topic along a time series, number of tweets specific to language from different parts of the world.

Keywords: Data Analysis, Sentimental Analysis, Twitter, Map-reduce, Time series analysis, Data visualization.

I. INTRODUCTION

This Sentiment Analysis also termed as opinion mining is a process of obtaining knowledge on a specific subject from a source content. Sentiment Analysis uses techniques from various disciplines such as machine learning, text mining and natural language processing. Sentiment Analysis (SA) is a current research trend which is resulted from multiple disciplines such as Natural Language Processing (NLP), Text Mining (TM) and Machine Learning (ML). The research in this area has enormously increased with the Web 2.0 development. With the advent of Internet, people can post the messages in the form of their views, sentiments and emotions on a particular topic, product, people and on a life. Hence Internet is a best source of textual data collection for extracting opinions, sentiments and data analysis. The applications of Sentiment Analysis range from identifying the opinions on contestants in the elections from the political campaigns, opinions for increasing the consumption of the products and decision making for smart business. The target of Sentiment Analysis is mainly to identify the text under study is of subjective type or objective type. For the subjective tweets find whether the text is negative or positive or neutral tweet.

Twitter is an on line micro blog service. It is a networking service for social cause. Using Twitter, people can post and view messages of short length with the length of 140 character. These messages are termed as tweets. Twitter is a best source for huge collection of data for sentiment analysis as there are more than 500 million active users on Twitter. This makes twitter the largest platform where people express their views, experiences and comment on a variety of topics, products, situations and scenarios. Performing analysis on the data retrieved from such sources provides unparalleled conclusions on the subject matter under concern. A sentiment from a Tweet can be defined as tuple with five features such as < O, F, V, H, T >, where O is a target object, F is a feature of the object O, V is the sentiment value of the opinion of the opinion holder H on feature F of object O at time T. The value of V can be positive, negative or neutral.

The current trends in research on Sentiment Analysis are towards constructing generic models. The generic models are used to identify complex relations from text. As the web contains a large collection of data on various subject opinions in various forms such as blogs, reviews, the unsupervised approaches are more useful for Sentiment Analysis by considering the term co-occurrence within the text. Sentimental Analysis can be viewed in terms of consumer's perspective, producer's perspective and societal perspective.

In this paper, an extensive data analysis has performed on data fields obtained from twitter about a given topic or subject. Various inferences are drawn from Twitter such as comprehend the opinions which are divided into three labels such as positive, negative and neutral, aggregate number of tweets generated by the users on a particular discourse along a time series, accumulate the reach of the subject under discussion and infer to the source of the tweets.

II. LITERATURE SURVEY

The importance of the data generated by the Web 2.0 phenomena is readily apparent. Pang et al. in [1] addresses the use of availability of huge amount of data (CGM) from

DOI 10.17148/IJARCCE.2016.5608

news papers, blogs and chatting rooms for profit and risk analysis for an organization. The complexity issue is still relevant even when narrowing the search space to a single source of information.

The useful data is just as plentiful as the irrelevant. There are endless amounts of data being produced in outlets across the Internet. As in Moore et al., 2011 [2], the data in the Internet is a important source for making decisions on marketing strategies, psychology of humans and others interesting topics opinions, views, moods, and attitudes.

The challenge that exists after the search space is established is to locate the relevant data. After the relevant data is established it can then be assessed for sentiment. These two stages are commonly referred to as subjectivity classification and sentiment classification. Classification based on the subjectivity is to identify the author opinion or to extract the fact. According to Pak et al., 2012 in [3], classification on subjectivity can stop the polarity of sentiment by considering not relevant or misleading text.

As in Kuffman et al., 2010 [4], depending on the application, contextual matching or similar may be applied to the resulting data that is already deemed subjective. Sentiment classification has some variation among designers of each approach but ultimately serves the same abstract purpose.

As in [1], Sentiment analysis categorizes the data from web comments into three types such as positive, negative and neutral categories. As in [5], there are two types of classifications on sentiment analysis. They are binary sentiment classification and multi-class sentiment classification. Multi-class sentiment approach assigns granularity to the text instead of assigns labels.

Human emotion spans a much more complicated spectrum than the simple black and white notions of positive and negative. Human have the strange capability to love and to hate something at the same time. This is easy for humans to decipher but much more complicated for a machine. A few different approaches have been developed to create more accurate results to address the mentioned problem. General polarity-based sentiment classification is a great step forward from the previous contextual only approaches [6].

There are more elaborate designs that break down the content into greater detail allowing for more results that are more specific. After the sentiments are established each sentiment analysis system will then use the results in ways appropriate to the application. Qiu 2010 in [4], developed an idea titled dissatisfaction-oriented Advertising Sentiment Analysis (DASA) that combines traditional sentiment analysis with basic keyword matching. In this approach the software detects the negative sentiment of certain products. The advertising on the web page that contains the text then displays a product that has the positive attributes that the original text complained about.

The example used in [4], is one in which the writer on the forum complains about the safety of a car. After the comment is posted and a new user loads the forum page, the advertisements are re-established based on the new comment. The new advertisements now have a Volvo ad that exemplifies new safety features and a history of safe production standards. The uses of sentiment analysis can be applied to many industries. Any company under the scrutiny of public opinion should be analysing all relevant data.

As in [7], the most popular and basic use of sentiment analysis involves extracting knowledge from the text written in the form of reviews by the customers on products or services and then categorize the reviews into the opinions of type positive, negative or neutral. It is this type of classification that has become one of the foci of recent research endeavours sponsored by companies that realize the potential value of sentiment analysis on their data.

According to [8, 9], manufacturer of a product easily gather marketing strategies from the competitors and benchmark information of the product then compare and takes strategic decisions. Sentiment analysis will allow businesses the ability to use their pre-existing text data in ways to benefit several departments within the traditional business structure. Businesses only require new software plus the necessary hardware to handle the new processing techniques and storage of the results. Marketing companies and advertising branches of businesses are easy benefactors of the resulting conclusions derived from sentiment analysis.

It is to the pure benefit of companies to implement sentiment analysis if these companies have the relevant information available for such a process. The branding and marketing aspects of businesses revolve around the consumer psychology. Sentiment analysis could reveal this psychology in a form that could be used for further analysis and study. Sentiment analysis is a useful tool for all users of the Internet. Emotional classification and organization of content will be a beneficial contribution to the vast reservoir of data the Internet holds.

III.PROPOSED SYSTEM

The research until now was focused mainly on the different ways to calculate sentiment of a tweet and various techniques and methods have been developed. There is a need to develop a comprehensive model which includes sentiment as well as data analysis on other fields to know the online community. It is proposed a new dimension to twitter analysis by not only calculating the sentiment but also by drawing the various inferences such as Tweet sources, Tweet languages, Tweet reach and Tweet analysis based on time series.

The integration of sentiment analysis with data analysis on other fields provides an exceptional knowledge in political scenarios, E-commerce applications, education and in many other domains. The process of sentimental analysis and data analysis has performed with various steps. Firstly the data is retrieved through twitter API. Then it is stored in the database. In the second step, the flow shifts into two processes such as parse the data through the natural language processor and mine the data for aggregation towards data analysis. The process is presented in Figure 1.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

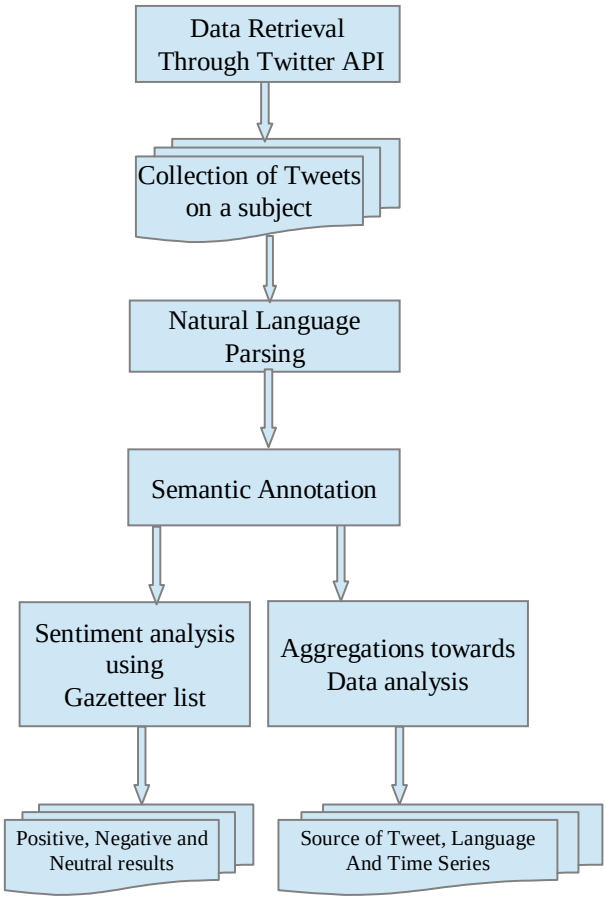

Fig 1: Flowchart of the proposed model

IV. RESULTS AND DISCUSSIONS

Tweets were collected on four topics such as Amazon, android, iphone6 and Obama. These tweets were collected by suing Twitter APIs. The REST APIs are used to access the data from Twitter and post data on the Twitter. It is also useful to read author profile and the details of the followers of the author. Twitter applications and users can be identified using REST Oauth API. Responses can be extracted from the JSON API.

New responses to REST API queries can be delivered from streaming APIs by establishing HTTP connection. Based on the search query matching, updates are received in sync with user profile updates. REST APIs the Streaming APIs is a better solution if the application is rate-limited for over-polling. Twitter offers several streaming endpoints such as public streams, user streams and site steams and each is customized to certain use cases.

The collected tweets were parsed using Natural Language Tool kit for stemming, tagging, syntactic parsing and term extraction. The terms are annotated with WordNet for grouping the related terms together. The positive, negative and neutral sentiments are identified using gazetteer list. The results obtained from the process are as follows in Table 1.

Table 1: Tweets on various subjects with sentimental analysis

S.No	Subject	Positive	Negative	Neutral	Total
1	Amazon	34558	4518	17804	56880
2	Android	59944	11429	30178	101551
3	iphone6	9937	2475	11871	24283
4	Obama	12846	13211	9685	35742

The source of tweets generated on each topic was identified using MongoDB, which is a cross-platform document-oriented database with an aggregation using source of post as a filter point. Table 2 contains the list of sources and the number of tweets from each source on the Amazon.

Table 2: Sources of Tweets on Amazon

Source	No. of Tweets	Source	No. of Tweets
Amazon Lab	80	Facebook.com/ twitter	357
BufferApp	311	ebayrt.co	368
Twitter/Android	1718	Feed140.net	129
Twitter/Iphone	2290	Google	180
Twitter.com	18016	dlver.it	12146
Twitbot	1484	IFTTT	2053
Twitter/Ipad	304	Instagram	264
Twitterfeed	2396	Linkis	269
Mobile.twitter.com	884	HootSuite	1512
SocialOoph	354	RoundTeam.co	668
SudoCart	1739		

Tweet generated in different languages for Amazon are listed in Table 3.

Table 3: Tweets from various languages

Language Codes in Twitter	No. of Tweets	Language	No. of Tweets
Ar	12	ko	2
Bn	3	lt	15
Ba	49	lv	2
De	537	nl	43
El	2	no	50
En	37551	pl	24
Es	1220	pt	121

DOI 10.17148/IJARCCE.2016.5608

Et	42	ru	6
fa	2	sl	13
fi	9	sv	13
fr	293	th	3
hi	2	tl	39
ht	139	tr	25
hu	4	uk	2
in	133	und	1506
it	596	vi	2
iw	1	zh	15
ja	14406		

Dynamic interpretation of number of tweets originating along a time period is shown Figure 2.

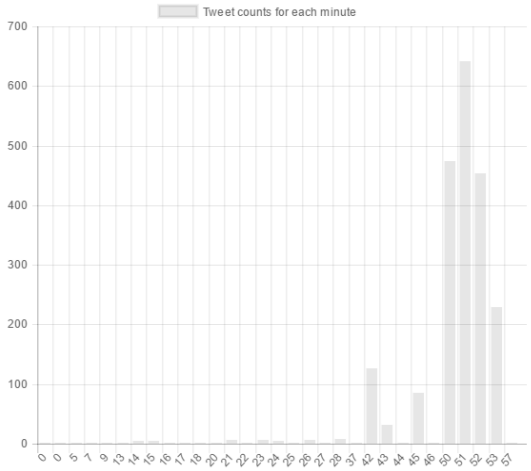

Fig 2: Language and Time series

V. CONCLUSIONS AND FUTURE SCOPE

Sentiment analysis methods till now have been used to detect the polarity in the thoughts and opinions of all the users that access social media. Analysis on the data to understand the behaviour, interest, opinions and thoughts of the people is of interest to the Researchers and Businesses. Organizations use this kind of data to analyse and evaluate their advertisement strategies and to improve products sales.

There is too much potential in machine learning, overtaking some of the manual labour of some lexicon based tasks that are labour intensive. Sentiment analysis is based on the fact that the feelings that are come from the inherently human beings, it becomes important for many business decisions in future. Improved accuracy and consistency in text mining techniques can help overcome some current problems faced in Sentiment analysis.

There is a lot of scope in analysing the video and images on the web. Now a day, with the advent of Facebook, Instagram and Video vines people are expressing their thoughts with pictures and videos along with text. Sentiment analysis will have to pace up with this change. The use of punctuation is an obstacle in Sentiment Analysis which is under research as well. The field will have to combine with effective computing, psychology and neuroscience to converge on a unified approach to understanding the sentiments better.

Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template.

REFERENCES

[1] Pang, Lee, **"** Opinion Mining and Sentiment Analysis ", Journal Foundations and trends in information Retrieval, Volume 2, Issue 1-2, January 2008,Pages 1-135.
[2] Efthymios Kouloumpis, Moore "Twitter Sentiment Analysis: The Good the Bad and the OMG", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 538-541,2011.
[3] Alexander Pak, Patrick Paroubek," Twitter as a Corpus for Sentiment Analysis and Opinion Mining", 1320-1326
[4] Willso, Kuffman,**"** Recognizing contextual polarity in phrase-level sentiment analysis", Proceedings of the conference on Human language technology and empirical methods in Natural Language Processing, 2010, pages 347-354.
[5] Pang, B. and Lee, L. 2008. "Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval". Pg 235-241.
[6] Clarence, Aravindh S, Shreeharsha A.B, "Comparative Study of the New Generation, Agile, Scalable, High Performance NOSQL Databases", International Journal of Computer Applications (0975 – 888) Volume 48– No.20, June 2012.
[7] Xiaolin Wang,Sch. of Software, Shanghai Jiao Tong Univ., Shanghai, China; Haopeng Chen;Zhenhua Wang." Research on Improvement of Dynamic Load Balancing in MongoDB".
[8] Argamon, Garg, "Using appraisal groups for sentiment analysis", Proceedings of the 14th ACM international conference on Information and knowledge management, pg 625-631
[9] D. M. Bruls, C. Huizing, and J. J. van Wijk, "Squarified treemaps", In Joint Eurographics and IEEE TCVG Symposium on Visualization, pages 33–42, 1999.

DOI 10.17148/IJARCCE.2016.5608