International Journal of Scientific Research and Management (IJSRM)
||Volume||5||Issue||07||Pages||5816-5820||2017||
Website: www.ijsrm.in ISSN (e): 2321-3418
Index Copernicus value (2015): 57.47 DOI: 10.18535/ijsrm/v5i7.05

Sentiment Analysis using Enhanced Data Dictionary

Neha Tyagi[1], Dr. Bhaskar Pant[2], Neelam Singh[3]

[1]Graphic Era University, M.Tech Scholar Department of CSE,
Dehradun, U.K, India
[2] Graphic Era University, Associate Professor, Department of Information Technology
Dehradun, U.K, India
[3] Graphic Era University, Assistant Professor, Department of Computer Science & ENG,
Dehradun, U.K, India

ABSTRACT: ***As utilization of internet business is expanding at a large domain consumers not only buy products but also give their input recommendations that will be very advantageous to different clients. Individual's opinion and experience are extremely important information for decision making process. Now a days many websites encourages clients to express and trade their ideas, opinions, views, suggestions related to particular product, its policies, services publically. Sentiment analysis is one of the most sought after technique to review information posted by user to gain insights for decision making process. Data collected for sentiment analysis from heterogeneous sources often comprises of missing values, noisy data etc. which needs to be preprocessed using data dictionary. In this paper we are aiming to provide enhanced data dictionary to handle data preprocessing more efficiently and accurately required for sentiment analysis. We worked on a case study based on twitter data to find the brand reputation of three popular mobile brands based on our enhanced data dictionary.***

Keywords: Sentiment Analysis, preprocessing, data dictionary.

I. INTRODUCTION

The web becomes a virtual space where to express and share people opinion, feelings, and views about a marketing behavior. Online networking is impacting clients by forming their state of mind, attitude and behavior. Sharing the opinion becomes extremely normal task on the social media. They express their sentiments in media, organizations web journals, discussions and so on. They can write review about the product that they have brought. These surveys are utilized by potential clients to choose whether to buy product or not or to select another option [1].Additionally utilized by the item makers to distinguish the item, popularity, brand reputation and to discover information about the competition. Social media is a decent approach to measure customer loyalty, keeping track on their supposition towards brands on items. Extracting the valuable information gives rise to new territory of research called opinion mining and sentiment analysis. Opinion mining is procedure of programed extraction of knowledge from the opinion of others. It is new zone of research that deals with data retrieval knowledge discovery from text mining and natural language processing. Sentiment analysis which is used to record the state of people mind towards specific subject or any related item.[20] It includes building a framework to gather and inspect opinions about the item made in blog posts, remarks, audits or tweets.[6] Sentiment mining refers to the utilization of common natural language handling, content investigation and computational linguistic to recognize and extricate subjective data in source materials.[21] Feeling investigation is generally connected to audits and online networking for an assortment of uses, running from promoting to client benefit. An essential undertaking in feeling examination is ordering the extremity of a given content at the archive, sentence, or highlight/viewpoint level. Basically sentiment analysis is about deciding the subjectivity, extremity (positive or negative) and extremity quality (pitifully positive, somewhat positive, firmly positive, and so on.) of a bit of content – at the end of the day: It can be helpful in many ways.[12] For instance, in marketing, it can be very helpful in judging the accomplishment of the advertisement or new brand in the market or services that are popular and even recognize which one is better and the particular feature of the product. The object of SA is the tem or an administration whose survey has been made open in the Internet. [11]Through media people can express their feelings, issues, problems and expert can use them to analyze it. In promoting and publicizing area opinion mining and sentiment analysis is being very important.so opinion mining and sentiment is needed to analyze it. [8] In this paper we are going to find out the sentiment analysis of brand by using enhanced data dictionary. Some of the effective pre processing approaches are used which gives more preprocessed data by using enhanced data dictionary.

1.1 Sentiment analysis has been investigated at three levels

1.1.1 Document level-The task at this level is to arrange whether an entire felling record a positive or negative review. The framework decide whether the survey communicates a positive or negative supposition about item. [17]

1.1.2 Sentence level-The task at this level goes to sentence and decides of each sentence communicated a positive, negative or unbiased conclusion.

1.1.3 Feature/Aspect level-Product feature are characterized as item qualities or components. Examination of such components for identifying sentence of document is called aspect or feature. [17]

II. RELATED WORK

In 2012, Balakrishnan Gokulakrishn et. al. [6] proposed an approach where a plugged stream of tweets from the Twitter microblogging webpage are preprocessed and grouped in light

of their emotional content as positive, negative and irrelevant; and investigates the execution of different ordering calculations in light of their precision and recall in such cases

In 2013, MichalMunket.al. [10] proposed a philosophy which is a plan and suggestions for dependable information procurement from e-documents in procedure of finding the sequence patterns. They suggest groupings recognizable proof in view of sentences when exchange/arrangement model is utilized, accepting that the arrangement of specific issue does not require other way to deal with succession distinguishing proof. Normally, this way to deal with arrangement distinguishing proof is by all accounts the most appropriate if there should arise an occurrence of issues explaining from the territory of quantitative linguistic structure examination

In 2013, Mohamed M. Mostafaet. al. [12] proposed that the examination creators utilized an arbitrary example of 3516 tweets to assess buyers' feeling towards well known brands, for example, Nokia, T-Mobile, IBM, KLM and DHL. They utilized a specialist predefined dictionary including around 6800 seed adjectives with known introduction to direct the investigation. Comes about demonstrate a by and large positive buyer sentiments towards a few renowned brands. By utilizing both a subjective and quantitative philosophy to break down brands' tweets, this review adds broadness and profundity to the level headed discussion over states of mind towards cosmopolitan brands.

In 2014,Calvinet. al.[2]proposed a model where sentiment extremity of Twitter surveys are measured utilizing Naive Bayes classifier strategy. The model demonstrates a promising come about on characterizing the ubiquity in light of consume satisfaction and along these lines characterizing the best supplier to be utilized.

In 2014, Zhao Jianqiang et. al. [4] proposed model that talked about the impacts of content pre-processing strategy on sentiment characterization execution in two sorts of classification task, and summed up the grouping exhibitions of six pre-processing techniques utilizing two feature models and four classifiers on five Twitter datasets. The outcomes demonstrated that the Naive Bayes and Random Forest classifiers are more sensitive than Logistic Regression and Support Vector Machine classifiers when different pre-processing techniques were connected.

In 2014, Aizhan Bizhanovaet. al. [8] proposed a model for naturally characterizing the opinion of Twitter messages toward item/mark, utilizing emoticons and by enhancing pre-processing steps keeping in mind the end goal to accomplish high exactness.

In2015, Aashutosh Bhatt et. al. [5] proposed a framework that plays out the classification of customer reviews after by discovering estimation of the surveys. A rule based extraction of item highlight assumption is likewise done. The outcomes demonstrated that characterization of reviews alongside sentimental investigation expanded the exactness of the framework turn provides accurate reviews to the user.

In 2015, Nur Azizah Vidya et. al. [11] proposed how to take care of the issue of how to improve the brand reputation of mobile suppliers in view of individuals reviews on their administrations quality. They are attempting to cover the issue by measuring brand reputation in view of consumer opinion through client's assumption investigation from Twitter information. This paper likewise talks about some associated business bits of knowledge in a telecommunication services industry. In view of the general correlation of these five items, the NBR scores for PT XL AxiataTbk, PT TelkomselTbk, and PT IndosatTbk are 32.3%, 19.0%, and 10.9% individually.

In 2016, Sandip D Mali et. al. [1] proposed another framework called SentiView which a vocabulary based approach for sentiment investigation. They have gotten high accuracy because of preprocessing and expulsion of non-opinion tweets from data.

In 2016, Sanjana Woonnaet. al.[3] proposed a framework that examinations tweets into three classifications which are positive, negative and neutral utilizing supervised learning approach After the execution, the outcomes demonstrated which viewpoints individuals like or aversion and how feelings on motion pictures changes over a timeframe

In 2017,Yue Guo et. al. [14] proposed the key measurements of client administration voiced by inn guests utilize an information mining approach, (LDA). The enormous informational collection incorporates 266,544 online surveys for 25,670 lodgings situated in 16 nations. LDA reveals 19 controllable measurements that are key for inns to deal with their connections with guests. They additionally discovered contrasts as indicated by statistic portions.

In 2017,Kai Yang et. Al [16] proposed a highly effective hybrid model combining different single models to overcome their weaknesses. They build the sentimental dictionary from exterior data. As single model have many limitations and weakness. That why they build a hybrid model by combing many single approaches to overcome those limitations of single model. The experimental results show that our hybrid model shows very great performance. In hybrid model 2 approaches that are SVM and GDBT (Gradient boosting decision tree) are combined together that are based on stacking approach.

Based on the above analysis we found that the existing dictionary (English language) being used for data preprocessing does not works well on slangs....So we in this paper has proposed an enhanced data dictionary to perform data preprocessing more effectively and efficiently.

III. METHOD DESCRIPTION:

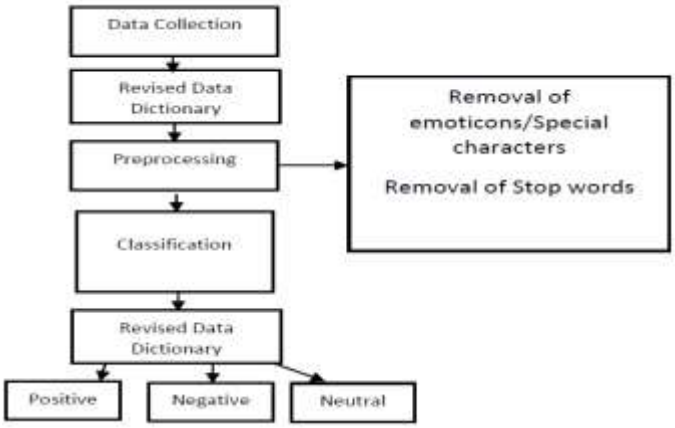

Figure 1: Work flow sentiment analysis

3.1 Data Collection- As social media is emerging day by day. User can get most of the information on the social media.[8]User opinion becomes a major issue for the quality of any service and improvement of any product. Millions of people post their opinions, views, feelings related to a specific. Twitter is a famous microblogging services that allows individual to post their messages on twitter.[1] we collected the data by writing program script in python that will collect the raw tweets by using the twitter API .[4] Twitter API is used to collect the data online. Streaming API is used to collect real time tweets. Data is gathered of 3 months from March to May, 2017. Data are extracted from twitter account of 3 mobile providers that are (Samsung, Motorola and IPhone).The collected data is stored into the .txt file

3.2 Preprocessing- Preprocessing the information is the way toward cleaning and preparing the content. Online text contain text with lot of informal and noisy data for example html, URL, hashtags other informal symbols, special characters.[4] Keeping those words makes the dimensionality of the issue high .so it becomes necessary to deal with such kind of data.

3.3 Removal of URL's /Hashtags- Twitter contains lot of data and information. In that data their contains a URL's links .As that URLs don't convey much data with respect to the estimation of the tweet. So URL are expelled from the tweets to refine the tweets as those URL does not convey any sentiment or any other useful information.[9]Hashtags are another substance that are usually present in tweets. Hash labels substances should be removed from the tweets as these entity does not play any significance role in sentiment.

3.4 Removal of emoticons and special character/Numbers- Tweets contains many informal language like emoticons and many other complex special characters.[3]These special character does not contain any important or meaningful sentiment.so to decrease the complexity we should remove these special characters that do not play any role in sentiment. Numbers are also of no use when the focus is on measuring the sentiment towards any product.so numbers also be removed from the tweets to refine tweets for better results.[9].

3.5 Removal of Stop words- Stop words are most commonly utilized words in any language.[4] They are the common words that are used commonly like 'the' 'at' 'of' 'is' 'was'. These words did not have any significant meaning in measuring the sentiment and only increases the complexity.[1] If we remove such kinds of words then we can focus on important words instead of these kinds of words so these words should be expelled from the text for better accuracy and to achieve great performance.

3.6 Revised Data Dictionary- The enormous advances in social media and their power to reflect and influence public opinion made them a domain of great interest for marketers, communication specialists and companies that want to advertise their products and services, or simply want to boost and monitor their brand name.[11] sentiment words are labeled to construct sentiment dictionaries. The online data dictionary available has some of the limitations. We have refined the data dictionary in our research. Proposed work will be focused on improving data dictionary by inserting (positive dictionary, negative dictionary, slangs dictionary, abbreviations) with respect to English language to enhance data preprocessing task so that we will have more accurate results.

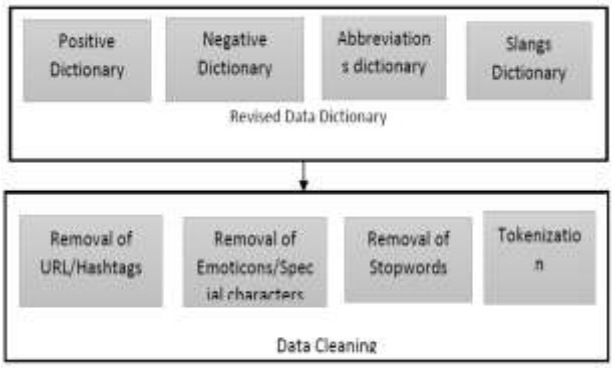

Figure 2: Preprossing with the help of enhanced data dictonary

3.7 Classifier- In classification stage we have applied machine learning classifier on the preprocessed data that is Naïve Bayes. It compares the tweets from given dictionary. Then the tweets are classified into the positive, neutral and negative.[5].

IV. IMPLEMENTATIONS AND THE RESULT

The usage and comparative efficiency of the enhancing data dictionary has been calculated by using a case study based on the analysis of brand reputation of 3 popular mobile brands.

4.1 Case Study- Analyzing brand reputation of three popular mobile brands 'Samsung', 'IPhone' and 'Motorola' using Enhanced data dictionary for preprocessing the data and applying naïve bayes classification for the same.

4.2 Methodology Implemented- The actual data is extracted from the twitter by using the twitter API. We collected data by writing program script in python that will automatically collects raw tweets from the twitter. Collected data of 3 brands that are (Motorola, Samsung, Iphone) from month of March to May 2017.Approximately 50000 raw tweets are collected and that tweets are saved in the .txt file.

Then the preprocessing is done on the raw tweets to refine them. Various steps that are taken in our approach are-

- Removal of URL/Hashtags
- Removal of Emoticons/Special Characters
- Removal of Stop words

4.3 Revised Data Dictionary- We have enhanced data dictionary by inserting (positive dictionary, negative dictionary, slangs dictionary, abbreviations) with respect to English language to enhance data preprocessing and data cleaning more efficiently and will provide better results than the previous ones which will give more good results for brand reputation.[5].

Last step is to pass the data by the classifier which will give results (Positive, Negative, Neutral)

V. RESULTS

After passing data into the classifier we get to know the proper results of all the brand.We have done analysis of 3 brands with existing data dictionary and one is by applying the enhanced data dictionary with which preprocessing becomes more efficient and appropriate then the previous one and gives better results from the existing dictionary.

Table 1 : Classification results for three brands using existing data dictionary

Existing Data Dictionary				
Brand	Positive	Negative	Neutral	Total words
Motorola	564	1894	47590	50048
Samsung	549	2230	37780	40559
Iphone	513	3867	42620	47000

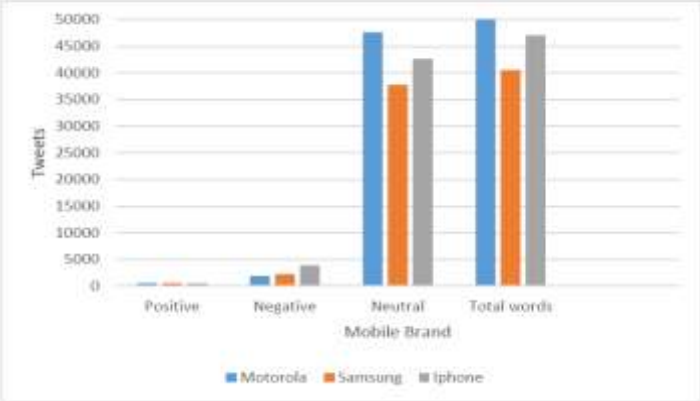

Figure 3: The overall reputation score of Motorola, Samsung and IPhone

Table 2: Classification results for three brands using existing data dictionary

Enhanced Data Dictionary				
Brand	Positive	Negative	Neutral	Total words
Motorola	4207	3345	42496	50048
Samsung	2178	2870	35511	40559
Iphone	3146	2090	41764	47000

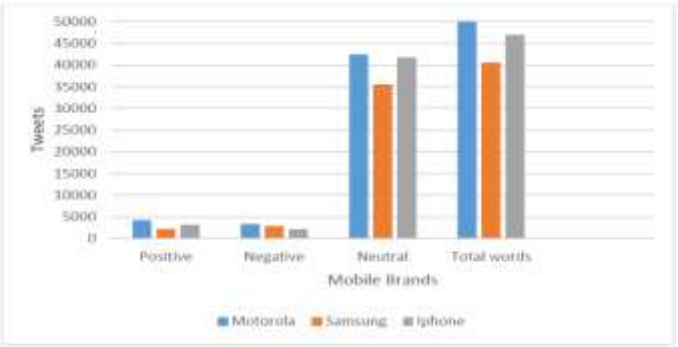

Figure 4: The overall reputation score of Motorola, Samsung, Iphone

From the above implementation it is found that the enhanced data dictionary performs well as compared to the existing dictionary giving more refined and accurate results. From the above study we came to the findings that Motorola is the most preferred mobile brand as compared to Samsung and IPhone.

VI. CONCLUSION

Social media was giving a new environment to the customers and the users to exchange their ideas and information and giving feedbacks and various aspects about the products and the services. Many of the organization, companies are using internet as a source to make their product popular and user having to know about the reviews of products. Extracting the valuable information gives rise to new territory of research called opinion mining and sentiment analysis. Opinion mining is procedure of programed extraction of knowledge from the opinion of others. We are finding the brand reputation (Motorola, Iphone, Samsung) by using the sentiment analysis using enhanced data dictionary. After enhancing dictionary we get the brand reputation more accurate than the previous ones. Preprocessing techniques provides more efficient results using the enhanced data dictionary.

References

[1] Sandip D Mali, Dr. Sachin N Deshmukh, Ashish A Bhalerao, "SentiView: A Lexicon Based Approach for Twitter Sentiment Analysis" Vol. 4, Issue 11, November 2016

[2] Calvin and Johan Setiawan, "Using Text Mining to Analyze Mobile Phone Provider Service Quality (Case Study: Social Media Twitter)" International Journal of Machine Learning and Computing, Vol. 4, No. 1, February 2014

[3] SanjanaWoonna and PriyankaGiri, "Sentiment analysis of twitter data" International Journal of Innovation and Technology, 2016

[4] Zhao Jianqiang, Gui Xiaolin "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis"DOI10.1109/ACCESS.2017.2672677, IEEE Access

[5] Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande "Amazon Review Classification and Sentiment Analysis" International Journal of Computer Science and Information Technologies, Vol. 6, 2015, 5107-5110

[6] BalakrishnanGokulakrishnan "Opinion Mining and Sentiment Analysis on a Twitter Data Stream"The International Conference on Advances in ICT for Emerging Regions - ICTer2012 : 182-188

[7] Ajay Deshwal, Sudhir Kumar Sharma, "Twitter Sentiment Analysis using Various Classification Algorithms"978-1-5090-1489-7/16/$31.00 ©2016 IEEE

[8] AizhanBizhanova, Osamu Uchida, "Product Reputation Trend Extraction from Twitter"Social Networking, 2014, 3, 196-202

[9] NER K. Jayamalini, "Tweet data Preprocessing and Segmentation to NER"International Journal of Scientific & Engineering Research, Volume 8, Issue 1, January-2017 2075 ISSN 2229-5518

[10] Michal Munk, Martin"Data Pre-Processing Evaluation for Text Mining: Transaction/Sequence Model"International Conference on Computational Science , 11. The Third Information Systems International Conference

[11] NurAzizahVidya, Mohamad Ivan Fanany, IndraBudi"Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers"Information Systems International Conference (ISICO2015) doi:10.1016/j.procs.2015.12.159

[12] Mohamed M. Mostafa"More than words: Social networks' text mining for consumer brand sentiments"Instituton Universitário de Lisboa, Business Research Unit, Avenida das Forças Armadas, Lisbon, Portugal
[13] Yue Guo, Stuart J. Barnes ,QiongJia"Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation"0261-5177/© 2016 Elsevier
[14] Kai Yang "An Effective Hybrid Model for Opinion Mining and Sentiment Analysis"465978-1-5090-3015-6/17/$31.00 ©2017 IEEE
[15] Tahura Shaikh "A Review on Opinion Miningand Sentiment Analysis"International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Science & Information Technology
[16] Dipak Gaikar ,BijithMarakarkandy"Product Sales Prediction Based on Sentiment Analysis Using Twitter Data"International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015,
[17] Federico Neri, Carlo Aliprandi"Sentiment Analysis on Social Media"2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining
[18] Apoorv Agarwal "Sentiment analysis of twitter data"Department of Computer Science Columbia University New York, NY 10027 USA
[19] XingFangandJustinZhan "Sentiment analysis using product review data"JournalofBigData (2015) 2:5 DOI10.1186/s40537-015-0015-2
[20] R.Madhvi "Analytical mapping of opinion mining and sentiment analysis research during 2009-2015" Information processing and mangment,2016
[21] Swati N. Manke "A Review on: Opinion Mining and Sentiment Analysis based on Natural Language Processing"InternationalJournal of Computer Applications (0975 – 8887) Volume 109 – No. 4, January 2015