Review on Sentiment Analysis of Twitter Data with Some Classifiers Ensembles

Smridhi Chawla[1], Priya[1], Anchal Garg[1], Bharti Jha[2]
[1]UG Students, [2]Assistant professor
[1,2]Department of Computer Science and Technology,
Manav Rachna University, Faridabad, Haryana, India
Email:*Smridhichawla22@gmail.com*
DOI:

Abstract

In the research paper, the focus is on the citizen's emotions towards different organization, brands, and on the different interests. By Sentimental Analysis on twitter, which give the attractive and speedy way to the people to enjoy the above mentioned different things. Apart from the Sentimental Analysis, the semantic approach is to increase the features and get more accurate results. These are just applying some techniques to differentiate the twitter analysis's data. By this, the result shows some harmonic score to investigate the positive and negative data.

Keywords: *Twitter, Social Media, Naïve Bayes, SVM*

INTRODUCTION

By the commencement of social media, people started communicating with each other, shared their thoughts and views and so on. In today's time, there are more than 2500 million messages exchanged on twitter and more than 600 users are there. As we know everything has some good and bad in it, although twitter has become the most usable app for the people but it has some problems in it. There is a problem related to its small length and irregular contents. In this the initial concernis of finding new methods to run such analysis such as performing sentiment label propagation on Twitter follower graphs [2]. The second is focused on identifying new sets of features to add to the trained model for sentiment identification, such as micro blogging features including hash tags, smiley [4], the presence of intensifiers such as all-caps and character repetitions [3].

The researchers in this paper worked with the second concept, by distinguishing someset of columns that are taken out from the well-formed illustration of the entries in tweets. The well- formed illustration contains the linguistic concept (e.g. people, cities, companies etc) that refers to the following (e.g. Bill Gates, America, Idea etc). The reason behind such things is that these entries and concepts make theco-relation with the positive and negative blogs. After having the ideas of this co-relation, they help in determining the same entries and hence increase the sentimental accuracy.

In the previous experiment, results shown are better, which are adding allowable aspects by infusion. Hence by consolidation the aspects may introduce to Naïve Baye's (NB) chart by using the infusion approach. By doing the several approaches and tries, there are three evidencesgrouped which are from the twitter, they are STS that stands for The Stanford Twitter Sentiment, OMD that stands for The Obama-McCain Debate and the last is (HCR) Health Care Reform. Basically at the end they wanted to show 3 types that are by applying the well – formed content in micro blogging which has an advantage on different techniques to avoid non-positive text analysis in big data.

The remaining research paper contains the matter as follows: Part second is all about their related work on the sentimental analysis over twitter. Part third tells about their three datasets to improve the micro bloggings by avoiding non-positive contents in the big data. Part fourth is all about the approaches. Part fifth is about the baseline and the consequences. In part sixth it is all about the consequences of the experiments. Part seventh is the future approaches and last is Part eighth which is conclusion.

Sentimental analysis is classified into two parts: first is filtering of text that transform unstructured text data into the intermediate manner and second is refining of knowledge that decreases the abilities from the intermediate form.Intermediate Form can be of two types: Semi-Structured like graph representing or structured like representing of relational data's. It can be Documented-based form or can be Conceptual-based form in which every entities display some logics or data in a particular domains. Documented-based form are clustering and categorizing. Predictive model and Associative Discovery are the features of text analysis.

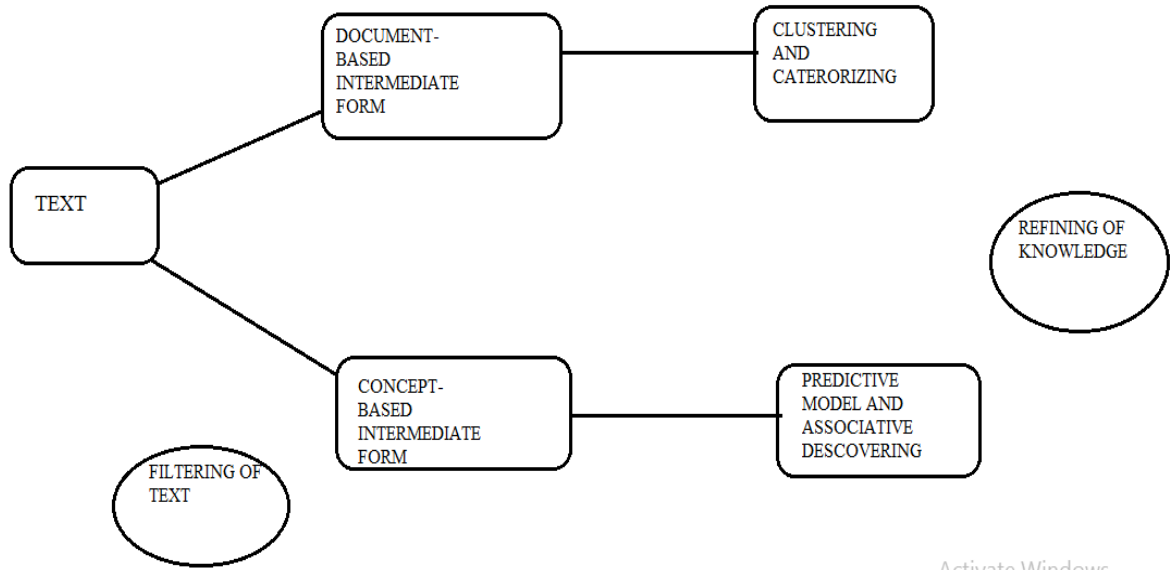

Figure 1:*Layout of Sentimental Analysis.*

In this, the unstructured data is transfer into Intermediate form by filtering of text. Intermediate form if further classified into Document-based Intermediate form and Concept-based Intermediate form.

Many tweets contain undesirable data, smiley's, pictures. So they are then preprocessed and changed in a desirable way which tells the correct public view. There are three ways of preprocessing the tweets: Tokenizing, Removing of Undesirable Words from messages, by using some Special Characters to enhance the messages or using hash tags and others tags. First is *Tokenizing*of tweets where these tweets are divided into individual words through spaces in between so that the undesirable symbols are removed like emotions. Second are Undesirable words where the wordswhich do not show any kind of emotion are the undesirable words. After splitting the words from tweets. For example "this is a good product" so the words like "is", "a", etc. are removed from the tweets. Third is using the Special characters in the tweets. Many tweets contain hashtags (#), @ tags etc.are replaced. For example #Windows is written as Windows, @ViratKohli is

written as a User. Tweets having prolonged words that show the emotions like “This is a verrrrycooooool product.” are then written as “This is a very cool product.” After these preprocessing ways tweets become ready for sentiment analysis.

RELATED WORK

The phenomenon of natural language processing of tweet is harder than conventional. This could be due to the small size of tweets information, or because of using informal language and the change in the typing language of tweets. The work is in progress to achieve the new-based approaches to enhance the text analysis. Go et al.[4].By Point of Sight different n- gram features are introduced under the supervision of Naïve Baye’s(NB), Maximum Entropy(MaxEnt) and Support Vector Machines (SVM);by this accuracy increases. There were two classifiers Barbosa and Feng, who concluded that by introducing more numbers of infrequent words in the tweets, may reduce its performance. Inspite of this, they suggested to introduce hash tags, replies, punctuation, etc.They said that using the above mentioned entities the accuracy for analyzing text may increased by some percentage. There was another researchers named Kouloumpis et al[6]. They said that by using smiley, abbreviation and intensifiers will shows the best result by using n-gram features may improve accuracy.

Many researchers have given their own suggestion in improving the accuracy of text analysis. Anotherresearcher named Sperious et al[14] has also given their own suggestion on improving the text analysis that they are constructing a graph which consist of hash tags and smiley where smileys are of tweets word unigram on nodes. A Label Propagation is passed through these nodes.It is the technique whose output performance which is trained from the noisy label gave the accuracy by 84% on the twitter sentimental test.

The research paper is about the sentiment analysis, which is identifying and then categorizing the opinion to judge the attitude towards particular thing. It is said that Sentiment analysis is a kind of Natural Language Processing in many levels. It can be classified as document level, sentence level, and phase level. In this, the researchers also use Naïve Bayes, maxEnt and SVM- Support vector machines. They uses emotions to describe the attitude towards particular topic by positive smileys ☺ and negative smileys ☹ . In this research paper, number of researchers had propose some of their methods to make more precise result towards Sentiment Analysis.[6]

This research paper is based on Graph-Based Hashtag Sentiment Classification Approach in which a hashtag graph HG=fH is used where the edge set consists E of links between hashtags and each edge eij represents an undirected link between hash tags hi and hj , which co-occur in at least one tweet. The baseline approach is developed on sentiment analysis, results of the tweets containing the hashtag through simple voting strategy. The performance of this intuitional approach is not encouraging. In order to improve the hashtag-level sentiment classification, the researchers proposed a graph model to boost the result from the voting baseline. For eg:- for a new product it is expected to present a list of related features together with typical sentiment expressions (negative or positive) .[7]

The research is done on sentiment analysis. In this a given sentence is determined on the basis of its positive or negative sentiment aspect. There are two kinds of approaches used this paper: one is lexicon approach and the other is the machine learning approach. In this research paper: First, sentiment analysis is been carried out at an entity level which is been done on a fine level. Second, sentiment analysis is been done on three

classes that is positive, negative and neutral. Therefore, positive and negative classes are been identified through lexicon-based approach and neutral are identified through actual opinionated.[8]

In this research paper, the authors discuss about the interest of people on social networking sites, sentiment analysis and blogging sites. The author discuss about the different author's works and basically did further modification. The first researches the author discussed is pang and lee, 2008, where they both worked on the existing approaches and techniques to retrieve the data. The second is Yang et al., 2007 where the author use web blogs ad emotions icons to indicate the user's attitude and also SVM. Basically the author discuss about the worked done by the different authors onto this and the do some further modifications.[9]

The work presented in this paper specifies a novel approach for sentiment analysis on Twitter data. To reveal the sentiment, we extracted the opinion words (a combination of the adjectives along with the verbs and adverbs) in the tweets. The corpus-based method was used to find the semantic orientation of adjectives and the dictionary-based method to find the semantic position of verbs and adverbs. The overall tweet sentiment was then calculated using a linear equation which combine emotion intensifiers too. This work is fact finding in nature and the prototype evaluated is a preliminary prototype. The initial results show that it is a motivating technique. Microblogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions. Survey through the literature confirm that the methods of spontaneously annotating sentiment at the word level which is categorized into two parts one is dictionary-based approaches and another one is corpus-based approaches.[10]

In research paper is based on SemEval-2014 systems which briefly uses supportvector machine(SVM) as explanation of algorithm. Features like lexicon uses three manual lexicon methods, two of them are automatically constructed. The lexicons which are manually constructed include NRC Emotion Lexicon (Mohammad and Turney, 2010; Mohammad and Yang, 2011)[11][12], the MPQA Lexicon (Wilson et al., 2005)[13], and the Bing Liu Lexicon (Hu and Liu, 2004)[14]. Computation of automatically constructed lexicons is done by PMI (pointwise mutual information) within positive or negative tweets and the terms by sentiment score: SenScore (w) =PMI (w,pos)−PMI(w,neg) where w is a term in the lexicons. PMI (w,pos) is the PMI score between w and the positive class, and PMI(w,neg) is the PMI score between w and the negative class.[15]

Table 1: *Tabular Form.*

Sno.	Paper	Author	Year	Techniques
1.	Twitter as a Corpus for Sentiment Analysis and Opinion Mining	Alexander Pak, Patrick Paroubek	2010	Naives Bayes
2.	A Graph-based Hashtag Sentiment Classification Approach	Xiaolong Wang , Furu Wei , Xiaohua Liu , Ming Zhou , Ming Zhang	2011	Sentiment Lexicon Based Method
3.	Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis	Lei Zhang, RiddhimanGhosh, Mohamed Dekhil, Meichun Hsu, Bing Liu	2011	Augmented Lexicon Based method, Opinionated Tweet Extraction, Unigram model
4.	Sentiment Analysis of Twitter Data	ApoorvAgarwal , BoyiXie , Ilia Vovsha, Owen Rambow, Rebecca Passonneau	2011	Tree kernel, Unigram model, Senti-features, Unigram plus Senti-features, Tree Kernel Plus Senti-features.
5.	Sentiment Analysis on Twitter	Teeja Mary Sebastin , Akshi Kumar	2012	Dictionary- Based Approach , Corpus Based Approach
6.	Recent Improvements in the Sentiment Analysis of Tweets	Xiaodan Zhu, Svetlana Kiritchenko and Saif M. Mohammad	2014	Improving Lexicons, Ngram features and Negation Models

DATASETS

In the research paper, the researcher have worked and did many experiments of different features to get more accurate result on text analysis. The Table 2 below gives the statistically behavior of the datasets of twitter with the help of three different datasets.

Table 2: *Statistical Behavior of the Datasets of Twitter.*

NAMES OF THE DIFFERENT DATASETS	THE TOTAL NO. OF TWEETS	THE POSITIVE TWEETS	THE NEGATIVE TWEETS
1. STANFORD TWITTER SENTIMENT CORPUS	60k	30k	30k
2. HEALTH CARE REFORMS	839	234	421
3. OBAMA- McCAIN DEBATE	1081	393	688

Stanford Twitter Sentiment Corpus (STS)

The STS (Stanford Twitter Sentiment) dataset consists of 60,000 tweets where 50% tweets are of sure smiley's and 50% are of unsure smiley's (suresmiley's are like ☺ , : -) and unsure smiley's are like :(,:-( ) In the real datasets there are millions of tweets which have some unsure and some sure tweets. There are training sets which are based on fixed smileys. The test set was collected by searching Twitter API with specific queries including product names, companies and people. During this test, 12 researchers are selected and each tweet is allocated to each researcher. So finally after this test set, general tweets were 60K, with total test set of 1,000 tweets. There were 527 unsure and 473 sure.

Health Care Reforms (HCR)

This dataset is build by the tweets having hash tag "#hcr". This corpus's subset was allocated by 3 lead labels (sure, unsure, neutral) and further split into testing sets and training. In their research paper, they have focus on searching the sure and unsure tweets and excluding the neutrals ones. Their future plans are to find out the neutral tweets too. There are 839 tweets in Health Care Reform and remaining 839 tweets are used for training purposes.

The Sentimental Of Obama-Mccain Debate

This debate contains 3267 tweets. This is conducted by the U.S. president. This debate was their first time. A survey was done to create the index of dataset as positive or negative. This analysis was taken by the Amazon Mechanical Turk, where votes were given to each and every tweet. According to this, good, bad, mixed, neutral votes are extracted. This results in 395 positive and 690 negative tweets, having total of 1092 tweets. This dataset is small in size so a new approach is chosen which Fold Cross Validation.

Linguistic Features of Sentimental analysis

This section is used to describe the linguistic features and their role in text

analysis technique. This technique is used to separate the linguistic entries from the entries in correlation group by applying some priorities. By having such +techniques, it becomes easy to differentiate the views of the users. For Example - Let us say there is a column of different products as IPad, iPod and Mac Book Pro. These products were mapped with Products/Apple. As a result, the tweet from the test set "Finally, I got my iPhone". What a product!"is more likely to haveapositivepolaritybecauseitcontainsthe entity"iPhone"whichisalsomappedto the concept

PRODUCT/APPLE

E. Baselines of Analysis

In the research paper, there are many comparisons on the presentation between semantic analysis and the points given below:

Features of Unigrams

This feature is very simple used for sentimental analyzing of datasets. In this, they have used the Naïve Baye's classification which is trained by the word unigram. It is their first baseline model. They had analyzed a data in which STS has 37050 word unigrams, HCR has 2060 word unigrams and OMD has 2364 word unigrams.

Features of POS

POS means Part of Speech, is a feature which is used widely in the sentimental analysis of the twitter datasets. In this, Naïve Baye's theoremis used which was further trained by the POS and by word Unigram. It is a baseline model. If POS is to be extracted, then it can be done by NLP POS tagger.

OPEN PROBLEMS AND FUTURE DIRECTIONS

The Intermediating Form

Intermediating form is commonly used in mining processes. To make the clear relationship between the objects or concept, it is important to have a correct syntactically form. It can be Semi-Structured and Structured. Both of them have different form of representation. Intermediate Form can be Document-Based IF and Concept-Based IF. However, Semantic techniques are often costly and work on few words per second.

Multilingual Text Refining

Text mining is not language dependent. It is necessary to develop a technique of sorting out the different language text. Most of the data refining gadgets are English based. Analysis on different language gives more information and it is a kind of good opportunity to make a gadgets which can find out the other language based opinions too.

Domain knowledge integration

It is used in distillation of the positive and negative views. It also help in predictive model task which helps in improving the mining technique.

DISCUSSION AND FUTURE WORK

In their research paper, they discuss the role of syntactic feature for finding out the negative and positive opinions of users in twitter. They have used the different ways for semantic analysis in sentimental analysis of twitter and they used them on AlchemyAPI for the better performance and more accuracy. The important factor that affects the result is abstraction of the concept from the entities extractor. These concepts are abstract in nature which are used to point out the persons in some places like for example "I wanted to go to India and wanted to meet the Prime Minister Narender Modi" So AlchemyAPI introduce the concept of person which refers to the Prime MinisterNarender Modi.

The identifiers concept comes to play when linguistic features are used in tweets. Semantic features in sentimental analysis may increase the accuracy in some

concepts like city, songs and sometimes decrease the accuracy in some concepts like persons, company. They are also working in finding out the neutral views from the tweets. Presently the researchers are able to find out the positive and negative opinions from the twitter.

CONCLUSION

They have given the way of using the semantic features in sentimental analysis of twitter. They have also given the three approaches which are applied on different analysis; by replacing the text, by augmenting some data, and by interpolation. They have also done an experiment on twitter data and compared them by using the semantic analysis such as Point Of Sight features, features of Word Unigram. Their experiment shows that by using both POS and Word Unigram, gives the better features and is more accurate in searching the positive and negative opinions.

By these, they have concluded that sentimental analysis is more comfortable with small sized datasets where as semantic analysis is comfortable with large sized datasets.

ACKNOWLEDGEMENT

We would like to express our thanks of gratitude to Accendere Knowledge Management Services for providing us the Platform & Opportunity to pursue the research.

REFERENCES

1. Hassan Saif, Yulan He and Harith Alani.: Knowledge Media Institute, The Open University, United Kingdom.
2. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with labelpropagationoverlexicallinksandthe followergraph.
3. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: Thegood the bad andthe omg!
4. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data.
5. Lei Zhang, RiddhimanGhosh, Mohamed Dekhil, Meichun Hsu, Bing Liu ” Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis” 2011
6. Alexander Pak, Patrick Paroubek “Twitter as a Corpus for Sentiment Analysis and Opinion Mining” 2010.
7. XiaolongWang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang “A Graph-based Hashtag Sentiment Classification Approach” 2011.
8. Lei Zhang, RiddhimanGhosh, Mohamed Dekhil, Meichun Hsu, Bing Liu “Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis” 2011.
9. Apoorv Agarwal,Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau “Sentiment Analysis of Twitter Data” 2011.
10. Teeja Mary Sebastin,Akshi Kumar “Sentiment Analysis on Twitter” 2012.
11. Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, LA, California.
12. Saif M. Mohammad and Tony (Wenda) Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In Proceedings of the ACL Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Portland, OR, USA.
13. Theresa Wilson, JanyceWiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phraselevel sentiment analysis. In Proceedings of

HLTEMNLP,HLT'05,pages347–354, Stroudsburg,PA, USA.
14. Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of KDD, pages 168–177, New York, NY, USA. ACM.
15. Xiaodan Zhu, Svetlana Kiritchenko and Saif M. Mohammad "Recent Improvements in the Sentiment Analysis of Tweets" 2014.

Cite this article as: