Senitiment Analysis and Data Mining In Twitter

Suryank Sajwan, Musheer Vaqur

Abstract— Now a days, Twitter has become a platform where people share their views regarding poltics, sports, movies, product review etc, so here I am trying to develop a new approch that can enhance the sentiment analysis result.There are thousands of research papers already published on sentiment analysis but there is still need of improvement in sentiment analysis result and this research paper is based on the extraction of a word (opinion) because it is a crucial stage for sentiment analysis. People generally use short handwriting (e.g, gud for good, awsm for awesome) and misspell them, so identifying their actual sentiments will enhance the sentiment result. So this research paper is based on using mining technique (extraction of hidden meaning of sentiments) in twitter platform which will improve sentiment analysis result. In this research, we work on English language.

Index Terms— Data Mining, Extraction, Microblogging, POS, Senitiment Analysis, Slang, Twitter .

1 INTRODUCTION

TODAY microblogging websites have evolved to become a source of varied kind of information where people post real time message about their opinions on a vareity of topics, discuss current issues, complain and express positive sentiment for products they use in daily life. Internet users tend to shift from traditional communication tools such as mail to microblogging services. As more and more people post about product and services they use or express their political and religious views, microblogging websites become valuable sources of people's opinions and sentiments.Such data can be efficiently used by marketing organization, poltical parties and any individual to get a view about their opinion, product and services.

Twitter is an efficient platform for sentiment analysis and mining because it consist maximum 140 letters and the people are usually straight to the point so sentiment analysis will give better result on twitter but people try to express more sentiments in fewer words thus why they use shorthand writing and these opinion words matter a lot due to their hidden meaning so extracting actual sentiments of a particular tweets is a crucial task.

Sometime people tweets gramatically incorrect English so it can change actual sentiments of tweet holder so parts of speech tagging (POS) (Pang Lee and Vithyanathan, 2002) is also important while extracting the sentiments of particular tweets. Adjective and adverb show the sentiments about any entity (Liu and Hui, 2004)shows the tweet holder sentiments so extracting sentiments are not only solution to analysis but also retriving entity (target) regarding that sentiment is also important.

- *Suryank Sajwan is currently pursuing M.Tech degree program in computer science engineering at Uttaranchal University, Dehradun, Uttrakhand. surya04.sath@gmail.com*
- *Musheer Vaqur is currently at the position of Assistant Professor in computer science engineering in Uttaranchal University, Dehradun, Uttrakhand. musheer77@gmail.com*

2 OBJECTIVE

My objective is to try to extract all the hidden information related to each tweet from a particular period of time after that we have to refine the tweets ,it is a most crucial part so this research paper is trying to find out most suitable way by which we can extracts or refine tweets in a better way.

3 RELATED WORK

The term sentiment analysis first appeared in (Nasukawa and Yi 2003) and the term opinion mining first appeared in (Dave, Lawrence and Pennock, 2003). However, the research on sentiments and opinions appeared earlier (Das and Chen, 2001; Pang Lee and Vithyanathan 2002, Turney, 2002). My research paper closely related to (Hu and Liu, 2004) in which they identify opinion sentences in each review but their work concentrated on dictionary based and they use the adjective as an opinion word but it was review based research work where author rarely used shorthand writing.

Extraction of entity task named entity recognition(NER) in information extraction (Hobbs and Riloff, 2010) used a technique to extract the names of entities and objects from the text and to identify the role that plays in event description and this system generally focus on a specific domain or topic. Mining Comparative Sentences and Relation (Jindal and Liu, 2006) studied about a text mining problems.

Most of the research paper studied about mining sentimental word and POS (parts of speech) using corpus-based techniques and dictionary-based techniques. Due to the absence of misspelling and shorthand, writing words they considered misspell word and short handwriting either words as a neutral sentiment or abandoned but if we get their hidden meaning it will worth it and will enhance the sentiment analysis result.

4 TWEET DEFINITION

A tweet consist of two components : a target g and a sentiment s on the target,i.e., (g,s), where g can be any entity or aspect of the entity about which a tweet has been expressed,and s is a

positive, negative or neutral sentiment, or a numeric rating score .But tweet change with time and trends.
Now we can define tweet as a quadruple.
A tweet is a quintuple, (e, a, s, h, t)
Where e is the name of an entity, a is an aspect of e,s is the sentiment on aspect a of entity e,h is the tweet holder and t is the time when the tweet is expressed by h.
This definition provides a framework to transform unstructured text into structured data.The quintuple above is basically a data base scheme.based on which the extracted tweets can be put into a database table.
So extraction itself a problem in natural language text,people often write the same entity in different ways e.g. (good may be written as gud and good).

5 TEXT MINING

Text Mining is an important step of Knowledge Discovery process. It is used to extract hidden information from non-structured or semi-structured data. This aspect is fundamental because most of the Web information is semi-structured due to the nested structure of HTML code, is linked and is redundant. Web Text Mining helps whole knowledge mining process in mining, extraction and integration of useful data, information, and knowledge from Web page contents. Web Text Mining process is able to discover knowledge in a distributed and heterogeneous multi-organization environment. In this paper, our basic focus is to study the concept of Text Mining and various techniques. Here, we are able to determine how to mine the Plain as well as Structured Text. It also describes the major ways in which text is mined when the input is plain natural language, rather than partially structured Web documents.
The Text mining processes unstructured information, extracts meaningful numeric indices from the text and makes the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted from the summarized words of the documentmining project. Text mining converts text into numbers which can then be included in other analyses such as predictive data mining projects, clustering etc. Text mining is also known as text data mining, which refers the process of deriving high-quality information from text. High-quality information is derived through the statistical pattern learning. Text mining includes the process of structuring the input text like parsing and other successive insertion into a database. TM derives patterns within the structured data, evaluates them and finally produces the output. Text mining takes account of text categorization, text clustering, sentiment analysis, document summarization, and entity relation modeling. Text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze these data objects.
So the words can be analyzed and also the similarities between words and documents can be determined or how they are related to other variables in the data.

6 OBSERVATION

We observe lots of tweets where people used shorthand writing and avoid to write full spelling to express their sentiments ,these shorthand writing words have a pattern ,generally, people try to avoid writing vowels due to limited words(140 max) e.g. Love –lv, Good-gd, awesome-awsm, etc.So retrieval their actual meaning and sentiments will enhance the result of sentiment analysis.
According to the definition of tweets target and sentiments are essential part and due to short handwriting recognize their actual meaning is so important.

7 PROPOSED WORK

We are proposing an algorithm, which will find the actual meaning of short handwriting words (slangs).
Now we observe that people try to avoid vowels when they express their sentiments and there are only five vowels in English language and without them meaningful word is not possible in the English dictionary.So we make an algorithm which will find the possible meaningful word from shorthand writing word. (Fig.1)
Step 1- First, we check the shorthand words in dictionary and if it is already present there then return true;
Step2- If shorthand word is not present their then return false;
Step3- If return true then stop the algorithm no need to extend the word;
Step 4- If return false then we insert each vowel after each consonant of shorthand word;
Step 5 – It will generate various string and each string will check in dictionary simultaneously and every string that will present in dictionary will store in temporary file if that string is not present in dictionary then drop those string and go to step 4;
Step 6 – Finally all meaningful words are present in a temp file which are closely related to shorthand word.

(Note- there are several words that need direct mapping due to limitation of above algo e.g. U for YOU and now a days people use slang language due to social networking culture e.g Sd for Sweet dreams.So we studied about several slang words that become fashion in social networking sites so we have done direct mapping of those words.)

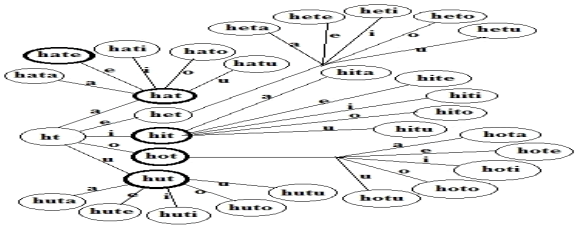

Fig 1

But we experiment that several slang words failed to find out closely related meaningful words due to the limitation of algorithm e.g; people use'luking' for 'looking', 'wt' for 'what' etc. Direct mapping is an option but it is a time-consuming process.

So we use "Edit distance algorithm" (Andoni and Onak) which will help to find out an appropriate letter in misspelling word to find out a meaningful word.

After we get all the meaningful word now finding the approprite word used by tweet holder we need part-of-speech Tagging (POS) because it will help us to find out which word may get a right sense or follow grammatically rules.The method in (Turney 2002) is such a technique. It performs classification based on some fixed syntactic patterns that are likely to be used to express opinions.

From Parts-of –speech (POS) algorithm, we will get appropriate word and sentiments of twitter holder.

E.g., I lv u.

I-I

Lv-Love, lava, lev, levo, live (these words we get from our proposed algortihm)

u-You (direct mapping)

Now we have to find out exact sentiments of tweet holder on a word "lv".

Now we will use parts of speech (POS) tagging algorithm. 'Love' is 'Verb', 'Live' is a 'Verb' ,'lev 'is Noun,'Lava' is 'Noun','Lev' is 'Noun '.

In given tweet if we see there are already two Nouns and without any verb no sentence completes so according to part of speech (POS) tagging algorithm "love"is giving sense to above tweets so we extract 'Love' for Lv.

8 SENTIMENT ANALYSIS APPROCH

Our approach is to use different machine learning classifiers and feature extractors. The machine learning classifiers are Naive Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM). The feature extractors are unigrams and unigrams with weighted positive and negative keywords. We build a framework that treats classifiers and feature extractors as two distinct components. This framework allows us to easily try out different combinations of classifiers and feature extractors.

9 EVALUATION

There are publicly available data sets of twitter messages with sentiment indicated by and we have used a combination of these two datasets to train the machine learning classifiers. For the test dataset, we randomly choose 4000 tweets, which were not used to train the classifier. The details of the training and test data as explained in Table 1.

The Twitter API has a parameter that specifies which language to retrieve tweets in. We always set this parameter to English (en). Thus, our classification will only work on tweets in English because the training data is English-only.

Table 1
Details of Training

Data Set	Positive	Negative	Neutral	Total
Training	9666	9666	2271	21603
Test	Randomly choosen Tweets			4000

10 CONCLUSION

Retrieval of the appropriate meaning of slangs and misspell words used by opinion holder give more accuracy in sentiment analysis result. Dictionary based method gives the adventage of adding further words in future and study of a several slang words which has become fashion in social networking sites e.g.lol for lots of laughter etc ,so extraction of these words and adding these words in dictionary using direct mapping reduce the complexity of analysis phase and enhence the sentiment analysis result.

11 ACKNOWLEDGMENT

This work has been guided by Assistant Professor "Musheer Vaqur" and I am thankful to him for his valuable guidence. I am also grateful for all the faculty member of Department of Computer Science (Uttranchal University) for their support and motivation.

REFERENCES

[1] Dave, Kushal, Steve Lawerence, and David M. Pennock.Mining the peanut gallery: Opinion extraction and semantic classification of product reviews in proceeding of International Conference on World Wide Web (WWW-2003). 2003.

[2] Jindal, Nitin and Bing Liu. Mining comparative sentences and relations. In Proceedings of National Conf. on Artificial Intelligence (AAAI – 2006). 2006b.

[3] Hobbs, Jerry R. and Ellen Riloff. Information Extraction, in in Handbook of Natural Language processing, 2nd Edition, N. Indurkhya and F.J. Damerau, Editors .2010, Chapman & Hall/CRC Press.

[4] Liu Bing. Sentiment Analysis and Subjectivity, in Handbook of Natural Language Processing, Second Edition, N.Indurkhya and F.J.Damerau, Editors. 2010.

[5] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In proceedings of conference on Empirical Methods in Natural Language Processing (EMNLP-2002).2002.

[6] Peter D Turney. Thumbs up or thumbs down? : Semantic orientation applied to unsupervised classification of reviews. In Proceeding of the 40th Annual Meeting on Association for Computational Lingusitic. 2002.

[7] Santorini Beatrice. Part-of –sppech Tagging guidelines for the Penn Treebank Project, 1990: University of Pennsylvania, School of Enginnering and Applied Science, Dept. of Computer and Information Science.

[8] Minquing hu, Bing Liu.Mining and summarizing cutomer reviews.In

proceeding of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining .2004.

[9] Agrawal, R, & Srikant, R. 1994. Fast algorithm for mining associates rules. VLDB'94, 1994.

[10] Bruce, R., and Wiebe, J. 2000. Recognizing Subjectivity: A case Study of Manual Tagging. Natural Language Engineering.

[11] Alexandr Andoni, Krzysztof Onak.Approximating edit distance in near-line time. SIAM Journal on Computing, 2012.