ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Analysing Twitter Reviews of Car Using SA Techniques

Ketan L. Thakare, Sachin N. Deshmukh

M.Tech Student, Dept. of Computer Science&IT, Dr. B.A.M.U, Aurangabad, India

Professor, Dept. of Computer Science&IT, Dr. B.A.M.U, Aurangabad, India

ABSTRACT: The importance of Text Mining applications has increased in recent years because of the large number of web-based applications which lead to the creation of such data. Now a days, newer aspects of Text Mining can be apply on emerging platforms such as Social Networks. Opinion Mining and Sentiment Analysis are one of the applications of Text Mining. Opinion Mining refers to the extraction of lines and phases from the social networks that contain some opinion. Sentiment analysis identifies the polarity of opinion being extracted. Daily huge amount of data is generated by these Social networks such as Twitter. Users not only use these social networks but also give their valuable feedback, thus generating additional information. Due large amount of users opinion, views, feedback and suggestion available through social networks, it's very much essential to explore, analyse and organize their views for better decision making. This paper focuses on existing approaches for opinion mining on social data especially for twitter data and also modifies techniques for sentiment analysis on social data in order to obtain better result that will helpful to user for better Decision Making.

KEYWORDS: Microblogging sites,Twitter, Opinion Mining, Sentiment Analysis, MPQA Lexicon, Tree Tagger,

I. INTRODUCTION

Social networks is fastest growing network. These networks contain Microblogging sites such as Twitter, Facebook. These Microblogging sites growing up very rapidly and they are become anorigin of various kind of information. This is due to nature of microblogs on which people post real time messages. The messages include user opinion on a different of topics, hash out on current issues, their displeasure, and express positive or negative views for products they use in daily life. One challenge is to build technology to detect and summarize an overall sentiment. Opinion Mining and Sentiment Analysis (OMSA) is the solution for that. OMSA is the Natural Language Processing (NLP) task that analyses the data (user reviews) from the sources and explore that whether it is negative or positive.Sentiment analysis is utilized to reviews and social media for a variety of applications, ranging from marketing to customer service.

In this paper we focus on popular Microblogging site Twitter. Twitter contain large number of text messages and number these text messages increase every day. Audience of twitter varies from regular users to celebrities, company representative to its client, politicians and even the country presidents. Therefore, it is possible to collect text posts of users from different social and interest groups. Using this data from twitter we propose two sentiment analysis methods. First classify the twitter messages as Subjective and Objective and further distinguish the subjective messages as Positive, Negative and Neutral.

The remainder of this paper is organized as follows. In section 2, we discuss the literature survey related to this paper. In section 3, we discuss about problem formulation. In section 4, we discuss the methodologies for twitter sentiment detection. In section 5, we show the results of experiment we done on different twitter datasets. In section 6, we discuss conclusion and future work.

II. RELATED WORK

Detecting sentiment from tweet data is considered as a much harder problem than sentiment analysis on conventional text such as review documents, mainly due to the short length of tweet messages i.e. limit of 140(Which is about to change) characters per tweet, the frequent use of informal and irregular words, the rapid evolution of language

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

in Twitter, and the data streaming paradigm that Twitter has. As twitter contain much more noisy data, so it is difficult to deal with that data while sentiment analysis.

The short length of twitter messages forces each user to express their opinion in very short text. Because of this Sentimentanalysis of twitter is ambitious task. For best result of sentiment analysis we use the supervised learning approaches such as Naive Bayes and Support Vector Machines, but supervised learning approach requires the manual labelling, which is very expensive. Some work has been done on unsupervised (e.g. [1]) and semi-supervised (e.g. [2, 3]).

The most widely used feature model for all text classification tasks is Bag-of-Words. The model classified textas collection of individual words with no word is depending on other. This model is also very popular in sentiment analysis and has been used by various researchers. One of the ways to use this feature model in our classifier is by using the unigram as feature. Unigram classifies our text into sequence of 'n' words which are completely independent of any other word. So unigrams is just a collection of individual words in the text to be classified, and this can be shown to provide good performance using the bag-of-words model [4, 5].

One simple way to use unigrams as features is to assign them a prior polarity, and take the average of the prior polarity of each word in text, which makes easy to calculate polarity of each text i.e. tweets.

There are three ways of using prior polarity of words as features:

1. The simpler un-supervised approach is to use publicly available online lexicons/dictionaries which map a word to its prior polarity. The Multi-Perspective-Question-Answering (MPQA) is an online resource with such a subjectivity lexicon which maps a total of 6886 words according to whether they are "positive" or "negative" and whether they have "strong" or "weak" subjectivity [6]. The SentiWordNet 3.0 is another such resource which gives probability of each word belonging to positive, negative and neutral Classes [7].
2. The second approach is to construct a custom prior polarity dictionary from our training data according to the occurrence of each word in each particular class. For example if a certain word is occurring more often in the positive labelled phrases in our training dataset.
3. The third approach is a middle ground between the above two approaches. In this approach we construct our own polarity lexicon but not necessarily from our training data, so we don't need to have labelled training data. One way of doing this as proposed by Turney, is to calculate the prior semantic orientation [1].

Other grammatical features like "Part of Speech" tagging are used in this sentiment analysis task. Using that tagging we identifies the conjunctions of adjectives from our data and then mark the data either positive or negative using the orientation of adjectives (positive or negative) [8].

III. TWITTER SENTIMENT DETECTION

Sentiment detection of twitter messages is the basic function needed by the various applications that are depend on twitter data. So our target is to label each tweet as positive, negative or neutral that contains some opinion about product, its quality, services provide etc. for this we use lexicon approach [10].

In First step we collect data from twitter using the twitter API. In second step we pre-process the data to remove noise from the data and make it suitable for further processing. In third step extract feature i.e. opinion tweets. In fourth step we classify the tweets as Subjective and Objective tweets this is because we only have to work with subjective tweets. At end we find the polarity of tweets. Following figure shows workflow for twitter sentiment detection.

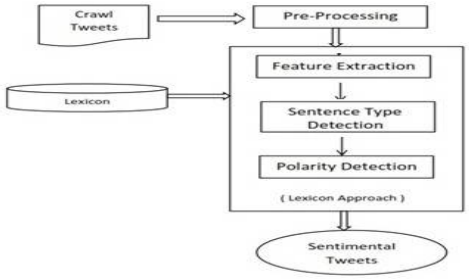

Fig. 1:- Workflow for Twitter Sentiment Detection

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

3.1 Data Acquisition

Many social networks and apps have their own interface that programmers can work with. These interfaces are called API's (Application Programing Interface). We acquire our dataset by using such a API's. We use Twitter API that allows us to interact with its data i.e. tweets. So using that API we request API for data and then API gives us data that is in JSON form that will easily read by our program.

3.2 Pre-Processing

The task of data pre-processing is to take raw data as input and convert it to the form that is suitable to the application as an input. For data pre-processing, we give raw data as a raw input vector and the transformed data output produce by the preprocessor is term as Preprocessed Input Vector.

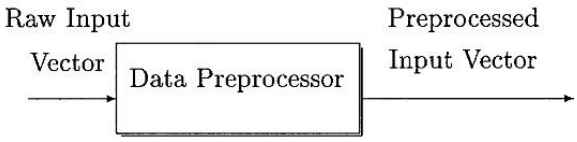

Fig. 2:- Data Pre-processing

Data pre-processing includes removing of 1) duplicate tweets, 2) Retweets, 3)Twitter usernames, which start with @ symbol, 4) Punctuation marks, 5) Numbers, 6) All URL, 7) Unnecessary space, 8) Twitter hashtags

3.3 Feature Extraction

In this context, our target is to find out the tweets that contain some opinion about some product, services etc. and find its polarity. Hence we use meta-information about words in tweets as feature.

3.3.1 Meta-Features

Firstly we map each word in tweet to its part of speech (POS) tagging using part of speech dictionary. These POS tags are good indicators for sentiment analysis [13, 14]. An opinion message mostly contains adjectives or interjections. These adjectives are the way for user to express his opinion or feelings about product and services. So in our experiments only those tweets as features that contains some adjectives or interjections. Adjectives and subjectivity has strong relationship among them. Further we map the word to its subjectivity to identify the subjective tweets.

3.4 Sentence Type Detection

Our target is to find polarity of twitter messages that contain opinion about product, service etc. for this purpose we first find out the tweets that contain some opinion, this can be done using subjectivity classification.

Subjectivity classification involves discrimination between subjective and objective utterances, like sentences, or even phrases. Subjective utterances reflect a private point of view, emotion or belief. Subjective sentences influenced by personal feelings, tastes, or opinions. So the recognize the subjectivity is important from several point of view. For this we use MPQA subjectivity lexicon [11] that contain about 6886 words with their prior polarity and subjectivity information.Using that we find subjective tweets from chunk of tweets.

3.5 Polarity Detection

Next task is to detect the polarity of subjective sentences, whether tweets given by user reflects positive or negative attitude of user towards product, services. We perform this using the MPQA lexicon that contains polarity information about each word in it. Using that we calculate the prior polarity of each word of tweet and then summing the polarity of all word to identify whether tweet is positive, negative or neutral.

3.5.1 Negation Handling

One of the major issue while polarity classification is handling the negation of tweets. Many tweets contain the negation word in it, these negation words shifts the polarity of tweets i.e. from Positive to Negative or vice versa.

Example:

1. Camera quality is not good

As shown in above example not word comes before adjective word GOOD that shifts the polarity of that sentence. That is user want to say that quality of camera is BAD. Such types of negation words are called Polarity Shifters.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Hence problem is to handle these polarity shifters. If we use the classifiers [3, 9], then we find that while performing classification task of negation sentences the classifier removes the negation words by considering it as a "stop word". This changes the label of sentence from Positive to Negative.

We have overcome this problem, idea behind handling these polarity shifters is, first find out all negation words [12] from tweets and replace all negation words with exclamatory mark "!". Next to this is shift the polarity of those tweets that contains the negation mark i.e. "!". While applying this logic we mainly focus on adjective words that indicate user feelings or opinion. So any of negation word comes before the adjective word then its shift the polarity of that tweet.

Example:

1. Camera quality is not good

In above example we replace word not with "!". Then it becomes

"Camera quality is !good"

So when the adjective words comes with such a mark then we consider its apposite polarity i.e. Positive to Negative or vice versa. So conclude that above sentence is of negative polarity.

IV. EXPERIMENT AND RESULTS

In this section, we show experiments and their results. We perform different task like acquire data from twitter, pre-process data to remove noise from it, feature extraction, subjectivity extraction and then polarity detection.

4.1 Experimental Setup

Data Sets: For Sentiment Analysis we had collected data from Twitter API. For this analysis, as the subject of interest we use "CAR" as query term to retrieve data from twitter API. The dataset contains 7543 tweets retrieved from twitter. After collecting data, we perform some data pre-processing task to clean data because dataset contain some noisy data. Table 1 show the result of pre-processing tasks.

Table 1:-Pre-processing Results

Sr. No	Pre-processing	% Reduction
1	Duplicate	0
2	Retweets	2.72
3	@People	0.83
4	Punctuations	7.73
5	Numbers	1.96
6	HTML Links	1.69
7	Unnecessary Space	3.80

4.2Feature Extraction

After pre-processing next task is feature extraction i.e. extracting those tweets that contain adjectives. For this we perform part of speech tagging on dataset to find out the adjective containing tweets. We use TreeTagger part-of-speech technique [15].Using that we obtain the tags for all words in each tweet. Based upon that result, we find out the list of adjectives and list of those tweets that contains those adjectives and use these tweets as feature. After performing Tree Tagger part of speech tagging we obtain the list of 1332 adjectives and we find the tweets that one or more adjectives from this list of adjectives. By performing this we obtain 6451 tweets as adjective containing tweets. These tweets we basically consider as the opinion tweets.

4.2.1 Subjectivity Classification

Next task is to detect subjective sentence. This is second feature that we focus that is to classify tweets as

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Subjective or Objective and use subjective tweets for further processing. For this we use tweets that contain adjectives as training data for subjectivity detection.To check the accuracy of our subjectivity classifier we use dataset of 5000 tweets as a testing dataset that are already labelled as subjective tweets. Table 2 shows result of subjectivity detection for training datasets, where we take 6451 tweets and classify it as Subjective and Objective.

Table 2:-Training Data Result

Training Data	Subjective	Objective
6451	5795	656

Following table shows the result of testing dataset where 5000 tweets taken as testing dataset, which classify as Subjective and Objective.

Table 3:-Testing Data Result

Testing Data	Subjective	Objective	Accuracy
5000	4921	79	98.42 %

4.3Polarity Detection

For polarity detection we take set of 4000 featured subjectivity tweets as training data and perform polarity classification on this training data(i.e. classify tweets as positive, negative and neutral), Table 4 shows the result of training data.

Table 4:-Training Data Result

Training Data	Positive	Negative	Neutral
4000	1322	824	1854

Further we take set of testing data, for testing data we has dataset of 1795 tweets which already labelled as positive, negative and neutral.Table 5 shows the result of testing polarity detection on testing dataset.

Table 5:-Result of Labelled tweets

	Opinion (1114)	Non Opinion (681)
Opinion	631	125
Non Opinion	483	556

As we performed polarity classification on both training and testing datasets we obtain above results. For the labelled tweets we performed polarity detection, we got 66% accuracy.

V. CONCLUSION AND FUTURE WORK

This is an effective sentiment detection approach for twitter messages. We obtained this performance because of our approach generates abstract representation of tweets and the performed sentiment analysis. We proposed method for negation handling of sentences that minimizes the problem of negation handling in sentiment analysis. The limitation of our approach is comparative and antagonistic type of sentences. As the future work we want perform more precise sentiment analysis of such a comparative and antagonistic sentences.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

REFERENCES

1. Peter D. Turney. "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews." In Proceedings of the Annual Meeting of the Association of Computational Linguistics,2002.
2. Alexander Pak and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In Proceedings of international conference on Language Resources and Evaluation (LREC),2010
3. Luciano Barbosa and Junlan Feng's "Robust Sentiment Detection on Twitter from Biased and Noisy Data." In Proceedings of the international conference on Computational Linguistics (COLING), 2010.
4. Alec Go, Richa Bhayani and Lei Huang. "Twitter Sentiment Classification using Distant Supervision." Project Technical Report, Stanford University, 2009
5. Efthymios Kouloumpis, Theresa Wilson and Johanna Moore. "Twitter Sentiment Analysis: The Good the Bad and the OMG!." In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2011.
6. Multi Perspective Question Answering (MPQA). Online Lexicon "http://www.cs.pitt.edu/mpqa/subj_lexicon.html".
7. Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani. "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". In Proceedings of international conference on Language Resources and Evaluation (LREC), 2010.
8. Hatzivassiloglou, V., & McKeown, K.R. "Predicting the semantic orientation of adjectives." In Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, 2009.
9. Shruti Wakade, Chandra Shekar, Kathy J. Liszka and Chien-Chung Chan. "Text Mining for Sentiment Analysis of Twitter Data".In Proceedings ofWorldcompProceedings, 2012.
10. Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, Wayne Niblack. "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques". In Proceedings of the ICDM's 03 proceedings of third IEEE International Conference on Data Mining, 2003.
11. Multi Perspective Question Answering (MPQA) Online Lexicon <http://www.cs.pitt.edu/mpqa/subj_lexicon.html>
12. Www.wikipedia.com/negation word list.
13. Wiebe, J. and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. Computational Linguistics and Intelligent Text Processing, pages 486–497, 2005.
14. Riloff, E., J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In Proceedings of the 7th Conference on Natural Language Learning, pages 25–32, 2003.
15. www.cis.uni-muenschen.de/~schmid-tools/TreeTagger/