International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 1, Issue 2, July – August 2012 ISSN 2278-6856

Preprocessing the Informal Text for efficient Sentiment Analysis

I.Hemalatha[1] Dr. G. P Saradhi Varma[2] Dr. A.Govardhan[3]

[1]Research Scholar JNT University Kakinada, Kakinada, A.P.,India.

[2]Professor & Head, Dept., of Information Technology, S.R.K.R. Engineering College, Bhimavaram, A.P.,India.

[3]Director of Evaluation JNT University, Hyderabad, A.P.,India.

Abstract: ***Social networking sites are increasing rapidly in this global world where communication plays a major role. There is drastic increase in the usage of social networking sites among all age groups. This can be used for business development, reviews about various social activities and acceptance of any new ideas by means of Sentiment Analysis. Thus for sentiment analysis, preprocessing is an essential task. The reviews by the user represents valuable approaching of sentiment analysis. This research paper focuses on the preprocessing techniques implemented on a specially designed algorithms in order to perform sentiment analysis.***

Keywords: Web mining preprocessing, sentiment analysis, classification.

1. INTRODUCTION

The rapid growth of the Communication and Information Technology has made information broadcast very critical. The social networks, in particular, continues to play a major role in the passing of information and as well as business intelligence. To get any information, we need a social networking sites. This sites can offer valuable information imminent into the Sentiment analysis of a particular product or a movie. It represent the action of many users over a product through positive and negative reviews. Most organizations identify these reviews as an important part of their decision making. Social networks reviews can be effectively applied to sentiment analysis of information. Properly identified reviews present a baseline of information that indicates ideal levels and supports the business intelligence. It also supports in business decisions. This research paper explains the necessary information to get preprocess the reviews in order to find sentiment and confirm its analysis whether it is positive or negative.

2. SOCIAL NETWORKING SITES

2.1 Background

Social networking sites are playing a crucial role in every aspect and in all corners of the world. Many social networking sites like facebook, twitter, myspace etc., are extensively used now a days. We used tweets from twitter to carry our research work. This is because tweets are generally confined to 140 words unlike other social networking sites where the review may contain large number of words.

3. TWITTER DATA

Twitter is a real-time information network that connects you to the latest stories, ideas, opinions and news about what you find interesting. Simply find the accounts you find most compelling and follow the conversations.

At the heart of Twitter are small bursts of information called Tweets. Each Tweet is 140 characters long, but don't let the small size fool you—you can discover a lot in a little space. You can see photos, videos and conversations directly in Tweets to get the whole story at a glance, and all in one place.

4. METHODOLOGY

4.1 Twitter Tweets

The below table shows reviews by different people about the apple product.

Table 1 Tweets by users

kvkruthika: im the lucky few to own brand new unlaunched (in India) **APPLE** I-PHONE 5S model...:) :) :) :) :) :) i'm loving it...!!!
YourLeader : Dear **Apple**, y tf is my iPhone's USB cord so short?! Signed, Irked
linderxlum4: CarrieSBitz I don't do **apple** so can't comment on that. I use cisco/Linksys and seem to replace route
MistahBungle: @jonfortt Yours is biased Jonny. You love **Apple**. Admit it.
LilMamaRollsUp: I don't have nobody to impress out at **apple** blossom , so I don't need to make myself all up *shrugs* lls
RichardZimmer: @iRyan77 Buy anything at **Apple**?

4.2 Preprocessing

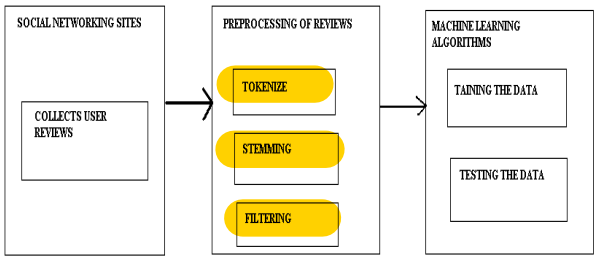

Figure1 Preprocessing of Informal Text

Data preprocessing is done to eliminate the incomplete, noisy and inconsistent data. Data must be preprocessed in order to perform any data mining functionality.

Data Preprocessing involves the following tasks

- Removing URLs

In general URLs does not contribute to analyze the sentiment in the informal text. For example consider the sentence "I have logged in to www.Ecstasy.com as I'm bored" actually the above sentence is negative but because of the presence of the word ecstasy it may become neutral and it's a false prediction. In order to avoid this sort of failures we must employ a technique to remove URLs.

- Filtering

Usually people use repeated letters in words like happyyyyy to show their intensity of expression. But, these word are not present in the sentiwordnet hence the extra letters in the word must be eliminated. This elimination follows the rule that a letter can't repeat more than three times hence can eliminate such letter.

- Questions

The question words like what, which, how etc., are not going to contribute to polarity hence in order to reduce the complexity such words are removed.

- Removing Special Characters

Special characters like.,[]{}()/' should be removed in order to remove discrepancies during the assignment of polarity. For example "it's good:" if the special characters are not removed sometimes the special characters may concatenate with the words and make those words unavailable in the dictionary. In order to overcome this we remove special characters.

- Removal of Retweets.

Retweeting is the process of copying another user's tweet and posting to another account. This usually happens if a user likes another user's tweet. Retweets are commonly abbreviated with \RT." For example, consider the following tweet: Awesome! RT @rupertgrintnet Harry Potter Marks Place in Film History http://bit.ly/Eusxi :).

5. SENTIMENT ANALYSIS BASED ON SENTIWORDNET

The approach described in this paper is based on SentiWordNet, a lexical resource for opinion mining. In SentiWordNet (http://sentiwordnet.isti.cnr.it/), to each synset of WordNet, a triple of polarity scores is assigned i.e., a positivity, negativity and objectivity score. The sum of these scores is always 1. For example the triple {0, 1, 0} (positivity, negativity, objectivity) is assigned to the synset of the term "bad". The sum of all scores of this synset is 1. SentiWordNet has been created automatically by means of a combination of linguistic and statistic classifiers. It has been applied in different opinion-related tasks, i.e. for subjectivity analysis and sentiment analysis with promising results.

5.1 Bigrams

Bigrams are used in order to increase the accuracy of the classifier. The effect of previous word on current word plays major role in sentiment analysis hence we consider bigrams rather than unigrams. In general preceding word will show more effect on the current word rather than the succeeding word hence we consider the polarity of preceding word.

For example consider the sentence "The art shows the culture and social issues prevailing at that point of time." Bigrams can be done as follows "The art", "art shows", "shows the", "culture and", "and social", "social issues", "issues prevailing" , "prevailing at", and so on.

6. SES ALGORITHM FOR PREPROCESSED DATA TO CALCULATE SENTIMENT

Step 1: Assign a weight to each word from the SentiWordNet dictionary

Step 2: Sentence level polarity is calculated as consider the sentences to calculate the average score

Step 3: check (sent_sentim_word + 3) and (sent_sentim_word - 3) for Modifier from modifier_dict if word found as modifier then calculate overall weight.

Step 4: If there is negation word (Not, Never, N't, Does'nt, Cannt, Nor, Don't, Would'nt, No) near the N, Check (N+3) and (N-3) then reverse its polarity. e.g. (OW=+0.8 → OM= -0.8)

Step 5: Check the modifier word in the sentence, if exists then recalculate the polarity referring the weightage dictionary the same process will be repeated that score of which opinion word will be effected. For e.g, in the sentence "the staff were very nice and cooperative", in this sentence the very is enhance the weight of the nearest opinion word i.e., nice

Step 6: Certain nouns affect the sentence polarity, so recalculate the polarity if such types of word occur. From

the dictionary of the weights of words/terms, assign weights to each sentence accordingly.

The steps of rule base system for contextual valance shifter is describes as below:

- if the modifier is a negation modifier then sentim_word_score:= Reverse the polarity of sent_sentim_word
- if the modifier is a intensifier then sentim_word_score:= sentim_word_score + modifier_weight
- if the modifier is a decelerator or enhancer or context shifter then sentim_word_score:= intensifying modifier_weight obtained from modifier_dict
- Calculate the final weights of each sentence and review to decide if it is positive, negative or neutral. So, the opinion strength for both sentence and feedback is calculated by assigning the combined opinion weight to the sentence and review using the Eq. 3 and 4:

$$\text{SentenceScore(Sen)} = \frac{\sum_{i=1}^{n} score(i)}{n} \quad \text{.....Eq.3}$$

- Where, Score (Sen), are the positive or negative score of sentence Sen, Score(i) is the positive, negative score of ith word in sentence S. n is the total no. of words in Sen:

$$\text{ReviewScore(Rew)} = \frac{\sum_{i=1}^{n} score(Sen)}{n} \quad \text{.....Eq.4}$$

Where, Rew(Score), are the positive or negative score of Review Rew, Score(Sen) are the positive, negative score of ith sentences in review. n is the total no. of sentences in the review.

7. EXPERIMENTS AND RESULTS

The sample of results obtained from each intermediate step of the proposed approach is given in this section. For providing sample results, we take 100 reviews from twitter. The reviews are initially subjected to preprocessing which includes Removing URLs, Filtering, Questions and Removing Special Characters. Preprocessing results in reviews represented. The reviews and preprocessing results are given in the figure 1.

Figure 1 Preprocessing window that displays results of corresponding tasks

8. CONCLUSION

In this paper, we have proposed an efficient method for preprocessing. Where it has to be done before applying any classification algorithm. We have performed three preprocessing tasks. One task to remove URLs from the input file next one to remove special characters, here we can also remove repeated letters from a word, the last task is to remove question words. Now the preprocessed document can be given as input to any Machine Learning algorithms.

REFERENCES

[1]. Document-Word Co-Regularization for Semi-supervised Sentiment Analysis by Vikas Sindhwani and Prem Melville, Business Analytics and Mathematical Sciences, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 {vsindhw,pmelvil}@us.ibm.com

[2]. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis by Hiroshi Kanayama Tetsuya Nasukawa, Tokyo Research Laboratory, IBM Japan, Ltd. 1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502 Japan {hkana,nasukawa}@jp.ibm.com

[3]. LargeScale Sentiment Analysis for News and Blogs Namrata Godbole? Manjunath Srinivasaiah? Steven Skiena_namratagodbole@gmail.com manj.blr@gmail.com skiena@cs.sunysb.edu?Google Inc., New York NY, USA}Dept. of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400, USA

[4]. Language processing Techniques - ǂ IBM Tokyo Research Lab, 1623-14, Shimotsuruma Yamato-shi, Kanagawa-ken 242-8502, Japan nasukawa@ip.ibm.com

[5]. Sentiment Elicitation System for Social Media Data - Kunpeng Zhang, Yu Cheng, Yusheng Xie, Daniel Honbo Ankit Agrawal, Diana Palsetia, Kathy Lee, Wei-keng Liao, and Alok Choudhary - *Department of Electric Engineering & Computer Science, Northwestern University, Evanston, IL 60208,*

[6]. Sentiment Analysis in Practice Yongzheng (Tiger) Zhang , Dan Shen*, Catherine Baudin

[7]. Sentiment Analysisin Short and Informal Text – Marco Veluscek with the supervision of Prof. Sune Lehmann, PhD

[8]. Text normalization in social media: progress, problems and applications for a pre-processing system of casual English - Eleanor Clarka* and Kenji Arakia Pre-processing very noisy text - Alexander Clark, ISSCO / TIM, University of Geneva, UNI-MAIL, Boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland.

I.Hemalatha received her M.Tech degree from Andhra University, pursuing Ph.D in computer Science Engineering. A member of CSI, Co-ordinator for Microsoft Student Education Academy, Member in Infosys Campus connect Programme. Working as Assistant Professor in S.R.K.R. Engineering College, China-Amiram, Bhimavaram.