International Journal of Engineering & Technology, 7 (4.5) (2018) 628-630

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Domain specific opinion mining

Surabhi Thorat [1] *, Dr. C. Namrata Mahender [1]

[1] Department of Computer Science & IT, Dr. Babasaheb Ambedkar University, Aurangabad(MS
**Corresponding author E-mail:*

Abstract

The manuscript Social media is a very promising platform of communication between the peoples. Remarkable work has been done recently focusing on the analysis of social media in order to analyze the people thinking and behavioral trends about current topics of interest but still many challenges are yet to be uncovered. In this paper, we focused on analyzing the domain specific tweets collected from social media. To improve the result accuracy firstly we had done the polarity test to find the polarity of tweets categorized in negative, positive and neural labels. Secondly we applied N-gram model that assigns probabilities to sentences and sequences of words started from unigram, bigram, and trigram up-to four gram. Lastly, we performed association mining on the tweets to find the association of do- main specific data with its back and forth paired text.

Keywords: *Social Media; Drought; Opinion Mining; Domain Specific; Polarity; N-Gram; Association Mining.*

1. Introduction

Opinion Mining or Sentiment Analysis is the field to extract the opinionated text datasets and summarize in understandable form for end user [1].In recent years as need of social media increased tremendously that gives motivation to researcher's to analyse individual experiences and opinions based on Domain specific area. It lead to the development of new technologies for automatically extracting or analysing personal opinions from web documents. That can be used as an alternative to traditional questionnaire- based social or customer research and would also benefit Web users who seek reviews on specific domain.

The opinion is the subjective expression which describes people's opinions, emotions and sentiments towards entities and their properties, particular topic, product or services. Opinion mining is to extract the positive, negative or neutral opinion summery from unstructured data [2].It is a tedious work to compile a list of opin- ion expressions, which will be equally applicable to different do- mains because some opinion phrases are used only in a specific domain while the others are context oriented [3]. Sentiment lexi- cons adapted to a particular domain or topic have been shown to improve task performance in a number of applications, including opinion retrieval[4] and expression level sentiment classification [5]. In addition there are several studies about context-dependent opinion expressions [4].In this paper, we propose a domain specif- ic mining approach on tweets collected from twitter to improve the accuracy.

2. Proposed model

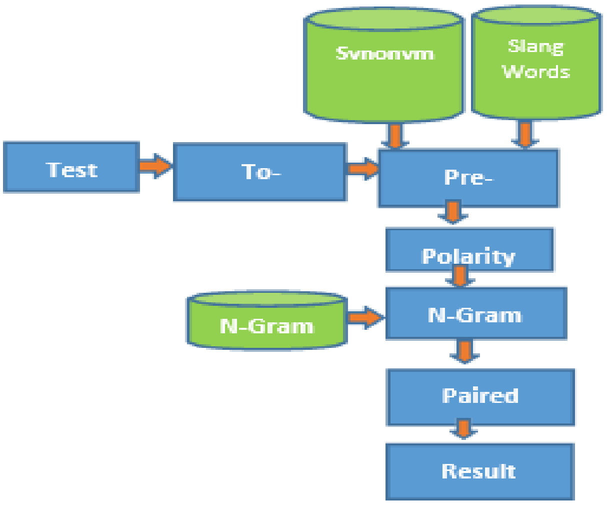

Fig. 1: Proposed Model.

In our proposed model we had done domain specific opinion mining on tweets collected from social media. In this model we started by collecting the tweets. Firstly we tokenized the tweets to as thetweets are highly unstructured and having ambiguity in sentence, we had refined them by pre-processing the tweets. In this stage we correct the individual words by applying the synonym and slang word corpus that we had created for our work. Once the tweets are smoothen up polarity test is performed. To find the probability of domain in tweets we had applied N-gram model started from unigram to trigram probability test. Lastly we had done paired association with our domain which works on previous and next word association. This improves the result accuracy in much extant.

3. Preprocessing domain specific tweets

For pre-processing the unstructured text, we had collected 1500 tweets available in twitter for drought tag.
Data Pre-processing involves the following tasks

3.1. Unique tweets selection

Tweets collected from twitter contains duplicate tweets that are coming again and again in the dataset. So we applied the function to eliminate the repetitive tweets and select only the distinct one to reduce the processing task.

3.2. Drought keyword selection

As we focused on drought tag only, we initially selected only tweets containing the drought keyword. We had focused on the keyword drought but we found that the word drought is used by the people differently in different context. It is used in many ways that's resemblance is denoting lack of humanity, financial crisis, lack of water. Like, people are tweeting "there is drought in hu- manity", "my country is facing drought to fulfilling basic needs" etc.

3.3. Remove slang expression

We had analysed that people are using the slang expression or non-word which are not found in English dictionary while tweet- ing the text. Like people are writing "n" instead of "and" or writ- ing "u" instead of "you". To solve this problem we had created a CSV file which contains the correct word for most of the non- words found in the tweets. After that we had replaced the non- word with correct words with the help of CSV file created by us. This will become a fruitful corpus for converting the slag or non- words into its correct format. This can be used for any sort of re- search work where the conversion from slang words to correct words is required.

3.4. Finding synonyms

We did not found any exact synonym for the word drought for which we are looking for in many of the word bank like Wordnet etc. We had created a file which has all possible synonym of the word drought and selected the tweets from twitter that matches with any of the synonym in the file.

3.5. Segmentation

After applying the initial task of filtering. We had done segmentation of tweets to divide a whole tweet in to individual segments. After applying this step we are able to do opinion mining on pre-processed social media data.
After pre-processing the 1500 tweets we got 686 tweets only which are having unique text with 0% missing or null value. All other are eliminated as duplicate tweets. At this stage we applied polarity test on the pre-processed data to analyze the sentiments ofthe people to find their opinions on drought disaster in three cate- gories positive, negative and neutral.

4. Polarity test

For polarity test we had created domain specific corpus .We had used "drought" as our domain for this work. Our corpus consist of all possible dictionary words related to the drought domain with its polarity as positive or negative. The pre-processed tweets is passed to the polarity test where tweets are first tokenized and then it finds the polarity value from our domain specific corpus containing the relevant words along with the polarity. After apply- ing the polarity test we got the following results.

Table 1: Domain Specific Corpus

Word	Polar-ity	Sense
drought	1	a shortage of rainfall
love	2	have a great affection or liking for
scarcity	1	a small and inadequate amount
water	2	A colourless transparent odourless liquid which forms the seas lakes rivers and rain and is the basis of the fluids of living organisms
suicide	1	person who kills himself intentionally
kill	1	destroy a vitally essential quality of or in
drop	1	go down in value
severe	1	causing fear or anxiety by threatening great harm
battle	0	an energetic attempt to achieve something
worsen	1	decline
fund	2	accumulate a fund for the discharge of a recurrent liability
crop	2	A cultivated plant that is grown on a large scale commer-cially

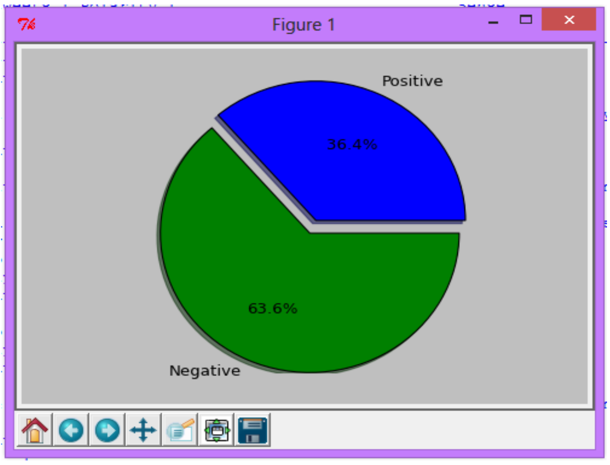

Fig. 2: Polarity Test Results.

5. N-gram model

An N-gram is a sequence of N-gram words. Like bigram is a two-word sequence of words like "drought kills", "drought again", and a trigram is a three-word sequence of words like "drought kills farmer", or "drought diverse effect". N-gram model used to estimate the probability of the last word of an N-gram given the previous words, and also to assign probabilities to entire sequences. N-gram is used to mean either the word sequence itself or the predictive model that assigns it a probability [6].
We had created separate corpus for unigram, bigram and trigram words based on our domain drought. Like in bigram we approximates the probability of a given all the previous words P (Wn | W1 n-1) by using only the conditional probability of the preceding word P (Wn | Wn-1). The general equation for this N-gram ap- proximation to the conditional probability of the next word in a sequence is

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$$

By applying this equation for unigram, bigram and trigram we got the following results

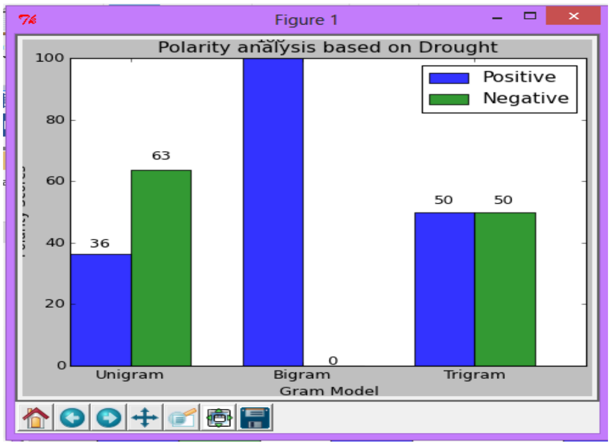

Fig. 3: N-Gram Results.

6. Paired association

In this phase we had done paired association with domain specific words. We had performed trigram paired association with drought domain .Two approaches where in first approach we had paired “drought” with next two bigram words and in second approach we had paired “drought” with previous two bigram words. This is the sample results we get after paired association.

Table 2: Next Word Association

Next _Word_Tri_Pair	Polarity
drought kill the	neg
drought scarcity kills	neg
drought free area	neg
drought scarcity kills	neg
drought going to	neg
drought conditions leave	pos
drought now over	neg

Table 3: Previous Word Association

Previous_ Word_ Tri_ Pair	Polarity
due to drought	neg
declaring as drought	neg
no more drought	pos
increases by drought	neg
effects of drought	neg
change induced drought	neg

As results, shows that the previous word association approach gives better result in comparison to the next word association ap- proach. Like it gives negative polarity for pair “drought free area” and “drought now over”.

7. Conclusion

This paper introduces an approach for mining tweets collected from social media using domain specific sentiment dictionary specially created for “drought”. This model shows that it is able to predict better sentiment analysis on short comment sentences in natural language. This paper high lighten the point that word asso- ciation plays a vital role in opinion mining.

References

[1] G. Eason, Hulth, A. and Megyesi, B.B., “A Study on Automatically Extracted Keywords in Text Categorization”, Proceedings of the 21st International Conference on Computational Linguistics in 2006.

[2] Surabhi Thorat,” Opinion Mining and Sentiment Analysis- Its Tools and Challenges”, International Journal of IT, Engineering and Applied Sciences Research (IJIEASR) ISSN: 2319-4413 Volume 3, No. 11, November 2014, p. 12-15.

[3] Lu Y., Castellanos M., Dayal U. and Zhai C. 2011.Automatic Con- struction of a Context-Aware Sentiment Lexicon: An Optimization Approach In Proceedings of the World Wide Web Confer- ence(WWW).

[4] Jijkoun V., de Rijke M., and Weerkamp W. 2010. Generating focused topic-specific sentiment lexicons. In ACL '10, pages 585–594.

[5] Choi Y. and Cardie C. 2009 adapting a polarity lexicon using inte- ger linear programming for domainspecific sentiment classification. In EMNLP '09, pages 590–598.

[6] N-grams,Speech and Language Processing. Daniel Jurafsky & James H. Martin. Draft of September 1, 2014.