Use Supervise - unsupervised methods for tweet level Contextual semantics

Shilpi Goyal[1], Nirupama Tiwari[2]

Computer Science and Engineering
Shri Ram College of Engineering and Management
Banmore, Morena

Abstract---*Identifying the sentiments at tweet level based on contextual feature comes under text classification and clustering, an evolving area of wide research with many algorithms had been already proposed. With the recent advancements in the field of text mining there are many new techniques consistently emerging that need to be implemented for text classification in search of a better classifier. In this paper, we propose a lexicon based method that can remove the limitation of Senticircles. In this paper, we modify senticircle in the sense that senticircle tend to assign or update pre-assigned positive sentiments to negative or neutral one. The tweets are first parsed using open NLP, using this part of speech tagging to assign prior polarities. We in turn use partitioning around medoids method to take a sentiment of a term instead of senti-median method. We run evaluations on three datasets, @narendramodi , DeMonetization and #FightAgainstCorruption and shown through results whether our implementation of bag of words with its update sentiment with its strength gives a better performance for text based classification. Results show that the proposed method can achieve an accuracy of 71.34%, 72.98% and 71.37% with SVM, CTree and J48 classifier on datasets.*

Keywords - *Sentiment Analysis, Text Mining, Contextual feature, Twitter, SentiCircles*

I. INTRODUCTION

As one of the most prosperous applications of text analysis and understanding emotions and short messaging text has recently received consequential attention especially during the past several years. This is corroborated by the emergence of Web 2.0, social networking services, micro blogging like Twitter, blogs, chats, online reviews, forums, discussions . Worldwide Twitter[1] is one of the leading social media, one of the ten most-visited websites and has been described as "the SMS of the Internet" [2] of 2016, Twitter had more than 319 million monthly active users. These tweets are free to post in the form of text that is limited to 140 characters. This include usage of slangs, emoticons, acronyms etc. to express most of our views in a limited free text.

Information Retrieval, refers to text mining, and natural language processing mainly used to predict the sentiment of the text data called sentiment analysis or emotion recognition. This mainly involves the classification of tweets to determine in which class it falls. This include the polarity of each tweet whether it is positive, negative, or neutral.

In this paper, the interest is to classify the tweets on the bases of polarity identification and its contextual features. We break this problem in three steps. In first step we prepare bag of words , initially having some prior sentiment polarity irrespective to its context based on some lexicon approach. Then in second step, this includes updation of strength and sentiment of those words that are not present in prior sentiment polarity considering the co-occurrence pattern used in different context to capture semantics and calculate its sentiment at tweet level. The last step involves training of model using some supervised technique and test the set to evaluate performance. This paper takes three datasets, @narendramodi , DeMonetization and #FightAgainstCorruption that are extracted from Twitter using Twitter API[3], apply an unsupervised technique to identify the bag of words with assigned sentiment with its strength and then apply supervised techniques to train and test the model using decision tree, support vector machine and J48.

The rest of the paper will be divided as follows: section II presents the related word. Proposed model is presented in Section III, and in Section IV we evaluate the model and prepare charts to show result analysis and finally conclusions are given in Section V .

II. RELATED WORK

[Thelwall et al.] proposed SentiStrength, a lexicon approach to overcome the ill-formed language by applying several rules. As tweet size in limited to 140 characters, it contains short messaging text, slangs, acronyms, etc. Thelwall develop lexicons for these unstructured language to compute the average sentiment strength of text. The only problem with sentistrength is its static prior sentiment values of words regardless of their context.

[Hassan Saif et.al] presents a new lexicon approach using contextual behaviour of words, called SentiCircles, that capture the latent semantics of words from its co-occurrence patterns and update the sentiment and its strength accordingly. They propose many methods to detect the sentiments at both tweet and entity level. In this, the only limitation is when we update the prior sentiment and strength using Senti-Median method, updated sentiment is almost equal to zero as it takes into consideration all the senticircles for each term present in tweet. And every term is not assigned a pre-assigned sentiment, so by default it gives the sentiment zero. Let us take an example:

ipad and iphone are amazing.

As ipad, and, iphone, are not present in dictionary, only amazing is present in dictionary, it gives the sentiment for amazing only, for others it gives its sentiment to zero. When calculate the sentiment of tweet using senti-median method, the tweet sentiment is almost zero. Most of the positive tweets tend to negative or neutral.

[Fajri Koto et. al.] propose part of speech as one of the feature to investigate pattern or word combination of tweets in two major areas: subjectivity and polarity. They investigated many combinations of it by incorporating AFINN with many variations of POS Sequence, and concluded that feature of POS sequence are able to boost the accuracy.

[Shilpi Goyal et.al.] presents a literature survey related to recognizing emotions. They presents a 20 year research work that enhances day by day. This gives us a boost to identify more and more about this topic.

III. PROPOSED FRAMEWORK

In this section , we first present the general architecture shown in Figure 1 and then explain each part.

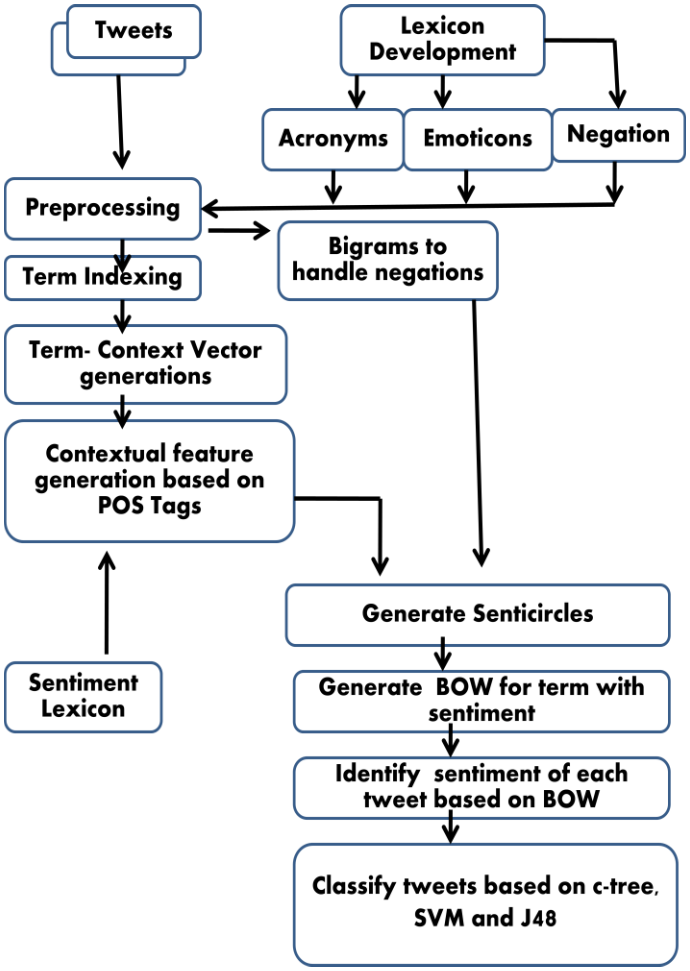
Tweets
Lexicon Development
Acronyms
Emoticons
Negation
Preprocessing
Bigrams to handle negations
Term Indexing
Term- Context Vector generations
Contextual feature generation based on POS Tags
Sentiment Lexicon
Generate Senticircles
Generate BOW for term with sentiment
Identify sentiment of each tweet based on BOW
Classify tweets based on c-tree, SVM and J48

Figure 1: Proposed Framework Architecture

A. ***Tweets Collection:*** Tweets are collected from Twitter API. We collect three datasets @narendramodi , DeMonetization and #FightAgainstCorruption with 17 attributes and 1000 tweets per dataset filtering non-english terms.

B. ***Lexicon Development:*** Various lexicons are developed inspired by Senti-Strength to remove ill-formed text, convert acronyms to its full form, convert emoticons , if any to its meaning while writing that tweet to express the mood of writer.

C. ***Sentiment Lexicon:*** Inspired by sentiwordnet to extract sentiment on contextual basis, we use SentiWordNet as one and uses AFINN as another lexicon based approach as our baseline.

D. ***Cleaning and Pre-processing:*** Each tweet is pre-processed so that acronyms, emoticons lexical tables are applied then remove numbers, remove hashtags, remove urls, remove retweets, lower case the tweets and stemming after that to remove stemmed words.

E. ***POS Tagging:*** Using openNLP each tweet is parsed ,then FW,CD,TO etc tags are removed from tweets. Considering only the noun, pronoun, adjective, adverb, verb as a feature extraction to find the necessary term that contribute to contextual semantics.

F. ***Negation:*** This can be identified by using bigrams. If a bigram first word is 'not' then negate the sentiment of second word in bigram. For instance 'amazing' is positive word but if a bigram 'not amazing' , a first word is 'not', negate the sentiment of 'amazing'.

G. ***Term-Indexing:*** This step creates an index of terms from a collection of tweets.

H. ***Term-Context Vector Generation:*** For each tweet(T) after preprocessed ,the term context vector of a term (m) is a vector ,c = (c1, c2, c3, ...) that occur with m in any tweet using term frequency-inverse document frequency method.

$$TCV(m,ci)= f(ci,m)* \log(N/Nci)$$

where N = total number of terms and Nc_i= total terms that co-occur with $c_{i.}$ where i means from 1 to length of context terms with each term m.

I. ***Generate senticircles:*** For each term(m), plot a senticircle with a term m as dark dot while all context term of that term are surrounding with light dot, labelled strength in x-coordinate axis and sentiment in y-coordinate axis.

J. ***Generate Bag of Words(BOW) for term with sentiment:*** The final update of sentiment with its strength after using partitioning around method is recorded and those terms are collected as bag of words.

$$\mathrm{BOW(term,g)} = \arg\min_{g\in R^2} \sum_{i=1}^{n} ||\mathrm{pi} - \mathrm{g}\,||^2$$

where BOW(term,g) is the bag of words of all terms with associated strength and sentiment, term is the word from a tweet, g comprises combination of strength and sentiment considering only those sentiments not equal to zero, pi is all context terms associated with term.

K. ***Tweet Classification:*** Using these BOW, identify the sentiment of each tweet.

L. ***Using supervised techniques:*** Using SVM, C- Tree and J48 techniques with these bag of words to identify the sentiments of tweets, we can test which model works best with our proposed model.

IV MODEL EVALUATION

This section presents the results and discuss the findings of proposed model. The training data consists of 500 tweets. The evaluation of the model was done at polarity of the tweet only.

Polarity Classification (Positive, Negative and Neutral)										
Methods	**Accuracy**	**Positive**			**Negative**			**Neutral**		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
AFINN +SVM	74.40	100.0	72.76	84.23	24.00	100.0	38.70	10.34	100.0	18.75
AFINN +CTree	73.60	96.49	73.33	83.33	41.37	87.5	24.13	14.00	70.58	52.17
AFINN +J48	60.00	67.25	92.00	77.70	22.00	30.55	25.58	82.75	26.97	40.67
SentiCircle+SVM	72.11	95.88	73.42	83.16	20.83	62.50	31.25	24.24	61.54	34.78
SentiCircle+CTree	68.13	100.0	68.00	80.95	0.00	0.00	0.00	3.03	100.0	5.88
SentiCircle+J48	71.37	93.60	73.18	82.14	15.38	42.10	22.53	33.33	88.88	48.48
Proposed+SVM	71.71	98.27	71.42	82.72	18.86	76.92	30.30	0.00	0.00	0.00
Proposed+CTree	72.98	99.41	72.15	83.62	0.00	0.00	0.00	41.66	90.90	57.14
Proposed+ J48	**73.39**	97.04	73.21	83.46	23.40	64.71	34.38	21.88	100.0	35.90

Table1: Tweet Level sentiment analysis results for @narendramodi tweets

V CONCLUSION AND FUTURE WORK

In this study, we discuss about the use of modified senticircle to form Bag of Words as a feature extraction to identify sentiments over polarity domain. It takes into account the co-occurrence pattern of words in different different contexts in tweets to capture their meaning to update the prior polarity and strength. Our approach allows for the detection of sentiment at tweet level. We evaluate our proposed approach on three datasets fetched from Twitter using AFINN and sentiwordnet with senticircles . Results show that our approach perform better than AFINN and sentiwordnet with senticircles by 2-11% in accuracy with J48 but falls marginally with SVM and CTree by 1-2%.

As one of the challenges was to cope up with multi-languages that are used in online social networking sites. As our datasets are related to Indian politics, most of the tweets are well expressed in Hindi language only, so our main focus will be to deal with Hindi language.

REFERENCES

[1] Yassine, Mohamed, and Hazem Hajj. "A framework for emotion mining from text in online social networks." *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*. IEEE, 2010.

[2] Thelwall, Mike, et al. "Sentiment strength detection in short informal text." *Journal of the American Society for Information Science and Technology* 61.12 (2010): 2544-2558.

[3] Saif, Hassan, et al. "Contextual semantics for sentiment analysis of Twitter." *Information Processing & Management* 52.1 (2016): 5-19.

[4] Koto, Fajri, and Mirna Adriani. "The use of POS sequence for analyzing sentence pattern in Twitter sentiment analysis." *Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on*. IEEE, 2015.

[5] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREc*. Vol. 10. No. 2010. 2010.

[6] Goyal, Shilpi, and Tiwari, Nirupama.”Emotion Recognition: A Literature Survey.” *International Journal For Technological Research In Engineering Volume 4, Issue 9, May-2017*