IJCTA, 9(10), 2016, pp. 4631-4638

Unsupervised Word Embedding Based Polarity Detection for Tamil Tweets

*** Nivedhitha E, Sanjay S P, Anand Kumar M, Soman K P**

Abstract : With the advent of technological advancements in the recent times, people, and Internet have become inseparable. People belonging to all categories access the internet. Micro blogs from twitter has real time information as on how the general audience feel. Extracting this information and finding out the intent of the general audience can be of great help for business and political organizations. Sentimental analysis is a sub-branch of Natural Language Processing which intends in finding out the polarity of contextual information. Tamil Tweets are collected and they are being manually tagged to develop a system that can identify the polarity. We have used word embedding and unsupervised methodology to identify the polarity of Tamil tweets. We have also evaluated our system using SAIL-2015 data set available for Tamil language and we were able to obtain state-of-art accuracy.

Keywords : Sentimental Analysis; unsupervised learning; Machine learning; sentiments for tweets; word vector ; Word Embedding; Opinion mining.

1. INTRODUCTION

We have always needed other person's opinion in the decision making process. With the humongous improvement in internet technology and web, it is now possible to find out the opinion of common people, who are neither our friends nor professional critics. They are just common people who happen to buy a particular product or opt a particular service. Lately, there has been a walloping development in micro blogging sites like twitter, tumblr, etc. These micro blogs contain real time information about what people feel. This information can be their opinion about current issues, or comment on products they purchased, reviews on the movies, comments on a particular service they opted and so on. Since these micro blogging sites give a direct connection to the common people, companies aim at collecting these posts and learn what the general audience feel about the product or service the company provides. And this is exactly what sentimental analysis aims at capturing, the sentiment of contextual information. Given a raw tweet, our classifying engine removes the unwanted information present in the tweet and then through our unsupervised learning algorithm, classifies the tweet if it is positive, negative or neutral.

2. RELATED WORK

Sentimental Analysis is a developing zone of Natural Language processing with exploration running from the entire document as input [1] or taking in words or sentences as the input [2][3]. Since the number of characters are limited in tweets , they are being explored in sentence level [4]. Opinion mining in terms of machine learning problem was experimented with SVM, Naive Bayes and Maximum Entropy. This was initiated and experimented by Pang et al [5]. Out of the three classifiers he experimented for a data set which had reviews about movies, the SVM classifier has outperformed the other two. The further research on the work aimed at improving the accuracy of the other two classifiers [6]. But there was only a marginal improvement in the other classifiers, compared to SVM. Then a number of research based on the paper bloomed, most of which aimed at the betterment of feature

* Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapeetham, Amrita University, India e.nivedhitha@gmail.com,manandkumar@cb.amrita.edu

set to improve the feature set. SVM and enhanced features set were experiment by Mullen and Collier [7]. All these were in terms of supervised learning. This current research work is based on an unsupervised approach and aims at polarity detection of Tamil data..There are several research papers published based on polarity detection of Tamil data. Document based classification based on random kitchen sink algorithm are being experimented [13]. Sentiment classification using regularized least square method was experimented with SAIL_2015 data [16]. Majority of the research work in NLP is done based on English and only very few research community works on regional languages. Some of the research work based on Tamil language worth mentioning are work done by Sanjana et al and Reshma et al [13] [14] for document classification. Twitter data classification in Tamil language is experimented by Arun selvan et al [18] , Vinithra et al [15] Polarity detection for Tamil movie reviews based on frequency count was performed by Arunselvan in his research work. [18]. One previous research which gave us the insight to work on unsupervised approach is Paltoglou, G. and Thelwall's research. They proposed a natural, unsupervised, dictionary based method which gauges the emotional value in the given text. Subjectivity identification and Polarity characterization is what they aimed at to provide a solution to sentiment classification [8].

3. POLARITY DETECTION FRAMEWORK

A. Data set Analysis:

In Micro-blogging the users communicate and exchange valuable information through a concise text message. Twitter is one such well-known example of micro-blogging for sharing information with other people through instant messages called “tweets”. Due to its short nature, tweets are restricted to only 140 characters. So people tend to use special characters, emoticons, short-forms,acronyms and even spelling mistakes to express the indented meaning.

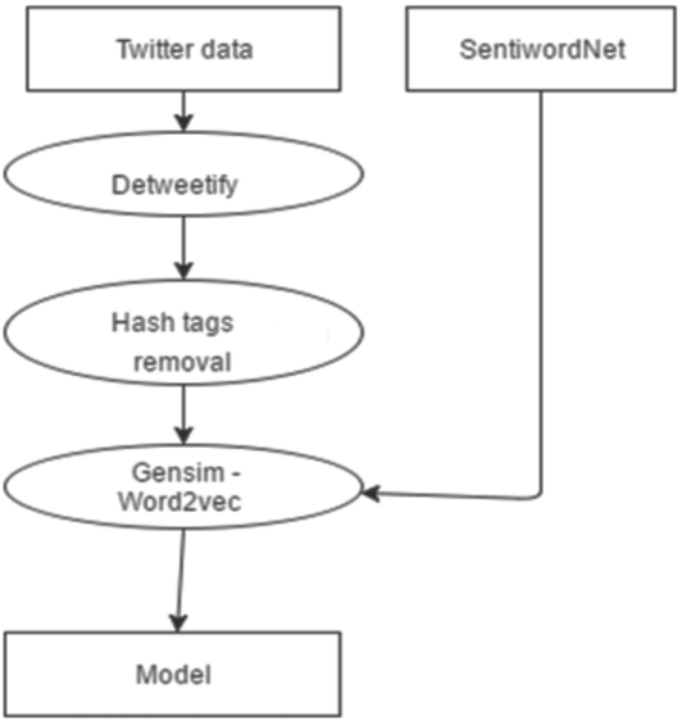

Fig. 1. An image depicting our preprocessing of data.

The figure 1 describes the preprocessing of data. Our input data is a raw twitter data which includes all the redundant information. In order to process these data, it is integral for us to remove these redundant information and keep only the necessary textual information. The raw twitter data is fed into the detweetify engine, where the intent is to remove the twitter format and extract only the necessary information. Soon after that, we remove the hash tags. This step happens in our hash tags removal block. After this, the data is sent to the next block, where the cleaned data is then converted to the vector form using GENSIM [11] and then fed into our model. Gensim is a

topic modeling toolkit implemented in python and supports a variety of implementations. Meanwhile, the dictionary data from SentiWordNet is also converted to its respective vector forms and fed into the model. SentiWordNet is a lexical database for Sentimental Analysis [9][10]. It assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. This public resource had 2279 positive words,4501 negative words and 415 neutral words. In addition to the available words in the dictionary, we have tried to add a few more words and emoticons in the dictionary to improve accuracy. These emoticons we added are twitter specific emoticons. We have separated out positive , negative and neutral emoticons and added them to the respective category. To evaluate the system, we are using SAIL-2015 [19] Tamil data set which contains 1103 tweets in training data and 560 in the test data. Our proposed data set contains 691 tweets which are tagged by us. These data sets are being used in our paper to obtain results and to cross validate its performance. We have collected around 74268 tweets for the training purpose. We have collected through twitter API and twitter archiver. Since the tweets contain other characters that are not needed for our experimental purpose, we remove these unwanted characters namely symbols like @,#, hyperlinks and other characters. The detweetifing engine takes care of this process and the resulting output is clean without the redundant information.This process recursively removes the unwanted data from the raw tweets and converts them into a file called model file. This file contains only the textual information of the tweets and their emoticons,which is needed for our sentiment classification.

Consider the following tweet,

Fig. 2. A sample Tamil tweet

In the above tweet, '@' represents the target user and '#' represents a particular topic and it may or may not contain a link. And 'RT' represents a tweet that is being re-tweeted. In most cases, re-tweets contain redundant data and hence in most cases are removed.

B. Model description

When a new tweet is given as input to the system, the very first step that happens in the system is to remove the unwanted information in the tweets. This is what happens in the detweetify module. We process the lexical database from SentiWordNet in parallel and feed the data into the model. This lexical database has words that are pre-categorized into positive, negative and neutral. They are converted to their vectored form after being fed into the model. We import word2vec algorithm for this conversion. Soon after this, we calculate the centroid vector for all the three categories. Now the distance vector is computed from the sentence vector which is our input sentence and the centroid vectors. This computed distance vector corresponds to the distance of positive, negative and neutral categories. Then we normalize the values for classifying. We have used zero-to-one normalization for straightlaced classification of data. After this the decision making algorithm predicts the results and generates the desired output.

C. Decision making algorithm

During the classification we have given unique identifying number for positive and negative classifier. We have assigned each polarity group with a unique id.

Accordingly identifier, '0', is assigned to positive polarity., if the identifier is '1', then it is assigned to negative polarity and we categorize to neutral polarity in other cases. If the sentence value from the identifier '0' equals to

one, then we assign it in the positive polarity. Alternatively if the sentence falls in the identifier ‘1 and equals to 1 , then we assign negative polarity. If the sentence falls in neither and valued to be one ,we assign neutral polarity. After this iteration we analyze values other than one. When the sentence value is greater than 0.5 and falls in the identifiers zero or one,we categorize them accordingly to positive or negative. Otherwise we set them to be neutral.

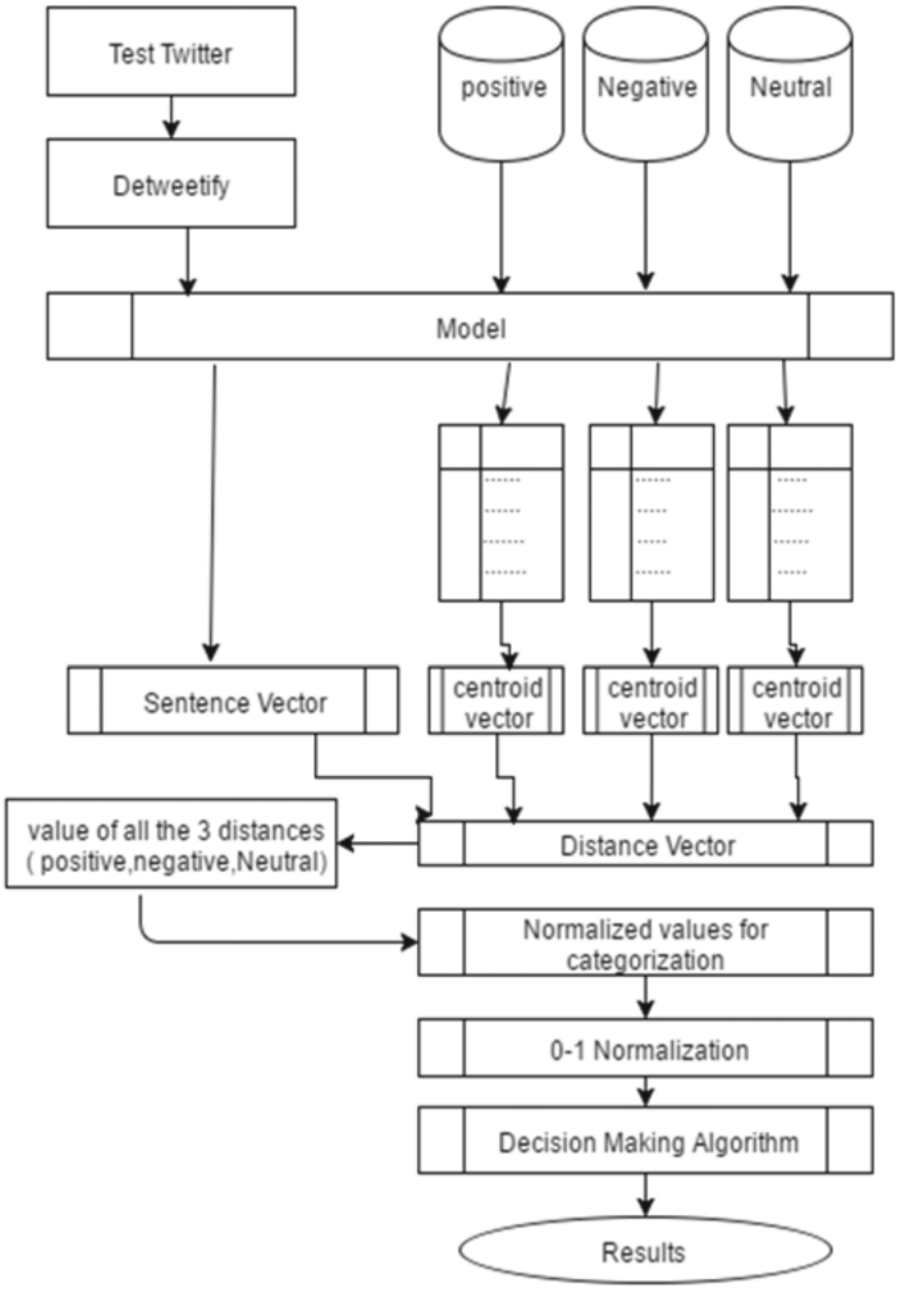

Fig. 3. Flowchart describing the entire system.

4. EVALUATION AND RESULTS

Our system is evaluated by two data sets. We have tagged 691 tweets and tested how the system classifies its polarity. Additionally we have used SAIL-2015 data set to test the performance of the system. Table 1, shows the scores generated of our proposed data set while Table 2 , shows the scores generated for the SAIL-2015 Tamil data set. Since there are three different data set involved, we have used . Our system predicted 388/560, 709/1103, 488/691 correct data from SAIL-2015 test data, train data and our proposed data respectively. We have

calculated the precision, recall and F_score values for all the three data set. The values in Table 1, 2 and 3 are calculated using the equations 1, 2,3. Precision is a measure of how accurate our prediction algorithm works. In the predicted items , how many are accurate to the query is calculated by recall measure. F-score is just a harmonic mean of precision and recall.

$$\text{Precision (P)} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall (R)} = \frac{TP}{TP + FN} \quad (2)$$

TP = True Positive
FP = False Positive
FN = False Negative

SAIL_2015 test data contains a totalof 560 tweets . Our system predicted 78 negative tweets out of 158, 80 neutral tweets out of 194 neutral tweets given and 180 positive tweets out of 297. And the average accuracy of this test data set is 60%. Training data of SAIL-2015 has 1103 tweets, out of which our system predicted 163 negative from 316 negative tweets , 274 neutral from 400 neutral tweets and 272 from 387 positive tweets.Accuracy acquired from this system is 64% and finally our proposed data has around 691 tweets. Our system was successful in predicting 178 negative out of 210 tweets ,98 neutral tweets out of 180 and 121 out of 301 positive tweets. The accuracy obtained for our proposed data is 70%. We have been able to achieve better accuracy for our data as we have updated and customized the dictionary data based on the tweets.

Table 1. Final result showing precision and recall scores of SAIL TRAINING data set

METRICS	***NEG***	***NEU***	***POS***
PRECISION	0.5532	0.6557	0.6061
RECALL	0.4937	0.4124	0.8654
F-SCORE	0.5217	0.5063	0.7129

Table 2. Final result showing precision and recall scores of SAIL TEST data set

METRICS	NEG	NEU	POS
PRECISION	0.6573	0.7117	0.5787
RECALL	0.5158	0.685	0.7028
F-SCORE	0.578	0.6981	0.6348

Table 3. Final result showing precision and recall scores of our proposed data set

METRICS	NEG	NEU	POS
PRECISION	0.6159	0.6853	0.8185
RECALL	0.8476	0.5444	0.7043
F-SCORE	0.7134	0.6068	0.7571

A. Micro and Macro Average

When we handle many classes, we often need a total measure that consolidates the measures for individual classes. In such situations we uses Micro and Macro averaging methods. Macro average is simply straightforward and it is just the average of all classes. So macro_precision is simply the average of all precision values and macro_recall is the average of all recall values. Whereas micro average sums up all the individual true positives and computes the average. Therefore, effectiveness of a system with huge collection of data can be computed using micro average values. The F_score for both micro and macro average is just the harmonic mean of precision and recall. We have calculated micro_precision and Micro_recall using equations 4 and 5. Macro_precision and macro_recall are calculated from equations 6 and 7.

Table 4. Micro and Macro Average scores

METRICS	*SAIL_2015 TEST DATA*	*SAIL_2015 TRAINING DATA*	*PROPOSED DATA*
MACRO_PRECISION	0.605	0.6492	0.7065
MACRO_RECALL	0.5905	0.6345	0.6987
MACRO_F-SCORE	0.5803	0.6369	0.6924
MICRO_PRECISION	0.6036	0.6427	0.7062
MICRO_RECALL	0.6036	0.6427	0.7062
MICRO_F-SCORE	0.6072	0.6427	0.7057

$$\text{Micro\_precision} = \sum_{i=1}^{n} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{4}$$

$$\text{Micro\_recall} = \sum_{i=1}^{n} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{5}$$

$$\text{TP} = \text{True Positive}$$
$$\text{FP} = \text{False Positive}$$
$$\text{FN} = \text{False Negative}$$

$$\text{Macro\_precision} = \frac{\text{P}_{\text{pos}} + \text{P}_{\text{neg}} + \text{P}_{\text{neu}}}{3} \tag{6}$$

$$\text{Macro\_recall} = \frac{\text{R}_{\text{pos}} + \text{R}_{\text{neg}} + \text{R}_{\text{neu}}}{3} \tag{7}$$

B. Accuracy

We have tested our system with three different data. SAIL_2015 test data and train data, then our proposed data set. Accuracy values obtained are 64% , 70% and 60% for SAIL-2015 training data and our proposed data and SAIL_TEST data respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{N}} \tag{8}$$

Where "TP" is the total number of true positives and "TN" the number of true negatives" and "N" total number of instances in the test set. We have provided a competitive table for the accuracy obtained for SAIL_2015 test data the system generated by AMRITA-CEN_@SAIL2015 and our proposed method. Comparison have been made with AMRITA_CEN-NLP_@SAIL2015 accuracy values too. A comparative table for accuracy values are shown in the table 5. On comparing the data, we conclude that our system is able to produce the state-of-art accuracy.

Table 5. Accuracy values of different data set

RUNS	*ACCURACY*
SAIL_2015_TRAINING DATA SET	0.6427
PROPOSED DATA	0.7062
SAIL_2015_ TEST DATA	0.6035
AMRITA-CEN_@SAIL2015	0.3928
Amrita_CEN_NLP_@SAIL2015	0.3232

5. CONCLUSION AND FUTUREWORK

Sentimental analysis for Tamil twitter data are being addressed in this paper.. We proposed a natural, unsupervised, dictionary based calculation which gauges the level of emotional quality in content so as to make a final decision. Once when tweet or short text is given into the application, the system predicts if the given data falls under positive ,negative or neutral polarity. The benefits of the methodology is that we don't have to train the system and therefore we have a great possibilities and can be connected to a situations of wide spectrum. We have been able to generate better accuracy with unsupervised learning compared to the state-of-art accuracy for existing data. Improving the dictionary database is observed to improve accuracy. In future, different advanced word embedding methods like Recurrent Neural Network, phrasal embedding etc can be experimented to improve the accuracy.

6. REFERENCES

1. Pang, Bo, and Lillian Lee. “Opinion mining and sentiment analysis.” Foundations and trends in information retrieval 2.1-2 (2008): 1-135.
2. Hatzivassiloglou, Vasileios, and Kathleen R. McKeown. “Predicting the semantic orientation of adjectives.” Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1997.
3. Esuli, Andrea, and Fabrizio Sebastiani. “Sentiwordnet: A publicly available lexical resource for opinion mining.” Proceedings of LREC. Vol. 6. 2006.
4. Kim, Soo-Min, and Eduard Hovy. “Determining the sentiment of opinions.” Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.
5. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. “Thumbs up?: sentiment classification using machine learning techniques.” Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
6. Pang, Bo, and Lillian Lee. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.” Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.
7. Mullen, Tony, and Nigel Collier. “Sentiment Analysis using Support Vector Machines with Diverse Information Sources.” EMNLP. Vol. 4. 2004.
8. Paltoglou, Georgios, and Mike Thelwall. “Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media.” ACM Transactions on Intelligent Systems and Technology (TIST) 3.4 (2012): 66.
9. Das, Amitava, and Björn Gambäck. ”Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality.” Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics, 2012.
10. Das, Amitava Opinion Extraction and Summarization from Text Documents in Bengali. Diss. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., Jadavpur University, 2011.
11. Rehurek, Radim, and Petr Sojka. “Software framework for topic modelling with large corpora.” In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010.

12. Das, Dipankar, and Sivaji Bandyopadhyay. “Labeling emotion in Bengali blog corpus-a fine grained tagging at sentence level.” Proceedings of the 8th Workshop on Asian Language Resources. 2010.

13. Sanjanasri, J.P., Anand Kumar, M. A computational framework for Tamil document classification using Random Kitchen Sink (2015) 2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015, art. no. 7275837, pp. 1571-1577.

14. Reshma, U., Barathi Ganesh, H.B., Anand Kumar, M., Soman, K.P. Supervised methods for domain classification of Tamil documents(2015) ARPN Journal of Engineering and Applied Sciences, 10 (8), pp. 3702-3707.

15. Vinithra, S.N., Arun Selvan, S.J., Anand Kumar, M., Soman, K.P. Simulated and self-sustained classification of Twitter Data based on its sentiment(2015) Indian Journal of Science and Technology, 8 (24), art. no. IPL0345,

16. Sachin Kumar, S., Premjith, B., Anand Kumar, M., Soman, K.P. AMRITA_CEN-NLP@SAIL2015: Sentiment analysis in Indian language using regularized least square approach with randomized feature learning(2015) Lecture Notes in Computer Science (including sub series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9468, pp. 671-683

17. .Se, S., Vinayakumar, R., Anand Kumar, M., Soman, K.P. AMRITA-CEN@SAIL2015: Sentiment analysis in Indian languages (2015) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9468, pp. 703-710

18. Arunselvan, S.J., Anand Kumar, M., Soman, K.P.Sentiment analysis of tamil movie reviews via feature frequency count(2015) International Journal of Applied Engineering Research, 10 (20), pp. 17934-17939.

19. A. Das and S. Bandyopadhyay. SentiWordNet for Indian Languages, In the 8th Workshop on Asian Language Resources (ALR), COLING 2010, Pages 56-63, August, Beijing, China