Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018)
IEEE Xplore Compliant Part Number: CFP18K74-ART; ISBN:978-1-5386-2842-3

Sentiment Analysis of Twitter Data

Radhi D. Desai
Department of Computer Engineering
Sardar Vallabhbhai Patel Institute of Technology,
Vasad, Gujarat- 388 306, INDIA
E-mail: desairadhi25@gmail.com

***Abstract*—Sentiment analysis of Twitter data became a research tread in the last decade. Among popular social network portals, Twitter had been the point to fascination to several researchers in important areas like prediction of democratic events, customer brands, movie box-office, stock market, reputation of personalities etc. The term sentiment refers to the feelings or opinion of person towards some particular domain. Analysis of sentiment (opinions) and its classification based on polarity is a challenging task. Other challenges are overwhelming amounts of information on one topic with all having different representation. Classification and clustering are two major methods applied to perform sentiment analysis of twitter data. We have used Possibilistic Fuzzy C-Means with SVM to improve accuracy on movie tweets and worked on upto 3-grams.**

Keywords*— *Sentiment analysis; Opinion mining; Clustering; Classification; Review analysis.

I. INTRODUCTION

Twitter has become very popular and has grown rapidly. As per the current report 974 million users and 500 million tweets per day that is considered as a valuable online source for opinions. And the length of the tweet is expanded to 280 characters since November 2017. Challenge is to extract important data from twitter reviews, since the nature of content is unstructured. Twitter need been the perspective of fascination will a few specialist on essential regions like prediction about equitable a few events, customer brands, motion picture box-office, stock market, Notoriety from claiming superstars and so on.[1]

"Sentiment Analysis is defined by the process of computationally recognizing and classifying opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral." Sentiment Analysis is generally carried out in three steps. First, liable towards which the subject is guided may be discovered. Then, the polarity of the sentiment is calculated and finally the degree of the polarity is assigned with the help of a sentiment score which denotes the intensity of the sentiment. Sentiments can be classified at various levels: Aspects or feature level, sentence level and document level. In Document level Sentiment polarity of whole document is determined whether it is positive or negative or neutral class. In Sentence level classifies each sentence based on their sentiment score for particular topic. In Aspect level classifies the sentiments based on the sentiments of each aspects or feature for particular topic. [2]

Fig.1: Review of Sentiments

Sentiment analysis can be done by Machine Learning Methods and Lexicon Method. In Machine Learning there are Supervised, Semi-supervised and Unsupervised approach, and in lexicon based there are Corpus based and Dictionary based approach. In supervised approach various types of classification (Decision tree classifier, Linear classifier, Probabilistic classifier, Rule based classifier); in unsupervised Clustering (Partition based clustering, Hierarchical clustering, Density based clustering, Hybrid clustering) and in semi-supervised the combination of both of them is used.

II. BACKGROUND

Liza et al [3] they suggest three phases of text mining i.e. pre-processing, processing and validation. After applying primary pre-processing, it performs weighting schemes and use Naïve Bayes as a classification algorithm. Then after in validation phase uses 10-fold cross validation testing.

Yunchao et al [4] present both unigram and bigram as feature extraction and cluster the texts using K-Means clustering. And after Naïve Bayes classification algorithm is applied.

Rishabh et al [5] proposed cluster-than-predict approach, first cluster the tweets using K-Means clustering and then perform classification using CART (Classification and Regression Trees) to improve the accuracy.

Gang Li et al [6] use TF-IDF scheme as feature extraction. Then for improve the result and detect neutral polarity they suggest voting mechanism and distance measure approach. After that apply K-Means clustering to find the review i.e. positive, negative or neutral.

Hima et al [7] a novel fuzzy clustering model and compare it with K-Means and Expectation Maximization algorithms. And the result is practicable for high quality twitter sentiment analysis.

Nagamma et al [8] applied TF-IDF for feature extraction and after Fuzzy C-Means clustering is used to improve the result of classification and after apply Support Vector Machine and Naïve Bayes. Then predict the revenue from the reviews about movie.

Yunchao et al [13] address to estimate the sentiment of unlabeled data, they use a two-step-merge method. They use clustering for sparsity problem and NB classifier for categories the text. It gives the better result than bag-of-words method.

III. METHODOLOGY

A. *Retrieve tweets:*

There are three methods to collect the tweets [9]. Data repositories such as UCI, SNAP and Friendster Automated tools: Premium tools such as Lithium, Simplify360, Sysmos, Radian6 and non-premium tools such as SocialMention, Tagboard, Keyhole, Topsy and are used for the tweet collection.

APIs: There are two types of APIs such as Stream API and Search API. Stream API is used to stream real time data from Twitter and Search API is used to collect Twitter data on the based on hashtags.

B. *Pre-Processing:*

Mining of Twitter data is a challenging task. The collected data is raw data. In order to apply classifier, it is essential to pre-process or clean the raw data. The pre-processing task involves removal of stop-words from raw data, removal of affixes, removal of other notation like hashtag, @, RT, emoticons, URLs, convert acronyms.[10]

C. *Sentiment Score:*

Sentiment score is assigned by the AFINN dictionary. Two versions of AFINN dictionary are AFINN-96 and AFINN-111. [12]

AFINN-96 consists 1468 words and AFINN-111 consists 2477 words including 15 phrases.

i. Very negative (-5 or -4)

ii. Negative (-3, -2, or -1)

iii. Positive (1, 2, 3)

iv. Very positive (4 or 5)

D. *Review of Sentiment:*

Using various supervised, unsupervised or combine of both of them techniques are used to find the polarity of the sentences.

In supervised method, there is machine learning methods are used with training and testing data i.e Classification, in that SVM and NB used. In unsupervised method, data are not pre-labeled i.e clustering, in that FCM and PFCM used.

IV. PROPOSED SYSTEM

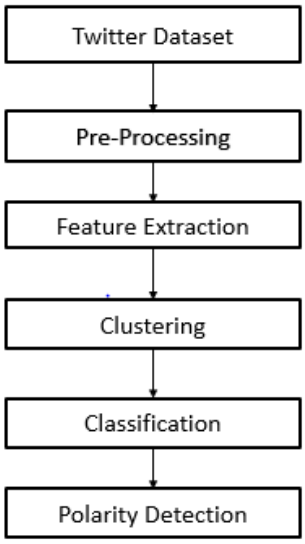

Fig. 2: Steps of Sentiment Analysis

A. Retrieve Tweets (Dataset): we are taking live tweets using twitter access tokens.

B. Pre-processing: Not remove special characters because it may belong to emoticons like :), :|. Furthermore, we attach word activated server to make Happyyy to its original word Happy.

C. Feature extraction: The pre-processed dataset has various discrete properties. We use three feature extraction schemes:
I. Emoticons: In this feature, entered the emoticons in the form text can also predict as pictorial emoticons. Like :) (☺) is predict as happy same as :((☹) is predict as sad. For that the lexicon dictionary is used. So we use combine (AFINN and lexicon) dictionary.
II. Synonyms: If the word is not present in dictionary, then score of that sentence will be 0, it leads to negative decision. So we implement that if the word is not found in dictionary, it will find it synonyms and replace with it.
III. 3-gram: N-gram is a contiguous sequence of n items from a given sample of text or speech. We use combine method of unigram, bigram and trigram. In that we try to improve the performance using trigram as NOT VERY BAD implies to the positive review.

D. Clustering: To improve the classification result, we use PFCM (Possibilistic Fuzzy C-Means) Clustering to overcome the disadvantages of K-Means i.e. cluster conclusion is responsive to the selection of the foremost cluster centroids and may converge to the local optima and Fuzzy C-Means i.e. constraint is each row in membership

matrix must sum to 1 and membership of a data points depends directly on the membership values of other cluster centers and noise points and outlier also accounted in the membership values.[11]

Implementing PFCM: The sentences are converted into data points and then the PFCM model is implemented on the data points. The constraint of FCM is neglected by PFCM. Cluster centers, membership matrix and typicality matrix are the output of PFCM. The optimization problem of PFCM is:

$$J_{m,\eta}(M,T,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c}(au_{ik}^{m} + bt_{ik}^{n}) \times \|x_k - v_i\|_A^2 + \sum_{i=1}^{c}\gamma_i \sum_{k=1}^{n}(1-t_{ik})^{\eta}$$

We have to minimize the J. uik >0 and tik >1, a>0, b>0, m>1, η>1 are user constant. A defines relative importance of fuzzy membership and b defines typicality values in the objective function.
The minimization problem can be achieved by calculating the membership matrix M, typicality matrix T, and centroids matrix V using following equations:

$$u_{ik} = \left(\sum_{j=1}^{c}\left(\frac{D_{ikA}}{D_{jkA}}\right)^{2/(m-1)}\right)^{-1}$$
$$1 \le i \le c;\ 1 \le k \le n$$

Distance between i^{th} cluster and kth data is represented by D_{ikA}
And represents typicality value as t_{ik}

$$D_{ikA} = \left[\sum_{j=1}^{s}(x_{kj} - v_{ij})^2\right]^{1/2}$$

$$t_{ik} = \frac{1}{1+\left(\frac{b}{\gamma_i}D_{ikA}^2\right)^{1/(\eta-1)}}$$
$$1 \le i \le c;\ 1 \le k \le n$$

To reduce the effect of outliers on centroids, we must use a upper value for b compare to a. And put the limit on the choice of η as per the values of m.

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta})\mathbf{x}_k}{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta})}$$
$$1 \le i \le c.$$

Calculate until the condition is not satisfied: $D_{ik} = \|x_k - v_i\|A < 0$.

E. Classification: After improving the feature sets using clustering, then apply the classification. Support Vector Machine classifier used for the final sentiment classification. Support Vector Machine is a classification technique uses hyper-plane formed during the training procedure to segregate one class from the other class.

V. Results and Analysis

A. Results

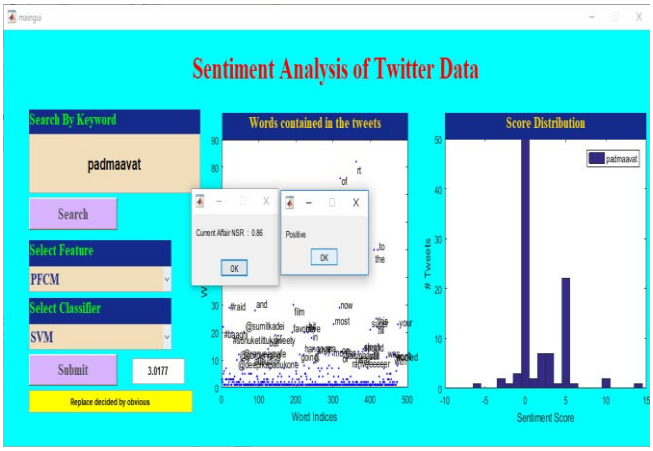

Fig. 3: PFCM Live tweet

As from above results we have taken 100 live tweets of "Padmaavat" movie, then we applied N-gram model upto 3-grams and PFCM clustering with SVM classifier. And get the word indices and NSR plot as in GUI. At last get the classifier decision in positive class and accuracy 88.89%.

```
SVM (1-against-1):
accuracy = 91.67%
Confusion Matrix in Persentage:
     6     0
     1     5


precision =

    0.8571
    1.0000


recall =

    1.0000
    0.8333

Predicted Query Belongs to Class = 1

Kappa_Index =

    0.8333
```

Fig. 4: Confusion Matrix

B. Analysis

No.	Clustering Type	Classification Type	Accuracy
1	FCM	NB	66.67%
2	FCM	SVM	83.33%
3	PFCM + upto 3-gram	NB	72.86%
4.	PFCM + upto 3-gram	SVM	91.67%

CONCLUSION

As reviews of individuals are extremely useful for people and company owner for making several decisions, is proposed in the literature for sentiment analysis. In this paper method applied for movie reviews, it is also applied for products, political parties, sport etc. To improve the results, paper uses new set of features extraction, using machine learning technique and hybrid dictionary (afinn, lexicon), this gives better accuracy than state-of-the-art techniques. From this research, conclude that PFCM gives better accuracy than FCM with SVM classifier.

REFERENCES

[1] Mitali Desai, Mayuri Mehta, "Techniques for Sentiment Analysis of Twitter Data- A Comprehensive Survey", IEEE, pp.149-154,2016

[2] Jatinder Kaur, "A Review paper on Twitter Sentiment Analysis Techniques", International Journal for Research in Applied Science & Engineering Technology, vol.4, pp.137-141, October-2016.

[3] Liza Mikarsa, Sherly Novianti Thahir, "A Text Mining Application of Emotion Classifications of Twitter's user using Naïve Bayes Method", IEEE, 2015

[4] Yunchao He, Chin-Sheng Yang et al, "Sentiment Classification of Short Texts based on Semantic Clustering", IEEE,2015.

[5] Rishabh Soni, K. James Mathai, "Effective Sentiment Analysis of a Launched Product using Clustering and Decision Tree", International Journal of Innovative Research in Computer and Communication Engineering, vol.4, pp.884-891, January 2016.

[6] Gang Li, Fei Liu, "Sentiment Analysis based on Clustering: A Framework in Improving Accuracy and Recognizing Neutral Opinions", Springer, September 2013.

[7] Hima Suresh, Dr.Gladston Raj. S, "An Unsupervised Fuzzy Clustering Method for Twitter Sentiment Analysis", IEEE,2016.

[8] Nagamma P, Pruthvi H.R et al, "An Improved Sentiment Analysis of Online Movie Reviews based on Clustering for Box-Office Prediction", IEEE, 2015

[9] "Three Cool and Inexpensive Tools to Track Twitter Hashtags", June 11,2013.[Online].Available http://dannybrown.me/2013/06/11/threecool-toolstwitterhashtags/ [Accessed: 19-Oct-2015].

[10] Roshan Fernandes, Dr. Rio D'Souza, "Analysis of Product Twitter Data though Opinion Mining", IEEE, 2016

[11] Nidhi Grover, "A Study of Various Fuzzy Clustering Algorithms", International Journal of Engineering Research, vol.3, pp.177-181, 2016

[12] Govin Gaikwad, Prof. Deepali Joshi, "Multiclass Mood Classification on Twitter using Lexicon Dictionary and Machine Learning", IEEE,2015

[13] Yunchao He, Chin-Sheng Yang, Liang-Chih Yu, K. Robert Lai, Weiyi Liu, "Sentiment Classification of Short Texts based on Semantic Clustering", IEEE, 2015