ISSN 2278-3091
Volume 7, No.5, September - October 2018
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse01752018.pdf
https://doi.org/10.30534/ijatcse/2018/01752018

A Review on Text Based Emotion Recognition System

Priyanka Thakur [1] Dr. Rajiv Shrivastava [2]

M.Tech Scholar[1]
Department of CSE, SIRT-Excellence
tpriyanka132@gmail.com

Professor & Director[2]
Department of CSE, SIRT-Excellence
drrajivsri@gmail.com

Abstract

With development of Internet and Natural Language processing, use of regional languages is also grown for communication. Sentiment Analysis is natural language processing task that mine information from various text forms such as blogs, reviews and classify them on basis of polarity. Sentiment analysis is a sub-domain of opinion mining where the analysis is focused on the extraction of emotions and opinions of the people towards a particular topic from a structured, semi-structured or unstructured textual data. In this paper, we try to focus our task of sentiment analysis on text data. We examine the sentiment expression to classify the polarity of the text review on a scale of negative to positive and perform feature extraction and ranking and use these features to train our classifier to classify the text data into its correct label.

Keywords—Text Data; Emotions; Feature Extraction; Classification; Emotion recognition;

1.INTRODUCTION

Emotions are an important aspect in the interaction and communication between individuals. The exchange of emotions through text messages and posts of personal blogs poses the informal kind of writing challenge for researches. Extraction of emotions from text will applied for deciding the human computer interaction that governs communication and many additional [1]-[3]. Emotions is also expressed by a person's speech, facial and text primarily based emotion respectively. Emotions are also expressed by one word or a bunch of words. Sentence level emotion detection technique plays a vital role to trace emotions or to look out the cues for generating such emotions. Sentences are the essential info units of any document. For that reason, the document level feeling detection technique depends on the feeling expressed by the individual sentences of that document that in turn depends on the emotions expressed by the individual words.

Globally, the emotions are divided into six types that are joy, love, surprise, anger, disappointment and worry [2]. Adequate amount of work has been done associated with speech and facial emotion detection however text based emotion recognition system still needs attraction of researchers. The short messaging language have the power to interrupt and falsify natural language processing tasks done on text data.

Human brain is trained with previous experiences. However once it involves natural language processing tools, they're trained and adopted to work properly with plain text. Mapping short messaging language words to plain text words are often terribly sensitive at some cases. A wrong mapping may result in alternations of the means or it's going to destroy semantics under the applied context.

The rapid growth of the World Wide Web has facilitated increased on-line communication, blog post and written content over the websites and opens the newer avenues to detect the emotions from that text data. This has led to generation of large amounts of online content rich in user opinions, emotions, and sentiments [4]. These needs computational approaches to successfully analyse this online content, recognize, and draw useful conclusions and detection of emotions. The existing techniques deals with the polarity recognition of sentiment. The sentiment maybe positive or negative [5].

Classification [6] is the process of classifying instances into their respective classes. Classification comprises of variables with known values to predict the unknown or future values of other variables. For example, a bank loan officer wants to analyze data in order to know which customer i.e. loan applicants are risky or safe. Some classification strategies are binary while the other classification strategies include ontology, neural networks, deep learning, etc. Multiclass classification [5]-[12] is classification of instances into more than two classes. Multiclass classification makes an assumption

that each sample is assigned to one and only one label i.e. a flower can be only rose or lotus not both at same time. In this paper we are using ontology to predict multiple classes of Hindi text. Ontology [12]-[18] is defined as 'Explicit specification of conceptualization'. As knowledge representation formalism, ontology's have found a wide range of applications in the areas like [4] knowledge management, information retrieval and information extraction. Sentiment analysis (SA) is natural language processing task that extracts the sentiments from various texts and classifies them accordingly into positive, negative or neutral classes. A classic example of SA is, shopping online for any product. A customer read reviews for that product.

2. RELATED WORKS

In the field of sentiment analysis, very limited amount of work has been done in Hindi language.[9] The very initial research work was done in Hindi, Bengali and Marathi language. Das and Bandopadhya[1] developed sentiwordnet for Bengali language using English-Bengali dictionary. 35,805 words were created by them.

Das and Bandopadhya[2] gave four strategies to predict sentiment of word. First strategy proposed by them was an interactive game which returned annotated words with their polarity. In second strategy, they use bi-lingual English and other Indian Language dictionaries to predict the polarity. In third approach, they use wordnet and synonym-antonym relation to predict the polarity. In fourth approach, polarity is determined by learning from pre-annotated corpora. Joshi et al. [3] proposed fall back strategy for Hindi Language. Their strategy follows three approaches: In-Language Sentiment Analysis, Machine Translation, Resource based sentiment analysis. They developed Hindi SentiWordnet(HSWN) by replacing words of English SentiWordnet by their Hindi Equivalents. Final accuracy achieved by them is 78.14.

Piyush Arora[4] proposed a graph based method to build a subjective lexicon for Hindi Language which is dependent on Wordnet. They initially build a small list of seedwords and expanded them by using wordnet, synonym, antonym. Every word in the seedlist is considered as node and is connected to their synonym and antonym. They achieved 74% accuracy on classification of reviews and 69% in agreement with human annotators.

Namita Mittal et al [5] developed an efficient approach based on negation and discourse relation for predicting sentiment. They improved HSWN by adding more opinion words to it. They proposed rules for handling negation and discourse that affected the prediction of sentiments. 80% accuracy was achieved by their proposed algorithm.

M. Farhadloo et. al. [6] proposed multiclass sentiment analysis for English language using clustering and score representation. The model used aspect level sentiment analysis. Bag of nouns was preferred instead of bag of words to enhance clustering results, score representation and more accurate sentiment identification.

Bhattacharyya et. al. proposed a fall-back strategy for sentiment analysis in Hindi. The three approaches [7] Machine Translation, In -language translation and resource based SA are used for Sentiment analysis in Hindi. To determine polarity SVM classifier was used. In machine translation, Google translator is used to translate Hindi data into English and they check polarity in terms of positive and negative. In resource based SA, the subset of EnglishSentiWordNet was used to build the subset of HindiSentiWordNet. They have achieved 78.14% as the best accuracy using in-language sentiment analysis for Hindi documents. Kisorjit, Bandyopadhyay proposed a verb based approach for Manipuri Sentiment analysis [8]. They used an unsupervised learning approach called CRF (Conditional Random Field). With the help of POS tagger the verbs are identified and polarity is notified. They also proposed the same model for Bengali language.

Aditya Joshi, Balamurli [9] proposed cross lingual sentiment analysis for Indian Languages. Machine translation is not possible for every pair of languages so they proposed a model for linkage of sentinets of two languages to overcome the language gap and provide better accuracy. An accuracy of 72% and 84% was achieved for Hindi and Marathi sentiment classification respectively.

Godbole, Manjunath and Stevens in their work [10] display a framework that measures positive or negative sentiment to each particular substance in the text corpus. Their framework comprises of two stages, a sentiment acknowledgment stage where opinion expressing elements are resolved and a scoring stage where a relative score for every substance is resolved. In the work by Annett and Kondark [11] it was determined that ML method of sentimental analysis on movies reviews is very fruitful and it was additionally watched that the sort of highlights that are picked dramatically affect on precision of the classifier. As there is an upper bound on the precision level that a reference based approach has as shown in lexical approach.

Pang & Lee work [12] is thought to be a standard in sentimental analysis of movie review. They consider the issue of ordering archives not by topic, but rather by overall sentiment, e.g. deciding if a review is good or bad. They inference, that traditional machine learning methods gives preferable outcomes over human-created baselines. In any case, the three machine learning techniques they utilized (Naive Bayes, Maximum entropy Classification, and Support vector machines) don't give as effective outcomes on sentiment grouping as on traditional classifier based classification. They additionally extricating these portions [13] and executing productive systems for discovering least cuts in

graphs; this significantly supports liberality of cross-sentence relevant imperatives, which gives an effective intends to coordinating inter sentence level logical data with customary dictionary of words features.

Singh et al. [14] presents experimental analysis on SentiWordNet approach for performance evaluation for document level sentiment arrangement of Movie audits and Blog posts. Researchers performed variation in semantic features, scoring schemes and thresholds of SentiWordNet approach alongside two most important machine learning approaches i.e. Naive Bayes and SVM. The similar execution of the methodologies for both movie as well as blog reviews is represented through standard execution assessment measurements of Accuracy, F-measure and Entropy.

Tirath Prasad Sahu et al. [15] extracted that features which are strongly effective in deciding the extremity of the movie reviews and used computation linguistic methods preprocessing of the information. Feature impact analysis is also performed by researchers in this paper by computing information gain for each feature to derive a reduced feature set. Six classification techniques are analyzed on this technique and found that Random Forest outperforms an accuracy of 88.95%.

Sruthi S et al [16] proposed an algorithm to eliminate the irrelevant information from the reviews in preprocessing stage. N-gram and lexicon approach are used to extract key terms and SVM classifier is used to classify sentiments. Performance of SVM is about 95% and proposed sentiment analysis system efficient in terms of time and cost.

3.PREVIOUS APPROACHES

English is the most popular language for research in Natural Language Processing. Most approaches used in this area are :-

- Subjective Lexicon
- Machine Learning

A. Subjective Lexicon Approach

Lexicon approach depends on finding opinion lexicon which analyzes sentiment of text. This approach has 2 methods:- Dictionary based and Corpus based. There are 3 main approaches in finding opinion list. Manual approach is very time consuming so it is combined with either of these two. Hindi language is scarce due to limited resources till now.

There are three popular methods for generation of subjective lexicon:-

- Use of Bi-lingual dictionary[2]
- Machine Translation[2]
- Use of Wordnet[4]

B. Machine Learning Approach

In such way total feature vector is generated for each text input using above features. These features are further classified by using classifiers. For each extracted features of emotional text classification algorithm is applied on different set of inputs. Different classifiers are discussed below[13]-[16]:

i. Support Vector Machine (SVM)

SVM, a binary classifier is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems, and under the conditions of limited training data, it can have a very good classification performance compared to other classifiers [14]. The idea behind the SVM is to transform the original input set to a high dimensional feature space by using kernel function. Therefore non-linear problems can be solved by doing this transformation.

ii. Hidden Markov Model (HMM)

The HMM comprises the first order Markov chain whose states are hidden from the observer so the inner behavior of the model remains hidden. The hidden states of the model capture the temporal structure of the information. Hidden Markov Models are statistical models that describe the sequences of events. HMM has the advantage that the temporal dynamics of the text features may be trapped owing to the presence of the state transition matrix. Throughout clustering, text is taken and therefore the chance for every text provided to the model is calculated. An output of the classifier is predicated on the maximum chance that the model has been generated this signal [15]. For the emotion recognition using HMM, 1st the information is prepared according to the mode of classification then the features from input waveform are extracted. These features are then further added to database. The transition matrix and confusion matrix will further created, that generates the random sequence of states and emissions from the model.

iii. K Nearest Neighbor (KNN)

A more general version of the nearest neighbor technique bases the classification of an unknown sample on the "votes" of K of its nearest neighbor rather than on only it's on single nearest neighbor. Among the various methods of supervised statistical pattern recognition, the Nearest Neighbor is the most traditional one, it does not consider a priori assumptions about the distributions from which the training examples are drawn. It involves a training set of all cases. A new sample is classified by calculating the distance to the nearest training case, the sign of that point then determines the classification of the sample. Larger K values help reduce the effects of noisy points within the training data set, and the choice of K is often performed through cross validation [16].

iv. AdaBoost Algorithm

AdaBoost algorithm is an adaptive classifier which iteratively builds a strong classifier from a weak classifier. In each iteration, the weak classifier is used to classify the data points of training data set. Initially all the data points are given equal weights, but after each iteration, the weight of incorrectly classified data points increases so that the classifier in next iteration focuses more on them. This results in decrease of the global error of the classifier and hence builds a stronger

classifier. AdaBoost algorithm is also used as a feature selector for training SVMs [15].

v. Neural Network Algorithm

In neural network input data and target data are loaded. Input data here is a matrix of the features extracted from the text inputs. Target data indicates the emotional states of these inputs. Next, the percentage of input data into 3 categories namely training, validation and testing is chosen randomly. The training set fits the parameters of the classifier i.e. finds the optimal weights for each feature. Validation set tunes the parameters of a classifier that is it determines a stop point for training set. Finally test set tests the final model and estimates the error rate. The default value sets training in 70 percent and 15 percent each for the rest. Initially the default values are used. Next, the number of hidden layers is chosen such that, more the number of hidden layers, more complicated the system, better the result. Lastly the network is trained several times. The mean square together with error rate will indicate how good the results are [13].

4.PERFORMANCE EVALUATION PARAMETERS

i. Recognition Accuracy

This measure signifies the recognition accuracy in percentage for each known test text input to the total trained emotional text data and is given by [20]-[25]:

$$\text{Accuracy} = \text{correct/predictions}$$

ii. Precision Rate

It is defined as the ratio of correctly recognized emotions for each class to the correctly recognized emotions for all the classes and is given by [25]:

$$\text{Precision} = \frac{\text{Correctly recognized emotions for a class}}{\text{Correctly recognized emotions for all class}}$$

iii. F-Measure

The F-Measure is the merit of combination of precision rate and recall. The performance of the implementation was evaluated from this factor to obtain the overall performance of the system in term s of correct results i.e. by not considering the wrong recognition observations and is given by [25]:

F-Measure= 2 * [(precision * recall) / (precision + recall)]

I. PROPOSED METHODOLOGY

A. Data Collection

The proposed algorithm first of all build a corpus of text data[24][25].

B. Preprocessing

Data preprocessing and cleaning step is important for subsequent analysis [11]. Preprocessing includes removal of extra symbols. Stemming is also done as a part of data preprocessing. Removal of stop words was done by stop word list created in text.

C. Negation Handling

There are certain words which are called negation words like- ”NA”, ” NAHI” These words can invert the polarity of the sentence. So, these words are also considered in finding polarity of text.

D. Classifcation

Then the proposed algorithm decides the threshold scoring scheme that will classify the given text into different class of emotions.

5.CONCLUSION

Sentiment analysis (or) text mining plays a significant role in business decision making. Many of the organization and enterprises will take their business decision only based on their customer review. In this study, the overview of different text emotion recognition methods are discussed for extracting text features from text sample, various classifier algorithms are explained briefly. English text Emotion Recognition has a promising future and its accuracy depends upon the combination of features extracted, type of classification algorithm used and the correct of emotional text database. This study aims to provide a simple guide to the researcher for those carried out their research study in the text emotion recognition systems.

REFERENCES

[1] Amitava Das, Sivaji Bandopadaya, “SentiWordnet for Bangla”, Knowledge Sharing Event -4: Task, Volume 2,2010.

[2] Amitava Das, Sivaji Bandopadaya, “SentiWordnet for indian language”, Workshop on Asian Language Resources, pp. 56-63, Beijing, China, 21-22 August 2010.

[3] Aditya Joshi, Balamurali AR, Pushpak Bhattacharya, “A fall back strategy for sentiment analysis in hindi”, International Conference on Natural Language Processing, 2010.

[4] Piyush Arora, Akshat Bakliwal, Vasudev Verma, “Hindi Subjective Lexicon Generation using WordNet Graph Traversal”, IJCLA vol. 3, no. 1, pp. 2539, 2012.

[5] Namita Mittal, Basant Aggarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek, “Sentiment Analysis of Hindi Review based on based on Negation and Discourse Relation”, International Joint Conference on Natural Language Processing, pp 45-50, 2013.

[6] Mohsen Farhadloo, Erik Rolland,” Multi-Class Sentiment Analysis with Clustering and Score Representation”, IEEE 13th International Conference on Data Mining Workshops, pp. 904-912, 2013.

[7] D. Das and S. Bandyopadhyay, “Sentence-level emotion and valence tagging,” Cognitive Computation, vol. 4, no. 4, pp. 420–435, 2012. https://doi.org/10.1007/s12559-012-9173-0

[8] Nongmeikapam, Kishorjit, Sivaji Bandyopadhyay, Dilipkumar Khangembam, Wangkheimayum Hemkumar,

Shinghajit Khuraijam, "Verb Based Manipuri Sentiment Aanalysis", International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, pp. 113-119, 2014. https://doi.org/10.5121/ijnlc.2014.3311

[9] Balamurali A R, Aditya Joshi, Pushpak Bhattacharyya, "Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets", COLING 2012, pp. 73-8, 2012.

[10] Godbole, Namrata, ManjaSrinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." ICWSM 7 (2007): 21.

[11] Annett, Michelle, and GrzegorzKondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs." Advances in artificial intelligence. Springer Berlin Heidelberg, 2008. 25-35.

[12] Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. "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.

[13] 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. https://doi.org/10.3115/1218955.1218990

[14] Singh, V. K., et al. "Sentiment analysis of movie reviews: A new feature based heuristic for aspect-level sentiment classification." Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on. IEEE, 2013.

[15] Tirath Prasad Sahu and Sanjeev Ahuja, "Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms", IEEE, 2016.

[16] Sruthi S, Reshma Sheik and Ansamma John, "Reduced Feature Based Sentiment Analysis on Movie Reviews Using Key Terms", IEEE, 2017.