IOSR Journal of Mobile Computing & Application (IOSR-JMCA)
e- ISSN: 2394-0050, P-ISSN: 2394-0042.Volume 7, Issue 3 (May - June 2020), PP 24-29
www.iosrjournals.org

Hindi Text Emotion Recognition based on Deep Learning

Akhilesh Sharma
M. Tech. Scholar Department of CSE
Patel Groups of Institutions (PCST) Bhopal, India

Abstract: *With development of Internet and Natural Language processing, use of regional languages is also grown for communication. In India people express their views by using mother tongue such as Hindi, Bengali, Kannada, Marathi etc. As Hindi is fourth most spoken language in the world therefore many researchers are working on Hindi Sentiment Analysis. 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 as positive, negative or neutral. A speech is combination of variety of topics. So there is requirement for classifying given Hindi speech document in to different classes and then extract sentiments in terms of positive, negative and neutral such as joy, love, surprise, anger, disappointment and worry. In this paper a study is presented on classification of Hindi text documents into multiple classes with the help of machine learning. Further study is given for sentiment analysis which may be carried out by using machine learning to determine the polarity of individual class.*

Keywords: *Hindi Text; Emotions; Feature Extraction; Classification; Emotion recognition;*

Date of Submission: 22-05-2020 Date of Acceptance: 09-06-2020

I. 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.

Emotions may be expressed by a person's speech, face and text. 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. We are classifying a document into multiple classes such as खेती, कृषि, संस्कृति, खुशी etc. Then, we are further extracting sentiments in a form of positive, negative and neutral from respective classes. 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.

II. 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.

Mewada et al. [9] proposed English text sentiment analysis of IMDB movie review. Author examined the sentiment expression to classify the polarity of the movie review on a scale of negative to positive and perform feature extraction and ranking and use these features to train our multilevel classifier to classify the movie review into its correct label.

K. M. Anil Kumar et al [10] proposed a model for retrieving user's sentiments from Kannada Web documents. Machine translation was used to translate the English reviews into Kannada, further POS tagger is used to implement adjective analysis and Turney algorithm which focuses on pair of POS. The polarity is considered as the difference between the positive and negative counts. If the value results more than zero then considered as positive, less than zero then negative else neutral.

Yakshi Sharma et al. [11] discussed and proposed a sentiment analysis using Hindi language based on an unsupervised lexicon method for classification.

Sumitra Pundlik et al. [12] proposed a model for classification of Hindi speech documents into multiple classes with the help of ontology. Further, sentiment analysis is carried out using HindiSentiWordNet (HSWN) to determine the polarity of individual class. To improve accuracy of polarity extraction result resercher have combined HSWN and LMClassifier.

Md Shad Akhtar et al. [17] proposed a novel hybrid deep learning architecture which is highly efficient for sentiment analysis. The selected features are optimized by selecting through a multi-objective optimization (MOO) framework. The optimized sentiment vector obtained at the end is used for the training of SVM for sentiment classification. The result analysis is performed on four Hindi datasets covering varying domains.

III. 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 audio signal using above features. These features are further classified by using classifiers. For each extracted features of emotional speech 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 speech features may be trapped owing to the presence of the state transition matrix. Throughout clustering, a speech signal is taken and therefore the chance for every speech signal 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. 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 speech 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].

IV. Deep Learning Approaches For Hindi Text Emotion Recognition

Deep learning is a branch of machine learning in which programs learn from experience and understand the world in terms of a hierarchy of concepts, where each concept is defined in terms of its relation to simpler concepts. This approach allows a program to learn complicated concepts by building them based on simpler ones.

The most used deep learning model here is long short-term memory (LSTM). LSTM is a special form of recurrent neural network (RNN) with the capability of handling long-term dependencies. LSTM overcomes the vanishing or exploding gradient problem common in RNNs.

Figure 1 outlines the main steps of LSTM for emotion recognition in text. First, text preprocessing is performed on the emotion dataset. The preprocessing steps may include tokenization, stop words removal, and

lemmatization. After that, the embedding layer is built and is fed into one or more LSTM layers. Then, the output is fed into a dense neural network (DNN) with units equal to the number of emotion labels and a sigmoid activation function to perform the classification.

Proposed Methodology

A. Data Collection

The proposed algorithm first of all build a corpus of hindi text [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 hindi.

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. For this BiLSTM with Random forest classifier is used.

One of the type of recurrent neural network (RNN) is Long short-term memory (LSTM). It is type of deep learning neural network approach composed of several neural network modules. The LSTM network is composed of four units: memory cell, input unit, output unit and forget unit.

The memory unit is the unit that stores the data values for some time intervals and remaining three units regulates the flow of data values for evaluation of output value. BiLSTM means bidirectional LSTM, which means the signal propagates backward as well as forward in time. At each time step t there is a set of vectors, including an input gate it, a forget gate f_t, an output gate o_t and a memory cell C_t.

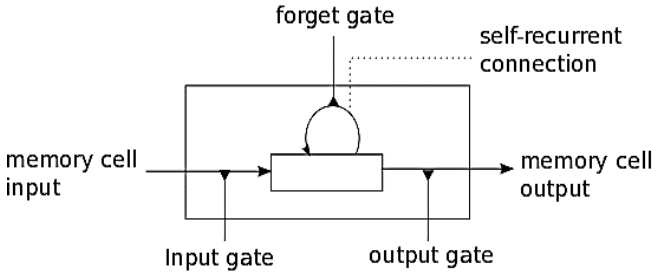

Figure 1: Long Short-Term Memory (LSTM) Units

All these together are used to compute the output of the hidden layer ht as follows:

$$f_t = \sigma(W_f * x_t + U_f * h_{t-1} + b_f) \quad \text{(i)}$$

$$i_t = \sigma(W_i * x_t + U_i * h_{t-1} + b_i) \quad \text{(ii)}$$

$$\hat{C}_t = tanh(W_c * x_t + U_c * h_{t-1} + b_c) \quad \text{(iii)}$$

$$C_t = i_t * \hat{C}_t + f_t * C_{t-1}) \quad \text{(iv)}$$

$$o_t = \sigma(W_o * x_t + U_o * h_{t-1} + b_o) \quad \text{(v)}$$

$$h_t = o_t * \tanh(C_t)) \quad \text{(vi)}$$

In this model, σ is the sigmoid activation function, tanh the hyperbolic tangent activation function, xt the input at time t, Wi , WC, Wf , Wo, Ui , UC , Uf , Uo are weight matrices to regulate the input and bi , bC , bf , bo are bias vectors. Four steps of the LSTM network are discussed below:

Step A

First, the model needs to determine what to throw away from the cell State. This is referred to as the forget gate values ft. The input in this step is the output of the previous step ht−1 and the input xt. A sigmoid activation function is used to give output values between 0 and 1, where 0 corresponds to "let nothing through" and 1 to "remembering everything".

Step B

The next step is to determine what information is going to be added to cell state. In this step again the inputs are h_{t-1} and x_t. The input layer gate it first applies a sigmoid layer over the input to determine which parts of the cell state will be updated. Then a tanh layer is used to create new candidate values $\tilde{C}_t$. In the next step, these two will be combined to update the cell state C_t.

Step C

Now the old cell state is multiplied by ft, to forget the things that are not needed anymore and the new information is added to the cell state memory.

Step D

In the final step, it is determined what the output h_t is. First, a sigmoid layer is applied to the previous output h_{t-1} and input x_t, to determine the output gate values o_t. This is a value between 0 and 1 indicating which parts of the cell state are going to be output. Then the cell state C_t is transformed by a tanh function to get values between -1 and 1.

When BiLSTM is used as a classifier then, at last layer softmax is used as a classifier. But in this research work, softmax layer is replaced with random forest classifier.

Random Forest machine learning algorithm is capable of acting each regression and classification tasks. In random forest technique many decision trees are shaped and algorithm combines the principles of those decision tree and produces an ensembled learning rules for prediction. By using ensemble of those decision trees the model produces the correct and precise results as a result of it's supposed with deep and totally different practiced learnings of many decision tree. As random forest is collection of various weak classifier and combinedly forms robust classifier which may turn out prediction and deep insight into the dataset. For training purpose the algorithm is given with random samples of dataset from that all weak decision trees learn and generates learning rules. any by combining these rules the random forest generates a powerful classifier that's combination of these weak classifiers.

The prediction is made on testing dataset. The algorithm predict the results by applying the learning rules and generates the output in form of class or label. After making different decision trees then voting is performed among them to generate strong learner. This process is termed as "bagging". In growing strong decision tree, exhaustive searches across all possible weak decision trees is conducted to find the possible path in the tree.

Before applying training data there is automatically available holdout data termed as "Out of Bag (OOB)" data. Every decision tree that are generated have different OOB because every tree has a different training sample. Keeping track of for which trees a specific record was OOB allow as to easily and effectively evaluate Random Forest performance.

V. 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 hindi text sample, various classifier algorithms are explained briefly. Hindi 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]. Joshi, B. A. R, and P. Bhattacharyya, "A fall-back strategy for sentiment analysis in Hindi: a case study", International Conference on Natural Language Processing, 2010.

[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.
[9]. Pradeep Mewada, R. R., "Sentiment Analysis of Movie Review using Machine Learning Approach", IJOSTHE, 5(1), 2017. Retrieved from https://ijosthe.com/index.php/ojssports/article/view/83. DOI: https://doi.org/10.24113/ojssports.v5i1.83
[10]. K. M. Anil Kumar, N. Rajasimha, M Reddy, A. Rajanarayana, K. Nadgir, "Analysis of Users' Sentiments from Kannada Web Documents", International Conference on Communication Networks, vol. 54, pp. 247-256, 2015.
[11]. Yakshi Sharma, Veenu Mangat and Mandeep Kaur, "A Practical Approach to Sentiment Analysis of Hindi Tweets", International Conference on Next Generation Computing Technologies, pp.677-680, 2015.
[12]. Sumitra Pundlik, Prasad Dasare, Prachi Kasbekar, Akshay Gawade, Gajanan Gaikwad, Purushottam Pundlik, "Multiclass classification and class based sentiment analysis for Hindi language", International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 512-518, 2016.
[13]. Joonatas Wehrmann; Willian Becker; Henry E. L. Cagnini; Rodrigo C. Barros, "A character-based convolutional neural network for language-agnostic Twitter sentiment analysis", International Joint Conference on Neural Networks (IJCNN), 2017.
[14]. A. M. Abirami; V. Gayathri, "A survey on sentiment analysis methods and approach, International Conference on Advanced Computing (ICoAC), 2017.
[15]. Rincy Jose; Varghese S Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach", International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016.
[16]. P. Kalaivani; K. L. Shunmuganathan, "An improved K-nearest-neighbor algorithm using genetic algorithm for sentiment classification", International Conference on Circuits, Power and Computing Technologies, 2014.
[17]. Md Shad Akhtar, Ayush Kumar, Asif Ekbal, Pushpak Bhattacharyya, "A Hybrid Deep Learning Architecture for Sentiment Analysis", International Conference on Computational Linguistics: Technical Papers, pp. 482–493, 2016.
[18]. H. Kang, S. Yoo, D. Han," Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Systems with Applications, 39 -2012.
[19]. Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu," An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", IEEE transactions on systems, man, and cybernetics, vol.42, no. 3, 2012.
[20]. Farhan Hassan Khan, Saba Bashir, Usman Qamar, "TOM: Twitter opinion mining framework using hybrid classification Scheme", Decision Support Systems, Elsevier, 2013.
[21]. N. D. Valakunde, Dr. M. S. Patwardhan," Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process", IEEE, International Conference on Cloud and Ubiquitous Computing and Emerging Technologies ,2013.
[22]. Rao, Delip, and Deepak Ravichandran. "Semi-supervised polarity lexicon induction.", Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.
[23]. Jain, Amita, and D. K. Lobiyal. "A new method for updating word senses in hindi wordnet." Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on. IEEE, 2014.
[24]. S. Veeramani, S. Karuppusamy, A survey on sentiment analysis technique on web opinion mining, International Journal of Science and Research(IJSR), Aug 2014
[25]. Richa Nigam, Shweta Sharma, Rekha Jain, Opinion Mining in Hindi Language: A Survey, International Journal in Foundation of Computer Science and Technology (IJFCST),Vol.4, March 2014.