7th National Conference on
"Role of Engineers in Nation Building" organized by VIVA Institute of Technology, Mumbai
(1st and 2nd March 2019)
International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

A Survey on Hinglish Code-Mixed Text for Sentiment Analysis

Ketan Chavan[1], Ajay Parmar[2], Vishakha A. Hirve[3], Sunita Naik[4]
Department of Computer Engineering, Mumbai University, India

Abstract— *In multi-lingual country like India, code mixed social media text comprises the majority of the Internet. Hinglish code mixed text is combination of English and Hindi text, in which it used English characters to write Hindi Text. For example, "Movie bohut accha tha", in that text, it used English characters to write Hindi text. Previous research in code mixed text (Hinglish) was done by using Lexicon and Machine Learning approach. This paper presents the survey of main approaches used in Hinglish code mixed text for sentiment classification.*

Keywords— Code-Mixed Text, Deep Learning, Neural Network, Polarity, Sentiment Analysis.

I. INTRODUCTION

With the rapid development of the Internet, more and more people are inclined towards the Internet so that they can express their views. Sentiment analysis helps business in advertising, marketing and making business decisions for better customer satisfaction [11].

The main focus of sentiment analysis is to decide the opinion of a writer or speaker with respect to some topic or the overall contextual polarity or emotional reaction or the behavior to a particular event [13]. While many of them have come across different tasks conducted on multi-lingual texts, the task of sentiment analysis, in particular, has not been explored for multilingual code-mixed texts. Social media texts on the other hand are largely informal [11].

The Indian pages on the websites such as Facebook, Twitter often contains Hindi-English mixed comments, and mostly the users write these comments in Roman script due to difficulty in writing Devanagari. Marketers can use this to know public opinion of their company and products, or to analyze customer satisfaction. One can know what's right and what went wrong in other words positive and negative aspects of their service/product.

In a multi-lingual nation like India, people review on social media in a language which is mixture of two languages hence there is a need of classifier which will classify code mixed text. Former approaches consists of Lexicon method and Machine learning method. As both the approaches cannot handle code mixed text (Hindi-English) efficiently. Due to which, it needed better deep learning classifier for better results in terms of accuracy and performance.

In the proposed system, deep neural networks are used for the sentiment analysis of code-mixed text. This neural networks are CNN (Convolutional Neural Network) and LSTM (Long Short Term Memory). CNN is used for better feature extraction and LSTM is used for its sequential results. In the, Fully Connected layer (FC) gives the output in the form of positive or negative sentiment of the text.

II. LITERATURE SURVEY

2.1 Text normalization of code mix and sentiment analysis [1].

In this paper, tokenization of the dataset takes place and then they are fed into English language Identifier. Tokens belonging to English language are tagged as E at the end of each token and the tokens belonging to Hindi language which is written in the Roman script are tagged as H at last.

The system mainly focuses on the lexical based approach for sentiment classification in which count of positive, negative and neutral words gives the sentiment analysis of senetences.

2.2 Sentiment analysis of mixed language employing Hindi-English code switching [2].

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

In this paper the system uses neural network for the classification purpose which deals with two languages namely English and Hindi language. Recursive Neural Tensor Network (RNTN) is used for sentiment classification.

The system mainly focuses on sentiment analysis at both phrases and sub phrases level of the sentences.

2.3 Deep Sentiment Representation Based on CNN and LSTM [3].

In this paper the first step is the tokenization and then word2vec tool is provided by Google. All these vectors are fed to CNN (Convolution Neural Network) with help of word vector embedding. Output of CNN is given to LSTM (Long Short-Term Memory) layers.

The system has introduced a novel model for deep sentiment representation by skillfully combining one-layer CNN and two-layer LSTMs.

2.4 Opinion Mining on Hindi-English Code-Mixed Data [4].

Code-Mixed data has been collected in this paper. These data collected is from Narendra Modi, Aam Aadmi Party, Garbage Bin, BBC Hindi, ABP News, Doordarshan, Dainik Jaagran and Dainik Bhaskar pages where comments are in code mixed language. N-gram method is used for dividing data in pairs of words. Senti score is assigned to all these pairs by using SentiWordNet. At final stage, libSVM is used for classification purpose.

The method mainly discussed the problem of determining sentiment in Hindi-English code-mixed data. Unigrams and bigrams were baseline and added some features for handling the sentiments in a more accurate way.

2.5 Sentiment Analysis on Arabic Text using CNN and LSTM [5].

This paper uses character level, character n-gram level and word level for preprocessing of text which converts sentences into tokens. After that the combine method Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM) network is used in the system. In preprocessing of texts, Lexicon approach is merged with CNN for better feature extraction. LSTM layer is added in CNN layer because it has ability to capture sequential data by considering previous data. At last, FC layer gives output in the terms of positive or negative sentiment of the text.

System gives better results in terms of accuracy and performance. LSTM network is the main advantage in the system that's why it gives better results.

2.6 Multi-lingual Sentiment Analysis: AN RNN-Based Framework for Limited Data [8].

In this paper the system uses Recurrent Neural Network (RNN) model that takes text and pre-trained word embedding as inputs and generates a classification result. Different languages are translating into single language i.e. English text. The numeric of word embedding is input for RNN network.

System uses Recurrent Neural Network (RNN) as a classifier to classify the corpora of the text. RNN has disadvantages of remembering sequential data because it doesn't consider previous data.

2.7 A Hybrid Method for Bilingual Text Sentiment Classification Based on Deep Learning [6].

In this paper, the hybrid method based on deep learning is introduce for understanding the meaning of text. It uses semi-supervised learning for training the data. The hybrid method includes the following methods: RNN which is composed by LSTM, Naïve Bayes (NB)-Support Vector Machines (SVM), word2vec and bag-of-words.

Hybrid method is better than general method for classification purpose. Hybrid methods give better performance and accuracy in terms of polarity.

2.8 Deep CNN-LSTM with Combined Kernels from Multiple Branches for IMDb Review Sentiment Analysis [7].

The ACL Internet Movie Database (IMDb) dataset is used for learning word vectors. The proposed model is used with combined kernels from multi-branch CNN with LSTM model.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

System uses combined kernels which give benefit for feature extraction. This method gives better results and surpasses the baseline CNN+LSTM model.

2.9 Domain Specific Sentiment Analysis Approaches for Code Mixed Social Network Data [9].

In this paper there are two different approaches. The first approach is the Lexicon based approach and the second approach is the Machine Learning approach. Lexicon approach uses dictionaries of words which are annotated with the polarity which helps to classify the sentiment of the text. The polarity detection algorithm is applied to determine the overall sentiment of the sentence. Pre-processing in Machine Learning includes removing or avoiding misspellings or slang words.

The proposed system uses two approaches in which Lexicon approach gives the better accuracy than that of Machine Learning approach.

2.10 An Ensemble Model for Sentiment Analysis of Hindi English Code Mixed Data [10].

In this paper combination of character Trigram based LSTM method and word N-gram based multi nominal Naïve Bayes is used. The LSTM neural network is used to recognize deep sequential patterns in the text and MNB method captures low level word merging of keywords.

In the proposed system, the combine approach of LSTM and Multi Naïve Bayes gives better accuracy.

2.11 Sentiment Identification in Code Mixed Social Media Text [11].

In this paper the process is divided into three parts: a) Pre-processing b) Feature identification and extraction c) Classification of sentiment. Pre-processing includes removal of punctuations, removal of multiple character repetitions. Features extraction includes no. of words matches with sentiwordnet, English-Hindi sentiment words, lastly, they have classified the sentence into sentiment of text.

System uses Multi-Layer perception model which removes noises and processes the text to normalize the irregular words. The method gives better result.

2.12 Multi-Channel LSTM-CNN Model for Vietnamese Sentiment Analysis [12].

In this paper CNN and LSTM neural networks are combined to construct two information channels which are expected to enhance classification performance of the Soft-max layer.

System generates word embedding represented by an embedding layer. Output is given to LSTM model for generating LSTM feature vector and then is fed to CNN model for generating CNN feature.

III. ANALYSIS TABLE

The Analysis table is the summary of the studied research papers and the different techniques used on Sentiment analysis of code-mixed text.

TABLE 1

Sr. No.	Title of Paper	Technique/ Methods	Accuracy
1.	Sentiment analysis of mixed language employing Hindi-English code switching [2].	Lexical, RNTN (Recursive Tensor Neural Network)	80%

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

2	Text Normalization of code mixed and sentiment analysis [1]	Tokenization, transliteration, SentiWordNet	83%
3	Opinion mining in Hindi-English code- mixed data [4]	N-gram, SentiWordNet,	60.73%
4	Deep sentiment representation based on CNN and LSTM [3].	CNN and LSTM word2vec.	87.2 %
5	Sentiment analysis on Arabic text using CNN and LSTM [5].	N-gram, CNN and LSTM.	88%
6	A Hybrid method for Bilingual text Sentiment Classification Based on Deep learning [6].	RNN with LSTM, Naive Bayes with SVM, word2vec and bag of words.	80%
7	Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis [7].	Multiple branches of CNN based LSTM.	89%
8	Multilingual Sentiment analysis: An RNN based Framework for limited data [8].	Word embedding and RNN.	75%
9	Domain Specific Sentiment Analysis Approach for Code Mixed Social Network Data [9].	Lexicon approach, Machine Learning.	86%.
10	An Ensemble Model for Sentiment Analysis for Hindi English Code- Mixed data [10].	LSTM model with Multi Naive Bayes Model.	70.8%
11	Sentiment Identification in Code-Mixed Social Media Text [11].	Multilayer perception model is used.	68.5%.
12.	Multi-Channel LSTM-CNN Model for Vietnamese Sentiment Analysis [12].	CNN and LSTM are used.	Using Multi-Channel LSTM gives the accuracy.

IV. CONCLUSION

Previous approach for Hinglish text was done by machine learning and lexicon method. Machine Learning gives less accuracy for classification of text. Hindi-English text for classification is not done using deep learning method. Two combine neural networks Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) gives us the better accuracy and performance as compared to former approaches like RNN in Hindi-English text. This paper concludes that sentiment classification for Hinglish text is still have better research and approaches in future. Therefore, more research is required in Hinglish text for sentiment analysis.

REFERENCES

[1] Shashank Sharma, PYKL Srinivas, Rakesh Chandra Balabantaray," Text Normalization of Code-Mixed and Sentiment Analysis", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015.

[2] Prof.Dinkar Sitaram, Ms.Savita Murthy, Debrah Ray, Devansh Sharma, Kashyap Dhar, "Sentiment analysis of mixed language employing Hindi-English code switching", Proceeding of the 2015 International conference on Machine Learning and Cybernetics, Guangzohu, 12-15 July 2015.

[3] Qiongxia Huang, Xianghan Zheng, Riqing Chen, Zhenxin Dong," Deep Sentiment Representation Based on CNN LSTM", International Conference on Green Informatics, 2017.

[4] Vrishank,Gagan,Lokesh," Opinion Mining on Hindi-English Code Mixed Data ", arXiv, March 2016.

[5] Abdulaziz M. Alayba, Vasile Palade, Matthew England and Rahat Iqbal, "A Combined CNN and LSTM Model for Arabic Sentiment Analysis", arXiv, July 2018.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

[6] Guolong Liu, Xiaofei Xu, Bailong Deng, Siding Chen, Li Li , "A Hybrid Method for Bilingual Text Sentiment Classification Based on Deep learning", IEEE SNPD 2016, May 30-June 1, 2016, Shanghai, China.

[7] Alec Yenter, Abhishek Verma, "Deep CNN-LSTM with Combined Kernels from Multiple Branches for IMDb Review Sentiment Analysis", IEEE, 2017.

[8] Ethem F. Can, Aysu Ezen-Can, Fazli Can, "Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data", arXiv June 2018.

[9] Pravalika A, Vishvesh Oza, Meghana N P and Sowmya Kamath S, "Domain Specific Sentiment Analysis Approaches for Code-mixed Social Network Data", 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6.

[10] Madan Gopal Jhanwar, Arpita Das, "An Ensemble Model for Sentiment Analysis for Hindi English Code- Mixed data", arXiv, June 2018.

[11] Ghosh, Souvick & Ghosh, Satanu & Das, Dipankar. (2017). "Sentiment Identification in Code-Mixed Social Media Text" arXiv, July 2017.

[12] Vo, Quan-Hoang & Nguyen, Huy-Tien & Le, Bac & Nguyen, Minh-Le, "Multi-channel LSTM-CNN model for Vietnamese sentiment analysis", 9th International Conference on Knowledge and Systems Engineering (KSE), 2017.