Bi-LSTM and Ensemble based Bilingual Sentiment Analysis for a Code-mixed Hindi-English Social Media Text

1st Konark Yadav
Dept. of Electronics and Communication Engineering,
The LNM Institute of Information Technology, Jaipur, India
k.yadav2704@gmail.com

2nd Aashish Lamba
Dept. of Communication and Computer Engineering,
The LNM Institute of Information Technology, Jaipur, India
17ucc002@lnmiit.ac.in

3rd Dhruv Gupta
Dept. of Computer Science and Engineering,
The LNM Institute of Information Technology, Jaipur, India
17ucs052@lnmiit.ac.in

4th Ansh Gupta
Dept. of Electronics and Communication Engineering,
The LNM Institute of Information Technology, Jaipur, India
anshgupta.y17@lnmiit.ac.in

5th Purnendu Karmakar
Dept. of Electronics and Communication Engineering,
The LNM Institute of Information Technology, Jaipur, India
ORCID- 0000-0001-8663-2225

6th Sandeep Saini
Dept. of Electronics and Communication Engineering,
The LNM Institute of Information Technology, Jaipur, India
ORCID- 0000-0002-8906-8639

2020 IEEE 17th India Council International Conference (INDICON) | 978-1-7281-6916-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/INDICON49873.2020.9342241

***Abstract*—India is a multilingual and multi-script country and a large part of its population speaks more than one language. It has been noted that such multilingual speakers switch between languages while communicating informally. The code-mixed language is very common in informal communication and social media, and extracting sentiments from these code-mixed sentences is a challenging task. In this work, we have worked on sentiment classification for one of the most common code-mixed language pairs in India i.e. Hindi-English. The conventional sentiment analysis techniques designed for a single language don't provide satisfactory results for such texts. We have proposed two approaches for better sentiment classification. We have proposed an Ensembling based approach which is based on hybridization of Naive Bayes, SVM, Linear Regression, and SGD classifiers. We have also developed a bidirectional LSTM based novel approach. The approaches provide quite satisfactory results for the code-mixed Hindi-English text.**

***Index Terms*—Sentiment Analysis, Social Media text, LSTM, Code-mixed text, Hindi-English, Bilingual Sentiment Analysis**

I. Introduction

The world is converging closer and closer daily with the latest emerging technologies. Social media has played a vital role in this process. With the mixing of cultures, languages are also getting mixed in different forms of communication. India has always been a multilingual country. As per the Census 2011, 255 million people in India have good proficiency over 2 languages, and 87 million speak 3 languages. This population can communicate freely in all the known languages and infrequently mix two languages while expressing their thoughts. Lots of words are more commonly utilized in one language and their translation within the other language is not very fashionable and thus whenever the speaker is using those words in any context, they like the more popular language regardless of this language is spoken. This generates a sentence which has words from two languages arranged the grammatical structure of one language.

978-1-7281-6916-3/20/$31.00 © 2020 IEEE

Code-mixing is described as the mixing of words or phrases of two languages to make a single sentence. In this work, we have focused on the most common code-mixed language pair in India, i.e. Hindi-English. Code-mixed communication is popular on social media and other similar informal modes of communication. There are various reasons for the code-mixing of two languages.

1) Words are borrowed from other the language when there was no adequate translation, when they were not sure of a word, and when the word was hard to pronounce [1].
2) In the early stages of life parents use other language words when teaching new words to their children. Thus, bilingual parents might use language mixing as a strategy to make sure their children learn words equally in both languages.
3) Some words are more catchy/fashionable in one language and often mixed in other languages.

For example, the word "routine" is translated as "dincharya" in Hindi, but it's not a very common word in informal communication. So a person even while speaking a complete Hindi sentence tends to replace "Dincharya" with "routine" in their communications. Thus, the sentence is spoken as "Arey yaar tera "routine" bahut sahi hai. Thus, the grammatical structure of the whole sentence is based on Hindi grammar while one word has been very conveniently replaced and fitted without changing the semantics. There are a lot of such words that are more difficult to pronounce in Hindi, but their translation in English are simple words. Thus, the English words are more commonly used. This convenience has led to the mixing of more and more popular words into the sentences of other languages. Not only single words, but also phrases are fitted into the sentences of other languages. For example, in the sentence, " Oh yaar baat sun, *just let it go*", the phrase " just let it go" is having the grammar of English while the initial half of the sentence is following the Hindi grammar structure.

We can observe that the problems associated with a code-mixed sentence are much more complex as compared to a single language sentence. We have to not only deal with the tokens of two languages but also the grammar structures of two different languages. These challenges motivate us to explore to develop language models for bilingual sentiment analysis. Hindi is the most popular language in the North Indian states of India. And, a lot of Indians outside the subcontinent use Hindi. Being in English speaking cultures, the code-mixing is also more popular among Indians outside India. English is the language of education in a majority of Indian schools [2] and the native languages are used to teach English in these schools. Thus, mixing English with native languages becomes a part of informal communication at different stages. Comprehending only one language might be possible with one model [3], [4] but these models do not perform well on the bilingual text. In such cases, we have to expore the partial or complete machine trnslation between two languages [5].

Social media is emerging as a very powerful platform to exchange information in recent years. These platforms provide an open and widely accessible platform for anyone to express their views. Social media platforms are also very informal tools for communication. Since there are no moderations (in most of the platforms), people who interact with a lot of content on such platforms often code-mix. A good number of international brands as well as code mix in their promotions to attract the local people. For example, Amazon India launched its campaigns with the slogan " Dhamaka Deals for you" and Dominos having a punchline "hungry kya?". These advertisements are often translated to other Indian languages as well. Analyzing sentiments of such content helps detect the actual context of the person. There are two main challenges in a code-mixed dataset.

1) **Words having multiple spellings:** In informal communication, spelling and grammar structures are often neglected by the users. A user can use the spellings "goood", "gud", "goooood" and many such variations for a single word. Wrong spellings are not only in excitement of expressing their sentiments but also because of the way different people pronounce and write a particular word. The word referring to "I" can be written as "main", "main", "mein", or some other variation as well. For human readers, all these variations are comprehensible while machine struggles with such variations.
2) **Equivocal words:** In their chosen language a few words with almost the same pronunciation and transliteration have entirely different meanings. Yet it is very difficult to recognize the meaning of those words in a code-mixed text and to derive the correct opinion from those words. For instance, in Hindi the translation for "I" is "main',' however the same word is already an English word.

In this work, we have focused on the sentiment analysis of the social media contents of Hindi-English code-mixed sentences. We have prepared our dataset for the same which is considerably larger than the existing datasets for the same problem. We have proposed a bidirectional LSTM based architecture to analyze the sentiment and compared our results with state-of-the-art systems. In section II, we have described the conventional and state-of-the-art tools and techniques in the area of bilingual sentiment analysis. We have listed the existing work related to Hindi-English code-mixed sentiment analysis in the same section. In section III, we have explained our system for bilingual sentiment analysis. We have explained the different techniques used by us for obtaining the proposed model in the same section. In section IV, we have focused on our experimental setup and results. We have discussed the results in the same section and why our proposed system works better than the compared methods. In the last section, i.e. section V, we have concluded our work and give an insight into the future work in this field.

II. Literature Review

Sentiment analysis is a well-researched and established area of interest in the field of Natural Language Processing. Bilingual sentiment analysis a developing area of interest in recent years. Bilingual/Code-mixed sentiment analysis for Indian languages is growing an exciting area of interest. In 2014, Bali et.al. reported the first work on Hindi-English Code-mixed sentences on Facebook comments [6]. In 2016, Bhargava et.al [7] attempted to develop a system for mining sentiments from code mixed sentences for English with a combination of four other Indian languages (Tamil, Telugu, Hindi, and Bengali). The approach was based on the Language Identification and Sentiment Mining Approach. Ghaleb et.al. focused on the existing systems for Indian languages Sentiment Analysis [8]. Manish Shrivastava et. al. at the Language Technologies Research Centre, International Institute of Information Technology, Hyderabad have worked on the development of corpus for various applications of code-mixed texts for Hindi-English pair. Their tweeter dataset consists of 5250 English-Hindi code-mixed tweets [9]. The dataset is used for sarcasm detection in the reported work. A similar dataset is used for humor detection in English-Hindi Code-mixed Social Media Content [10] and Gender Prediction in English-Hindi Code-Mixed Social Media Content [11]. The code-mixed dataset can also be used for Curriculum Learning Strategies for Hindi-English code-mixed teaching [12].

In 2017, International Conference on Natural Language Processing (ICON-2017) started an online contest on Sentiment Analysis for Indian Languages (SAIL) [1]. The main task was to identify the sentence level sentiment polarity of the code-mixed dataset of Indian language pairs (Hi-En, Ben-Hi-En) collected from Twitter, Facebook, and WhatsApp. In this contest, 6 teams from all over India participated and developed bilingual sentiment analysis systems using different approaches [13]. The best performer was the IIIT-NBP team. They used features like GloVe word embeddings with 300 dimensions and TF-IDF scores of word n-grams. Team JU_KS [14] used n-gram and sentiment lexicon-based features. BIT Mesra team used SVM and Naive Bayes classifiers on unigram and bi-gram features to classify the sentiment of the code-mixed HI-EN dataset. NLP_CEN_AMRITA team had used different distributional and distributed representations. Bilingual sentiment analysis is evolving for other language pairs around the globe as well. Tao et. al. proposed a CNN based classifier for the task [15]. Based on the results obtained by these systems, we have worked on two aspects.

- Developing a large code-mixed Hindi-English dataset from social media contents.
- Proposed a Bidirectional-LSTM based sentiment analysis system for better accuracy.

III. Proposed Model

In developing a bilingual sentiment analysis systems, the foremost requirement is the availability of a suitable dataset. The public datasets available for the task were not large enough for proper training. We have developed our dataset for the purpose and explained the details in section IV. Once the dataset is available, then we have developed the bilingual sentiment analysis system with Baseline models and by observing the limitations of these models, we have proposed our approach for the task. We have considered the following baseline models for our study.

1) **Multinomial Naive Bayes:** is a well established text classifier and applied frequently for sentiment analysis as well [16], [17]. The multinomial model uses vectors of integer features (word counts) to represent documents.
2) **Support Vector Classifier:** is based on a supervised machine learning algorithm that can be used for both classification or regression challenges. Classification is performed by finding the hyper-plane that differentiates the multiple classes very well [18], [19].
3) **Stochastic Gradient Descent (SGD):** classifier works on the principle of finding the lowest point of a function. SGD starts from a random point on a function and travels down its slope in steps until it reaches the lowest point of that function. SGD is used for sentiment analysis in various studies [20].
4) **Extreme Gradient Boosting (XGB):** produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. XGB is used for Indian language sentiment analysis as well [21] and [22].
5) **Long and Short Term Memory (LSTM):** is a well-established model for text classification and prediction. It contains both keep and forget gates to learn from the past findings as well. LSTM based classification are used for code-mixed sentiment analysis [23], [24].

We have classified our dataset using the five baseline models and the obtained results are tabulated in table II. Baseline models work with inconstant accuracy for different types of sentiments. For example, the Multinomial Naive Bayes model works more efficiently on Positive and Negative sentiment classifications (F score 0.8 and 0.77) but fails to classify the Neutral sentiments (F score 0.23). The XGB classifier also performs with a similar trend. The LSTM and SGD based classifiers are more consistent with every type of sentiments, but provide a lesser average accuracy for the overall process.

A. *Ensemble model*

Considering the limitations and strengths of each such classifiers, we have chosen 4 best performing classifiers and Ensemble them. The proposed model is depicted in Figure 1.

The code-mixed data is pre-processed and tokenized. After the tokenization, the dataset contains multiple occurrences of a lot of words. Multiple words occurrences in the text are handled with by converting them into TF-IDF values. TF is Term Frequency and defined as the number of times a word appears in a document divided by the total number of words in the document.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{1}$$

Inverse Data Frequency (IDF) is definied as the log of the number of documents divided by the number of documents that contain the word w.

$$idf(w) = log\frac{N}{df_t} \tag{2}$$

And, TF-IDF is defnied as the product of TF and IDF values.

$$w_{i,j} = tf_{i,j} * log\frac{N}{df_t} \tag{3}$$

TF-IDF values are fed to 4 classifiers in parallel and the values of Precision (P), Recall (R), F score (F), and

[1] http://www.dasdipankar.com/SAILCodeMixed.html

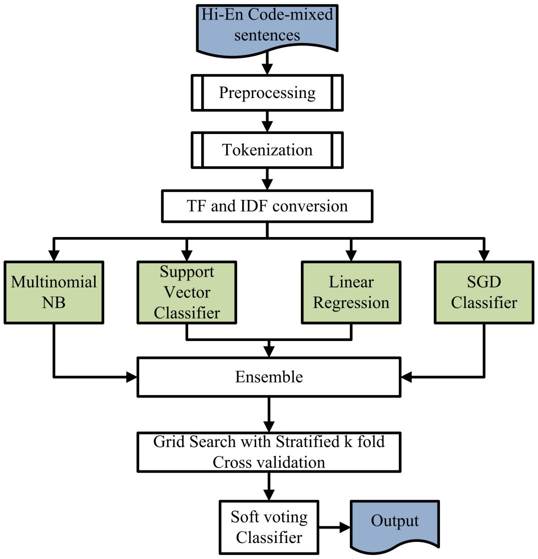

Fig. 1. Proposed model with Ensemble based approach.

average accuracies are obtained. These values are ensembled together and passed through a Grid search with Stratified 3 fold cross-validation. In the end, we have used a soft voting classifier to generate the final decision.

B. Bi-LSTM model

Secondly, we have worked on a model that can provide consistent classifications for each type of sentiment. As depicted in Table II, LSTM based classifier is the most consistent one. We have developed a Bidirectional Long and Short Term Memory (Bi-LSTM) based sentiment classifier. Huang et.al. [25] had proposed Bi-LSTM Models for Sequence Tagging. Bi-LSTM is an extension of traditional LSTMs which improves the performance of the model in terms of sequence classification. For a code-mixed data, our focus is on sequence classification and thus, the Bi-LSTM model is chosen. The system designed with the Bi-LSTM approach is shown in Figure 2. The Bi-LSTM structure is shown inside the block in Figure 2.

In the Bi-LSTM model, we have trained two sequences. The first input sequence and the second is a reversed copy of the input pattern. This can continue giving the network additional context and lead to quicker and even more detailed learning on the actual problem. In the given block X_0, X_1, … X_i are the word embeddings. The forward sequence is trained from S_0 to S_i. The reverse sequence is $S_0^{'}$ to $S_i^{'}$. This structure allows the networks to have both backward and forward information about the sequence at every time step. The output sequence Y_0, Y_1, … Y_i is given to the classifier for further information extraction.

IV. Results and Discussions

The proposed models are trained on a suitable dataset. Thus, the first task was the dataset generation for satisfactory results to be obtained.

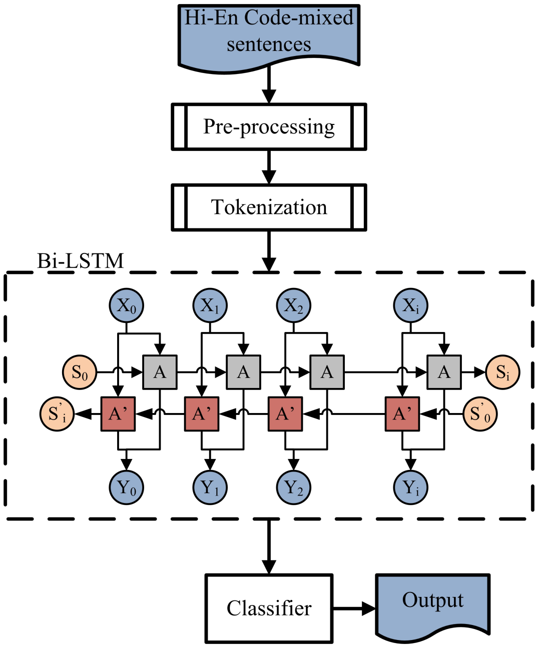

Fig. 2. Proposed model with Bidirectional LSTM based approach.

A. Dataset preparation

Dataset is the foremost requirement for bilingual sentiment analysis. The code-mixed dataset for Hindi-English text is not very commonly available and also the available datasets had a limited size. We referred to IIIT Code-mixed dataset [2] and found that this had around 5000 sentences. Although the number of sentences is satisfactory there are a very large number of sentences which are in a single language. Thus, we decided to prepare our dataset with better code-mixing. We have worked on code-mixed sentence extractions from social media comments and transcripts of informal celebrity interviews. We have collected a total of over 10,000 sentences and reduced them to 6,357 so that a better code-mixing is achieved. The prepared dataset details are depicted in Table I. The dataset is made public and available at GitHub repository[3] under open source license.

TABLE I
Details for the code-mixed Hindi-English dataset generated from social media text.

Data	Total Sentences	Positive Sentences	Neutral Sentences	Negative Sentences
Total Dataset	6357	1825	2895	1637
Train Dataset	5403	1551	2460	1392
Test Dataset	954	274	435	245

[2] https://github.com/drimpossible/SubwordLSTM/blob/master/Data/
[3] https://github.com/gptansh/Bilingual_Sentiment_Analysis_HI-EN

B. Pre-processing

Social media text is best suited for code-mixed analysis but it is also a challenging task to clean this data. The highly informal communication includes short forms, wrong spellings, multiple valid transliterations, and wrong grammatical structures. We have taken the following steps in the pre-processing process.

- All the text is converted to the small letter and capital letters are converted to small letters.
- We have removed special characters and numerical data from the sentences.
- We have created a list of stop words which contained unwanted words along with the proper nouns and some common nouns.
- We have created our own lemmatization corpus which helped us in generalising multiple words with the same meaning
- We have used Text blob to obtain words that decide the polarity of our sentences.

C. Results

As we have discussed in section III, we have considered 5 baseline models. Initially, we have tested all these baseline models with both of our proposed models. All the models are trained and tested on the dataset mentioned in Table I. We have measured the Precision (P), Recall (R), F score (F), and the best accuracy for each model on Positive, Negative, and Neutral sentiments. The generates scores are tabulated in Table II. The proposed Ensemble-based model provides the best average accuracy among all. The proposed Bi-LSTM based model is the most consistent in providing a similar P, R, and F scores for each type of sentiments.

Next, we have compared our results with the state-of-the-art systems for Hindi-English code-mixed sentiment analysis. We have considered the 3 best contestants from the SAIL competition [13] of ICON 2017 along with more recently reported systems. The comparison results of our proposed systems with these reported systems are shown in Table III.

We observe that the Ensemble-based model performs well for each type of sentiment classification. It also provides the best accuracy. It is also observed that the system outperforms the baseline models as well as the state-of-the-art systems for Hindi-English sentiment analysis. The second model is designed for a consistent classification. We observed that Bi-LSTM based system was performing better for sentences with longer length due to its ability to capture sequential information. The overall accuracy is lower because most of the sentences in our datasets are having less than ten words. The model can be used for text classification of longer sentences.

V. Conclusions

Bilingual sentiment analysis for a code-mixed text is a challenging task and the conventional systems designed for single language sentiment analysis have their limitations for such texts. In this work, we have proposed an ensemble-based approach for better average accuracy of sentiment classification. The proposed model takes the decision based on the best results from 4 different classifiers. The proposed model outperforms the state-of-the-art systems for Hindi-English code-mixed sentiment analysis. The model will be tested further on different popular code-mixed pairs in India. And, we also look forward to designing for three language code-mixed text. Hindi-English-Punjabi and Hindi-English-Bengali is are two such combinations which will be explored. The second model is designed for a consistent analysis of each category of sentiments. We can achieve precision, recall, and F score of 059 using this model while delivering an average accuracy of 0.73. So, this system is on par with the earlier one in terms of accuracy. A hybrid model will be explored to provide the highest values for each evaluation metric.

References

[1] Elizabeth Lanza. *Language mixing in infant bilingualism: A sociolinguistic perspective*. Oxford University Press on Demand, 2004.

[2] Jon Allan Reyhner. *Teaching the Indian child: A bilingual/multicultural approach*. Eastern Montana College, School of Education, 1986.

[3] Sandeep Saini and Vineet Sahula. Cognitive architecture for natural language comprehension. *Cognitive Computation and Systems*, 2(1):23–31, 2020.

[4] Sandeep Saini, Nitin Gupta, Shivin Bhogal, Shubham Sharma, and Vineet Sahula. Bayesian learner based language learnability analysis of hindi. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2089–2093. IEEE, 2016.

[5] Sandeep Saini and Vineet Sahula. Neural machine translation for english to hindi. In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 1–6. IEEE, 2018.

[6] Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, 2014.

[7] R. Bhargava, Y. Sharma, and S. Sharma. Sentiment Analysis for Mixed Script Indic Sentences. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 524–529, 2016.

[8] Osamah Ali Mohammed Ghaleb and Anna Saro Vijendran. Survey and Analysis of Recent Sentiment Analysis Schemes Relating to Social Media. *Indian Journal of Science and Technology*, 9(41):1–16, 2016.

[9] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A corpus of English-Hindi Code-mixed Tweets for Sarcasm Detection. *arXiv preprint arXiv:1805.11869*, 2018.

[10] Ankush Khandelwal, Sahil Swami, Syed S Akhtar, and Manish Shrivastava. Humor Detection in English-Hindi Code-mixed Social Media Content: Corpus and Baseline System. *arXiv preprint arXiv:1806.05513*, 2018.

[11] Ankush Khandelwal, Sahil Swami, Syed Sarfaraz Akhtar, and Manish Shrivastava. Gender Prediction in English-Hindi Code-Mixed Social Media Content: Corpus and Baseline System. *arXiv preprint arXiv:1806.05600*, 2018.

[12] Anirudh Dahiya, Neeraj Battan, Manish Shrivastava, and Dipti Mishra Sharma. Curriculum Learning Strategies for Hindi-English Code-mixed Sentiment Analysis. *arXiv preprint arXiv:1906.07382*, 2019.

[13] Braja Gopal Patra, Dipankar Das, and Amitava Das. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*, 2018.

TABLE II
AVERAGE ACCURACY, PRECISION (P), RECALL (R) AND F SCORES (F) OBTAINED FOR POSITIVE, NEGATIVE AND NEUTRAL CODE-MIXED SENTENCES USING BASELINE MODELS AND PROPOSED APPROACHES

Algorithm	**Best Accuracy**	**Positive**			**Negative**			**Neutral**		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Multinomial Naive Bayes	0.71	0.78	0.82	0.80	0.68	0.89	0.77	0.60	0.23	0.33
Support Vector Classifier (SVC)	0.72	0.71	0.64	0.67	0.71	0.69	0.70	0.50	0.60	0.55
Stochastic Gradient Descent(SGD) Classifier	0.72	0.71	0.66	0.69	0.72	0.72	0.72	0.51	0.57	0.54
Xtreme Gradient Boosting(XGB) Classifier	0.59	0.71	0.79	0.75	0.71	0.77	0.74	0.50	0.32	0.39
LSTM	0.59	0.60	0.57	0.59	0.70	0.61	0.65	0.45	0.57	0.50
Ensemble_Classifier (Proposed)	0.74	0.81	0.79	0.80	0.73	0.87	0.79	0.58	0.41	0.48
BI-LSTM (Proposed)	0.73	0.62	0.58	0.60	0.68	0.62	0.65	0.49	0.58	0.53

TABLE III
COMPARISON OF THE PROPOSED SYSTEM WITH STATE-OF-THE-ART SYSTEMS FOR CODE-MIXED HINDI-ENGLISH SENTIMENT ANALYSIS

Authors/Team	**Architecture**	**Year**	**Accuracy**	**Precision**	**Recall**	**F Score**
IIIT-NBP [13]	GloVe Words Embeddings	2017	0.61	0.59	0.56	0.57
BIT-Mesra [13]	SVM and MNB	2017	0.60	0.57	0.55	0.56
JU-KS [14]	MNB	2017	0.60	0.57	0.55	0.56
Jhanwar et. al. [26]	Ensemble	2018	0.70	**0.71**	0.61	0.66
Kumar et al., [27]	Vowel Consonent based	2018	0.62	0.65	0.52	0.58
Joshi et al., [23]	Sub-word composition	2016	0.69	0.68	0.62	0.65
Proposed model -1	Enseble	2020	**0.74**	0.70	**0.69**	**0.69**
Proposed model -2	Bi-LSTM	2020	0.73	**0.59**	**0.59**	**0.59**

[14] Kamal Sarkar. Ju_ks@ sail_codemixed-2017: Sentiment analysis for indian code mixed social media texts. *arXiv preprint arXiv:1802.05737*, 2018.

[15] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. Improving Sentiment Analysis via Sentence type Classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, 72:221 – 230, 2017.

[16] Jiang Su, Jelber S Shirab, and Stan Matwin. Large scale text classification using semi-supervised multinomial naive bayes. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 97–104. Citeseer, 2011.

[17] Kamal Sarkar and Saikat Chakraborty. A sentiment analysis system for indian language tweets. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 694–702. Springer, 2015.

[18] Nurulhuda Zainuddin and Ali Selamat. Sentiment analysis using support vector machine. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pages 333–337. IEEE, 2014.

[19] Parul Sharma and Teng-Sheng Moh. Prediction of indian election using sentiment analysis on hindi twitter. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1966–1971. IEEE, 2016.

[20] P Impana and Jagadish S Kallimani. Cross-lingual sentiment analysis for indian regional languages. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 1–6. IEEE, 2017.

[24] Abdulaziz M Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. A combined cnn and lstm model for arabic sentiment analysis. In *International cross-domain conference for machine learning and knowledge extraction*, pages 179–191. Springer, 2018.

[21] Shukrity Si, Anisha Datta, Somnath Banerjee, and Sudip Kumar Naskar. Aggression detection on multilingual social media text. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2019.

[22] Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, 2018.

[23] Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. Towards Sub-word Level Compositions for Sentiment Analysis of Hindi-English Code-mixed Text. *arXiv preprint arXiv:1611.00472*, 2016.

[25] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging, 2015.

[26] Madan Gopal Jhanwar and Arpita Das. An ensemble model for sentiment analysis of hindi-english code-mixed data. *arXiv preprint arXiv:1806.04450*, 2018.

[27] Upendra Kumar, Vishal Singh, Chris Andrew, Santhoshini Reddy, and Amitava Das. Consonant-vowel sequences as subword units for code-mixed languages. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.