# Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus: A Comprehensive Review

Gazi Imtiyaz Ahmad[1]
School of Computer Applications, Lovely Professional University, Punjab, India

Jimmy Singla[2]
School of Computer Science and Engineering, Lovely Professional University, Punjab, India

Anis Ali[3]
Department of Management
College of Business Administration
Prince Sattam Bin Abdulaziz University
Al-Kharj 11942, Saudi Arabia

Aijaz Ahmad Reshi[4]*
Department of Computer Science
College of Computer Science and Engineering
Taibah University
Al-Madinah Al-Munawarah, Saudi Arabia

Anas A. Salameh[5]
Department of Management Information Systems
College of Business Administration
Prince Sattam Bin Abdulaziz University
Al-Kharj 11942
Saudi Arabia

***Abstract*—A comprehensive review of sentiment analysis for code-mixed and switched text corpus of Indian social media using machine learning (ML) approaches, based on recent research studies has been presented in this paper. Code-mixing and switching are linguistic behavior shown by the bilingual/multilingual population, primarily in spoken but also in written communication, especially on social media. Code-mixing involves combining lower linguistic units like words and phrases of a language into the sentences of other language (the base language) and code-switching involves switching to another language, for the length of one sentence or more. In code-mixing and switching, a bilingual person takes one or more words or phrases from one language and introduces them into another language while communicating in that language in spoken or written mode. People nowadays express their views and opinions on several issues on social media. In multilingual countries, people express their views using English as well as their native languages. Several reasons can be attributed to code-mixing. Lack of knowledge in one language on a particular subject, being empathetic, interjection and clarification are some to name. Sentiment analysis of monolingual social media content has been carried out for the last two decades. However, during recent years, Natural Language Processing (NLP) research focus has also shifted towards the exploration of code-mixed data, thereby, making code mixed sentiment analysis an evolving field of research. Systems have been developed using ML techniques to predict the polarity of code-mixed text corpus and to fine tune the existing models to improve their performance.**

***Keywords*—*Sentiment analysis; code mixing; corpus; deep learning; machine learning; NLP; social media text***

## I. INTRODUCTION

People communicate in their native language or any other natural language having official, national or international status. In a bilingual or multilingual community, people use more than one language simultaneously as their medium of communication. These bilingual people often prefer to use mixed language constructions on the internet and social media platforms to communicate with their friends and relatives informally. The utilization of more than one language in a piece of text, whether through code-mixing or switching (or both), for effective communication, is the hallmark of the social media based text-corpus. With the advent of computers and advancements in technologies people used to analyze and process monolingual text-corpus, using various NLP techniques. NLP is the automatic manipulation of natural language text to decipher useful information. As an area of Artificial Intelligence (AI), NLP deals with training a machine for processing the text for human-computer interaction possible in natural languages [1]. NLP involves the use of computers to process natural language data. The process tends to be just about as straightforward as checking word frequencies to look at changed composing styles or as intricate as understanding total human expressions [2].

Now-a-days, people use social media for various purposes, ranging from daily news and update about the current political and social events, sports, business, entertainment, communicating with family and friends, product/service reviews and opinions and many more [3]. In a bilingual community, people often use more than one language for communicating their perspectives, reviews, opinions and proposals on various subjects. Online Indian language users are exponentially growing and as reported by an investigation by KPMG UK and Google, it is assessed that by 2021, 73% of Indian users would prefer to use native Indian languages. Researchers in the field of NLP have found it quite interesting

*Corresponding Author.

to analyze and decipher information from the text collected from well-known social networking platforms. However, the task is challenging because of a number of reasons. The text present on these platforms is characterized by having spelling errors, Meta tags (hash tags), creative spellings (*f9 for fine*) abbreviations (BTW for by the way) phonetic typing (becoz for because), word plays (*gooood for good*) and so on [4]. All these constraints make it challenging for an NLP researcher to deduce valuable information from the text. Therefore, a considerable percentage of text available on these sites is in languages such as Spanish, Chinese, Arabic, Hindi, Urdu, etc. In the recent past people, especially in bilingual countries like India, not only use a native script to write in their own languages, they also write in the Roman script to express their feelings.

Therefore, people write in code- mixed or code-switched form. Code in communications refers to the rule for converting a piece of information into another form of representation. Code mixing and code- switching are used in bilingual communities where people prefer their native language and a second language in different domains. Although code-switching and code- mixing are usually interchangeable terms in their usage, there are few differences between the two. While code- switching is actually the process of shifting from one language to another, code-mixing on the other hand means the mixing of different phonetic units such as words, phrases, morphemes, clauses, affixes and modifiers of some different language into the expressions of some other language. Thus, the code-switching is inter-sentential, while as code-mixing is intra-sentential which is constrained by grammatical principles.

Example of Code-mixing

Principal appki application ko reject karega. Likh kay leylo.

Translation: The principal will reject your application. Take it from me.

Example of Code-Switching

The principal will reject your application. Likh kay leylo.

Translation: The principal will reject your application. Take it from me.

There are many reasons why people use a multilingual approach while expressing themselves on the web and social media sites. Code mixing and code- switching occurs in informal communication and are used by multilingual speakers. In [5] a list of a number of reasons why code mixing occurs. Bilingualism, speaker and partner speaker, social community, the situation, vocabulary and prestige are the main reasons for code- mixing on social media platforms.

The main reason for code-mixing or switching can be the absence of a specific word or a phrase in a language that necessitates a person to use a word or a phrase from his/her native langue to make the receiver understand it better. The detailed motivation and reasons for code-mixing and code-switching are explained in [5]. On Social media platforms, in a multilingual society, people often mix multiple languages to express their feelings. However, they do not use native language scripts; rather they prefer the roman script to compose non-English words. Automatic language detection in such a scenario is a herculean task. Sentiment Analysis also referred to as opinion mining or emotion analysis, is the identification, recognition or categorization process of people's views and reviews for a service, a product, social issue, an event or a moment into 'positive', 'negative' and 'neutral' classes [6]. Sentiment analysis of dataset containing the data with code-mixed text is a laborious process, ranging from preprocessing of data, language identification to classification. The challenges which need to be addressed before assigning sentiments are posed mainly by unstructured sentences, mixed language constructs, spelling variants, grammatical mistakes, etc0 [7]. Also because of the noisy nature of code-mixed data and the non-availability of annotated resources, sentiment extraction from a code-mixed text has become a challenging task [8]. Therefore, sentiment analysis of the multilingual text has become increasingly an important research area [9]. The general workflow of the Code-Mixed text data Sentiment analysis process is shown in Fig. 1.

A comprehensive review of ML techniques for sentiment analysis of the code-mixed text is presented in this paper. Techniques and approaches of ML and Deep Learning (DL) for bilingual or multilingual text Sentiment Analysis are described along with their corresponding results in different scenarios and using different types of datasets.

The key research highlights of this study are:

- To present the results of recent literature on sentiment analysis of Code-Mixed and Switched languages.
- To provide a systematic review of studies performed on sentiment analysis of Code-Mixed and Switched English with Indian languages.
- To explore and report the current state of research in Code-Mixed and Switched languages using various machine learning and deep learning techniques.
- To present the results of various machine learning models in terms of their performance metrics used by the recent studies in code-Mixed and Switched English with Indian languages.

The paper is organized as follows: Section II and its subsections provide the Machine Learning and deep learning methods used in sentiment analysis of code-mixed social networking data. Section III presents the results of Sentiment Analysis of Code-mixed Indian languages; Section IV presents a discussion of the study. The conclusion is presented in Section V.

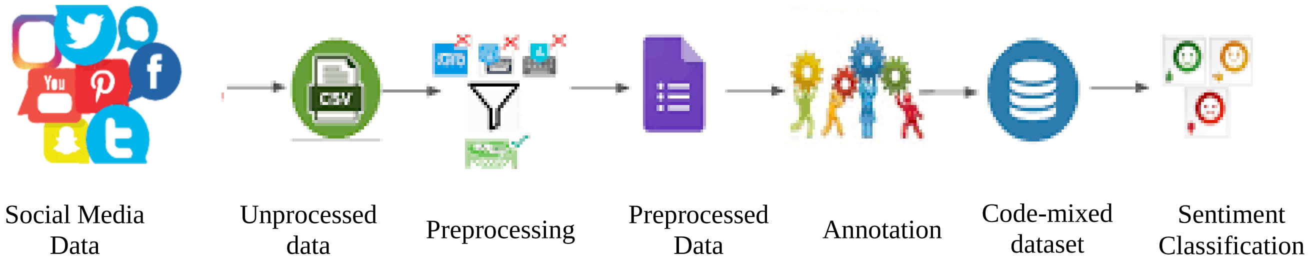

Fig. 1. Sentiment Analysis Process of Code-Mixed Data.

## II. Machine Learning and Deep Learning Approaches for Sentiment Analysis

Machine Learning allows computers to seek new tasks without being explicitly programmed to perform them. In Sentiment analysis, ML can be used to analyze text for polarity. Sentiment analysis models have been trained to analyze and understand complex natural language such as human patterns of speech, the context of the sentence, sarcasm, idioms, negation, metaphors, etc. with reasonable and accepted accuracy [10]. Researchers have successfully proposed various approaches for sentiment analysis of English language data using Machine Learning and Deep learning models [11] [12].

Deep Structured Learning commonly known as Deep Learning has acquired a lot of consideration from the recent past in the Machine Learning approach of research [13] Deep learning uses multiple layers to mine higher-level features from the given input data. It is used for a number of applications viz. text analysis, pattern analysis, classification, image processing, etc. and uses non-linear information for feature extraction and transformation in the supervised and unsupervised domain [14]. Deep Learning techniques permit computational models that manage various processing layers to learn representations of data with multiple layers of abstraction. Deep in Deep Learning denotes the layer numbers that form the Neural Network in traditional methods neural networks were of three layers viz. input, output and hidden. The maximum the number of hidden layers, the deep is the neural network [15]. Sentiment Analysis Approaches using Machine Learning and Deep Learning approaches have been illustrated in Fig. 2.

### A. Support Vector Machine

Support Vector Machine (SVM), designed by Vladimir Vapnik in 1995 [16], is a non-linear classifier and is a popular and robust classification and regression algorithm for data analysis and pattern [17]. The goal of SVM is to find the best and ideal hyper-plane that maximizes the gap between data points of two unique classes. If the data is un-labeled, Support Vector clustering is used [18]. SVM data classification concept has been illustrated in the plot given in Fig. 3. The support vectors represent the data points which are closest to the hyper-plane with a distance equivalent to margin.

A word-level classification of English-Nepali and English-Spanish code-mixed public network data was proposed in [19]. The authors performed experiments with linear kernel SVM classifier using word and character n-gram features. The model achieved an accuracy of 77.5% for Nepali- English and 80% accuracy for Spanish-English using basic features and applying a 6-way SVM classifier. The authors suggested that the features of Neural Network may improve the accuracy.

A Code mixed Language identification system for social communication text of Tamil-English and Malayalam-English was proposed in [20]. The system identifies the language on the basis of words. By using the character embedding approach, the system used trigram and n-gram features. For training and testing of the model SVM has been used. The proposed model achieved 93% and 95% accuracy for Malayalam-English and Tamil- English data. The authors suggested that availability of more code-mixed data and using trigram features shall be sufficient for the development of a language identification system.

A Hindi-English Sentiment Analysis system for Twitter data to forecast the sentiment present in the data has been proposed in [21]. Researchers have used tf-idf vector and GloVe Vector features along with the Support Vector Regression (SVR) model. The model achieved an f-score of 0.662.

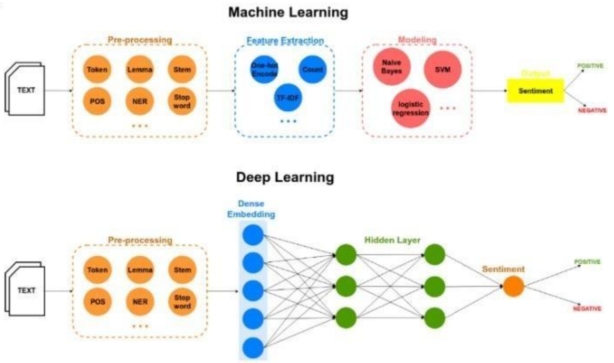

Fig. 2. Sentiment Analysis ML and DL Models.

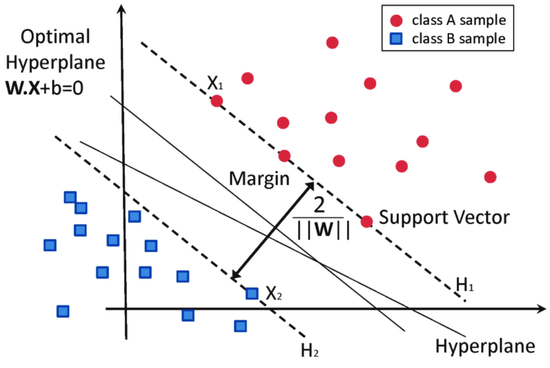

Fig. 3. Classification of Data by SVM.

Shared tasks on Sentiment Analysis of Indian Languages (SAIL) have been organized to identify sentiments in code-mixed datasets collected from media platforms like Twitter, Facebook and other social media platforms of Indian languages, especially language pairs of Hindi-English and Bengali-English [22]. Details of the shared task held during ICON-2017 (the International Conference on Natural Language Processing-2017) were presented by [23]. The goal of the shared task was to identify sentence-level sentiment polarity of code-mixed datasets of language pairs Hindi-English and Bengali-English. The authors presented a detailed overview of problem definition, dataset collection, participant systems and the evaluation process of the shared task. The SVM classifier achieved the best results. Word and character n-grams features were used and applied to SVM classifier for sentiment identification. Thus f-score of 0.569 were achieved for Hi-En and 0.526 for Bi-En datasets.

### B. *Naïve Bayes*

Naïve Bayes (NB), a data mining algorithm [24] is a probabilistic ML classification approach derived from the application of Bayes Theorem with a vast scope in real world applications [25]. The approach assumes that a new object is categorized to a class on the basis of the supposition that all features are independent given in the class [26]. The theorem can be written as in equation 1 and illustrated in Fig. 4.

$$\mathrm{P(A|B)} = \mathrm{P(B|A)P(A)} \tag{1}$$

Using the probability concept given by Bayesian theorem the equation can be represented as:

$$\text{Posterior} = \frac{\text{Prior x Likelihooh}}{Evidence} \tag{2}$$

NB classifier has been derived from the concept of Bayes Theorem with assumptions of having strong independence between the features.

A system to prepare, collect, filter and identification of sentiment of Twitter data was presented in [27]. The authors applied various supervised ML algorithms viz. Gaussian NB, Bernoulli NB and Multinomial NB for annotation and classification of English-Bengali code-mixed data. The system also applied Code-Mixed Index (CMI), Code-Mixed Factor (CMF) and other language aspects of sentiment classification.

A system to classify Hindi-English and Marathi-English tweets and comments on YouTube using a number of ML algorithms such as NB, SVM and KNN for performance evaluation of each algorithm was designed by [28]. The results reveal that NB and SVM performed better than KNN.

An automatic POS tagging system was proposed by [29]. The authors used coarse-grained and fine-grained social media text collected from Twitter and Facebook for experimentation purposes. Machine learning algorithms such as NB, Conditional Random Forest (CRF), and random forest along with Sequential Minimal Optimization (SMO) were applied for performance comparison. Various features were used in the process which was done on the word context information. The CRF based model thus attained the f1-score of 0.716.

Authors in [30] carried out experiments to construct an English-Punjabi text sentiment classification system. The data was collected from Facebook posts in the agricultural domain. Two classifiers viz. SVM and NB were applied for sentiment identification. Features like unigram and n-gram were applied to the model. The model achieved best accuracy of 85.5% using Naïve Bayes classifier.

A binary sentiment classification model was proposed by [31]. The model used English-Bengali data collected for movie reviews from social networking sites. For the classification and identification of positive and negative sentiments two supervised ML algorithms, NB and SVM were used. The experimental results reveal that if the test and train data are of similar type that is both language data is in Roman script, SVM gives better results. However, overall Naïve Bayes achieved the best accuracy.

### C. *Decision Tree*

Decision tree (DT) is referred to as a non- parametric ML technique of data mining. Decision Tree is commonly used in regression and classification problems such as marketing, sentiment analysis, scientific discovery, fraud detection, etc. [38] is one of the famous supervised ML classification algorithms. The decision tree splits data into two or more sets and important features that create the best split are used and calculated by the algorithm as illustrated in Fig. 5.

An essential part of NLP is POS Tagging. For the English language data, POS tagging is a complex task. However, for code-mixed text data, this is more challenging and is a focused research area in which still needs a significant amount of work to be done for Indian languages code mixed data. An approach for three code-mixed Indian language texts in language pairs (Hindi-English, Hindi-Bengali and Hindi-Telugu) POS tagging was presented by [32]. The authors used ICON-2015 code-mixed data and applied the Decision Tree ML algorithm for code mixed text POS tagging.

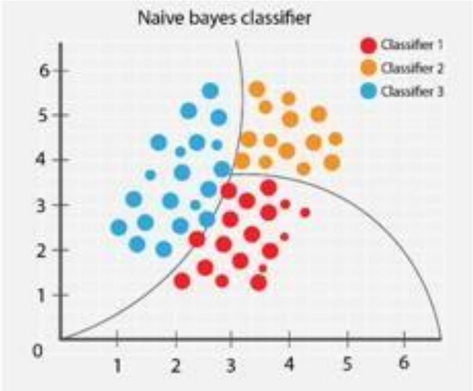

Fig. 4. Naïve Bayes Classification.

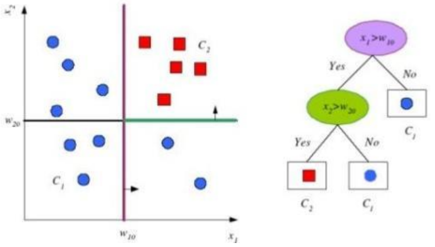

Fig. 5. Decision Tree.

Study on Hinglish (Hindi-English) code-mixed tweets sentiment analysis was done in [33]. Datasets were provided in SemEval-2020 (International Workshop on Semantic Evaluation-2020). The system used J48 Decision Tree as a training classifier and Weka as a tool for the classification. Performance evaluation of the model was done and f-score of 0.53 was achieved.

### D. *Random Forest*

Random forest (RF) is a supervised ML approach used both in for classification and regression problems [34]. It is an ensemble learning algorithm developed by [35]. The algorithm combines DTs and collects their results using averaging. Being a type of supervised learning algorithm, RF has been influenced by [36]. The algorithm works on divide-and-conquer rule. A Random Forest has generally shown excellent performance in scenarios in which the number of observations are less than the number of variables [37]. The general workflow of the technique has been shown in Fig. 6.

For detection of sarcasm, in Hindi-English code- mixed dataset consists of tweets, a baseline supervised classification approach was proposed by [38]. The authors perform 10-fold cross validation using Random Forest classifier. The proposed system also uses Linear SVM classifier and RBF Kernel SVM for the same dataset. However, the RF classifier achieved the better f-score of 78.4.

In collaboration with Forum for Information Retrieval Evaluation (FIRE), a shared task was organized for Code-Mixed Entity Extraction process in Indian Languages (CMEE-IL) in Kolkata, India [39]. Datasets were collected for Hi-En and Ta-En code-mixed social networking data in the said shared task and an Entity Recognition Model was developed. Random Forest Tree Classifier was used for classification. Conditional Random Field Entity Recognition with hybrid features were experimented on the collected corpus. The model achieved 95% of accuracy on training data and a reasonable performance on testing data.

The researchers in [40] have proposed a POS tagger for three Indian code-mixed language pair's viz. Hi-En, Bi-En and Telugu-English. A RF classifier along with a dictionary was applied for fine-grained and coarse-grained datasets consists of tweets, Facebook comments and WhatsApp chats collected from ICON-2016 for the three language pairs. The proposed model achieved best f-score of 78.744 in fine-grained model consisting of Hi-En tweets and 77.944 in coarse-grained model consisting of Bi-En Facebook posts.

### E. *Artificial Neural Network (ANN)*

The concept derived from the human brain in which numerous neurons are interconnected to process data in parallel. ANNs are non-linear mathematical models that show an intricate connection among information sources in order to get a new pattern. ANN can be applied in a range of tasks, including text analysis, image processing, speech recognition, machine interpretation and clinical determination. An ANN has an input layer of neurons or nodes, one or two hidden layers of neurons (or even three), along with a final output layer of neurons. A typical architecture of an ANN is shown in Fig. 7. In a Neural Network the lines connecting nodes (or neurons) are associated with weight. In Fig. 8, a transfer function computes the weighted sum of the inputs while the activation function obtains the result.

The authors of [41] introduced a model for sentiment analysis of Hindi-English text using sub-word level LSTM. The data was collected from Facebook posts and used a 3-class scale of 'positive', 'negative' and 'neutral'. The proposed sub-word level LSTM model achieved higher accuracy than the character-level LSTM model, SVM (Unigram) and Naïve Bayes techniques of machine learning. The overall accuracy of 69.7% was achieved by the proposed system.

Authors in [42] proposed a model in Hindi-English Twitter data for humor detection. Based on models like, Word2Vec and FastText an approach for bilingual word- embedding's applied to BilSTM system for the detection of humor in the text. The proposed approach achieved an accuracy of 73.6%.

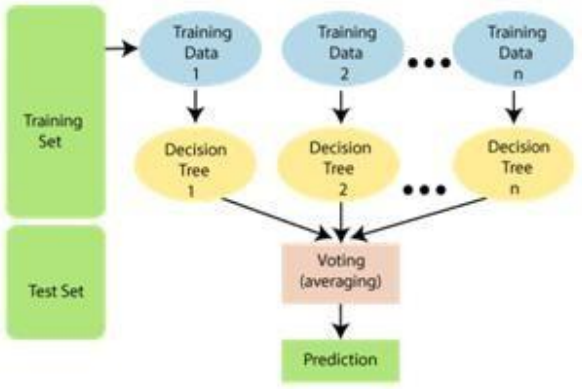

Fig. 6. Random Forest.

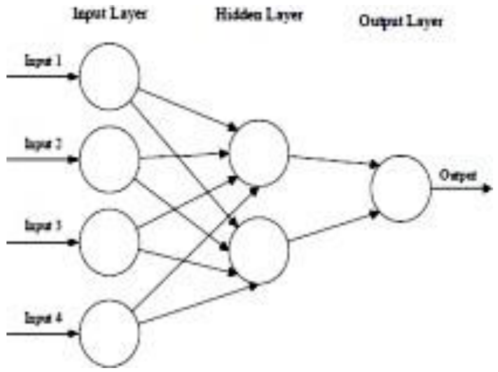

Fig. 7. General Architecture of Artificial Neural.

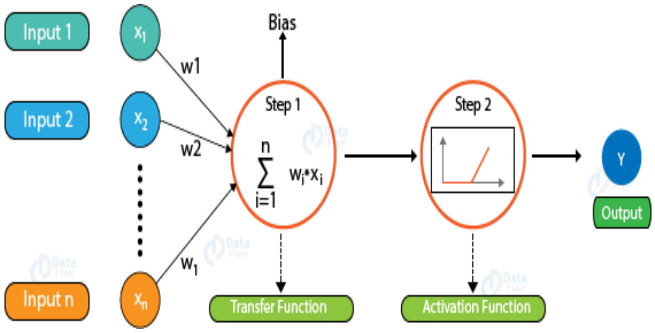

Fig. 8. Transfer Function.

Automatic extraction of sentiments from Hindi- English and Bengali-English Facebook posts was proposed by [43]. The corpus was manually created and annotated. Several preprocessing steps have been employed in order to remove unwanted data from the corpus. A Multilayer Perceptron Model was used for the detection of the sentiment polarity. The proposed model achieved anaccuracy of 68.5%.

### F. Convolutional Neural Network

Convolutional Neural Networks (CNN) in recent years have achieved ground-breaking results in a number of pattern recognition fields, ranging from image processing to voice recognition. The most advantageous feature of CNNs is that they reduce the number of parameters in ANNs. This accomplishment has prompted researchers and developers to tackle broader models in order to solve more difficult problems. CNNs are similar to conventional Artificial Neural Networks (ANNs), consisting of neurons that learn to optimize themselves [44]. The neurons obtain inputs to perform operations like the scalar product and non-linear functions, which acts as a foundation for countless Artificial Neural Networks. The complete neural network exhibit a single observant score function from raw input vectors to the final classification output. The general architecture of the CNN for the classification has been illustrated in Fig. 9.

A CNN based system for the sentiment identification of Hindi-English data was proposed by [45]. The sentiment analysis has been done using three class classifications. The classes included 'positive', 'negative' and 'neutral'. The classification of the classes have been done using word-level representations. Since tweets contain informal text, memorization of aspects of the word orthography in a word-level representation was done using CNN. The model achieved an f-score of 0.324 for Hindi- English data.

To compare ML and DL approaches researchers in [46] have used three code-mixed datasets viz. Hindi-English, Bengali- English and Kannada-English. The datasets used in the study have been sourced from Facebook posts and SAIL-2017. A number of Machine and Deep Learning techniques were applied on code-mixing datasets for sentiment analysis. The techniques used include Doc2Vec, SVM, CNN and Bi-LSTM. The experimental results showed that CNN performs better for Kannada-English dataset and achieved an accuracy of 71.5%. The BiLSTM performs better for Hindi-English and Bengali-English datasets with accuracies of 60.22% and 70.20% respectively.

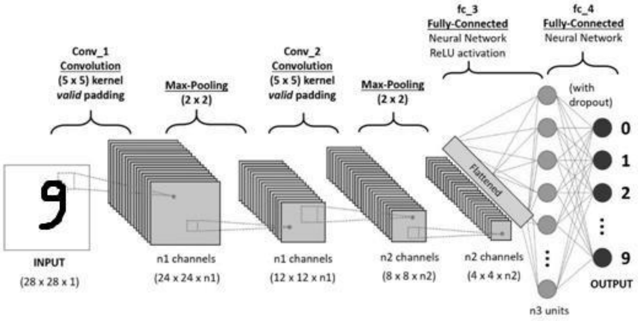

Fig. 9. A Convolutional Neural Network Architecture.

Authors in [47] presented a hybrid model for sentiment analysis of English-Hindi code-mixed data. The method used CNN architecture for generating sub-word level representation for the sentences. Two BiLSTMs, collective encoder and specific encoder are fed with the sub-word level representation. Finally, a Feature Network consists of orthographic features has been combined with the BiLSTMs to achieve an accuracy of 83.5%. The hybrid approach, therefore, combines surface features with Attention-based Recurrent Neural Networks to produce a single representation that can be trained for sentiment classification.

For the identification of emotions in Hindi-English Twitter and Facebook data, authors in [48] proposed a Deep Learning-based system. Several Deep Learning techniques such as 1D-CNN, LSTM, Bi-LSTM, CNN-LSTM and CNN-BilSTM were used to predict the polarity of the sentence. To generate feature vectors, the pre-trained bilingual model was used. The experimental results showed that CNN- BiLSTM model achieved the best accuracy of 83.21%.

The authors of [49] used Facebook comments of Hindi-English code-mixed dataset provided by Trolling, Aggression & Cyber bullying-I (TRAC-I) and apply machine and deep learning models for the classification of text data into a 3-class scale such as, 'Covertly Aggressive', 'Overtly Aggressive' and 'Non-Aggressive' classes. CNN model worked best with an f-score of 0.58 and accuracy of 73.2% as shown in the experimental results.

The study in [50] explores hate speech detection in tweets written in Hindi-English. The authors have used DL models, CNN-ID, LSTM and BiLSTM the semantics detection of hate speech along with the context. The embedding's were generated using Word2Vec. The experimental results were compared with the contemporary approaches. The CNN-ID model outperforms the other two and achieved an overall accuracy of 82.62%.

### G. Recurrent Neural Network (RNN)

RNNs are being used by researchers since 1990s. RNN is a neural network with feedback connections is known as a recurrent net [51]. RNN is a form of ANN that works with time series or sequential data. Techniques based on RNNs have been used to solve a broad range of problems. Machine Translation, Speech Recognition, Video Tagging, Text Analysis and Image processing are some examples where RNN algorithms are used. The general architecture of an RNN has been shown in Fig. 10. Each hidden state has hidden nodes also called hidden units.

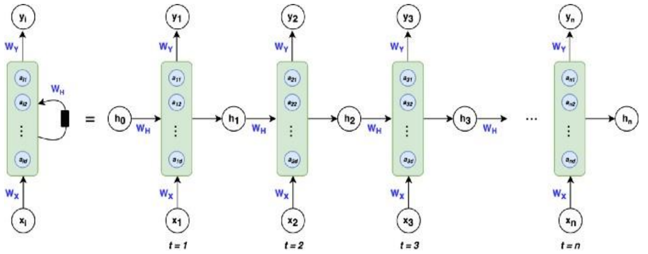

Fig. 10. The General Architecture of an RNN.

An automatic sentiment prediction system of Hindi-English code-mixed dataset consists of tweets was proposed by [52]. The model used the Recurrent Convolutional Neural Network approach to capture the semantics of the text and classify them into three scale classification. The dataset was collected from SemEval- 2020 shared task. An f1-score of 0.69 was achieved by the proposed approach.

The authors in [53] [54] proposed a part-of-speech tagger for Hindi-English, Bengali-English and Telugu-English datasets. The datasets were collected from social networking platforms such as Facebook, Twitter and WhatsApp. The proposed model used Recurrent Neural Network (RNN) to predict word-level part-of-speech tags.

A Sentiment Analysis model of Hindi-English data, based on RNNs, was proposed by [55][56]. Public Facebook pages of popular personalities of Indian Politics and Cinema were used to collect data. The model combines two different BiLSTMs for the identification of sentiment at the sub-word level as well as at the sentence level. The proposed approach used orthogonal features achieved accuracy of 83.5% and f1-score of 0.827.

## III. Results of Code-mixed Text Sentiment Analysis for Indian Languages

On Social media sites, netizens in India often use English and their native language such as Hindi in a mixed form to express their opinions on a wide range of topics. Over the years, researchers in the field of NLP have shown keen interest in this new form of text which is often informal and challenging. However, with the advent of NLP tools and techniques, the research related to the analysis of code- mixed textual data has also gained momentum. The significant research studies with their description and the results given for code-mixed text data analysis and sentiment analysis in Indian Languages is given in Table I.

TABLE I. Text Analysis and Sentiment Classification of Code Mixed Text of Indian Languages Gathered from Social Media

<table>
<tr><th>S#</th><th>Paper/Study</th><th>Language</th><th>Objective(s)</th><th>Dataset(s)</th><th>ML/DL Approach</th><th>Performance Evaluation</th></tr>
<tr><td>01</td><td>Patra, Braja Gopal, et al [25]</td><td>Hindi-English<br>Bengali-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>SVM</td><td><table><tr><td>Language</td><td>F1-Score</td></tr><tr><td>Hi-En</td><td>0.569</td></tr><tr><td>Bi-En</td><td>0.526</td></tr></table></td></tr>
<tr><td>02</td><td>Ansari &amp; Govilkar [30]</td><td>Hindi-English<br>Marathi-English</td><td>Sentiment Analysis</td><td>Tweets<br>Facebook posts<br>YouTube comments</td><td>NB<br>SVM</td><td><table><tr><td>Model</td><td>Language</td><td>F1-Score</td></tr><tr><td rowspan="2">NB</td><td>Hi-En</td><td>0.60</td></tr><tr><td>Ma-En</td><td>0.46</td></tr><tr><td rowspan="2">SVM</td><td>Hi-En</td><td>0.60</td></tr><tr><td>Ma-En</td><td>0.59</td></tr></table></td></tr>
<tr><td>03</td><td>Jamatia, Anupam et. al. [31]</td><td>Hindi-English</td><td>Development of annotated corpus,<br>POS tagging,<br>Sentiment Analysis</td><td>Tweets,<br>Facebook posts</td><td>CRF<br>SMO<br>NB<br>RF</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>CRF</td><td>0.716</td></tr><tr><td>SMO</td><td>0.397</td></tr><tr><td>NB</td><td>0.458</td></tr><tr><td>RF</td><td>0.706</td></tr></table></td></tr>
<tr><td>04</td><td>Singh, M et. al.[32]</td><td>English-Punjabi</td><td>Sentiment Analysis</td><td>Tweets,<br>Facebook posts,<br>YouTube comments</td><td>NB<br>SVM</td><td><table><tr><td>Model</td><td>Accuracy</td></tr><tr><td>NB</td><td>85.5%</td></tr><tr><td>SVM</td><td>85%</td></tr></table></td></tr>
<tr><td>05</td><td>Mandal &amp; Das [33]</td><td>English-Bengali</td><td>Sentiment Analysis</td><td>Movie reviews</td><td>NB<br>LR<br>SVM</td><td><table><tr><td>Model</td><td>Accuracy</td></tr><tr><td>NB</td><td>59%</td></tr><tr><td>LR</td><td>55%</td></tr><tr><td>SVM</td><td>57%</td></tr></table></td></tr>
<tr><td>06</td><td><i>Pimpale &amp; Patel [35]</i></td><td>Hindi-English<br>Hindi-Telugu</td><td>POS tagging,<br>Sentiment Analysis</td><td>Tweets<br>Facebook posts</td><td>NB<br>DT<br>RF</td><td>F1-measure<table><tr><td>Approach</td><td>Hi-Eng</td><td>Tal-Eng</td></tr><tr><td>NB</td><td>40.4</td><td>46.3</td></tr><tr><td>DT</td><td>44.6</td><td>50.3</td></tr><tr><td>RF</td><td>43.0</td><td>47.0</td></tr></table></td></tr>
<tr><td>07</td><td>Ghosh et al [46]</td><td>Hindi-English<br>Bengali-English</td><td>Sentiment Analysis</td><td>Facebook posts</td><td>MP</td><td>Accuracy: 68.5%</td></tr>
</table>

<table>
<tr><td>08</td><td>Sasidhar, T. T et. al.[51]</td><td>Hindi-English</td><td>Development of annotated dataset, classification of emotions</td><td>Tweets<br>Facebook posts<br>Instagram comments</td><td>CNN-BiLSTM</td><td>Accuracy: 83.21%</td></tr>
<tr><td>09</td><td>Kumar &amp; Dhar [57]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Facebook posts</td><td>BiLSTM</td><td>Accuracy: 83.54%<br>F1-score: 0.827.</td></tr>
<tr><td>10</td><td>Baroi, Subhra Jyoti, et al [58]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>Ensemble<br>LSTM<br>LSTM + Convolution Layer<br>BiLSTM<br>CNN</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>LSTM</td><td>0.5640</td></tr><tr><td>LSTM+Conv</td><td>0.5747</td></tr><tr><td>BiLSTM</td><td>0.576</td></tr><tr><td>CNN</td><td>0.5737</td></tr></table></td></tr>
<tr><td>11</td><td>Veena et. al. [59]</td><td>Hindi-English</td><td>Language Identification</td><td>Facebook posts,<br>Tweets,<br>WhatsApp chats</td><td>SVM</td><td>Facebook data (f-score =98.70)<br>Tweeter data (f-score=93.94)<br>WhatsApp data (f-score=77.60)</td></tr>
<tr><td>12</td><td>Si, Shukrity, et al. [60]</td><td>Hindi-English</td><td>Aggression detection</td><td>Tweets</td><td>SVM<br>GBM<br>LR<br>Adaboost<br>DT<br>KNN<br>LSTM</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>SVM</td><td>0.5349</td></tr><tr><td>GBM</td><td>0.5410</td></tr><tr><td>LR</td><td>0.5045</td></tr><tr><td>Adaboost</td><td>0.5030</td></tr><tr><td>DT</td><td>0.4938</td></tr><tr><td>KNN</td><td>0.4316</td></tr><tr><td>LSTM</td><td>0.4039</td></tr></table></td></tr>
<tr><td>13</td><td>Soman, K. P. [61]</td><td>Hindi-English<br>Bengali-English<br>Telugu-English</td><td>POS) tagging</td><td>Tweets,<br>Facebook posts</td><td>SVM</td><td><table><tr><td>Language</td><td>Accuracy</td></tr><tr><td>Hi-En</td><td>81.57%</td></tr><tr><td>Bi-En</td><td>76.18%</td></tr><tr><td>Te-En</td><td>68.85%</td></tr></table></td></tr>
<tr><td>14</td><td>Lakshmi &amp; Shambhavi [62].</td><td>English-Kannada</td><td>Word level Language Identification</td><td>Twitter<br>Facebook posts</td><td>MNB<br>BNB<br>SVM<br>RF<br>LR</td><td><table><tr><td>Model</td><td>Accuracy</td></tr><tr><td>MNB</td><td>85%</td></tr><tr><td>BNB</td><td>80%</td></tr><tr><td>SVM</td><td>87%</td></tr><tr><td>RF</td><td>78%</td></tr><tr><td>LR</td><td>80%</td></tr></table></td></tr>
<tr><td>15</td><td>Vijay Deepanshu, et al. [63]</td><td>Hindi-English</td><td>Emotion classification</td><td>Tweets</td><td>SVM</td><td>Accuracy : 58.2%</td></tr>
<tr><td>16</td><td>Pravalika et al.[64].</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Movies posts on Facebook</td><td>NB<br>SVM<br>DT<br>RF<br>MP</td><td>Precision &amp; Recall<table><tr><td>Classifier</td><td>Precision</td><td>Recall</td></tr><tr><td>NB</td><td>0.725</td><td>0.735</td></tr><tr><td>SVM</td><td>0.718</td><td>0.77</td></tr><tr><td>DT</td><td>0.701</td><td>0.723</td></tr><tr><td>RF</td><td>0.752</td><td>0.755</td></tr><tr><td>MP</td><td>0.695</td><td>0.702</td></tr></table></td></tr>
<tr><td>17</td><td>1) <i>Vijay, Deepanshu, et al [65]</i></td><td>Hindi-English</td><td>Sarcasm detection</td><td>Tweets</td><td>SVM<br>RF</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>SVM</td><td>0.77</td></tr><tr><td>RF</td><td>0.72</td></tr></table></td></tr>
<tr><td>18</td><td>2) <i>Wu, Wang &amp; Huang [66]</i></td><td>Hindi-English<br>Spanish-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>BiLSTM</td><td>F1-score: 0.730</td></tr>
<tr><td>19</td><td>Bhange &amp; Kasliwal [67]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>Ensemble (NB-SVM)</td><td>Accuracy: 0.667<br>F1-score :0.673</td></tr>
</table>

<table>
<tr><td>20</td><td>Sharma, &amp; Motlani, [68]</td><td>Hindi-English<br>Bengali-English<br>Tamil-English</td><td>POS Tagging</td><td>Tweets</td><td>CRF</td><td><table><tr><td>Language</td><td>Accuracy</td></tr><tr><td>Hi-En</td><td>80.68%</td></tr><tr><td>Bi-En</td><td>79.84%</td></tr><tr><td>Ta-En</td><td>75.48%</td></tr></table></td></tr>
<tr><td>21</td><td>Sarkar, K [69]</td><td>Hindi-English<br>Bengali-English<br>Tamil-English</td><td>POS Tagging</td><td>Text from social networking sites</td><td>Hidden Markov Model</td><td>Accuracy : 75.60%</td></tr>
<tr><td>22</td><td><i>3) Choudhary, Nurendra, et al. <b>[</b>70<b>]</b></i></td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>siamese network with twin character level Bi-LSTM networks</td><td>Accuracy: 78%<br>F1-score: 0.767</td></tr>
<tr><td>23</td><td>Singh, K. et. al. [71]</td><td>Hindi-English</td><td>Language Identification, Entity Recognition</td><td>Tweets</td><td>CRF<br>LSTM</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>LSTM</td><td>0.693</td></tr><tr><td>CRF</td><td>0.767</td></tr></table></td></tr>
<tr><td>24</td><td>Jamatia, Anupam, et al. [72]</td><td>Hindi-English<br>Bengali English</td><td>Sentiment Analysis</td><td>Tweets</td><td>BiLSTM CNN<br>Double BiLSTM<br>Attention Based Model</td><td><table><tr><td>Language</td><td>Model</td><td>F1-Score</td></tr><tr><td rowspan="3">Hi-En</td><td>BiLSTM</td><td>0.566</td></tr><tr><td>D-BiLSTM</td><td>0.595</td></tr><tr><td>ABM</td><td>0.604</td></tr><tr><td rowspan="3">Bi-En</td><td>BiLSTM</td><td>0.623</td></tr><tr><td>D-BiLSTM</td><td>0.659</td></tr><tr><td>ABM</td><td>0.675</td></tr></table></td></tr>
<tr><td>25</td><td>Raha, Tathagata, et al. [73]</td><td>Bengali-English</td><td>POS Tagging</td><td>Tweets</td><td>LSTM</td><td>Accuracy: 75.29%</td></tr>
<tr><td>26</td><td>Parikh, A et. Al[74]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>Ensemble Model (LR,RF, BERT)</td><td>F1-score : 0.693.</td></tr>
<tr><td>27</td><td>Pratapa, A et. al.[75]</td><td>Hindi-English</td><td>POS Tagging, Sentiment Analysis</td><td>Tweets</td><td>LSTM</td><td>F1-score : 0.56</td></tr>
<tr><td>28</td><td><i>4) Kumar, Vaibhav, et al. [76]</i></td><td>Hindi-English</td><td>Language modelling</td><td>Social media blogs<br>Facebook Comments</td><td>LSTM</td><td>Accuracy: 58.9%</td></tr>
<tr><td>29</td><td>Bohra, Aditya, et al [77]</td><td>Hindi-English</td><td>Dataset Creation,<br>Hate speech detection</td><td>Tweets</td><td>RF</td><td>Accuracy: 69.9%</td></tr>
<tr><td>30</td><td>Prabhu, Ameya, et al .[78]</td><td>Hindi-English</td><td>Corpus creation, Sentiment Analysis</td><td>Facebook posts</td><td>LSTM</td><td>Accuracy 69.7%</td></tr>
<tr><td>31</td><td>Dahiya, Anirudh, et al. [79]</td><td>Hindi-English</td><td>Language Identification, POS Tagging, Sentiment Analysis</td><td>Facebook Posts</td><td>BiLSTM</td><td>Accuracy 72.51%</td></tr>
<tr><td>32</td><td>Gopal &amp; Das [80]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Facebook posts</td><td>Ensemble ( LSTM MNB)</td><td>Accuracy 70.8<br>F1-Score 0.661</td></tr>
</table>

<table>
<tr><td>33</td><td>Singh &amp; Lefever [81]</td><td>Hindi-English</td><td>Sentiment Analysis</td><td>Tweets</td><td>BiLSTM<br>Transfer Learning</td><td><table><tr><td>Model</td><td>F1-Score</td></tr><tr><td>LSTM</td><td>0.616</td></tr><tr><td>TL</td><td>0.556</td></tr></table></td></tr>
<tr><td>34</td><td>Santosh &amp; Aravind, [82]</td><td>Hindi-English</td><td>Hate speech detection</td><td>Tweets</td><td>SVM<br>RF<br>LSTM</td><td><table><tr><td>Model</td><td>Accuracy</td><td>F1-Score</td></tr><tr><td>SVM</td><td>70.7%</td><td>0.429</td></tr><tr><td>RF</td><td>65.1%</td><td>0.292</td></tr><tr><td>LSTM</td><td>66.6%</td><td>0.487</td></tr></table></td></tr>
<tr><td>35</td><td>Sreelakshmi et. al. [83]</td><td>Hindi-English</td><td>Hate speech detection</td><td>Facebook Posts</td><td>SVM<br>RF</td><td><table><tr><td>Model</td><td>Accuracy</td></tr><tr><td>SVM</td><td>63.75%</td></tr><tr><td>RF</td><td>64.15%</td></tr></table></td></tr>
</table>

## IV. Discussion

Social networking has emerged as an essential part of our lives. It has not only become a platform for individuals to communicate with each other, it also acts as a news media, a platform to connect with people and develop a relationship. It gives an individual the opportunity to express their views on a particular product, service, social movement, government policy etc. Social media thus helps in business and governance tasks. India being the second-largest populous country of the world has a wide range of linguistic diversity. People often express their views in English as well as in their native language resulting in the proliferation of code-mixed data. Mixing of languages or language varieties either in oral or in written form is known as code-mixing.

Sentiment evaluation of social media data analysis plays a crucial role in modern commerce and governance. Classical sentiment analysis systems were developed for dealing with product reviews. With the advancement in NLP tools and technologies, sentiment analysis systems were developed for other tasks as well. Code-mixed text data sentiment analysis is a relatively challenging task right from data gathering to classification. Various studies have been accomplished on “Cross-Lingual Information Retrieval” (CLIR), “Multilingual Information Retrieval” (MLIR) and “Mixed Script Information Retrieval” (MSIR) [84]. In CLIR, a user queries in one language and retrieves desired information in more than one language. In MLIR, a person can query in one or more languages and retrieve information in more than one language. However, the task of retrieval becomes more difficult when dealing with MSIR, due to Romanized text of non-English languages. Also, the social media text contains many non-standard forms such as misspellings, improper use of grammar, letter substitutions, non-standard abbreviations and other ambiguities which makes preprocessing a necessary step in the code-mixed scenario. Various tools for POS tagging, language identification as well as named entity recognition (NER) have been developed for the analysis of code-mixed data over the recent years. However, due to limited datasets particularly annotated datasets for some language pairs and the non-availability of these resources for majority of native Indian languages, and the linguistic catalogues for informal code-mixed text, the automatic text analysis tool development is challenging.

Code-mixed text data analysis in multilingual societies like India has become a vital linguistic research area more specifically for social media content. However, processing such type of data for linguistic analysis is a challenging task due to inherited linguistic complexity and the presence of spelling and grammar variations [85] Therefore, to promote research in code-mixed text, MSIR workshops were organized at FIRE since 2008 [86] various workshops have been conducted on linguistic code-switching computational procedures for language identification and NER textual data for in code- mixing scenarios [87]. SemEval workshops (International Workshop on Semantic Evaluation) have also been conducted. SemEval-2020 was aimed to encourage research in code-mixed Sentiment Analysis of Twitter data.

This paper provides the results of a review study for the sentiment classification of code-mixed Indian languages. The adopted languages, ML/DL/ANN approaches, data sets and challenges in sentiment analysis of code-mixed text data have been highlighted. The results also show that various studies have been carried out in different application domains, thus each of the domains requires different analysis approaches to achieve better performance.

The results show that the most used ML classifier for the sentiment classification of code-mixed Indian language text is SVM followed by NB and RF. Ensemble approaches are also used to classify the code-mixed text. The study also showed that in terms of accuracy and f1- measure, Neural Network approaches perform better than the traditional models. Typically LSTM and BiLSTM algorithms are being used by the researchers for the classification of sentiment in code-mixed datasets. The study reveals that Twitter is the first choice of data collection followed by Facebook and movie/product reviews. Also, appreciable research has been carried out in the Hindi-English public networking site’s text followed by Bengali-English. Research has also been carried out in other code-mixed Indian languages such as Punjabi- English, Marathi-English, Telugu-English and Malayalam- English. However, limited or no annotated datasets, text analysis tools and SentiWordNets are not available in most of the code-mixed Indian language text.

## V. Conclusion

A comprehensive study of Machine Learning techniques for code-mixed Indian language text collected from popular media platforms has been carried out in this paper. Among traditional Machine learning approaches, SVM is the first choice of most researchers. In the case of Deep Learning approaches, BiLSTM dominates the research. Twitter data is used for most of the systems and code-mixed social media text for Hindi-English is most researched. Annotated datasets, text and language analysis tools and other lexical recourses are trivial while dealing with code-mixed datasets. In our future work we are going to present a statistical review of Machine Learning approach for Sentiment Analysis of code-mixed social- media text.

## REFERENCES

[1] Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2017). Supervised sentiment analysis in multilingual environments. Information Processing & Management, 53(3), 595-607.

[2] Mujahid, M.; Lee, E.; Rustam, F.; Washington, P.B.; Ullah, S.; Reshi, A.A.; Ashraf, I. Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19. *Appl. Sci.* **2021**, *11*, 8438. https://doi.org/10.3390/app11188438.

[3] Kapoor, K.K., Tamilmani, K., Rana, N.P. et al. Advances in Social Media Research: Past, Present and Future. Inf Syst Front 20, 531–558 (2018). https://doi.org/10.1007/s10796-017-9810-y.

[4] Das, A., & Gambäck, B. (2013). Code-Mixing in Social Media Text. The Last Language Identification Frontier? Trait. Autom. des Langues, 54, 41-64.

[5] Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language, 28(1), 56-75.

[6] Kim, E. (2006). Reasons and motivations for code-mixing and code-switching. Issues in EFL, 4(1), 43-61.

[7] Purba, Y. H., & Suyadi, N. F. (2018). An Anlysis Of Code Mixing On Social Media Networking Used By The Fourth Semester Students Of English Education Study Program Batanghari University In Academic Year 2017/2018. JELT: Journal of English Language Teaching, 2(2), 61-68.

[8] Srivastava, V., & Singh, M. (2020). IIT Gandhinagar at SemEval-2020 task 9: code-mixed sentiment classification using candidate sentence generation and selection. arXiv preprintarXiv:2006.14465.

[9] Kumar, R., & Kaur, J. (2020). Random forest-based sarcastic tweet classification using multiple feature collection. In Multimedia Big Data Computing for IoT Applications (pp. 131- 160). Springer, Singapore.

[10] Nankani, H., Dutta, H., Shrivastava, H., Krishna, P. R., Mahata, D., & Shah, R. R. (2020). Multilingual Sentiment Analysis. In Deep Learning-Based Approaches for Sentiment Analysis (pp. 193-236). Springer, Singapore.

[11] Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. arXiv preprint arXiv:2006.00210.

[12] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631– 1642, 2013.

[13] Reshi AA, Ashraf I, Rustam F, Shahzad HF, Mehmood A, Choi GS. 2021. Diagnosis of vertebral column pathologies using concatenated resampling with machine learning algorithms. PeerJ Computer Science 7:e547.

[14] Furqan Rustam, Aijaz Ahmad Reshi, Wajdi Aljedaani, Abdulaziz Alhossan, Abid Ishaq, Shabana Shafi, Ernesto Lee, Ziyad Alrabiah, Hessa Alsuwailem, Ajaz Ahmad, Vaibhav Rupapara, "Vector mosquito image classification using novel RIFS feature selection and machine learning models for disease epidemiology",Saudi Journal of Biological Sciences, Volume 29, Issue 1,2022,Pages 583-594.

[15] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[16] Aijaz Ahmad Reshi, Furqan Rustam, Arif Mehmood, Abdulaziz Alhossan, Ziyad Alrabiah, Ajaz Ahmad, Hessa Alsuwailem, Gyu Sang Choi, "An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification", Complexity, vol. 2021, Article ID 6621607, 12 pages, 2021.

[17] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[18] A.-Z. Ala'M, A. A. Heidari, M. Habib, H. Faris, I. Aljarah, M. A. Hassonah, Salp chain based optimization of support vector machines and feature weighting for medical diagnostic information systems, in: Evolutionary Machine LearningTechniques, Springer, 2020, pp. 11–34.

[19] K. P. Bennett and C. Campbell, "Support vector machines: Hype or hallelujah?," ACM SIGKDD Explorations Newsletter, vol. 2, no. 2, pp. 1–13, 2000.

[20] Barman, U., Wagner, J., Chrupała, G., & Foster, J. (2014, October). Dcu-uvt: Word-level language classification with code- mixed data. In Proceedings of the First Workshop on Computational Approaches to Code Switching (pp. 127-132).

[21] Veena, P. V., Kumar, M. A., & Soman, K. P. (2017, September). An effective way of word-level language identification for code-mixed facebook comments using word- embedding via character-embedding. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1552-1556). IEEE.

[22] Garain, A., Mahata, S., & Das, D. (2020, December). JUNLP at SemEval-2020 Task 9: Sentiment analysis of hindi-english code mixed data using grid search cross validation. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1276- 1280).

[23] Patra, B. G., Das, D., Das, A., & Prasath, R. (2015, December). Shared task on sentiment analysis in Indian languages (sail) tweets-an overview. In International Conference on Mining Intelligence and Knowledge Exploration (pp. 650-655). Springer, Cham.

[24] Patra, B. G., Das, D., & Das, A. (2018). Sentiment analysis of code-mixed Indian languages: an overview of SAIL_Code-Mixed Shared Task@ ICON-2017. arXiv preprint arXiv:1803.06745.

[25] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[26] Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. Knowledge-Based Systems, 192, 105361.

[27] Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis (Vol. 3, pp. 731-739). New York: Wiley.

[28] Mandal, S., Mahata, S. K., & Das, D. (2018). Preparing Bengali-English code-mixed corpus for sentiment analysis of indian languages. arXiv preprint arXiv:1803.04000.

[29] Ansari, M. A., & Govilkar, S. (2018). Sentiment analysis of mixed code for the transliterated hindi and marathi texts. International Journal on Natural Language Computing (IJNLC) Vol, 7.

[30] Jamatia, A., Gambäck, B., & Das, A. (2015). Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. Association for Computational Linguistics.

[31] Singh, M., Goyal, V., & Raj, S. (2019, November). Sentiment analysis of english-punjabi code mixed social media content for agriculture domain. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON) (pp. 352- 357). IEEE.

[32] Mandal, S., & Das, D. (2018). Analyzing roles of classifiers and code-mixed factors for sentiment identification. arXiv preprint arXiv:1801.02581.

[33] Brijain, M., Patel, R., Kushik, M. R., & Rana, K. (2014). A survey on decision tree algorithm for classification.

[34] Pimpale, P. B., & Patel, R. N. (2016). Experiments with POS tagging code-mixed Indian social media text. arXiv preprint arXiv:1610.09799.

[35] Singh, G. (2020). Decision Tree J48 at SemEval-2020 Task 9: Sentiment Analysis for Code-Mixed Social Media Text (Hinglish). arXiv preprint arXiv:2008.11398.

[36] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[37] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[38] Ho, T. K. (1998). The random subspace method for constructing

decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844.

[39] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.

[40] Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A corpus of english-hindi code-mixed tweets for sarcasm detection. arXiv preprint arXiv:1805.11869.

[41] HB, B. G., Kumar, M. A., & Soman, K. P. (2016). Conditional Random Fields for Code Mixed Entity Recognition. In FIRE (Working Notes) (pp. 309-312).

[42] Bhargava, R., Tadikonda, B. V., & Sharma, Y. (2016, December). BITS_Pilani_Team2@ POS Tagging for Code Mixed Indian Social Media. In International Conference on Natural Language Processing.

[43] Prabhu, A., Joshi, A., Shrivastava, M., & Varma, V. (2016). Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. arXiv preprint arXiv:1611.00472.

[44] Sane, S. R., Tripathi, S., Sane, K. R., & Mamidi, R. (2019, June). Deep learning techniques for humor detection in Hindi- English code-mixed tweets. In Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 57-61).

[45] Ghosh, S., Ghosh, S., & Das, D. (2017). Sentiment identification in code-mixed social media text. arXiv preprint arXiv:1707.01184.

[46] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[47] Aparaschivei, L., Palihovici, A., & Gifu, D. (2020, December). FII-UAIC at SemEval-2020 Task 9: Sentiment analysis for code-mixed social media text using cnn. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 928-933).

[48] Shalini, K., Ganesh, H. B., Kumar, M. A., & Soman, K. P. (2018, September). Sentiment analysis for code-mixed indian social media text with distributed representation. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1126-1131). IEEE.

[49] Lal, Y. K., Kumar, V., Dhar, M., Shrivastava, M., & Koehn, P. (2019, July). De-mixing sentiment from code-mixed text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 371-377).

[50] Sasidhar, T. T., Premjith, B., & Soman, K. P. (2020). Emotion Detection in Hinglish (Hindi+ English) Code-Mixed Social Media Text. Procedia Computer Science, 171, 1346-1352.

[51] Singh, Vinay, Aman Varshney, Syed Sarfaraz Akhtar, Deepanshu Vijay, and Manish Shrivastava.(2018) ”Aggression detection on social media text using deep neural networks.” Proceedings of the 2nd Workshop on Abusive Language Online (ALW2): 43-50.

[52] Kamble, Satyajit, and Aditya Joshi.(2018) ”Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models.” arXiv preprint arXiv:1811.05145.

[53] Patel, R. N., Pimpale, P. B., & Sasikumar, M. (2016). Recurrent neural network based part-of-speech tagger for code- mixed social media text. arXiv preprint arXiv:1611.04989.

[54] Fausett, L. (1994). Fundamentals of Neural Networks Prentice Hall. Englewood Cliffs, NJ, 7632.

[55] Banerjee, S., Ghannay, S., Rosset, S., Vilnat, A., & Rosso, P. (2020). LIMSI_UPV at SemEval-2020 Task 9: Recurrent Convolutional Neural Network for Code-mixed Sentiment Analysis. arXiv preprint arXiv:2008.13173.

[56] Kumar, V., & Dhar, M. (2018). Looking Beyond the Obvious: Code-Mixed Sentiment Analysis (CMSA).

[57] Baroi, S. J., Singh, N., Das, R., & Singh, T. D. (2020, December). NITS-Hinglish-SentiMix at SemEval-2020 Task 9: Sentiment Analysis for Code-Mixed Social Media Text Using an Ensemble Model. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1298-1303).

[58] Veena, P. V., Anand Kumar, M., & Soman, K. P. (2018). Character embedding for language identification in Hindi-English code-mixed social media text. Computación y Sistemas, 22(1), 65-74.

[59] Si, S., Datta, A., Banerjee, S., & Naskar, S. K. (2019, July). Aggression detection on multilingual social media text. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[60] Soman, K. P. AMRITA_CEN@ ICON-2015: Part-of-Speech Tagging on Indian Language Mixed Scripts in Social Media. In 12th International Conference on Natural Language Processing.

[61] Lakshmi, B. S., & Shambhavi, B. R. (2017, December). An automatic language identification system for code-mixed English- Kannada social media text. In 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS) (pp. 1-5). IEEE.

[62] Vijay, D., Bohra, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). Corpus creation and emotion prediction for hindi-english code-mixed social media text. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (pp. 128-135).

[63] Pravalika, A., Oza, V., Meghana, N. P., & Kamath, S. S. (2017, July). Domain-specific sentiment analysis approaches for code-mixed social network data. In 2017 8th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-6). IEEE.

[64] Vijay, D., Bohra, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A Dataset for Detecting Irony in Hindi- English Code- Mixed Social Media Text. EMSASW@ ESWC, 2111, 38-46.

[65] Wu, Q., Wang, P., & Huang, C. (2020). MeisterMorxrc at SemEval-2020 Task 9: Fine-tune bert and multitask learning for sentiment analysis of code-mixed tweets. arXiv preprint arXiv:2101.03028.

[66] Bhange, M., & Kasliwal, N. (2020). HinglishNLP: Fine-tuned Language Models for Hinglish Sentiment Detection. arXiv preprint arXiv:2008.09820.

[67] Sharma, A., & Motlani, R. (2015, December). POS tagging for code-mixed Indian social media text: Systems from iiit-h for icon NLP tools contest. In International Conference On Natural Language Processing.

[68] Sarkar, K. (2016). Part-of-speech tagging for code-mixed Indian social media text at ICON 2015. arXiv preprint arXiv:1601.01195.

[69] Choudhary, N., Singh, R., Bindlish, I., & Shrivastava, M. (2018). Sentiment analysis of code-mixed languages leveraging resource rich languages. arXiv preprint arXiv:1804.00806.

[70] Singh, K., Sen, I., & Kumaraguru, P. (2018, July). Language identification and named entity recognition in hinglish code mixed tweets. In Proceedings of ACL 2018, Student Research Workshop (pp. 52-58).

[71] Jamatia, A., Swamy, S., Gambäck, B., Das, A., & Debbarma, S. (2020). Deep Learning Based Sentiment Analysis in a Code- Mixed English-Hindi and English-Bengali Social Media Corpus. International journal on artificial intelligence tools, 29(5).

[72] Raha, T., Mahata, S. K., Das, D., & Bandyopadhyay, S. (2020). Development of POS tagger for English-Bengali Code- Mixed data. arXiv preprint arXiv:2007.14576.

[73] Parikh, A., Bisht, A. S., & Majumder, P. (2020, December). IRLab_DAIICT at SemEval-2020 Task 9: Machine Learning and Deep Learning Methods for Sentiment Analysis of Code-Mixed Tweets. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1265-1269).

[74] Pratapa, A., Choudhury, M., & Sitaram, S. (2018). Word embeddings for code-mixed language processing. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 3067-3072).

[75] Kumar, V., Pasari, S., Patil, V. P., & Seniaray, S. (2020, July). Machine Learning based Language Modelling of Code Switched Data. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 552-557). IEEE.

[76] Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). A dataset of Hindi-English code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people’s opinions, personality, and emotions in social media (pp. 36-41).

[77] Prabhu, Ameya, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. (2016) "Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text." arXiv preprint arXiv :1611.00472.

[78] Dahiya, A., Battan, N., Shrivastava, M., & Sharma, D. M. (2019, August). Curriculum Learning Strategies for Hindi-English Code-Mixed Sentiment Analysis. In International Joint Conference on Artificial Intelligence (pp. 177-189). Springer, Cham.

[79] Gopal Jhanwar, M., & Das, A. (2018). An Ensemble Model for Sentiment Analysis of Hindi-English Code-Mixed Data. arXiv e-prints, arXiv-1806.

[80] Singh, P., & Lefever, E. (2020, May). Sentiment Analysis for Hinglish Code-mixed Tweets by means of Cross-lingual Word Embedding's. In Proceedings of the The 4th Workshop on Computational Approaches to Code Switching (pp. 45-51).

[81] Santosh, T. Y. S. S., & Aravind, K. V. S. (2019, January). Hate speech detection in Hindi-English code-mixed social media text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (pp. 310- 313).

[82] Sreelakshmi, K., Premjith, B., & Soman, K. P. (2020). Detection of Hate Speech Text in Hindi-English Code-mixed Data. Procedia Computer Science, 171, 737-744.

[83] Chakma, K., & Das, A. (2016). Cmir: A corpus for evaluation of code mixed information retrieval of hindi-english tweets. Computación y Sistemas, 20(3), 425-434.

[84] Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014, October). "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 (pp. 116-126).

[85] Somnath Banerjee, Kunal Chakma, Sudip Kumar Naskar, Amitava Das, Paolo Rosso, Sivaji Bandyopadhyay, and Monojit Choudhury. 2016. Overview of the mixed script information retrieval (MSIR). In Proceedings of FIRE 2016. FIRE, December.

[86] Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., ... & Fung, P. (2014, October). Overview for the first shared task on language identification in code-switched data. In Proceedings of the First Workshop on Computational Approaches to Code Switching (pp. 62-72).

[87] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M.,. & Eryiğit,(2016, January). Semeval-2016 task 5: Aspect based sentiment analysis. In International workshop on semantic evaluation (pp. 19-30).