International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

"Sentiment Analysis of Code-Mixed Social Media Text in Indian-Languages"

Gazi Imtiyaz Ahmad, Syed Ishfaq Manzoor

Submitted:02/11/2024 **Revised:**18/12/2024 **Accepted:**28/12/2024

Abstract: The arrival of web 2.0 platforms and increasing usage of social networking sites have proliferated social media content on the web. These platforms also provides multilingual interface to allow people to write freely in their native language.Over the past few decades, a new phenomenon called "code-mixing" has been observed in social media data which has attracted attention of researchers in sociolinguists and Natural Language Processing domains. However, due to informal nature of the text present in code-mixing phenomenon, there are a number of challenges ranging from data extraction to summarization. Sentiment Analysis of code-mixed data is one of the key research field which has emerged in recent past. Sentiment Analysis is the combination of application areas such as Natural Language Processing, Statistical methods and linguistics that classify a document or a sentence into positive, negative and neutral categories. These classes represent opinions/views of a person about a product, service, an event, a social movement, a political issue or a government policy. Extracting such useful information from unstructured data has applications in business, marketing, commerce, travel, finance, healthcare, politics etc. The goal is to provide natural language processing (NLP) tools that can collect, analyzed, evaluate, and summarize CM ("code-mixed") data. The researchers had to deal with dataset construction, preprocessing, annotation, language identification, feature extraction and feature selection, and sentiment classification when it came to sentiment analysis of CMSMT ("Code-mixed Social Media Text"). This paper provides a detailed overview of work carried out in these challenges which will help researchers of this field in their future directions.

Key words: *Natural Language Processing, Machine Learning, Sentiment Analysis, Code-mixed, Datasets, Language Identification, Named Entity Recognition.*

1. Introduction

Internet Social Networking (ISN) tools and web 2.0 technologies have revolutionized the world in the communication and information technology. These tools and technologies have provided a new way of information sharing and knowledge dissemination. [1] Traditional digital communications such as email has been replaced by online social networking platforms which makes the communication more informal [2]. People now use social networking sites to express their opinions, views, ratings regarding products, services, social and political issues, government policies and much more. Therefore a data repository consists of user opinions and views has been created on the Internet and the size of this repository is increasing exponentially [3]. Social Networking Sites provides cost-effective and time saving environment for exchanges of ideas and suggestions and allows for global exposure instantaneously. However, users around the world speak different languages and wish to express their views and propositions in their mother language. Therefore, there is an immense rise of multilingualism in social networking platforms. Translation features introduced by Facebook, Instagram etc. display text in different language formats. On social media platforms netizens also prefer to use roman script instead of Unicode to express their views or frequently insertion English word or phrases within their native language script, hence use mix of two or multiple languages within a sentence or sentences. This phenomenon gave rise to the code-mixing data on social networking sites. Natural Language Processing (NLP) deals with derivation of meaningful information from natural languages text by enabling computers to use knowledge of computer science, AI and linguistics. Over the past three decades Natural Language Processing (NLP) of social media text in Standard English language has been a popular research area.

This involves information extraction, categorization, indexing, summarization and translation [4]. Researchers are working to build tools and approaches that will allow computers to interpret and manipulate natural languages in order to accomplish useful activities [5]. Sentiment analysis is a sub-task of natural language processing that deals with identifying and extracting subjective information from text. The aim is to identify polarity of a sentence or a document into positive, negative and neutral. [6][7].However, over the last decade, NLP research has also been shifted to develop systems for code-mixing social media text. Although the research is in its infancy stage, considerable work has been carried out throughout the globe, especially in India, where a large volume of code-mixed text is generated on social media networks because India is a home of surfeit of languages.

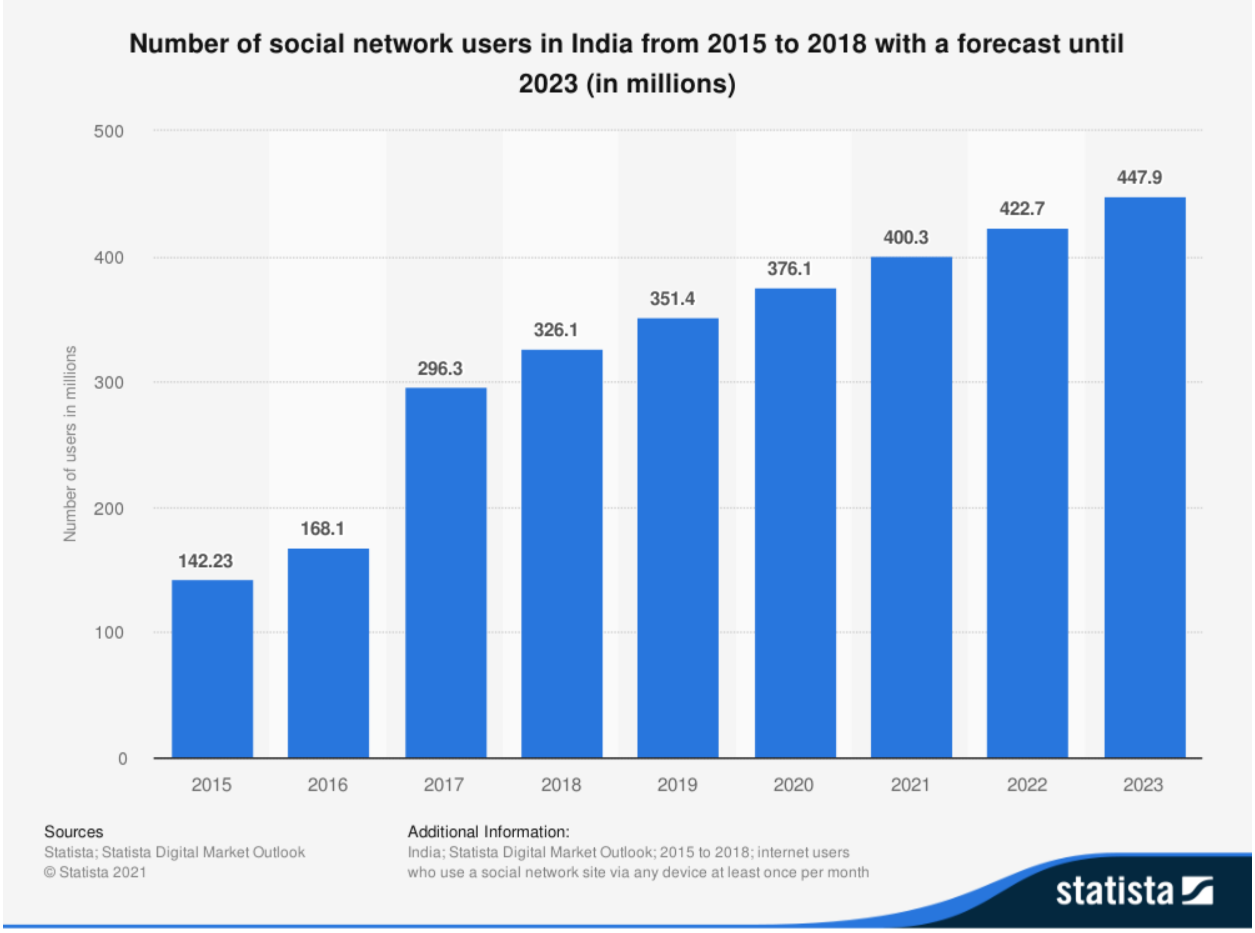
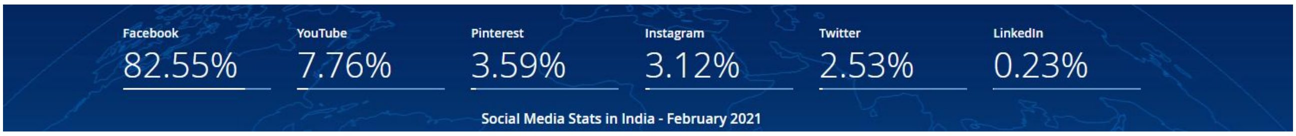


Figure 1: Social Media Users in India (Source www.statista.com)

Social Networking Sites (SNS) plays an important role in everyone's life. These sites allow users to connect, share ideas, build relationships, share pictures and videos and collaborate on social issues. India is the second largest populous country in the world and as per the report published by Digital 2021: India "(https://datareportal.com/reports/digital-2021-india)" as on January 2021, there are 624 million Internet Users in India as against 577 million Internet users in January 2020 which is an increase of 8.2% (47 million). Also there are 448 million social media users as on January 20201 which is an increase of 21% (78 million) from the last year. Therefore, number of Social Media users in India is about 32.3% of the total population. Therefore about two-third of Indians are using social networking sites like Facebook, Twitter, YouTube, Pinterest etc.



Facebook	YouTube	Pinterest	Instagram	Twitter	LinkedIn
82.55%	7.76%	3.59%	3.12%	2.53%	0.23%

Social Media Stats in India - February 2021

Figure 2: Top Social Media Sites in India (Source: gs.statcounter.com/social-media-stats/all/india)

Researchers have carried a lot of work on analysis of social media content throughout the world. English is the dominant language on the social media platforms, however, as analyzed by Hong et al [8], its dominance is declining. The authors have identified top ten popular languages of social media used in Twitter and apply their automatic language detection algorithm on more than sixty million tweets to show that only 51% of tweets are in English language..Automatic understanding of social media text in English language has been well studied and a number of linguistic and AI tools are available. As for as sentiment analysis of social media text in English language is concerned, a number of linguistic tools such as datasets, dictionaries, WordNets and SentiWordNets have been developed over many decades. Also NLP tools, supervised, unsupervised and semi-supervised ML approaches and DL techniques have been developed with reasonable accuracy.

Social media content is characterized as informal or noisy because of number of reasons including spelling mistakes, use of Meta tags, abbreviations, and use of more than one language or using roman script instead of Unicode. Also people use code-mixing in their conversations on SNS. In India people of different regions and cultures also use mixing languages on social media platforms which results in accumulation of large volume of code-mixed data on these sites.Therefore, mining of social media code-mixed content to deduce useful information is an important field of research. Code-mixing refers to the mixing of two or more languages or language units in a single verbal or written communication. On social media platforms, English is used with other native language(s) to express opinions. The non-English language is written in roman script. In case of written communication, people use script of only one language and write words or phrases of other language(s) in the script of dominant language. In bilingual or multilingual societies especially in India code-mixing is a natural phenomenon on the social media platforms. People often use code-mixing for a number of reasons including restricted vocabulary, for role and style identification, inability of expression, to convey special attitude, emotional arousal, ethnic identity, showoff etc. Specific mixed languages have been given specific names in code-mixed scenario. For example Hinglish (for Hindi-English), Poglish (Polish-English), Spanglish (Spanish-English), Tanglish (Tamil –English) etc. The availability of large volume of code-mixed data on social networking sites especially code-mixed data in Indian languages has provided researchers to explore the research to infer useful information in this field. This paper presents a detailed review of approaches used in sentiment analysis of code-mixed data.

2. Motivation for Research

People especially in India share their views almost on everything including products and services, movies and film stars, politicians and elections, government and social events etc. it has been observed that more than half of the tweets in India are code-mixed containing more than one language [9]. English-Hindi code-mixed text a.k.a. Hinglish had become the lingua franca for Indian netizens [10]. Therefore, a large volume of text in code-mixed form is available on social networking sites which encourages us to explore the research in this area.

The SA of CMSMT is an emerging field of research. Therefore this study presents various datasets, preprocessing linguistic and lexical resources, code- tools and techniques for SA of Indian language CMT. The paper also presents a summarized study of comprehensive and methodical search in the area of sentiment analysis of Indian language code-mixed text.

3. Article Structure

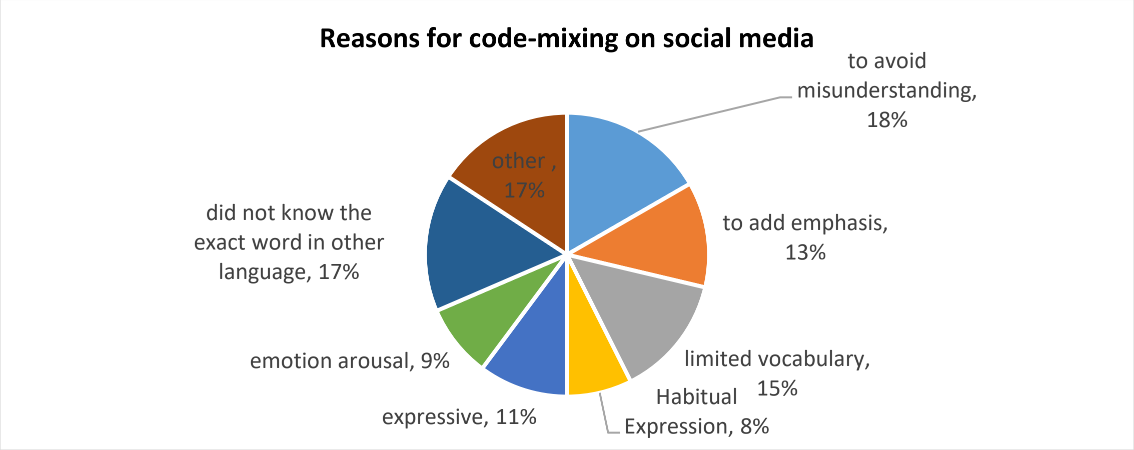
This structure of the paper is as follows: section 4 provided background and related work. The review methodology is presented in section 5. Section 6 and section 7 discuss about the sentiment analysis process and sentiment analysis approaches respectively. A detailed overview of status of SA for CMSMT in the context of Indian language pairs is present in section 8 and section 9. Section 10 discusses the survey findings. The conclusion ins presented in Section 11.

4. Background and Related Work

Having competence in communicating in more than one language broadens the prospects of individuals to express their feelings and thoughts and shape their identity. It also helps them to fulfill their social as well as individual needs in the context of languages used. Research scholars have examined how code-mixing and code-switching occurs and have been fascinated by the phenomenon [11][12].Code-mixing is defined as linguistic items and grammatical features from more than one language in a single sentence. [13] defines

code-mixing as the mixing of multiple linguistic elements (morphemes, words, modifiers, phrases, clauses, and sentences) from two participating grammar systems in a single sentence. Researchers have assumed that one of the languages in code-mixing acts as a base or native language also known as matrix language which provides morpho-syntactic structure, and a foreign language also known as embedding languages whose semantic units are inserted into the matrix language. Code-mixing is motivated by social and psychological factors. Although code-switching and code-mixing are used interchangeably, code-mixing is intra-sentential and controlled by grammatical principles, whereas code-switching is inter-sentential. Code-mixing process involves insertion when lexical items of one language are incorporated in a structure of another language, alternation when a speaker switches languages or language units in the context of single conversation and congruent lexicalization occurs when a grammatical structure is shared by two or more languages and can be lexically filled with elements from either language. [14] provided a detailed account on the causes for code mixing and code switching, as well as their motives.

In bilingual and multilingual societies, netizens often prefer to use more than one language while communicating on Internet especially on social networking platforms as code-mixing is prominent in informal communications than formal one. The study carried out by [15-22] suggested that code-mixing often occurs in short messages and mixing often takes place through simple insertions and there are a number of reasons for code-mixing which include social needs, to avoid misunderstanding, to add emphasis, easier to write in native language, limited vocabulary, habitual expressions, expressive, emotion arousal, gender (female tend to use mixed language more often than male) age group etc.



is difficult to provide exact reasons of code-mixing that happens in the world. It has been estimated that approximately 4% of the tweets on Twitter are code-mixed. The above pie-chart show some reasons for code-mixing on social media.

Social media text, therefore, contains considerable amount of code-mixed text. Sentiment Analysis systems for monolingual text usually ignore code-mixed text and often categorize it as noise. Sentiment analysis systems for monolingual text has been carried out from past three decades and new models are being developed to improve the efficiency of existing one. Cross-lingual models have also been developed and used. However, these models cannot be applied on code-mixed text. Therefore, in recent past SA of CMSMT has been an interesting topic of research for the scholars, Industry people and business units.

SA of monolingual text has been well studied by researchers over the last three decades. Medhat et.al. in[23] presented a detailed survey on sentiment analysis and applications. Every year numerous articles are presented in the field of SA. The interest in non-English languages in the field of sentiment analysis is also increasing. Rani & Kumar in[24] presented detailed survey about sentiment analysis of Indian languages. The paper introduced insights concerning approaches, type and size of corpora, lexical assets, instruments utilized and assessment measures for almost every Indian language. The paper provides systematic survey of 59 research papers published between 2010-2017on SA of Indian languages. In the recent past, SA of CMSMT has also gained momentum as the interest of researchers is persistently growing in this field. As a result, a systematic review is required to appraise, improve, and absorb the state-of-the-art research offered in this domain. Tho, Cuk, et al. in [25] presented a systematic literature review of SA of CMSMT using ML approach. However, they did focus on classifiers only. Other reviews include [26] which focus on cross-domain sentiment analysis and [27]presented review of soft computing approaches of sentiment analysis.

5. Review Methodology

The review was carried out by searching electronic databases such as Google Scholar (www.scholar.google.com), Science Direct (www.sciencedirect.com), ACM Digital Library (www.acm.org/dl), and IEEE Explore (www.ieeexplore.ieee.org) for relevant research publications. The details of papers selected for review are given in Table 1.

Table 1: Papers selected for the review

Journal papers	Conference Papers	Workshop Papers
35	45	20

Task	No. of Papers
Sentiment Analysis	47
Language Identification	17
Language Translation	09
POS Tagging	12
Named Entity Recognition	04
Corpus Development	11

6. Sentiment Analysis (SA) Process

SA is a NLP task involving analysis of natural language to identify emotion with respect to the topic expressed in a piece of text. It is the task of opinion extraction about specific entities [28]. In a piece of text each sentence may or may not contain sentiment. A sentence may be objective or a subjective. Analysis of subjectivity is the main idea behind sentiment analysis. Sentiment analysis of single language text has been studied thoroughly. However, on social media platforms, the text involves more than one language especially in multi-lingual societies. People also prefer to use roman script for non-English languages make the task of Sentiment Analysis more complicated. SA of CMSMT is a growing field of study that has gotten a lot of interest in recent years. Sentiment Analysis of Text entails a number of steps, beginning with data collection, preprocessing, sentiment detection, and classification, and ending with the display of results.

Data collection for text analysis is the primary phase of the sentiment analysis. The data collection tools and techniques plays an important role in the quality of dataset. In case of code-mixed data, collection process is difficult as compared to uni-language text. The code-mixed data is not available for all Natural languages. Datasets mostly are collected from social media either scrapped or crawled. For sentiment analysis task the data is usually annotated by subject experts. Tweets, Facebook posts, WhatsApp messages, YouTube comments and movie reviews are usually used as datasets in code-mixed sentiment analysis. The researchers also use manual annotation of text at word level. Table 2 provides an overview of datasets developed for code-mixed Indian Languages.

Table 2: Code-mixed datasets in Indian languages

Languages	Domain	Corpus size	Reference
Hindi-English	Tweets	4575	[102]
Hindi-English	Tweets	5250	[103]
Hindi-English	Facebook posts	1000	[104]
Hindi-English	Tweets Facebook posts Whatsapp chat	1077	Singh K. et. al. [105]
Hindi-English Bengali-English Gujarati-English Tamil-English	DSTC2 Restaurant reservation dataset (The Second Dialog State Tracking Challenge)	Hi-En 386 Be-En 360 Gu-En 387 Ta-En 424	Banerjee, Suman, et al.[106]
Hindi-English Bengali-English	NLP Tool Contest @ICON-2017	Hi-En 10995 Bn-En 2125	Mishra, P., et. al. [107]
Hindi-English Bengali-English Telugu-English	Tweets Facebook Comments WhatsApp chat	Hi-En: 1000 Bei-En: 1000 Te-En: 1000	Jamatia, A et. al. [108]
Tamil-English	YouTube Comments	15744	Chakravarthi, Bharathi Raja, et al [109]
Malayalam-English	YouTube comments	6738	Chakravarthi, Bharathi Raja, et al [109]
Punjabi-English	Social media posts	4812	Yadav, Konark, et al[110]

The data collected from social media platforms is often noisy and therefore, need to be cleaned before using for system development. Text preprocessing plays an important role in sentiment analysis techniques and application. Processing of text reduces file size, remove stop words and stem words to their root word. The most common text processing steps include tokenization, stop word removal and stemming.

Tokenization: A common task in NLP, tokenization involves separating a piece of text into smaller units such as words, sub-words or characters called tokens. Tokenization is must for ML models as they process text at the token level. It is useful in linguistics and in computer science to identify expressive words [29].

"Stop Word Removal": Stop words are most common words found in a text

document which do not contribute to the content or context of the textual document and are filtered out in an NLP task [30]. These words such as ‘a’, ‘an’, ‘the’, ‘this’ etc. are usually used to join words together in a sentence and does not bear meaning. In a sentiment analysis system these words are removed as they are not useful in classification. There is no universal list of stop words that are common in all NLP tasks.

Stemming: Stemming is the process of extraction of base form of a word by removing affixes and suffixes from it. As a result, the various versions of a word are merged into a single representation. For Example the words consult, consults, consultant, consultants, consulting, consultative have the common root word ‘consult’. Stemming is usually used in search engines for information retrieval. Stemming is used to reduce the dimensionality of data. Algorithms like porter stemmer and snowball stemmer are used in NLP tasks. Lemmatization is also used to chop the word to its root form but unlike stemming, lemmatization takes context of the word in the sentence into consideration before processing. A detailed information about stemming and lemmatization has been provided in [31] and [32].

An important task of NLP in SA of CMSMT is the identification of language. The technique of determining the natural language of a word or a document is known as language identification. A number of approaches have been proposed for automatic language identificationin monolingual documents. A detailed survey on automatic language identification of text is being provided in [33]. However automatic LI of code-mixed text is an incomplete problem. Various methods have been proposed for identification of languages in a code-mixed textual data.

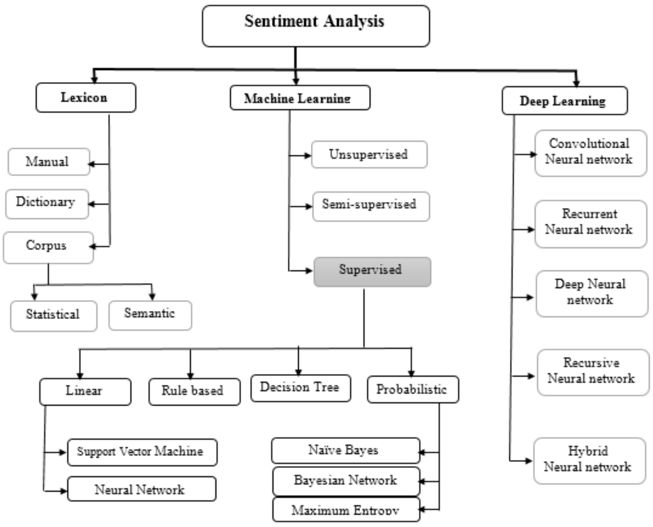


Figure 4: “Sentiment Analysis (SA) Techniques”

7. Approaches of Sentiment Analysis.

SA Approaches can be categorized as:

Lexicon based Approach

Also called rule base approach where rules are being followed to predict the sentiment of a textual document using sentiment lexicons. Sentiment lexicon is a collection of words along with their sentiment orientations

[34.]. The sentiment of words in an unknown document is calculated by comparing the words with the words in a sentiment lexicon and using an algorithm the overall sentiment orientation of the document or sentence is computed by aggregation of sentiment of individual words [24]. For construction of lexicon, manual, dictionary based and corpus based approaches are followed. While manual construction is time consuming and costly, dictionary and corpus based lexicons provide better accuracy. While the lexicon approach is easy to implement, the over-fitting problem persist and is unavoidable. Many open source sentiment lexicons for languages other than English are available. In the context of Indian languages many approaches such as dictionary, corpus and WordNet based approach have been used to develop SentiWordNets [35]. Hindi WordNet, created in 2006 and made available for research, paved way for creation of WordNets for major Indian languages. IndoWordNet, for example, is a linked framework comprising 19 main Indian languages belonging to the Indo-Aryan, Dravidian, and Sino-Tibetan families.

Table 3: SentiWordNets for Indian Languages

Language	No. of Synsets	Lexical Resource Used	Reference
Bengali	35805	English Sentiment Lexicon and English-Bengali Bilingual dictionary	Das &Bandyopadhyay[92]
Hindi	8061	HidniWordNet	Sharma & Bhattacharyya [93]
Hindi	16000	English SentiWordNet and English-Hindi WordNet	Bakliwal et. al. [94]
Hindi	22708	English-Hindi Electronic dictionaries (SHABDKOSH5 and Shabdanjali)	Das &Bandyopadhyay[95]
Telugu	27182	"Charles Philip Brown English-Telugu Dictionary , Aksharamala English-Telugu Dictionary and English-Telugu Dictionary developed by Language Technology Research Center (LTRC), International Institute of Hyderabad (IITH)"	Das &Bandyopadhyay[95]
Tamil	9495	English SentiWordNet AFINN111 lexicon, Subjectivity Lexicon, Opinion Lexicon and Google Translate.	Kannan et.al. [96]
Odia	4747	SentiWordNets Bengali, Tamil and Telugu	Mohanty et. al.[97]
Punjabi	19,011	Punjabi WordNet, Hindi SentiWordNet, Punjabi-Hindi Dictionary, English SentiWordNet, and English-Punjabi Dictionary	Diksha et. al.[98]
Urdu	9398	"Bing Liu's List of Opinion Words, English SentiWordNet, English to Urdu bilingual dictionary and Urdu Opinion Lexicon"	Asghar, Muhammad Zubair, et al.[99]
Kannada	5043	Hindi Sentiwordnet	Deepamala&Ramakanth[100].
Malayalam	2000	Hindi Sentiwordnet	Anagha, M., et al. [101]

Machine Learning (ML) Approach

ML is a form of AI ("Artificial Intelligence") that learns from data, experience and patterns to make better decisions with minimal human intervention [36]. Machine learning allows us to build systems that improves their performance through experience. Machine learning is an important discipline that solves the fundamental problems of science and engineering [37]. A diverse array of ML algorithms has been developed over the years to solve problems across different domains using a wide variety of data. Opinion mining and sentiment analysis has seen rapid growth over the past two decades. Sentiment analysis is a machine learning tool to identify polarity viz. 'positive', 'negative' and 'neutral' in a piece of text. Machine learning tools automatically learn how to identify sentiment in text using examples of emotions present in the text without human intervention.

ML algorithms and systems have dominated the field of sentiment analysis. Machine learning models provide better accuracy as compared to tradition lexicon based approach. However, performance of machine learning methods for sentiment analysis is constrained by how effectively features have been engineered [38] ML algorithms are classified as supervised, unsupervised and semi-supervised learning algorithms based on the nature of training data.

Supervised learning algorithms works on the principle of labelled training data. Given a collection of training data in the form of f(x, y), the goal of supervised learning algorithm is to predict y in response to an input x using a learned mapping of f(x). Supervised learning algorithms are commonly used for classification problems such as spam detection, face recognition, sentiment analysis, medical diagnosis etc. Algorithms like Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression, Decision Tree, Random Forest, Bagged and Boosted trees, Neural Networks and Memory based learning are some example of supervised learning algorithms[39].

In unsupervised learning, the training data is unlabeled and the machine learning algorithm based on the patterns, structures and other properties analyze the input data. [40]. Given a set of observations x1,x2,x3….xn, the goal is to deduce the properties of these observations without any external help and provide correct relation of each observation [41]. The study of how systems should learn to understand specific input patterns in a way that represents the overall statistical structure of the input patterns is known as unsupervised learning. Clustering, Neural networks, KNN (K-nearest neighbor) are some examples of unsupervised learning.

The study of how machines and natural systems, such as people, learn in the presence of both labeled and unlabeled data is known as semi-supervised learning. The goal of semi-supervised learning is to figure out how a mix of labeled and unlabeled data affects learning behavior and to create algorithms and models that take advantage of it. [42] Semi-supervised learning is common in machine learning and data mining because it can enhance supervised learning tasks with readily accessible unlabeled data when labelled data is limited or costly. Semi-supervised learning can be self-training, co-training and active learning depends on how the data is labelled.

Deep Learning (DL) Approach

DL is a type of ML that employs artificial neural networks to learn multiple levels of data representations automatically. [43] DL is a form of ML that allows computers to learn from their experience and comprehend the world as a hierarchy of concepts. [44]. DL is a representation learning technique that use nonlinear neural networks to learn many layers of representation, each of which turns the representation at one level into a higher, more abstract representation. DL is a form of machine learning that employs multiple algorithms in a sequential sequence of events to solve complex problems. It enables us to process large quantities of data accurately and with minimal human intervention. [45.] Deep learning employs CNNs (Convolutional Neural Networks), RNNs (Recurrent Neural Networks), Recursive Neural Networks, DBNs (Deep Belief Networks), and a variety of other networks. Neural networks are used to produce text creation, vector representation, word representation estimation, sentence classification, sentence modeling, and feature display[46]. In recent years, deep learning methods have shown promising results on sentiment analysis tasks, because it is influential in both unsupervised and supervised learning. Using deep learning sentiment analysis can be done in more efficient and effective way. Deep learning models have shown better accuracy than traditional machine learning models. A visible benefit of deep learning over traditional machine is automatic feature generation which saves both time and cost.

8. “Sentiment Analysis of Code-mixed Indian Social Media Text” (SA of CMSMT)

8.1 Code-mixing

Code-mixing refers to the use of two or more languages in a sentence while communicating via speech or text [47].People who know more than one language often use one or more words from one language and introduce it while communicating in another language. In bilingual societies such as in India, people often know more than one language and on social networking platforms they use words of one language and introduce it in their native language while expressing their views. The reason behind using more than one language via oral or written communication is the language dependency which is categorized as development of new mixed codes of communication [48]. On social networking platforms, code-mixing is usually dominated by one language and linguistic units from one or more languages are inserted to give emphasis on certain event or social movement or for clarification without changing the topic of the statement.

Example

"Bhai! wake up"

Translation: Brother wake up.

Example

"Main aaj friend kaysath movie dekhnayjawonga"

Translation: I will go to see movie along with my friend.

9. Status of work on code-mixed Indian Languages

In multilingual societies people prefer to use more than one language while expressing their ideas or opinions on some social or political movement, product and service reviews etc. From the last decade a new phenomenon called code-mixingemerged on social networking sites where in people often use one language (preferably English) and mix words or phrases of their native language or other language(s) in roman script in an english sentence. This new form of text on social media platforms and websites has attracted attention of NLP researchers to mine useful information from it. The task becomes more difficult as the code-mixed text is informal in nature and having a number of other issues such as spelling variations, use of acronyms and other shortcut words etc. In the field of SA for CMSMT, the unavailability of lexicon and language resources, the task of identification of polarity becomes a herculean one. With the development of NLP resources, the work on SA of CMSMT has also gained momentum. The work carried out in SA of code-mixed Indian language text is given as follows.

9.1. Hindi-English

Hindi is the official and most spoken language in India. Therefore noteworthy research has also been done in Hindi. Bali, Kalika, et al. in [49] provides a detailed analysis of Facebook data generated by bilingual users of Hindi and English. The study showed that a substantial amount of data on Facebook is code-mixed as people in bilingual societies prefer to write their feeling in more than one language. It has also been observed that while the mixing of Hindi words in English followed fixed patterns of Nouns and Particles, mixing of English words in Hindi occurs at different levels in the sentence.

One of the initial tasks for SA of Hindi-English CMSMT was performed by Joshi et al.[50]. The authors proposed a sub-word level LSTM model for Hindi-English code-mixed text collected from Facebook pages of popular personalities of India. Using sub-word level architecture, the model uses polarity value of morphemes as well. The model achieved an accuracy of 69.7%.

A lexicon based approach for SA of English-Hindi CMSMT was presented by Sharma et alin [*51*] The data was obtained from FIRE 2013 and FIRE 2014 shared task. The authors separated English and Hindi words and translated the Hindi words into Devanagari script. For assigning the polarity to words the authors used Opinion lexicon, Hindi SentiWordNet and AFFIN. The proposed model claimed an accuracy of 85%.

Gaurav S [52] performed sentiment analysis on Hindi-English code-mixed Tweets collected from SemEval-2020 into 'positive', 'negative' and 'neutral' classes. The techniques used include SVM, NB, DT, LR and RF. The experimental results showed that ensemble voting classifier outperforms the other classifiers by achieving an F1-score of 0.6907.

Mishra P et. al [53]used an ensemble model of ML approach and Neural Networks for SA of Hindi-English CMSMT. The model used TF-IDF feature vectors of character n-grams to achieve best F1-Score of 0.58.

A SA system for Hindi-English code-mixed Twitter data was designed by Baroi, SubhraJyoti, et al. [54]. The data was collected from SemEval-2020 shared task and achieved an accuracy of 0.617. The researchers used four individual models viz. LSTM, LSTM+ConV, BiLSTM ,CNN+Dense and an ensemble model to perform experiments. The experimental results showed that ensemble model performs better than the individual models.

Das and Gamback [55] presented a study of characteristics of code-mixed text on social media and also developed an automatic language identification system. The dataset used for the purpose was obtained from Facebook in mixed Hind-English and Bengali-English. The annotation work was done manually and machine learning approach viz. SVM using character n-gram features was used for the experimentation. The proposed system achieved an accuracy of 80%.

Vijay, Deepanshu, et al. in [56]presented an emotion prediction system for Hindi-English CMSMT obtained from microblogging site Twitter. The system used supervised machine learning approach comprises of a number of ML algorithms for detection and classification of emotions into a 6-class classification by employing character level, word level and lexicon based features.

A hybrid system that includes lexicon based approach and ML approach for SA of Hindi-English code-mixed twitter data was proposed by Pravalika, A., et al in [57] The proposed system take linguistic code-switching and grammatical transitions between Hindi and English languages into account for better performance.

Kumar &Dhar [58] proposed a hybrid approach for SA of Hindi-English CMSMT. The data was collected from popular Facebook pages of India. The hybrid system used twoBiLSTM networks for overall sentiment of the sentence as well as for individual words and sub-words that bear sentiments. Orthogonal features and word embedding features were used for better performance of the system. The proposed system achieved an accuracy of 83.5% and F1-score of 0.827

A novel approach for SA of CMSMT for classification of sentences into 3-class classification of 'positive', 'negative' and 'neutral' was proposed by Choudhary, Nurendra, et al in [59]. The proposed model uses Siamese networks for mapping the code-mixed and standard language sentences into a common sentiment space as well as basic clustering based preprocessing method for capturing of variations of code-mixed transliterated words. The proposed model achieved an accuracy of 77.3% and F-score of 0.759.

Ansari &Govikar [60] designed system to classify Hindi and Marathi Romanized text using SVM, Naïve Bayes and ontology based classification. The proposed system achieved an accuracy of 90% for code-mixed Marathi-English and 80% for Hindi-English CMSMT.

One of the challenging task in SA of CMSMT is the unavailability of standard annotated datasets and classification models. For this Sasidharet. al. [61] created a Hindi-English code-mixed dataset and annotated with Happy, Sad and Anger. Bilingual pre-trained models were used for generation of feature vectors and deep neural networks were used for classification. The authors achieved an accuracy of 83.21% with CNN-BiLSTMclassification model.

An ensemble model comprising of character-trigram based LSTM model and word-ngram based Multinomial Naïve Bayes model for sentiment analysis of Hindi-English code-mixed social media text was proposed by Jhanwar& Das [62.]. While the LSTM model encrypts the deep sequential patterns in text, Multinomial

Naïve Bayes works for grammatical inconsistencies by capturing word combinations of keyword. The proposed model achieved an accuracy of 70.8%.

A study for development of Hindi-English CMSMT of sentiment words has been proposed by [63]. The authors classify the text into 5-class categories viz. Highly Positive, Positive, Negative, Highly Negative and Neutral. The annotated dataset created also handles inconsistent and misspelled words.

A sentiment analysis system for Hindi-English code-mixed text using machine learning algorithms also with SVR and Grid Search was proposed by Garg& Sharma [64]. The dataset comprising of tweets was provided in SemEval-2020 shared task. Feature extraction algorithms such as bag-of-words and GloVe vectors were used. The proposed model achieved an f1- score of 0.662.

Text normalization of Romanized Hindi-English CMSMT was proposed by Sharma et. al. [65]. The authors proposed the identification and normalization of text which also include addressing of various inconsistent and misspelled and slang words. The authors also used lexicon approach for classification of normalized text sentences into positive and negative. The proposed model achieved an accuracy of 85%.

9.2. Bengali –English

SA of code-mixed Bengali-English Facebook text was presented by Ghosh et. al in [66]. The authors collected Facebook posts and annotated them manually into positive, negative and neutral. Extensive preprocessing techniques were employed to remove unwanted words and normalize the text. Multilayer Perception model using word based semantic and style base features was used to predict the polarity of each post. The proposed model achieved an accuracy of 68.5%.

Various supervised machine learning classification models were used by Mandal& Das in [67] to predict the polarity of code-mixed Bengali-English dataset of movie reviews. The models uses language related features to classify the statements into 'positive', 'negative' and 'neutral'. The experimental results suggested that support vector machine (SVM) model achieved highest accuracy of 72.50%.

A POS tagger for English-Bengali code-mixed Twitter data was developed by Raha, Tathagata, et al. in [68]. The authors first develops a language identification and segmentation module using binary LSTM classification approach. Two different POS tagging systems were developed for Bengali and English. A language identification system tags the individual words into English, Bengali and unknown tags. English tokens were fed to a English POS Tagger and Bengali words into a Bengali POS tagger. The unknown words were discarded. The individual taggers are then joined to develop a combined tagging system. The overall accuracy of the model was recorded as 75.29%.

Mandalet. al in [69] collected Bengali-English code-mixed data from Twitter using Twitter API and developed a sentiment classification system using combination of lexicon and supervised machine learning approaches. The authors also developed a language identification algorithm with known target languages. The data collected was refined to develop a gold standard Bengali-English code-mixed data having language and polarity tags. The language identification system achieved an accuracy of 81% and the sentiment classification system showed an accuracy of 80.97%.

Singh et. al. [70] studied the types of different spelling variations of words in three code-mixed language pairs viz. Hindi-English, Bengali-English and Tamil-English and develops a system for automatic capturing and substitutions of word variations using unsupervised machine learning approach. The authors suggested that such type of preprocessing improves the performance of a POS tagger and Sentiment analysis system.

SA of CMSMT in Hindi-English and Bengali-English using ML and neural network approach was conducted by Mishra et. al. [71]. Two models were developed for the task and four runs were conducted on the code-mixed text. The first model is an ensemble one consists of three classifiers viz. linear SVM, Logistic Regression and Random Forest. The second model uses linear SVM. Both moelds used TF-IDF feature vectors of character n-gram. The best f-score of 0.569 was achieved for Hindi-English and f-score of 0.526 was achieved for Bengali-English CMSMT.

A word level language identification system based on deep learning based approach for English-Hindi-Bengali code-mixed data was developed by Jamatiaet. al [72]. The data for the experimentation was collected from popular social media platforms like Facebook, Twitter and WhatsApp. The performance of the proposed approach was compared with feature based learning approach using Conditional Random Forest (CRF). The experimental results showed that the Deep learning based approach comprising of two Recursive Neural Network techniques (LSTM and BiLSTM) outperformed the CRF approach.

9.3. Tamil-English

Chakravarthi, Bharathi Raja, et al. [73] presented SA of code-mixed Tamil-English and Malayalam-English social media data. The dataset was provided during FIRE2020 shared task. The authors used pre-trained NLP models and deep learning based transfer learning to classify the code-mixed text into positive, negative, neutral, mixed emotions and unknown state using FastAi library. The proposed system achieved an F1-score of 0.6 for both Tamil-English and Malayalam-English code-mixed data.

SA of Tamil-English and Malayalam-English CMSMT was presented by Banerjee, Shubhankeret. al. [74]. The authors used an auto-regressive XLNetmodel for achieving better performance on large datasets. The datasets were provided by FIRE2020 shared task. However, the researchers did not achieve reasonable accuracy particularly for Tamil-English code-mixed data due to noise nature of the informal text. An accuracy of 49% was achieved for Malayalam-English and only 35% for Tamil-English code-mixed text.

A detailed process of creation and annotation of Tamil-English CMSMT collected from YouTube has been presented by Chakravarthi, Bharathi Raja, et al. [75]. The dataset contains 15744 annotated posts developed for sentiment analysis. The annotation process was done manually by language experts and three annotators were used to annotate each sentence. Machine learning approaches such as SVM, NB, LR, DT, and RF were used for annotation validation and polarity identification.

Ranjan, Prakash, et al. [76] presented data collection and a bootstrap process of LI for Tamil-English code-mixed text. The authors used RNNLM model for validation of the hypothesis that code-mixing usually occur in informal text and is less similar to formal text. For language identification, Random Forest approach provides better accuracy of 86.02% than SVM and Naïve Bayes.

9.4. Telugu—English

A sentiment classification system for Telugu-English CMSMT was presented by Padmaja et.al in [77]. The movie review data was collected from Twitter. The authors used two approaches for sentiment extraction. The first, based on the lexicon approach, sentiment bearing word in the sentence is extracted to identify the sentiment of the whole word. The second, machine learning approach, developed a sequential minimal optimization using uni-grams, bi-grams and skip-grams as feature vectors. The machine learning model with accuracy of 76.33% outperforms the lexicon based approach which achieved an accuracy of 66.82%.

A LI model for Telugu-English code-mixed social media text was developed by Gundapu & Mamidi [78]. The authors used four machine learning approaches Naïve Bayes, Random Forest, Conditional Random Forest and Hidden

Markov Model for identification of languages in Telugu-English code-mixed text. The experimental results showed that the CRF model with an f1-scoreof 0.91 outclasses the other three models.

Kovida at el. [79] developed a shallow parser for CMSMT in English-Telugu. The system performed word level language identification, normalization and POS tagging and also segments text into syntactically related units. The authors applied SVM, NB, MNB and RDF to develop a baseline sentiment analysis system for English-Telugu code-mixed social media data using n-gram, POS and language identification features. The experimental results showed that SVM performs better than other ML approaches.

9.5. Urdu-English

Mukund & Srihari [80] proposed a SA system for Urdu blog data written in Latin script. Using structural corresponding learning method. A number of issues such as spelling variations, pattern switching and script differences have been addressed in details. The authors also proposed POS tagger for word level identification of code-mixing behavior. Supervised learning method SVM was used to validate the proposed system.

A sentiment analysis system for Roman Urdu using deep learning approach of Recurrent Convolutional Network model was developed by Mahmood, Zainab, et al [81]. The authors collected from various websites and social networking sites to mine the opinions of people about a number of viz. politics, sports, entertainment, food, movies, software and electronic gadgets. The annotation process was done manually into three classes of positive, negative and neutral. Two models, a rule based model and RNN model were used to predict the sentiments of text. The experimental results revealed that RNN model with accuracy of 57.2% outperforms the rule based model.

Rafique, Ayesha, et al. [82] presented a SA system for Roman Urdu using three supervised machine learning approaches viz. SVM, Naïve Bayes and Logistic regression with Stochastic Gradient Descent. The authors developed a dataset collected from different websites and microblogging sites. The experimental results showed that among the supervised machine learning methods SVM with accuracy of 87.22% performs better than the other two.

9.6. Punjabi-English

Singh et. al. [83] developed a SA system for English-Punjabi code-mixed data collected from various agricultural related posts. The authors have created an

A SA system for English-Punjabi code-mixed data was presented by Yadavet. al. [*84*]. The authors used LSTM approach for sentiment identification and achieved reasonable results.

A statistical techniques used by Singh et. al. [*85*] for sentiment analysis to predict political elections in Punjab. The twitter data was collected during the election period and an English-Punjabi dictionary was created for opinionated words. These was classified as positive words and negative words were stored in a gazetteer list for identification of sentiment towards a particular political party.

English-Punjabi code-mixed dictionary after normalization of social media posts. The supervised machine learning algorithms used include SVM and NB using unigram and n-gram as features.

Bansal et. al. [86] presented a baseline approach for LI of English-Punjabi CMSMT using ML approaches viz. Gaussian NB, LR and DT classifiers. The data was collected from Facebook and Twitter and annotated manually using pipeline dictionary vectorizer and n-gram approaches. Experimental results revealed that Logistic regression perform better than other approaches with accuracy of 86.63%.

9.7. Other code-mixed Indian Languages

Mohammed Arshad et al. [87] presented a sentiment analysis system for code-mixed Marathi-English documents using supervised machine learning algorithms such as SVM, NB and RF. The system achieved best f1-score of 0.72 while applying linear SVM approach. The authors also identified various issues which needed to be addressed for achieving better performance while dealing with code-mixed text. Lakshmi &Shambhavi [88] presented a

An automatic LI system for Manipuri-English code-mixed text is presented by Lamabam and Chakma [89]. The authors collected code-mixed posts from Twitter and Facebook and applied Trigram-based and conditional Random Field based

word level language identification system by appending a dictionary module to a number of classification methods such as SVM, MNB, Bernoulli NB , Random Forest and Logistic Regression for Kannada-English code-mixed social media posts. The features used include n-grams and TF-IDF. The best accuracy of 89.6% without dictionary and 94.3% with dictionary module was achieved for SVM using TF-IDF features.

models. The experimental results showed that CRF with F1-socre of 0.90 outperforms the other approach.

Language Pairs	Pre-processing	Language Identificatio	POS Tagging	Dataset development	Sentiment analysis	Reference
Hindi-English	√	√	√	√	√	[57], [60], [66], [72], [102-108], [111-112],[114], [116-121]
Bengali-English	√	√	√	√	√	[66-67], [106-108], [118]
Telugu-English	√	√	√	√	√	[77-79], [108], [120]
Punjabi-		√		√	√	[83-86] [110]

The status NLP research of code-mixed Indian language pairs is given in Table 4.

English						
Tamil-English		√			√	[73-76], [106], [109], [113] [115]
Malayalam-English		√			√	[109], [113] [115]
Urdu-English	√				√	[122]
Marathi-English					√	[60] [106]
Gujarati-English				√		[106]
Kannada-English		√				[88] [115]
Manipuri-English		√				[89]

Table 4: Research Focus Area for code-mixed Indian language pairs

10. Findings of the survey

SA is the process of identification of polarity in a piece of text. SA, also known as opinion mining, appraisal extraction, polarity detection etc. is an promising field of research in the field of NLP especially non-English languages. Sentiment analysis of English language is a well-studied problem. However, for non-English languages the task become more challenging because of non-availability of language and linguistic resources. India is a multilingual and multicultural nation. The number of Internet users and web surfers are increasing every year because of cheaper data plans and availability of high speed Internet access at reasonable rates. Also the technologies provides Internet users options to communicate in their native languages as well. Multilingual websites provides content in more than one language and translators such as Google translate enables us to translate the content of a webpage in more than 100 languages. Therefore number of non-English web resources are increasing rapidly. This has encouraged people to write their feeling in their native languages. The proliferation of non-English content on web has also attracted the NLP researchers because primarily NLP resources have been developed for English language. The researchers are developing multilingual models for a number of NLP tasks such as machine translation, sentiment analysis, automatic speech recognition, information extraction, language identification, entity extraction and many more. The linguistic resources available in Indian languages is presented in Table 5 and Table 6.

From the recent past especially in bilingual and multilingual societies, people also prefer to use code-mixing, a phenomenon where a lexical unit from one language is inserted into a structure while following the grammatical rules of other language, to express themselves on the web particularly on social networking sites. Manihuruket. al. [90] and Syafaatet. al. [91] have provided detailed analysis of use of code-mixing on Facebook and Twitter. SA of CMSMT has emerged a new filed of research. In India many shared tasks like International Conference on Natural Language Processing (ICON), Sentiment Analysis of Indian Languages (SAIL), Forum for Information Retrieval (FORUM) have been organized for a number of NLP tasks on code-mixed data such as language identification, named entity recognition, corpus creation, question-answering challenge, machine translation, POS tagging and polarity detection.

SA of code-mixed textual data is a difficult task. A number of approaches have been used by researchers to develop NLP tools for code-mixed textual data. For SA of CMSMT, lexicon based approach;

ML approach and DL based approaches have been used by the researchers. Among these approaches, machine learning based approach has been used by most of the researchers. However, deep learning approach has gained momentum from the recent past as well. Twitter and Facebook data has been used for corpus creation, POS tagging and for other NLP tasks followed by product and movie reviews. Most of the researchers have performed sentence level sentiment analysis. Annotated datasets are being created and linguistic resources are being developed to facilitate researchers in this field of NLP. However, there are a number of challenges in SA of CMSMT. These include:

1. Generally code-mixed data is informal and noisy in nature and include spelling variations, slang words, abbreviations, out of vocabulary words which makes the NLP task more herculean.
2. Non-availability of lexical and linguistic resources.
3. Non-availability of annotated datasets.
4. Identification of language at word level.
5. Non-availability of POS taggers and SentiWordNets

Table 5: WorldNet for Indian Languages (Developed under IndoWordNet Consortium https://www.cfilt.iitb.ac.in/)

Language	No. of Synsets	Developed By	Reference/Source
Assamese	14958	Gauhati University GauhatiAssam	https://gauhati.ac.in/
Bengali	36346	Indian Statistical Institute, Kolkata, West Bengal	https://www.isical.ac.in/
Bodo	15785	Guwahati University, Guwahati, Assam	https://gauhati.ac.in/
Gujarati	35599	Dharamsinh Desai University Nadiad Gujarat	https://www.ddu.ac.in/
Hindi	40371	IIT Bombay, Mumbai, Maharashtra	https://www.iitb.ac.in/
Kannada	22042	Mysore University, Mysore, Karnataka	https://uni-mysore.ac.in/
Kashmiri	29469	Kashmir University, Srinagar, Jammu and Kashmir	https://www.kashmiruniversity.net/
Konkani	32370	Goa University, Taleigao, Goa	https://www.unigoa.ac.in/
Malayalam	30140	Amrita University, Coimbatore, Tamil Nadu	https://www.amrita.edu/campus/coimbatore
Manipuri	16351	Manipur University, Imphal, Manipur	https://www.manipuruniv.ac.in/
Marathi	32226	IIT Bombay, Mumbai, Maharashtra	https://www.iitb.ac.in/
Nepali	11713	Assam University, Silchar, Assam	http://www.aus.ac.in/
Odiya	35284	Hyderabad Central University, Hyderabad, Andhra Pradesh	https://uohyd.ac.in/
Punjabi	32364	Thapar University and Punjabi University, Patiala, Punjab	http://www.thapar.edu/
Sanskrit	37907	IIT Bombay, Mumbai, Maharashtra	https://www.iitb.ac.in/
Tamil	25419	Tamil University, Thanjavur, Tamil Nadu	https://www.tamiluniversity.ac.in/
Telugu	21091	Dravidian University, Kuppam, Andhra Pradesh	https://www.dravidianuniversity.ac.in/
Urdu	34280	Jawaharlal Nehru University, New Delhi	https://www.jnu.ac.in/

Table 6: Indian Language Linguistic Resources

Resource	Type	Language(s)
Dictionary	Trilingual	Hindi-Assamese-English, Hindi-Bangla-English, Hindi-Gujarati-English, Hindi-Kannada-English, Hindi-Kashmiri-English, Hindi-Malayalam-English, Hindi-Marathi-English, Hindi-Punjabi-English, Hindi-Sindhi-English, Hindi-Tamil-English, Hindi – Bodo – English,

		Tamil-Hindi-English
	Bilingual	Hindi-Assamese, Hindi-Bangla Hindi-Gujarati , Hindi-Kashmir, Hindi-Marathi, Hindi-Malayalam, Hindi-Punjabi, Hindi-Sindhi, Hindi-Tamil, Hindi-Telugu, Hindi-Urdu, Hindi-Oriya, Hindi-Maithili, Hindi-Dogri Urdu to Hindi, Marathi to Hindi, Punjabi to Hindi, Malayalam to Tamil, Kannada to Hindi, Bangla to Hindi, Sanskrit–Hindi Apte Dictionary, Tamil-English
	Multilingual dictionary	Assamese, Bangla, Dogri, Gujarati, Kashmiri, Kannada, Konkani, Marathi, Malayalam, Manipuri, Maithili, Nepali, Oriya, Punjabi, Telugu, Tamil, Sindhi, Sanskrit, Urdu
POS Tagger	CRF	Bengali, Gujarati, Hindi, Konkani, Marathi, Malayalam, Punjabi, Kannada, Tamil, Telugu, Urdu,
Shallow Parser		Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kashmiri, Konkani, Marathi, Maithili, Manipuri, Nepali, Odia, Punjabi, Santhali, Tamil, Telugu, Urdu
Morphological Analyzer		Bengali, Hindi, Kannada, Marathi, Punjabi, Tamil, Urdu

11. Conclusion

SA of CMSMT in the context of Indian languages is an emerging area of research. A number of approaches have been presented in the survey paper. Researchers in this field of NLP are developing linguistic and lexicon resources which include wordents, sentiwordnets, dictionaries, POS taggers, Shallow parsers, Machine translators etc. For sentiment identification in code-mixed textual data mostly researchers prefer to use machine learning approach. However, due to limited or no lexicon and linguistic resources available, researchers develop their own corpus and annotated the dataset manually. Identification of language in a code-mixed data is also a challenging task due to informal nature of the text and no standard rule for mixing of linguistic units of other language. Among code-mixed language pairs sentiment analysis of Indo-Aryan languages such as Hindi-English and Bengali English has been explored most followed by Dravidian languages such as Tamil-English and Telugu-English. Many other language pairs have been unexplored or less researched because of limited resources.

It has been observed that majority of the papers in this field of research have been published in conferences followed by journals. Researchers have also mainly used ML approach followed by lexicon based and deep learning based approached. However to achieve better performance, deep learning approach has also gained attention. Sentiment analysis work has also been carried at sentence level with two sentiment classes ‘positive’ and ‘negative’. A number of domains have been used for polarity detection such as tweets, Facebook posts, Whatsapp chats, YouTube comments, movie reviews etc. this paper also presented an overview of online lexicon and linguistic resources for sentiment analysis of code-mixed text which can help researchers to perform sentiment analysis on these language pairs. However, the research in this filed is in early stages and a lot of work needs to be done to improve the performance of sentiment analysis systems. Lack of annotated datasets and other linguistic resources such as language identification systems obstructs the research, because creation of annotated datasets and other linguistic resources is a time consuming process. Therefore, there is a need to encourage researchers in this field to develop annotated corpus and other resources and make them available online for further investigation and for improvement of existing sentiment analysis systems. The linguistic resources such as POS taggers needs to be improved for better performance.

12. References

1. Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of computer‐mediated Communication*, *13*(1), 210-230.
2. Kumar, V., &Dhar, M. (2018). Looking Beyond the Obvious: Code-Mixed Sentiment Analysis (CMSA).
3. Singhal, S., &Garg, N. (2018). Web Page Representation Using Backtracking with Multidimensional Database for Small Screen Terminals. In *Innovations in Computational Intelligence* (pp. 299-307). Springer, Singapore.
4. Farzindar, A. A., &Inkpen, D. (2020). Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, *13*(2), 1-219.
5. Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.
6. Barkur, G., &Vibha, G. B. K. (2020). Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. *Asian journal of psychiatry*, *51*, 102089.
7. Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135.* doi:10.1561/1500000011.
8. Hong, L., Convertino, G., & Chi, E. (2011, July). Language matters in twitter: A large scale study. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1).
9. 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.
10. Barnali, C. (2017). Code-Switching and Mixing in Communication− A Study on Language Contact in Indian Media. In *The Future of Ethics, Education and Research* (pp. 110-123). ScientiaMoralitas Research Institute.
11. Muysken, P., &Muysken, P. C. (2000). Bilingual speech: A typology of code-mixing. Cambridge University Press.
12. Wei, L. (2005). "How can you tell?" Towards a common sense explanation of conversational code-switching. Journal of Pragmatics, 37(3), 375-389.
13. Ritchie, W. C., & BHATIA, T. (2004). 13 Social and Psychological Factors in Language Mixing. The handbook of bilingualism, 46, 336
14. Kim, E. (2006). Reasons and motivations for code-mixing and code-switching. Issues in EFL, 4(1), 43-61
15. Sotillo, S. (2012). Ehhhhutedehacen plane sin mi???:@ im feeling left out:(form, function and type of code switching in SMS texting. In ICAME (Vol. 33, pp. 309-310).][Bock, Z. (2013). Cyber socialising: Emerging genres and registers of intimacy among young South African students. Language Matters, 44(2), 68-91.
16. Zarate, A. L. X. (2010). Code-mixing in Text Messages: Communication Among University Students. Memorias del XI EncuentroNacional de Estudios en Lenguas. Retrieved on, 26.
17. Shafie, L. A., &Nayan, S. (2013). Languages, code-switching practice and primary functions of Facebook among university students. Study in English Language Teaching, 1(1), 187-199.
18. NegrónGoldbarg, R. (2009). Spanish-English codeswitching in email communication. Language@ internet, 6(3)
19. Li, D. C. (2000). Cantonese-English code-switching research in Hong Kong: A Y2K review. *World Englishes*, *19*(3), 305-322.
20. San, H. K. (2009). Chinese-English code-switching in blogs by Macao young people]
21. Hidayat, T. (2012). An analysis of code switching used by facebookers (a case study in a social network site). *Unpublished BA thesis. SekolahTinggiKeguruandanIlmuPendidikanSiliwangi, Bandung*
22. Das, A., &Gambäck, B. (2015). Code-mixing in social media text: the last language identification frontier?

23. Medhat, W., Hassan, A., &Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113
24. Rani, S., & Kumar, P. (2019). A journey of Indian languages over sentiment analysis: a systematic review. *Artificial Intelligence Review*, *52*(2), 1415-1462
25. Tho, C., Warnars, H. L. H. S., Soewito, B., &Gaol, F. L. (2020, November). Code-Mixed Sentiment Analysis Using Machine Learning Approach–A Systematic Literature Review. In *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 1-6). IEEE
26. Al-Moslmi, T., Omar, N., Abdullah, S., &Albared, M. (2017). Approaches to cross-domain sentiment analysis: A systematic literature review. *Ieee access*, *5*, 16173-16192
27. Kumar, A., &Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, *32*(1), e5107
28. Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89
29. Vijayarani, S., Ilamathi, M. J., &Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16
30. Gurusamy, V., &Kannan, S. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16
31. Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performance
32. Jivani, A. G. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, *2*(6), 1930-1938
33. Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., &Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, *65*, 675-782
34. Kaity, M., &Balakrishnan, V. (2020). Sentiment lexicons and non-English languages: a survey. *Knowledge and Information Systems*, 1-36
35. S Thavareesan (2018). Review on Sentiment Lexicons of Indian Languages. Progress in Nonlinear Dynamics and Chaos 6(2), 65-69 http://www.researchmathsci.org/PINDACArt/PINDAC-v6n2-1.pdf
36. Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning
37. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260
38. Hasan, A., Moin, S., Karim, A., &Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11
39. Caruana, R., &Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168).
40. Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, *1*(3), 295-311
41. Ghahramani, Z. (2003, February). Unsupervised learning. In Summer School on Machine Learning (pp. 72-112). Springer, Berlin, Heidelberg
42. Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, 3(1), 1-130
43. Rusk, N. (2016). Deep learning. *Nature Methods*, *13*(1), 35-35.
44. Goodfellow, I., Bengio, Y., Courville, A., &Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press
45. Tang, D., & Zhang, M. (2018). Deep learning in sentiment analysis. In *Deep Learning in Natural Language Processing* (pp. 219-253). Springer, Singapore
46. Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., &Rehman, A. (2017). Sentiment analysis using deep

learning techniques: a review. *Int J AdvComputSciAppl*, *8*(6), 424

47. Ho, J. W. Y. (2007). Code-mixing: Linguistic form and socio-cultural meaning. *The International Journal of Language Society and Culture*, *21*(7), 1-8.
48. Kachru, B. B. (1978). Toward structuring code-mixing: an Indian perspective
49. 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)
50. Joshi, A., Prabhu, A., Shrivastava, M., &Varma, V. (2016, December). Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2482-2491)
51. *Sharma, S., Srinivas, P. Y. K. L., &Balabantaray, R. C. (2015, August). Text normalization of code mix and sentiment analysis. In 2015 international conference on advances in computing, communications and informatics (ICACCI) (pp. 1468-1473). IEEE*
52. Singh, G. (2021). Sentiment Analysis of Code-Mixed Social Media Text (Hinglish). *arXiv preprint arXiv:2102.12149*
53. Mishra, P., Danda, P., &Dhakras, P. (2018). Code-Mixed Sentiment Analysis Using Machine Learning and Neural Network Approaches. *arXiv preprint arXiv:1808.03299*
54. 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)
55. Das, A., &Gambäck, B. (2015). Code-mixing in social media text: the last language identification frontier?
56. 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)
57. 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
58. Kumar, V., &Dhar, M. (2018). Looking Beyond the Obvious: Code-Mixed Sentiment Analysis (CMSA)
59. Choudhary, N., Singh, R., Bindlish, I., &Shrivastava, M. (2018). Sentiment analysis of code-mixed languages leveraging resource rich languages. *arXiv preprint arXiv:1804.00806*
60. 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
61. 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
62. Jhanwar, M. G., & Das, A. (2018). An ensemble model for sentiment analysis of Hindi-English code-mixed data. *arXiv preprint arXiv:1806.04450*
63. Garg, N., & Sharma, K. (2020). Annotated corpus creation for sentiment analysis in code-mixed Hindi-English (Hinglish) social network data. Indian Journal of Science and Technology, 13(40), 4216-4224
64. Garg, N., & Sharma, K. (2020). Annotated corpus creation for sentiment analysis in code-mixed Hindi-English (Hinglish) social network data. *Indian Journal of Science and Technology*, *13*(40), 4216-4224
65. Sharma, S., Srinivas, P. Y. K. L., &Balabantaray, R. C. (2015, August). Text normalization of code mix and

sentiment analysis. In *2015 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 1468-1473). IEEE

66. Ghosh, S., Ghosh, S., & Das, D. (2017). Sentiment identification in code-mixed social media text. *arXiv preprint arXiv:1707.01184*
67. Mandal, S., & Das, D. (2018). Analyzing roles of classifiers and code-mixed factors for sentiment identification. *arXiv preprint arXiv:1801.02581*
68. 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
69. 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*
70. Singh, R., Choudhary, N., &Shrivastava, M. (2018). Automatic normalization of word variations in code-mixed social media text. *arXiv preprint arXiv:1804.00804*
71. Mishra, P., Danda, P., &Dhakras, P. (2018). Code-Mixed Sentiment Analysis Using Machine Learning and Neural Network Approaches. *arXiv preprint arXiv:1808.03299*
72. Jamatia, A., Das, A., &Gambäck, B. (2019). Deep learning-based language identification in English-Hindi-Bengali code-mixed social media corpora. *Journal of Intelligent Systems*, *28*(3), 399-408
73. Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Suryawanshi, S., Jose, N., Sherly, E., & McCrae, J. P. (2020, December). Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In Forum for Information Retrieval Evaluation (pp. 21-24)
74. Banerjee, S., Jayapal, A., &Thavareesan, S. (2020). NUIG-Shubhanker@ Dravidian-CodeMix-FIRE2020: Sentiment Analysis of Code-Mixed Dravidian text using XLNet. arXiv preprint arXiv:2010.07773
75. Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., & McCrae, J. P. (2020). Corpus creation for sentiment analysis in code-mixed Tamil-English text. arXiv preprint arXiv:2006.00206
76. Ranjan, P., Raja, B., Priyadharshini, R., &Balabantaray, R. C. (2016, December). A comparative study on code-mixed data of Indian social media vs formal text. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I) (pp. 608-611). IEEE
77. Padmaja, S., Bandu, S., & Fatima, S. S. (2019, March). Text Processing of Telugu–English Code Mixed Languages. In *International Conference on Emerging Trends in Engineering* (pp. 147-155). Springer, Cham
78. Gundapu, S., &Mamidi, R. (2020). Word level language identification in englishtelugu code mixed data. *arXiv preprint arXiv:2010.04482*
79. Nelakuditi, K. (2017). Towards Building a Shallow Parsing Pipeline for English-Telugu Code Mixed Social Media Data (Doctoral dissertation, Master's thesis, International Institute of Information Technology, Hyderabad)
80. Mukund, S., &Srihari, R. K. (2012, June). Analyzing Urdu social media for sentiments using transfer learning with controlled translations. In *Proceedings of the Second Workshop on Language in Social Media* (pp. 1-8)
81. Mahmood, Z., Safder, I., Nawab, R. M. A., Bukhari, F., Nawaz, R., Alfakeeh, A. S., ...& Hassan, S. U. (2020). Deep sentiments in Roman Urdu text using recurrent convolutional neural network model. Information Processing & Management, 57(4), 102233
82. Rafique, A., Malik, M. K., Nawaz, Z., Bukhari, F., &Jalbani, A. H. (2019). Sentiment analysis for roman urdu. Mehran University Research Journal of Engineering & Technology, 38(2), 463
83. 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

84. *Yadav, K., Lamba, A., Gupta, D., Gupta, A., Karmakar, P., & Saini, S. (2020, October). Bilingual Sentiment Analysis for a Code-mixed Punjabi English Social Media Text. In 2020 5th International Conference on Computing, Communication and Security (ICCCS) (pp. 1-5). IEEE*
85. *Singh, M., Goyal, V., & Raj, S. (2021). Sentiment Analysis of English-Punjabi Code-Mixed Social Media Content to Predict Elections. In Advances in Information Communication Technology and Computing (pp. 81-90). Springer, Singapore*
86. *Bansal, N., Goyal, V., & Rani, S. (2020). Experimenting language identification for sentiment analysis of englishpunjabi code mixed social media text. International Journal of E-Adoption (IJEA), 12(1), 52-62*
87. 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*
88. 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
89. Lamabam, P., &Chakma, K. (2016, March). A language identification system for code-mixed English-Manipuri Social Media text. In 2016 IEEE International Conference on Engineering and Technology (ICETECH) (pp. 79-83). IEEE
90. Manihuruk, L. M. E. (2016). An Analysis Of Code Mixing In Facebook Status. *The Episteme Journal of Linguistics and Literature*, *2*
91. Syafaat, P. M. F., &Setiawan, T. (2019, April). An Analysis of Code Mixing in Twitter. In *International Conference on Interdisciplinary Language, Literature and Education (ICILLE 2018)* (pp. 276-281). Atlantis Press
92. Das, A., &Bandyopadhyay, S. (2010). Sentiwordnet for bangla. *Knowledge Sharing Event-4: Task*, *2*, 1-8
93. Sharma, R., & Bhattacharyya, P. (2014, December). A sentiment analyzer for hindi using hindisenti lexicon. In *Proceedings of the 11th International Conference on Natural Language Processing* (pp. 150-155).
94. Bakliwal, A., Arora, P., &Varma, V. (2012, May). Hindi subjective lexicon: A lexical resource for hindi polarity classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)* (pp. 1189-1196).
95. Das, A., &Bandyopadhyay, S. (2010, August). SentiWordNet for Indian languages. In *Proceedings of the eighth workshop on Asian language resouces* (pp. 56-63).
96. Kannan, A., Mohanty, G., &Mamidi, R. (2016, December). Towards building a SentiWordNet for Tamil. In *Proceedings of the 13th International Conference on Natural Language Processing* (pp. 30-35).
97. Mohanty, G., Kannan, A., &Mamidi, R. (2017, September). Building a sentiwordnet for odia. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 143-148).
98. DikshaGoyal&Gurpreet Singh Josan (2018, June) Automatic Sentiment Lexicon Construction For Punjabi in An International Journal of Engineering Sciences Issue June 2018, Vol. 30, Web Presence: http://ijoes.vidyapublications.com
99. Asghar, M. Z., Sattar, A., Khan, A., Ali, A., MasudKundi, F., & Ahmad, S. (2019). Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. *Expert Systems*, *36*(3), e12397.
100. Deepamala, N., & Kumar, R. (2015, June). Polarity detection of Kannada documents. In *2015 IEEE International Advance Computing Conference (IACC)* (pp. 764-767). IEEE

101.Anagha, M., Kumar, R. R., Sreetha, K., Rajeev, R., & Raj, P. R. (2014). Lexical resource based hybrid approach for cross domain sentiment analysis in Malayalam. *Int J EngSci*, *15*, 18-21.

102.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).

103.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*

104.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.

105.Singh, K., Sen, I., &Kumaraguru, P. (2018, July). A twitter corpus for Hindi-English code mixed POS tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media* (pp. 12-17).

106.Banerjee, S., Moghe, N., Arora, S., &Khapra, M. M. (2018). A dataset for building code-mixed goal oriented conversation systems. *arXiv preprint arXiv:1806.05997*.

107.Mishra, P., Danda, P., &Dhakras, P. (2018). Code-Mixed Sentiment Analysis Using Machine Learning and Neural Network Approaches. *arXiv preprint arXiv:1808.03299*.

108.. Jamatia, B. Gambäck, and A. Das. Collecting and Annotating Indian Social Media Code-Mixed Corpora. In the proceeding of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING), April 3–9, 2016, Konya, Turkey.

109.Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., & McCrae, J. P. (2020). Corpus creation for sentiment analysis in code-mixed Tamil-English text. *arXiv preprint arXiv:2006.00206*.

110.Yadav, K., Lamba, A., Gupta, D., Gupta, A., Karmakar, P., & Saini, S. (2020, October). Bilingual Sentiment Analysis for a Code-mixed Punjabi English Social Media Text. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)* (pp. 1-5). IEEE.

111. 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.

112. 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).

113. 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.

114. Bora, M. J., & Kumar, R. (2018, May). Automatic word-level identification of language in assamese english hindi code-mixed data. In 4th Workshop on Indian Language Data and Resources, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (pp. 7-12).

115. Jauhiainen, T., Ranasinghe, T., &Zampieri, M. (2021). Comparing Approaches to Dravidian Language Identification. arXiv preprint arXiv:2103.05552.

116. , R., & Joshi, R. (2020). Evaluating Input Representation for Language Identification in Hindi-English Code Mixed Text. arXiv preprint arXiv:2011.11263.

117. Sinha, N., &Srinivasa, G. (2014). Hindi-English Language Identification, Named Entity Recognition and Back Transliteration: Shared Task System

Description. In Working Notes os Shared Task on Transliterated Search at Forum for Information Retrieval Evaluation FIRE'14.

118. 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

119. 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.

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

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

122. Younas, A., Nasim, R., Ali, S., Wang, G., & Qi, F. (2020, December). Sentiment Analysis of Code-Mixed Roman Urdu-English Social Media Text using Deep Learning Approaches. In 2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE) (pp. 66-71). IEEE.