International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.8 January - February – 2024
DOI:10.46647/ijetms.2024.v08i01.032 ISSN: 2581-4621

A Technical Review on Emotion Detection from Informal Text for Indian Regional Languages: Approaches and Challenges

Kajal Patil[1], Jitendra Nasriwala[2] , Rakesh Savant[3]
[1] Assistant Professor, Faculty of Computer Science, Uka Tarsadia University, Bardoli, India
[2] Associate Professor, Faculty of Computer Science, Uka Tarsadia University, Bardoli, India
[3] Assistant Professor, Faculty of Computer Science, Uka Tarsadia University, Bardoli, India
Corresponding Author Orcid ID: 0009-0008-8148-2377, 0000-0002-5585-1668, 0000-0003-3497-8026

ABSTRACT
Effective communication is a crucial element in the existence of a human being. Emotions are a common component of human communication and can appear in a variety of ways, including spoken words, written messages, and non-verbal indicators like gestures and facial expressions. Textual communication which is the common way of interaction has been rapidly increasing day-by-day. People nowadays are using social media to convey their thoughts and beliefs in the form of comments, posts and stories with associated emotions. As a result, there is a growing need to detect and understand the emotions conveyed in texts. Although the emotions expressed in text are readily understood by the human brain, teaching a machine to do the same is a challenging task. However, when dealing with Indian languages, this process becomes especially difficult. As India has plenty of vernacular languages, people opt to use monolingual or code-mix languages. These types of languages are generally informal way which creates an obstacle to convey the emotions. The shortage of annotated corpus especially for regional languages is also one of the challenges. This paper aims to analyze and review emotion detection from textual data specifically for Indian regional languages and its challenges.
Keywords— emotion detection, text-based emotion detection, natural language processing, corpora, regional language

1. Introduction

Emotion detection is a branch of artificial intelligence and natural language processing that focuses on recognizing and interpreting human emotions expressed in various forms of communication, such as text, speech, or facial expressions. The primary goal is to train machines with the ability to understand and respond appropriately to the emotional shades inherent in human interactions. As the world is moving towards digitalization, the usage of online applications is growing higher. Nowadays people are using online platforms for various purposes such as purchasing products, digital entertainments, chat applications, etc. to express their thoughts, beliefs or feelings in an informal way in the form of short messages or comments. Apart from this, the written text is not following any standard or syntactical structure of grammar. Consequently, recognizing emotion from the text is also a challenging task due to the fact that textual expression does not always have emotion directly referenced in the written material, instead it could interpret the meaning of concept in text.

Emotional expression is the significant form of communication where it can be expressed in six basic components of emotion such as happiness, anger, fear, sadness, disgust, and surprise. There is a growing demand for machines to comprehend human emotions during natural conversations. Emotion recognition in an informal text message, conversation or comment can be also useful to recognize physiological state of an individual. While emotion can be conveyed in any language, English language is much more investigated research field as compared to non-English language for recognizing emotion in textual data. Apart from this, specifically Indian language lacks behind in this

DOI:10.46647/ijetms.2024.v08i01.032 ISSN: 2581-4621

and less amount of work is reported. However, the content on internet is growing day by day where non-English and multilingual contents are used in majority.[1]

India is a multilingual country which has 22 constitutional languages. Due to this reason, usage of local languages by people in expressing their emotions and thoughts in the form of stories, comments, or other conversation modes becomes the need of the hour. Additionally, Gujarati is also the official language of the Indian state of Gujarat, which has sizable immigrant populations all over the world. Most research on emotion recognition has been carried out in other Indian languages like Hindi [2][3][4],Punjabi [5][6], Bangla [7], Marathi [8],etc. nonetheless Gujarati language lags behind in this field. In addition to this, when communicating informally, code mixed language is frequently used to convey one's feelings and ideas, especially in multilingual nations like India[9].

The structure of the paper is as follows. In Section 2, overview of emotion detection. In Section 3 briefs an overview of Regional language and their related work on emotion recognition. In Section 4 we discussed background study about Gujarati language and the challenges. In section 5 we discussed related work along with challenges and corpus. In the last section, we conclude our paper followed by references.

1.1 Emotion Detection

Emotion detection from text involves analysing and identifying emotions that are expressed by an individual in a written content. However, the emotions have different shades of emotion where the analysis includes a detailed level. The goal is to understand the emotional or physiological state of an individual and to extract relevant information from their language or behaviour. However, it's important to note that emotion detection is not always accurate, as human emotions are complex, and can be influenced by various factors. There are two categories of emotion namely, Categorical and Discrete. In the categorical model, six basic emotions such as fear, anger, joy, sadness, disgust, and surprise. Whereas, discrete model categories the emotion based on valence, provocation and power. In the following section, we will delve into detecting emotions in regional languages and explore the existing research.

1.2 Regional language

Emotion recognition in regional language has become a need of an hour as people nowadays feel comfortable to express themselves in their own native language. Extraction of emotions from text is more difficult especially from the multilingual, like in social media posts, comments or during chat conversation. There are numerous ways an individual can express the emotion in written form such as using monolingual, multilingual or code-mix. Nevertheless, code-mixing language is highly used because it reflects the natural way of communication in multilingual and multicultural environments. People seamlessly switch between their known language to convey their feelings or shades of emotions. Majority of the research has been carried out using a pair of languages such as Urdu - English [10], Hindi-English [9][11][12][13][14], Bangla-English[7]. Still, other Indian languages, especially Gujarati – English code-mix pair lack behind in this field of research.

Besides this, one obstacle to emotion recognition in regional languages is a deficiency of resources. The development of efficient emotion detection is made difficult by a lack of labeled datasets, pre-trained models, and linguistic resources in Gujarati. Efforts have been made by researchers to create their own corpus and annotation with associated emotion is done manually by annotator [9][6][15].

In addition to this, for regional languages like Hindi or Gujarati researcher[6][15] are using “Navarasa “to classify emotional states. Where Nava means nine and rasa means emotional state. The nine emotional states or Navarasa originated from the Natyashastra given by Bharata Muni. Nine emotions are Shringara (love/beauty), Hasya (laughter), Karuna(sorrow), Raudra (anger), Veera (heroism/courage), Bhayanaka (terror/fear), Bibhatsa (disgust), Adbutha (surprise/wonder), Shantha (peace or tranquility)[6][15]. In the upcoming section, we will explore the discussion on Indian regional languages, focusing specifically on Gujarati.

1.3 Background study: Gujarati Language

DOI:10.46647/ijetms.2024.v08i01.032 ISSN: 2581-4621

India is a rich linguistic diversity, with numerous languages spoken across its vast and culturally varied landscape. Gujarati is one of the prominent languages spoken in India, primarily in the western state of Gujarat. There are sizable Gujarati-speaking populations in other Indian states as well as foreign countries, particularly in areas where there is a sizable Gujarati diaspora. The Gujarati language utilizes the Gujarati script, which originates from the Devanagari script. It belongs to the Indo-Aryan branch of the Indo-European language family.

As there is rapid growth in expressing emotion or thoughts in native language, though Gujarati language is a rich and expressive language, it does face certain challenges in terms of its Grammar.

Noun Gender: There are three Genders in Gujarati, namely, narajaati (masculine), naarijaati(feminine) and naanyatar (neutral).

For instance, "□□□□□" (boy) is masculine gender, "□□□□□" (girl) is a feminine gender and "□□□□□□" (child) is a neutral. Thus, assigning the correct gender to nouns is essential for accurate parsing and understanding.

Sentence structure: Gujarati uses a different sentence structure than English: Subject-Object-Verb (SOV). Thus, it can be challenging for models to adapt this structure.

Variation in spelling is also poses challenges as a word in Gujarati language may be written in different forms, for an instance: -□□□□□□□□□, □□□□□□□□□ ,□□□□□□□□□ and □□□□□□□□□. Additionally, variations occur in the way vowels and diacritical marks are used. Similar-sounding vowels, such as □ – □ (hasva-i), □ – □ (dirgha - ee) and □ – □ (hasva-u), □- □ (dirgha - ee), are frequently utilized interchangeably in texts.

In the following section, we will examine the existing research on emotion detection from textual data in Indian regional languages and highlight the challenges that have been observed.

2. Related work

In Punjabi language , work presented by [5] considered Punjabi textual data for emotion recognition using Hybrid approach Keyword Based Approach and Machine Learning Approach. They have experimented using standard Punjabi textual dataset HC corpora which consists of various online Punjabi websites of news, social networking, blogs, etc. The outcome of the emotion is based on Ekman's basic six emotions (happy, fear, anger, sadness, disgust and surprise).

For code-switching text , in [11] Emotion detection in code-switching text researches have mainly focused on analyzing emotions in monolingual text. Due to the lack of publicly available resources for Hindi-English data, the author created linguistic resources for this pair of languages from social media. The base system gave an overall accuracy of 83.54% for their dataset. Pre-processing was performed on the code-mix social media text. Apart from this, manual annotation was done for language identification. The Factor Graph Model (FGM) which is a probabilistic graph model was used to learn both monolingual and bilingual information from each post from social media.

In[9] authors created Hindi-English code-mix corpus from twitter from different domains like politics, social events and sports using Web scraping due lack of publicly available resources. During corpus creation, the author removed tweets which were either purely in English or Hindi. Apart from this, they further removed those tweets which do not express any emotions. Annotations of emotions were based on six standard emotions of Ekman, namely, Happiness, Sadness, Anger, Fear, Disgust and Surprise. For annotation, inter-annotation agreement between two annotators was done using the Kappa coefficient. Their experiment showed that punctuation marks and emoticons show better accuracy. Apart from this they were able to achieve 58.2% accuracy with features trained with SVM classifiers.

According to Plutchik's wheel, this paper[1] categorizes tweets into eight groups of emotions: joy, trust, fear, surprise, sadness, anticipation, rage, and disgust. The dataset consists of manually annotated English, Gujarati, and Hindi tweets about Indian politics. The paper demonstrates the effectiveness of the Hybrid technique in emotion classification by utilizing both Supervised Learning and a Hybrid approach, using tf-idf for primary features and introducing two SenticNet-based

DOI:10.46647/ijetms.2024.v08i01.032 ISSN: 2581-4621

algorithms for secondary feature generation. By assigning tweets to emotion categories, multilabel classification reveals feelings inside the Indian political discourse.

In [12]this study presents a deep learning method to identify emotions in Hindi-English code-mixed language on social media like Twitter and Facebook. They gathered and refined 12,000 code-mixed sentences, expressing emotions like happiness, sadness, and anger. Utilizing a bilingual pretrained model, retrained with their dataset, the CNN-BiLSTM model achieved an 83.21% detection accuracy. Along with, the CNN created more meaningful information from the word embedding which is passed as an input to Bi-LSTM which captures semantics of the sentences.

In [6] the 'Kāvi' Punjabi poetry corpus was manually annotated, encompassing nearly 1000 poems in Gurmukhi script and reflecting nine emotional states from 'Navrasa'. Linguistic and poetic features in each poem were assessed and weighted using TF-IDF. Naïve Bayes and Support Vector Machine models were trained and tested on these features where SVM improved by giving accuracy of 70.02% for classifying emotion.

In [10] the authors implemented a multi-classification emotion model for English-Urdu code-mix textual data. Samples include English-Urdu code-mixed text with Romanized Urdu. They used XLM-RoBERTa and Indic-BERT for classification. The data were translated into English. The findings show that XLM-Roberta attained the highest F1 score among the evaluated models, registering at 0.60. In contrast, Indic-BERT yielded the lowest F1 score among the tested models, scoring at 0.54.

[8] focuses on hate speech detection, sentiment and simple text classification for Marathi language. The evaluation was done in a monolingual and multilingual BERT model. The dataset was considered as L3CubeMahaHate which consist of hateful and non-hateful categories, HASOC-2021 consist of 25000 tweets with 4 major classes namely hate, offense, profane and not, News articles from sports, entertainment and lifestyle domain and a collections of Marathi headlines. They concluded that the monolingual dataset outperformed the standard multilingual models.

In [7] detected emotion from multilingual text and multi-emotional sentences. They have used Countvectorizer as a feature extraction model which was not used in detection of emotions. For the dataset, they have utilized WASAA- 2017 which contained 7000 textual data with different emotions such as Anger, fear, joy and sadness. Apart from this, they have managed to gather 600 Bangla texts for detecting emotions. The model was successfully able to detect multiple emotions from complex sentences.

References	**Language**	**Type of text**	**Dataset**	**Input**	**Pre-processing**	**Feature extraction**	**Approach**	**Classifier**	**Output**	**Accuracy**
[5]	Punjabi	Monolingual	Standard HC Corpora	Paragraph	Segmentation, Tokenization, Stemming, Stop-word	-	Hybrid (Keyword and ML)	SVM, NB	happy, fear, anger, sadness, disgust and surprise	Not mentioned
[11]	Hindi-English	Code-switching	Own dataset from social media	Sentence	Tokenization, Language identification, POS tag and Shallow parsing	-	Joint Factor Graph Model	Maximum entropy	-	83.54 % for dataset.
[9]	Hindi-En	Code-mix	Own corpus creation	Sentence	Text cleaning: removal of URLs,	N-gram,BOW,	Machine	SVM	happy, fear, anger,	58.2%

DOI:10.46647/ijetms.2024.v08i01.032 ISSN: 2581-4621

	glish		from tweets		replacing user names, replacing emoticons, removal of punctuation	Emoticons, Punctuations, Repetitive Characters, Intensifiers	learning		sadness, disgust and surprise	
[1]	English, Gujarati, Hindi	-	Creation of annotated corpus tweets	sentences	Stop word removal Cleaning and tokenization	TF-IDF, SentiNet	Supervised and Hybrid	Logistic RegressionMultinomial Naïve Bayes and SVM	joy, trust, fear, surprise, sadness, anticipation, anger, disgust	-
[12]	Hindi-English	Code-mix		Sentences	Word Tokenization, Text cleaning include removal of URLs, additional spaces, transferring in lower case	Word2Vec, CBOW	Deep Learning	CNN-BiLSTM	Happy, sad and anger	83.21 %
[6]	Punjabi	Monolingual	948 Punjabi poetries	(sentences) Title of poetry	Symbols removed, (|, ||,?,',',!) Word tokenization Stop word removal, Frequency count (passed for feature extraction)	Linguistic feature: POS Poetic feature: Statistic feature: TF-IFD	Machine learning	Naïve Bayes, SVM	"Navrasa" Karuna, Shringar, Hasya,Raudra, veer, bhayanak, vibhata,adbhut and shanti	70.02 %
[15]	Gujarati	Monolingual	300+ poem	Poem (paragraph)	Zipf's law for identical token, Tokenization	-	Deep learning	-	"Navrasa" : Karuna, Shringar, Hasya,Raudra, veer, bhayanak, vibhata,ad	87.62 %

DOI:10.46647/ijetms.2024.v08i01.032 ISSN: 2581-4621

									bhut and shanti	
[10]	English-Urdu	Code-mix	11914 code-mix SMS messages from [16] manually created corpus	Short text messages	Already pre-processed according to [16]	-	-	XLM-RoBERTa and Indic-BERT	12 emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral	F1 score: 0.57
[8]	Marathi	Monolingual and multilingual	L3Cube MahaHate, HASOC-2021, Articles and Headlines	Sentences	Tokenization	-	Transformer based	mBERT, indicBERT, and xlm-RoBERTa	major classes namely hate, offensive, profane, and not.	-
[7]	English and Bangla	Multilingual	WASSA-2017 English(tweet)600 data – Bengali	Sentence	Tokenization	TF-IDF, Count vectorizer	Machine learning	LSVM	Anger, Fear, Joy, and Sadness	

From the mentioned literature review it has been observed that certain emotions may not have equivalent words in other languages, or if they do, they may not be communicated with the same intensity. There is a degradation of accuracy due to translation of one language to another and due to Colloquial or slang words during informal communication. Furthermore, for code-mix textual data the syntactic structure is changing and due to this reason recognition emotion from this type of text becomes difficult. Difficulty in handling multiple emotions in a complex sentence. The mis-classified categories are also observed. Additionally, due to scarcity of corpus researcher are creating their own corpus which is discussed in next section.

Corpora are essential for training and testing language models in computational linguistics and natural language processing (NLP). They are used by researchers to build and improve algorithms for various applications such as information retrieval, speech recognition, machine translation or emotion detection and sentiment analysis. For identifying emotion from human written text which may be in monolingual, multilingual or code-mix form, no standard dataset is available publicly. Therefore, for this reason, researchers are creating their own corpus. The process includes collecting text data from different platforms as per targeted application. Annotation is done manually by tagging associated emotions by different annotators who are known to native languages. Inter Annotator Agreement is made between annotators using Kappa coefficient.

DOI:10.46647/ijetms.2024.v08i01.032 ISSN: 2581-4621

References	Language	Type	Nature of Dataset	Manual annotation	Discussion
[9]	Hindi-English	Code-mix	Own corpus creation	Inter Annotator Agreement between two using Kappa coefficient	2866 tweets were expressing any emotion
[1]	English, Gujarati, Hindi	Multilingual	Creation of annotated corpus (twitter)	Three annotator using Kappa coefficient	Tweet discarded which with no emotion or not related to political
[6]	Punjabi	Monolingual	'Kavi' 980 Punjabi poetries written in Gurumukhi script only collected from online source.	Three annotator (Punjabi) using Fleiss Kappa index to capture Inter Annotator Agreement	Out of 980 poetry, 948 poetries were successfully labeled according to 'Navarasa'
[15]	Gujarati	Monolingual	'Kavan' 300+ poem from different Gujarati literature	-	-

In the concluding section of the paper, the discussion centers around the conclusions drawn regarding emotion detection in regional languages.

CONCLUSION

There is a tremendous usage of social media nowadays and therefore the majority of people are addicted to this platform. Individuals are utilizing online platforms to express their thoughts, engage in discussions, chat conversation, or share stories or posts relating to their perspectives and beliefs on diverse subjects. In this paper, the main aim is to understand the recent work that has been carried out in Emotion detection from textual data for Indian regional languages. In India, individuals employ a combination of English and their regional language, known as code-mixed language, to convey their emotions. Automatically identifying these emotions within code-mixed languages poses a challenging task due to the presence of features from two or more distinct languages. Dataset is also one of the challenges due to scarcity of adequate annotated dataset. Additionally, it has been observed that less amount of work is witnessed for Gujarati language. To overcome these challenges, future research should prioritize the creation of comprehensive and representative datasets specific to the Gujarati language and other regional languages.

References

[1] L. Gohil and D. Patel, "Multilabel classification for emotion analysis of multilingual tweets," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1, pp. 4453–4457, 2019, doi: 10.35940/ijitee.A5320.119119.
[2] Y. Kumar, "BHAAV (□ □ □) - A Text Corpus for Emotion Analysis from Hindi Stories," 2017.
[3] A. Sharma, "Hindi text emotion recognition based on deep learning," *IOSR J. Mob. Comput. &Application (IOSR …*, vol. 7, no. 3, pp. 24–29, 2020, doi: 10.9790/0050-07032429.
[4] M. A. Ali and S. B. Kulkarni, "Preprocessing of Text for Emotion Detection and Sentiment Analysis of Hindi Movie Reviews," *SSRN Electron. J.*, no. Icicnis, pp. 848–856, 2021, doi: 10.2139/ssrn.3769237.
[5] S. Grover and A. Verma, "using Hybrid Approach," *Des. Emot. Detect. Punjabi Text using Hybrid Approach*, vol. 2, pp. 1–6, 2016.

DOI:10.46647/ijetms.2024.v08i01.032 ISSN: 2581-4621

[6] J. R. Saini and J. Kaur, "Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on 'Navrasa,'" *Procedia Comput. Sci.*, vol. 167, pp. 1220–1229, 2020, doi: 10.1016/J.PROCS.2020.03.436.
[7] S. I. Khan, FaisalBinAziz, and M. Uddin, "Emotion Detection from Multilingual Text and Multi-Emotional Sentence using Difference NLP Feature Extraction Technique and ML Classifier," *Int. J. Adv. Netw. Appl.*, vol. 14, no. 03, pp. 5429–5435, 2022, doi: 10.35444/ijana.2022.14303.
[8] A. Velankar, H. Patil, and R. Joshi, "Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13739 LNAI, no. November, pp. 121–128, 2023, doi: 10.1007/978-3-031-20650-4_10.
[9] D. Vijay, A. Bohra, V. Singh, S. S. Akhtar, and M. Shrivastava, "Corpus creation and emotion prediction for hindi-english code-mixed social media text," *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Student Res. Work.*, vol. 2018-Janua, pp. 128–135, 2018, doi: 10.18653/v1/n18-4018.
[10] B. H. Vedula, "PrecogIIITH @ WASSA2023 : Emotion Detection for Urdu-English Code-mixed Text," pp. 601–605, 2023.
[11] V. M. Rao, "Emotion Detection in Code-Switching Text," vol. 4, no. 12, pp. 988–992, 2017.
[12] T. T. Sasidhar, B. Premjith, and K. P. Soman, "ScienceDirect Emotion Emotion Detection Detection in in Hinglish (Hindi Hinglish (Hindi + Code-Mixed Social Social Media Text Media Text," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1346–1352, 2020, [Online]. Available: https://doi.org/10.1016/j.procs.2020.04.144
[13] A. Wadhawan and A. Aggarwal, "Towards Emotion Recognition in Hindi-English Code-Mixed Data: A Transformer Based Approach," *WASSA 2021 - Work. Comput. Approaches to Subj. Sentim. Soc. Media Anal. Proc. 11th Work.*, pp. 195–202, 2021.
[14] G. S. S. N. Himabindu, R. Rao, and D. Sethia, "A self-attention hybrid emoji prediction model for code-mixed language: (Hinglish)," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, 2022, doi: 10.1007/s13278-022-00961-1.
[15] B. Mehta and B. Rajyagor, "Gujarati Poetry Classification Based on Emotions Using Deep Learning," *Int. J. Eng. Appl. Sci. Technol.*, vol. 6, no. 1, 2021, doi: 10.33564/ijeast.2021.v06i01.054.
[16] I. Ameer, G. Sidorov, H. Gomez-Adorno, and R. M. A. Nawab, "Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods," *IEEE Access*, vol. 10, pp. 8779–8789, 2022, doi: 10.1109/ACCESS.2022.3143819.