International Journal on Recent and Innovation Trends in Computing and Communication A Comprehensive Review of Sentiment Analysis on Indian Regional Languages: Techniques, Challeng...

Article *in* International Journal on Recent and Innovation Trends in Computing and Communication · August 2023

DOI: 10.17762/ijritcc.v11i9s.7401

Sunil Digamberrao Kale
Veermata Jijabai Technological Institute

Rajesh Prasad
Bharati Vidyapeeth Deemed University

Girish P. Potdar
Pune Institute Of Computer Technology

Deepak T. Mane
Vishwakarma Institute of Technology

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
Article Received: 03 May 2023 Revised: 30 June 2023 Accepted: 21 July 2023

A Comprehensive Review of Sentiment Analysis on Indian Regional Languages: Techniques, Challenges, and Trends

Sunil D. Kale[1], Rajesh Prasad[2], Girish P. Potdar[3], Parikshit N. Mahalle[4], Deepak T. Mane[5], Gopal D. Upadhye[6]
[1, 4] Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Information Technology, Pune, India
kalesunild@gmail.com, aalborg.pnn@gmail.com
[2] School of Computing, MIT Art, Design and Technology University, Pune, India
rajesh.prasad@mituniversity.edu.in
[3] Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India
gppotdar@pict.edu
[5] Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, India
dtmane@gmail.com
[6] Vishwakarma Institute of Technology, Pune,
gopalupadhye@gmail.com

Abstract— Sentiment analysis (SA) is the process of understanding emotion within a text. It helps identify the opinion, attitude, and tone of a text categorizing it into positive, negative, or neutral. SA is frequently used today as more and more people get a chance to put out their thoughts due to the advent of social media. Sentiment analysis benefits industries around the globe, like finance, advertising, marketing, travel, hospitality, etc. Although the majority of work done in this field is on global languages like English, in recent years, the importance of SA in local languages has also been widely recognized. This has led to considerable research in the analysis of Indian regional languages. This paper comprehensively reviews SA in the following major Indian Regional languages: Marathi, Hindi, Tamil, Telugu, Malayalam, Bengali, Gujarati, and Urdu. Furthermore, this paper presents techniques, challenges, findings, recent research trends, and future scope for enhancing results accuracy.
.

Keywords- Sentiment Analysis, Machine learning, Indian Regional Languages, Opinion Mining, Text Mining

I. INTRODUCTION

Sentiment analysis (SA), also known as opinion exploration, is used to study and identify emotion or opinion given by a piece of literature, such as reviews, comments, feedback, news, facts, or simple text. This process comes under the field of natural language processing, text analysis, and machine learning [41]. The text is classified into three categories for sentiments - positive, negative, and neutral [4]. The significance of sentiment analysis is since a myriad of businesses rely on the feedback and reviews received from the public about their products and services. SA provides a way to analyze customer experiences, thereby improving customer service and satisfaction, which benefits hospitality, travel, retail, and other service-based companies. SA is closely linked with brand marketing and business intelligence which is why many product companies rely on it to give various insights for better decision-making. It is also now popularly used for stock market research and even in the healthcare industry where opinions and responses from people are essential. Even though most of such texts are in the English language on social media, many texts are written using English alphabets (transliteration) [10], but the actual language is different. Mixed Language Text (MLT) is also very common [6] [13]. The native language is region-specific and has different grammar, rules, and composition than English, which is why language-specific SA is becoming important [11]. This survey focuses on predominant Indian languages like Marathi, Hindi, Tamil, Telugu, Bengali, Gujarati, and Urdu. We studied the existing research papers on sentiment analysis in these to understand the general approach and techniques used. This paper is a comprehensive overview of the overall trend, results received, challenges faced, and the possible future scope of this topic.

A. *Importance of Study*

Plenty of research is available on sentiment analysis for the English language. Now a day's, researchers are working on various other languages for sentiment analysis around the globe. Figure 1 to 4 is the bibliometric analysis from Scopus on the search for sentiment analysis on Indian regional languages from 2010 to 2022. The result shows that a total of 169 documents

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

are available. Research on Indian regional languages has sped up from 2017, as per figure 1. Figure 2 shows the Scopus documents available from various sources. Figure 3 shows the documents by subject area and figure 4 by type with percentage.

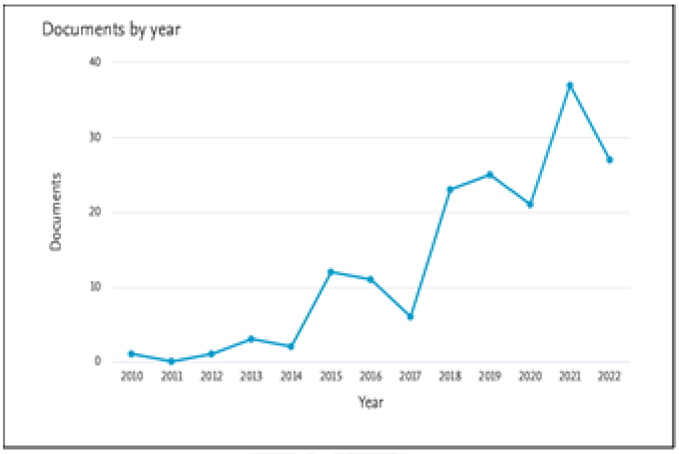

Figure 1. Scopus documents by year

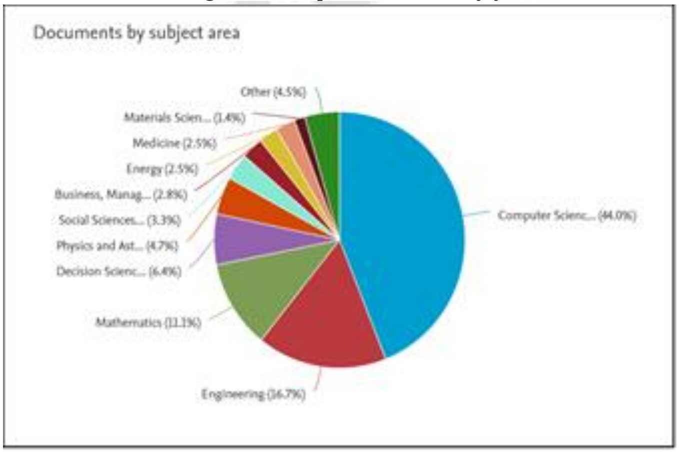

Figure 2. Scopus documents from various sources

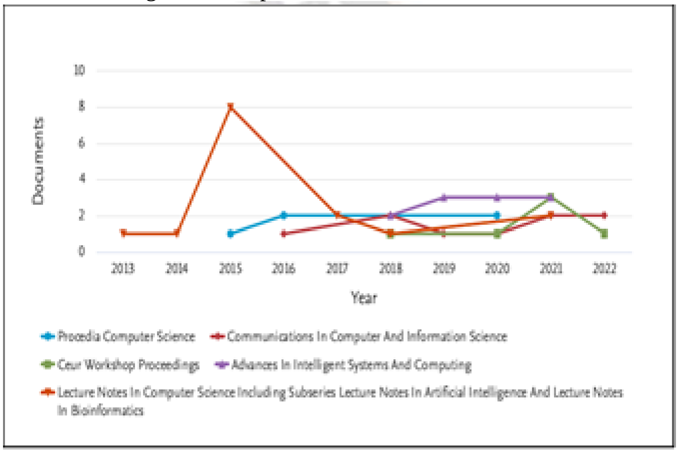

Figure 3. Scopus by subject area

B. *Approaches to Sentiment Analysis*

There are primarily two types of approaches used for analyzing sentiments [1]

1. Machine Learning Approach:

The text is classified using machine learning methods. Whether the data is labeled or not, supervised or un/semi-supervised algorithms are used. A model is trained using these algorithms on a training dataset usually containing the same text category as the test data. The accuracy is then calculated by standard measures like precision, F-score, etc.

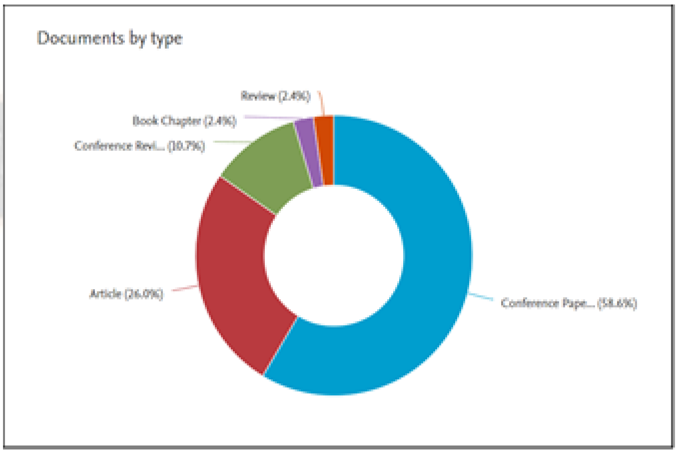

Figure 4. Scopus document by type

2. Lexicon-Based Approach:

In this approach, an already prepared lexicon is used as a reference to classify the text. A lexicon is a collection of words or phrases allocated a polarity score, i.e., +1 for positive, -1 for negative, and 0 for neutral. These individually assigned polarities are then aggregated to get the absolute contradiction. This has two different categories - corpus based and dictionary-based

C. *Components of Sentiment Analysis*

In the studied papers, based on the approach and input data used, there was a slight variation in the process and system architecture. However, the major components were the same, which are shown in figure 5.

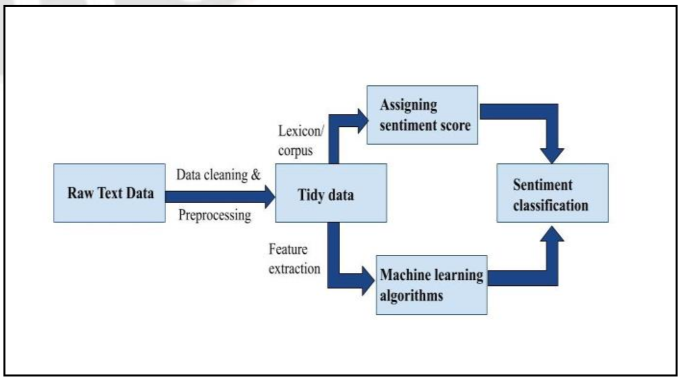

Figure 5. Components of sentiment analysis

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

The initial stage is to gather the raw data that will be used in the analysis. Generally, web scraping is used to get a variety of data from websites like comments, reviews, tweets, etc. In the Data Cleaning and Preprocessing step, this raw input data is cleaned by removing stop words, stemming, POS tagging, etc., and a proper tidy dataset is created. A lexicon-based approach uses a lexicon or corpus with a predefined set of words along with their polarities. The data in the clean dataset is compared with the dictionary, and accordingly, sentiment scores are assigned. Usually, +1, -1, and 0 indicate positive, negative, and neutral polarities.

The cumulative sentiment score is generated by summing the individual ratings to conclude overall polarity. The feature selection step is essential for a machine learning-based approach, where the data is first vectorized. The specific features (words) that might contribute to the sentiment are selected. The classifiers are created using machine learning techniques that predict the sentiment category based on the given features. After the final step of sentiment classification, the accuracy of results is calculated along with the performance evaluation.

II. LANGUAGE-WISE RESEARCH ON SENTIMENT ANALYSIS

A. Marathi

Marathi is the official language of the state of Maharashtra. With roughly 99 million speakers, it is India's third most popular language. The script used in Marathi is called Devanagari. Table 1 shows category of sentiments, some of the sentiment words used in Marathi and its Marathi pronunciation representation in English.

Sentimental features -

Positive: उत्तम (Excellent), सुंदर (Beautiful), आनंदी (Happy),अभिनंदन(Congratulations), प्रगती(Progress), प्रामाणिक(Honest)

Negative: दुःखी(Sad), गंभीर(Serious), नाही(No), बंद(Off), वाईट(Bad)

Neutral: माहिती(Information), निर्णय(Decision), प्रयत्न(Try), नीरस(Disinterested), चर्चा(Discussion)

TABLE I. SENTIMENT WORDS IN MARATHI

Category	Marathi Word	Marathi Pronunciation Representation in English
Positive	उत्तम, सुंदर, आनंदी	Uttam, Sundar, Anandi
Negative	दुःखी, गंभीर, नाही	Dukhi, Gambhir, Nahi
Neutral	माहिती, निर्णय, प्रयत्न	Mahiti, Nirnay, Prayatna

[1] Presented a Lexicon-based approach in which the system uses a predefined lexicon to classify feelings accurately. They concluded that to get efficient results, one requires a richer database. [2] Used a Corpus-based (Lexicon) approach and created a feasible corpus for the Marathi language from English SentiwordNet. The system mainly focused on resource creation and found that framing of the sentences, limited scope of English SentiWordNet, and special characters affected the accuracy. The first significant dataset for Marathi Sentiment Analysis - L3CubeMahaSent, publicly available, was created by [3] and had 16000 tweets. They employed deep learning algorithms to conduct SA on the dataset, with IndicBERT and CNN composed through Indic fastText word embeddings achieving the most incredible accuracy. Chitra [4] did Marathi SA with Marathi WordNet. They used General Architecture for Text Engineering (GATE) data processing and classification tools. It was concluded that more Marathi words must be incorporated to enhance the system, for which the synset replacement algorithm is better. It was also found that an NLP-based approach may perform poorly for grammatically incorrect text. [5] Performed Sentiment Analysis of Marathi text using Unsupervised Learning composed through Word Cloud visualization. The system summarised the clusters of the Marathi corpus using unsupervised learning, which is the first of its kind. Among the algorithms, Fuzzy K-means clustering was found to have better accuracy.

In his thesis, [6] performed SA on Mixed Language Text (MLT) containing Marathi+English and the exact text written in the Marathi Devanagari script. He concluded that if collected data is translated and then classified, it shows better change than working on the original text. They observed that the random Forest classifier had the best accuracy. They also created an open-source dataset for MLT and Devanagari script. [7] Studied the SA of Marathi tweets using various machine learning algorithms. They concluded that the Bag-of-words (BOW) method is the most often utilized way to model text in statistical machine learning approaches to sentiment analysis. [8] Studied various sentiment classification techniques to perform a comparative study of Marathi sentiment analysis. They emphasized the opportunities and challenges researchers encountered regarding Marathi SA and developed a shaded-based approach that falls under the problem of sentiment categorization using a semantic corpus.

[9] Represented machine Learning methodology by performing polarity-based sentiment analysis on Marathi using the deep learning algorithm LSTM (Long Term Short Memory). [10] [11] Researchers proposed features, [12] shown an effect of imbalanced data and pattern style matching presented in [text analysis] for author identification on Marathi. A detailed review on author identification in Marathi is provided in [13] [14] [15]

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

B. Hindi

Hindi is the official language of the Government of India, majorly spoken in Northern states. It is the most predominant language in India, with more than 690 million speakers. Like Marathi, it has also written in Devanagari script. Table 2 shows category of sentiments, some of the sentiment words used in Hindi and its Hindi pronunciation representation in English.

Sentimental features :

Positive: अच्छा(Good), नेक(Noble), बढ़िया(Excellent), सेहतमंद(Healthy), बढ़ोतरी(Growth), आभारी(Thankful), सुगंधित(Fragrant)

Negative: नाराज़(Angry), असफल(Fail), अवैध(Illegal), बेकार(Waste), नाख़ुशी(Unhappiness)

Neutral: जीव(Life), आजकल(Nowadays), खोज(Search), एकमात्र(Only), सोचना(Think), अनुमति(Permission)

TABLE II. SENTIMENT WORDS IN HINDI

Category	**Hindi Word**	**Hindi Pronunciation Representation in English**
Positive	अच्छा, नेक, बढ़िया	Accha, Nek, Badhiya
Negative	नाराज़, असफल, बेकार	Naraaj, Asafal, Bekaar
Neutral	जीव, आजकल, खोज	Jeev, Aajkal, Khoj

[16] Used a Machine learning approach using WordNets and algorithms like SVM, etc., to perform SA on Hindi text. It built upon existing methods by integrating their best and concluded that results were better than individual methods. [17] Examined current work in SA that focuses on multilingual text with indigenous language. The review mentioned the trends in the SA field, concluding that around 67% of researchers used the machine learning approach, and about 29% used the lexicon. For Indian languages, most work has been done for Hindi compared to other languages. [18] Presented an approach aimed at cross-lingual SA for Marathi and Hindi that utilizes features as WordNet synset identifiers for a supervised classifier where SVM was used for classification. The study hopes to perform the same in a multilingual setup. [19] Wrote various approaches for language identification in multilingual texts for sentiment analysis tasks. Their survey concluded that the N-Gram algorithm outperformed others for multilingual texts containing Marathi, Hindi, and English. [20] Thesis focused on Independent Subjective Lexicon creation and sentiment analysis of Hindi using different machine learning algorithms.

[21] Created an annotated dataset for testing and devised rules for handling negation and discourse relations using HindiSentiWordNet for improved performance. The HSWN was expanded to include all modulated words of present root words, improving its scope. [22] Manually created an annotated dataset and trained a Deep Belief Network model for SA in Hindi. It was concluded that DBN shows good performance with very little annotated data, and this semi-supervised approach is easy and quick to set up.

[23] Here researchers focused on mixed language sentiment analysis which was generally used on social media. The authors proposed a strategy by using pseudo label built BERT and TF-IDF. The authors prepared an ensemble model with the Bert model and word frequency information to classify the sentiments of Hindi -English (Hi-En) code-mixed tweets. The reported F1 Score is 0686 and with Ensemble being 0.725 and 0.731. [24] This paper reviews "sentiment analysis of English text." Sentiment analysis, an extension of text mining, includes research on mining sentiments and emotions from real-world data. Traditional sentiment analysis of text, dependent on correct language syntax, can be useful for numerous decision-making procedures. On the other hand, social media comments do not always adhere to strict grammar requirements, with many instances of social media posts written in non-original scripts. Many people in India utilize Hinglish (a mix of Hindi and English) as a colloquial language in their WhatsApp messages, Facebook posts, and Instagram reels, among other things. The latest breakthroughs in sentiment analysis from not lone English text but also code mixed text, as well as various difficulties related to the same, are given in this paper. Selecting a suitable classification model is critical for overcoming the barrier of aspect-based sentiment analysis. In this paper, the validity of several classification approaches, as well as various types of aspects of emotional text data and extraction methodologies, has been examined. This research concludes that most studies commonly utilize TF-IDF for feature extraction because it outperforms other strategies. The Classifier is the next crucial component in sentiment analysis. The most common machine learning approach for handling sentiment analysis problems is Naive Bayes. Future scopes provided as ambiguity issues and aspect-based sentiment analysis in Hinglish language.

[25] This paper emphases on the issue of most Sentiment Investigation approaches identifying only the complete polarity of a sentence in its place of the polarization of all aspect stated in the sentence. It focuses on the Aspect-Based Sentiment Analysis (ABSA) methodology, which classifies the features inside the specified sentence, and the sentiment specified for each feature. Using pre-trained models such as BERT has recently produced state-of-the-art outcomes in natural language processing. This paper proposes two ensemble models grounded on multilingual BERT, namely, mBERT-E-MV and mBERT-E-AS. Using different methods, they constructed a supplementary sentence by this aspect and converted the ABSA problem into a sentence-pair classification task. Then,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

fine-tuning of dissimilar pre-trained BERT models is done, followed by producing an ensemble for a final prediction based on the proposed model. The results indicated that overall, BERT-grounded models performed much better than the other models. [26] Here authors presented the polarity of sentiments on Hindi and Bengali tweets data. In the proposed method, three separate classifiers are constructed, each with a unique feature set. Model performance was evaluated and compared using the 10-fold cross-validation performance metric. The authors prepared a dataset containing 1500 tweets from SIAL 2015 Bengali training and test data, and the average accuracy of each model is computed over ten folds. Experiments have revealed that ensembles of classifiers with a variety of features are effective and achieve 63.5% of accuracy.

[27] The work presented in this paper was inspired by the large-scale improvement of opinion differences and conflicts concerning the new assessment system. The proposed approach is to analyze the responses of community sentiment on Twitter grounded on general arguments either straight or indirectly connected to GST. The dataset collected here includes 200K tweets exclusively about GST from June 2017 to December 2017 in two stages. The authors prepared a topic-sentiment significance model to guarantee the relevance of the scraped tweets. The result of accuracy with the LSTM model is 84.51% reported.

C Tamil

Tamil is a Dravidian language vocal in several southern Indian states. The Tamil script is called the abugida script, which comprises combined units of consonants and vowels. The total count of Tamil speakers is above 77 million, making it the fifth most spoken language in India. Table 3 shows category of sentiments, some of the sentiment words used in Tamil and its Tamil pronunciation representation in English.

Sentimental features:
Positive: நல்ல(Good), வெற்றி(Success), அற்புதமான(Awesome), கைத்தட்டல்(Applause), நன் றாக(Thank You)
Negative: வெறுப்பு(Dislike), சலிப்பு(Boring), குறைபாடு(Defect), வெறுக்கிறேன்(Hate), குற்றம்(Offence)
Neutral: உரையாடல்(Conversation), காலநிலை(Climate), சந்தித்தல்(Meeting), படம்(Picture),
உணவு(Food)

TABLE III. SENTIMENT WORDS IN TAMIL

Category	**Tamil Word**	**Tamil Pronunciation Representation in English**
Positive	நல்ல, வெற்றி, அற்புதமான	Nalla, Verri, Arputhamana
Negative	வெறுப்பு, சலிப்பு, குறைபாடு	Veruppu, Calippu, Kuraipathu
Neutral	உரையாடல், காலநிலை, சந்தித்தல்	Uraiyaadal, Kalanilai, Cantittal

[28] developed a Tamil lexicon corpus and proposed various classification methods for feature subset selection. It was found that Ensemble based classification techniques were the best of the lot. Sajeetha Thavareesan et al. [29] studied the objectives, corpus, features, techniques, and challenges along with the accuracy or F-measure presented in Tamil SA literature. SVM and RNN classifiers utilizing TF-IDF and Word2vec characteristics of Tamil text were more efficient than other classifiers. Vallikannu Ramanathan et al. [30] used enhanced Tamil SentiWordNet, and TF-IDF feature to perform domain-specific ontology and contextual semantic sentiment analysis.

[31] The author suggested an augmented dictionary in Tamil that intends to construct contextual links between terms in multi-domain datasets while minimizing the feature mismatch problem. The original dictionary used pointed, mutual information to generate contextual weights. The final dictionary calculated the rank score based on the importance of phrases across all reviews. This research uses a unified lexicon with an extensive vocabulary to classify reviews from several target domains in Tamil. This extendible dictionary considerably enhanced the accuracy of DA conducted between several domains, yielding an accuracy of 70.5%, which was extremely high considering the multi-domain datasets in Tamil.

[32] Sentiment analysis is much more complex with class imbalance problems. A significant challenge in sentiment analysis is 'Code-Mixing,' which involves using multiple languages in a text or sentence. The author proposes a solution to both challenges using sampling techniques combined with Levenshtein distance metrics. The system was built in three stages: data preprocessing, feature extraction, and in the end, classification. In data preprocessing, SMOTE and ADASYN, these resampling techniques helped improve the F-1 Score by 50%. Using the combination of sampling techniques and the Levenshtein distance metrics helped improve the code-mixing problem, but much more work must be done to improve the class imbalance problem.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
Article Received**: 03 May 2023* ***Revised: *30 June 2023* ***Accepted:*** *21 July 2023*

[33] In this research, the author has proposed a solution to classify the general polarity of feelings expressed in Tamil and Malaya languages over twitter in three categories: positive, negative, and neutral. Dataset is prepared for each language, and preprocessing is performed to label the data. The author has undertaken the deep learning approach for each dataset's classification problem. In conclusion, among Recurrent NN, Recursive NN, and LSTM, LSTM performed best in both the Tamil and Malaya languages, with an average accuracy of 97%.
[34] In the case of sentiment analysis and offensive language identification, the author polled to see if multi-task learning training models are more advantageous than single-task learning training models. The survey's data set consists of code-mixed YouTube comments in Tamil, Malayalam, and Kannada. Experiments have shown that the multi-task learning model outperforms the single-task learning model and also decreases the space and time restrictions for training individual models. Results provided for Sentiment Analysis and Offensive Language Identification are F-1 Score (66.8% and 90.5%), (59% and 70%), and (62.1% and 75.3%) for Kannada, Malayalam, and Tamil, respectively.
[35] Much work has been done on the code-mixed dataset sentiment analysis, but most implementations use traditional methods, LSTM, CNN, and transformer models. Here, the author has explored graph CNN on CMSA. The dataset used for this work was taken from the Dravidian CodeMix FIRE 2020. Initially, the author transliterated the data to build a word document graph for the entire dataset. Then a three-layer GCN with multi-headed attention on CMSA presented promising results by outstripping numerous traditional approaches. Researchers reported the best results of a weighted F1 of 0.75, and an accuracy of 0.73 was obtained on the Malayalam-English CM dataset.
[36] Sentiment analysis of code-mixed datasets is one difficulty, but the dimension of low-resourced languages adds a new level of complexity. For this problem, the author formed a standard Tamil-English code-mixed, a sentiment-interpreted dataset containing 15,744 commentaries from several posts on YouTube. The authors concluded as out of many different classifiers, Random Forest, Logistic Regression, and Decision Tree classifiers performed comparatively better on sentiment analysis. At the same time, the SVM model had low performance on the same. Using deep learning methods also did not help achieve better performance in three automatic metrics. In conclusion, the author has achieved a higher inter-annotator contract in terms of Krippendroff alpha collected from the voluntary annotators through Google forms.

D. Telugu

Similar to Tamil, Telugu is also a Dravidian language with abugida as its written script. It is spoken by over 94 million in southern and some central states. It is the fourth most spoken language in India. Table 4 shows category of sentiments, some of the sentiment words used in Telugu and its Telugu pronunciation representation in English.

Sentimental features:
Positive: ఆనందం (Happiness), గెలుపు(Win), దయ(Kindness), అద్భుతం(Awesome), తీపి(Sweet), పరిపూర్ణమైనది(Perfect)
Negative: నష్టం(Loss), చేదు(Bitter), భీభత్సం(Terror), కూలిపోతుంది([will] Collapse), బాధించింది(Hurt)
Neutral: నిర్ణయం(Decision), సంఘటన(Incident), మార్పు(Change), రాత్రి(Night), సూచన(Reference)

TABLE IV. SENTIMENT WORDS IN TELUGU

Category	**Telugu Word**	**Telugu Pronunciation Representation in English**
Positive	ఆనందం, గెలుపు, దయ	Anandam, Gelupu, Daya
Negative	నష్టం, చేదు, భీభత్సం	Nastam, Cedu, Bheebatsam
Neutral	నిర్ణయం, సంఘటన, మార్పు	Nirnayam, Sanghatana, Marpu

[37] created a manually annotated corpus and word embedding model for Telugu text. He suggested a hybrid method of query choice approach with active learning techniques. He found that the extreme gradient boosting (XGBoost) classifier in combination with the Hybrid query choice method gave the best accurateness. Reddy Naidu et al. [38] performed sentiment analysis with a two-phased approach, subjectivity classification and sentiment classification using Telugu SentiWordNet. [39] A sarcastic statement is difficult to detect since it comprises solely good words that convey a negative attitude. Existing sarcasm detection systems can only detect sarcasm in the English language. To carry out the same in other less resourceful languages like Hindi and Telugu is challenging as valuable datasets are scarce. The author has manually collected data from various resources and prepared a fundamental dataset for sentiment analysis in sarcasm detection using the Telugu language. The authors have used algorithms grounded on hyperbolic features, for example, Intensifier, Interjection, Question mark, and exclamation symbol, and achieved an accuracy of 94%.
[40] Opinion about a movie can be considered a short description of the film and a review. It can be positive, negative,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

or neutral. The author has proposed a framework to look through film surveys utilizing transliteration plots. In short, from Telugu to English into positive, negative, and neutral for better classification. The method is prepared with rule-based NLP and a machine learning approach. The author's main focus was generating a seed list of words based on the type. The author implemented an opinion extraction on Telugu movie analysis using NLP with the proposed system. Since the opinion extraction is created on verbs, adjacent, and adverbs, the proposed model successfully provided good results as high as 96%. [41] Sentiment analysis can also be used for opinion extraction. Since a vast amount of data is generated on e-commerce websites, the reviews on the products can be used to analyze the overall ratings of the same product. Initially, the information is collected and preprocessed to convert unstructured data into structured data. In the next step, the author used a weighted XGBoost classifier to categorize the product analyses and created ratings—5-star meaning excellent and 1-star rating meaning terrible review. For the feature level rating method, the accuracy was 81.02%, f-measure was 87%, precision was 78%, and recall was 89%.
[42] The author proposed a precisely designed framework for sentiment analysis in the Telugu language in this paper. The system combined the Word2Vec word embedding model, a language translator, and deep learning techniques such as RNN and Naive Bayes. The data was collected manually and preprocessed to construct the dataset before applying it to the framework. In conclusion, the accuracy was measured at 80.45%, specificity at 76.57%, and precision at 82.46%. [43] Automatic text summarization under NLP can also recognize the sentiment behind the text. In this paper, the researchers utilized the ADABooster classifier. In conclusion, the results obtained in terms of accuracy with different algorithms were as follows: 78.0 for Hybrid Query Selection Approach, 73.2 for Convolutional Neural Networks, and 80.56 for the proposed Adabooster classifier.

E. Malayalam

Malayalam is a Dravidian language with Vatteluttu as its written script. It is spoken by over 35 million people, primarily in Kerala and the union territories of Lakshadweep and Puducherry. It is the tenth most spoken language in India. Table 5 shows category of sentiments, some of the sentiment words used in Malayalam and its Malayalam pronunciation representation in English.

Sentimental features:
Positive: സന്നമായ(cheerful), അഭിനന്ദിക്കുക(Congratulations), സ്വാദിഷ്ടമായ (Delicious), രസകരം (Fun), ഉൾക്കാഴ്ചയുള്ള (Insightful)
Negative: കഠിനമായ (Harsh), അപമര്യാദയായ (Rude), വിരസത (Boring), ശല്യപ്പെടുത്തുക (Annoy)വേദനിപ്പിച്ചു (Hurt)
Neutral: സംസാരിക്കുക (Speak), ഭക്ഷണം (Food), ജോലി (Work), വാർത്ത (News), വസ്തു (Object)

[44] Given the lack of satisfactory datasets for code-mixed Malayalam-English, the author has created a new gold standard benchmark dataset annotated by voluntary contributors. The author has provided this dataset for the research community to use as a benchmark dataset. In this approach, the author used the following classifiers: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Multinomial Naive Bayes (MNB), Decision Tree (DT), Dynamic Meta-Embeddings DME, K-nearest neighbors (KNN), Contextualized DME (CDME), Bidirectional Encoder Representations for Transformers (BERT), 1D Dimensional Convolution (1DConv), Bidirectional Encoder Representations for Transformers (BERT). In conclusion, 1DConv scores better on recall, precision, and F-1 Score, while BERT fails to identify some classes. DME and CDME are successful in recognizing all of the categories.

TABLE V. SENTIMENT WORDS IN MALAYALAM

Category	**Malayalam Word**	**Malayalam Pronunciation Representation in English**
Positive	പ്രസന്നമായ, അഭിനന്ദിക്കുക, സ്വാദിഷ്ടമായ	Prasannamāya, abhinandikkuka, svādiṣṭamāya
Negative	കഠിനമായ, അപമര്യാദയായ, വിരസത	kaṭhinamāya, apamaryādayāya, virasata
Neutral	സംസാരിക്കുക, ഭക്ഷണം, ജോലി	Sansārikkuka, bhakṣaṇaṁ, jēāli

[45] As sentiment analysis lacks work in regional languages, the author has proposed an approach to achieve sentiment analysis for Malayalam tweets using deep learning approaches. The author has used CNN and LSTM as primary classifiers and compared the results with traditional classifiers such as SVM. The output is categorized into three categories of opinion, i.e., namely positive, negative, and neutral. Further, it is concluded that deep learning classifiers outperformed traditional classifiers on the developed dataset. It was also observed that CNN and ReLU, ELU, and SELU activation functions gave much better results.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

[46] To contribute to the research community, the author has conducted a sentiment analysis on the code-mixed language dataset 'Dravidian-CodeMix-FORE2020'. The approach used for this experiment was the AWD-LSTM model and the ULMFiT framework incorporating the FastAi library for classifying the input into the following categories: positive, negative, neutral, mixed emotions, and not Malayalam. Using this approach, the author has successfully implemented the approach and achieved the weighted F1 Score of 0.6 equally for Malayalam-English and Tamil-English languages. The author also concludes that results can be further improved by handling data imbalances.

F. Bengali

In India, the Bengali language is frequently verbal in the eastern region and is a native language of the Bengal state. With over 100 million speakers, it is India's second most verbal language. Written Bengali is an abugida script. Table 6 shows category of sentiments, some of the sentiment words used in Bengali and its Bengali pronunciation representation in English.

Sentimental features :
Positive: উত্তেজিত(Excited), ভাল(Good), মনোযোগী(Attentive), অসাধারণ(Awesome), সুন্দর(Beautiful)
Negative: খারাপ(Bad), পরাজয়(Defeat), ব্যাথা(Pain), সমস্যা(Problem), বিরাগ(Disgust)
Neutral: গান(Song), তত্ত্ব(Theory), কারণ(because), পড়া(read), গমন(going)

TABLE VI. SENTIMENT WORDS IN BENGALI

Category	Bengali Word	Bengali Pronunciation Representation in English
Positive	উত্তেজিত, ভাল, সুন্দর	Uttejito, Bhalo, Shundor
Negative	খারাপ, পরাজয়, ব্যাথা	Kharap, Porajay, Betha
Neutral	গান, তত্ত্ব, কারণ	Gana, Tattva, Karan

[47] Worked with a deep learning model for their sentiment classification task. Along with two classes, they also created a 3 class sentiment dataset and built a classifier, which is a first for Bengali text. A multilingual BERT algorithm which was enhanced by adding three more network layers - Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM), and Convolutional Neural Network (CNN), was shown to have greater accuracy than existing models. The results also concluded that political and sports-related news contained more negative comments while religious news gathered more positive sentiments from people. [48] Performed sentiment analysis on both the original Bengali text and its machine-translated English version. The class balancing of datasets with Synthetic Minority Over-sampling Technique and machine-translated text showed improved classification performance. The authors concluded that lack of resources was a significant factor and that work needs to be done so that the scope of the study could be extended from bilingual to multilingual classification. [49] Worked towards creating a rich Bengali and English code-mixed dataset. To save the efforts for manual tagging, they used a hybrid approach by building their language tagging system with lexicon-based and supervised learning modules and a sentiment tagging system combining rule-based and supervised methods. Out of all the supervised methods, stochastic gradient descent had higher accuracy.

[50] analyzed sentiment with numerous machine learning algorithms. They classified the Bengali language data into five sentiments which are happy, angry, sad, excited, and surprised. The data was also classified as belonging to vicious and religious categories. The authors conclude that the SVM algorithm provides good accuracy. [51] Word Sense Disambiguation (WSD) is the method of recognizing the real sense of a word grounded on what context it is used. Even though much work is done in this field, there is a discontinuity when working on Bengali WSD. In this paper, the author has surveyed various approaches to Bengali WSD and the existing work of Bengali WSD. The author has classified existing datasets into four modules: raw corpora, WordNet, semi-annotated, and MRD. In an unsupervised approach, the author surveyed two methods and attained average correctness of 58.5%. In the knowledge-based approach, the author surveyed two techniques and achieved an average accuracy of 75%. In the supervised approach, the author surveyed three methods and attained an average accuracy of 71%.

[52] Even though Bengali is the seventh-highest ranked language globally, the work done on sentiment analysis for the Bengali language is minimal. Due to the lack of datasets, the author has also constructed a Bengali Sentiment Analysis Dataset (BSOD) of 8122 text expressions. The author has surveyed eight popular traditional classifiers, from Logistic Regression to AdaBoost, with TF-IDF and BoW features for sentiment analysis on three datasets (BSaD and two benchmark datasets). Further, the author has developed four ensemble methods by combining the three best-performing traditional classifiers: LR, RF, and SVM. In conclusion, the author has observed that the ensemble approach, along with TF-IDF features, performed significantly better than traditional classifiers, providing the uppermost correctness of 82% on BSOD. [53] Dataset was constructed using the 1469 available text sentences in Bengali by 20 male and 20 female individuals who spoke the Bengali language fluently. Data were categorized into three stimuli: evaluation, potency, and activity. This study used pan-respondent component analysis to examine

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

the respondent's usage of the EPA scale. Interesting patterns were found when looking at the data collected from the respondents who used the scale correctly. Potency scores had a curvilinear nature with evaluation for the respondents of both genders. For evaluation, gender correlations are as high as 0.93 but low for potency scores at 0.55 and even low at 0.30 for activity scores. In conclusion, the two cultures are very much similar in the case of evaluations, less similar in potency, and barely similar in activity ratings.

[54] For sentiment analysis of Bengali and Hindi tweets, the author has used the SAIL 2015 dataset as a benchmark dataset. The proposed approach combines a Multinomial NB classifier through character n-gram features, a Multinomial NB classifier through word n-gram features, and SVM through unigram features in an ensemble. The author has also incorporated the sentiment lexicon in the model. The author has observed that the 'majority voting' rule gave better results aimed at sentiment analysis of Bengali tweets and the 'average of probabilities' rule gave better results for sentiment analysis of Hindi tweets. In conclusion, Multinomial NB with character n-gram feature and sentiment lexicon performed the best out of all three base classifiers.

[55] Here authors used an attention-based deep learning model. Experimentation results show as even though the attention mechanism in CNN (A-CNN) performed better than A-LSTM at 96.55%, it fails to achieve the overall result of accuracy (66.06%), precision (66.04%), recall (65.66%) and F-measure (66.02%). Further, the author used the CGAN network to produce artificial data, which resulted in a performance improvement of A-CNN by nearly 6%.

G. Gujarati

Gujarati is the native language of the state of Gujarat. It ranks seventh in the list of languages spoken in India, with over 60 million speakers. The Gujarati script is a variant of Devanagari and is an Abugida. Table 7 shows category of sentiments, some of the sentiment words used in Gujarati and its Gujarati pronunciation representation in English.

Sentimental features:
Positive: વિશાળ(Huge), પ્રેમ(Love),સમર્થન(Support), સુવિધા(Convenience), ઉત્સવ(Festival),વિવિધતા(Variety)

Negative: મર્યાદિત(Limited), વિરોધી(Anti), હારી(Lost), તૂટી(Broken), ગેરકાયદેસર(Lawless), જોખમ(Danger)
Neutral: સામાન(Baggage), પદાર્થ(Substance), પ્રશ્ન(Question), રસ્તો(Way), બનાવટ(Creation),પ્રસંગ(Event)

TABLE VII. SENTIMENT WORDS IN GUJARATI

Category	**Gujarati Word**	**Gujarati Pronunciation Representation in English**
Positive	વિશાળ, પ્રેમ, સમર્થન	Vishal, Prem, Samarthan
Negative	મર્યાદિત, વિરોધી, હારી	Maryadit, Virodhi, Hari
Neutral	સામાન, પદાર્થ, રસ્તો	Saman, Padarth, Rastro

[56] Compared the accuracy of K-Nearest Neighbor and multinomial Naive Bayes when paired with two feature selection methods - TF-IDF and CountVectorizer. They found that MNB and TF-IDF had better accuracy than with CountVectorizer. The accuracy of KNN was the same in both cases. [57] combined POS tagging with an SVM classifier to improve its performance and achieved 92% accuracy for sentiment analysis. They concluded that preprocessing data increases the performance of classifiers, and other feature extraction techniques can be studied to test their performance with such classifiers.

[58] Studied how stemming algorithms affect the categorization of Gujarati web pages. The focus was mainly on the Gujarati STEmmeR (GUJSTER) algorithm and syllable tokenizer and dynamic stop words identification as part of preprocessing. They concluded that stemmer algorithms considerably influence supervised algorithms, improving their accuracy, specifically for problem statements of web page categorization.

[59] Created a Gujarati SentiWordNet (G-SWN) with the help of Hindi SentiWordNet and IndoWordNet. A lexical approach was used to classify sentiments on manually annotated corpora using the above G-SWN. They also stated that the accuracy could be improved in the future using Word Sense Disambiguation.

[60] For sentiment analysis in the Gujarati language, the dataset is organized by collecting reviews of numerous products in the Gujarati language. In this paper, the author has utilized two machine learning classifiers on the collected dataset, K-nearest neighbors (KNN) and Multinomial Naive Bayes (MNB). The author has also considered TF-IDF and Word Level Count Vectorizer to attain good results. In conclusion, the author has observed that the MNB classifier with TF-IDF improved performance as related to Word Count Vectorizer, and the KNN classifier had the same result in the case of TF-IDF as well as Count Vectorizer. [61] In this paper, the author has used sentiment analysis to classify tweets into eight categories: joy, surprise, fear, trust, sadness, anticipation, anger, and disgust. Two datasets are organized by gathering tweets linked to Indian politics and are marked manually for each language: English, Gujarati, and Hindi. The experiment is conducted for each language using basic machine learning classifiers and a hybrid classifier with SN and CS-SN for feature generation algorithms.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

In conclusion, it was observed that the machine learning approach gave better results than the hybrid approach for the Gujarati language. The author has also mentioned various reasons for the low performance of the hybrid approach.

[62] In this paper, the author has proposed a strategy to achieve sentiment analysis in the Gujarati language on tweets. The author has focused on classifying the data into positive and negative polarities. From the literature survey, the author understood that traditional machine learning classifiers gain performance by using preprocessing techniques such as POS tagging. In conclusion, the author used POS tagging for feature extraction and further applied the dataset to the SVM classifier. Results obtained by this approach gave an accuracy of 92%.

H. Urdu

Urdu has speakers across many states in India. Even though not an official language of any state, it is given some form of official recognition in certain states. It is spoken by over 62 million people making it the sixth most spoken language In India. Urdu is written from right to left and is closely related to Persian script. Table 8 shows category of sentiments, some of the sentiment words used in Urdu and its Urdu pronunciation representation in English.

Sentimental features :

Positive:,(Brave)بہادر ,(Fresh)تازہ ,(Better)بہتر ,(Interesting)دلچسپ (Intelligent)ذہین ,(Celebration)جشن

Negative:,(Expensive)مہنگا ,(Controversy)بحث ,(Disappointed)مایوس ,(Lack)کمی ,(Upset)پریشان

Neutral: (Environment)ماحول ,(Occupation)پیشہ ,(Location)جگہ ,(Process)عمل ,(River)دریا

TABLE VIII. SENTIMENT WORDS IN URDU

Category	Urdu Word	Urdu Pronunciation Representation in English
Positive	بہتر ; دلچسپ ; بہادر	Bahadur, Dilchasp, Behtar
Negative	بحث ; مایوس ; مہنگا	Mehnga, Behas. Mayors
Neutral	پیشہ ; جگہ ; ماحول	Mahol, Pesha, Jagah

[63] Focused on extracting the SentiUnits, sets of words/phrases which contributed to the sentiment of the text, and used a lexicon-based approach for the sentiment analysis. They also studied how various Urdu adjectives and their alternates add to the sentiment. It was concluded that work is needed to expand the lexicon for better results. [64] performed sentiment analysis using Word2Vec as a text vectorizer and LSTM neural network model as a classifier. An activation function called SoftMax, which assigns the polarity using a probabilistic approach was included in LSTM. It was concluded that followed approach is having better accuracy.

[65] Studied aspect-based sentiment analysis focusing on individual words/entities/aspects within the sentences. The team created a corpus for it containing information related to the aspect, its polarity, its category, and the polarity of the category. This is a benchmark in ASBA of Urdu text, and more machine learning algorithms and feature extraction methods could be tried out in the future for this task. [66] Worked towards creating a dataset containing Urdu tweet texts. By adding emojis and their polarity, the dataset is made useful for sentiment analysis purposes. The authors proposed that this large dataset can be used in fields like machine learning, NLP, and information retrieval. Tooba Tehreem et al. [67] Performed sentiment analysis on roman-Urdu text and did a comparative study of five classifiers. It was found that SVM performed the best when paired with Bag of Words feature extraction. The paper also concluded that various feature extraction methods could be used in the future to improve accuracy.

[68] Sentiment Analysis for YouTube Comments in Roman Urdu: Even though sentiment analysis is a vast area in machine learning, most work is done considering the English language. Roman Urdu is the language that is spoken by a majority of the population in Pakistan. For this case study, the author used people's comments on various Pakistani dramas and TV shows on YouTube. The author mainly focused on classifying the comments into positive, negative, and neutral categories. Dataset was tested using the following supervised learning algorithms: linear regression, SVM, Naive-Bayes, Multilayer Perceptron, and KNN classifier. Out of these algorithms, SVM had the highest accuracy of 64%.

[69] A survey on sentiment analysis in Urdu: A resource-poor language: Although the volume of studies on sentiment analysis is increasing rapidly, English is the primary language of concern. In this paper, the author mainly focused on describing three dimensions used for Urdu sentiment analysis: text preprocessing, lexical resources, and sentiment classification. In the survey's conclusions for recognizing the progress and shortcomings of Urdu sentiment analysis, the author has proposed guidelines for future work to acknowledge six critical points for rectification.

[70] Sentiment Analysis of Roman-Urdu Tweets about Covid-19 Using Machine Learning Approach: Due to the complex morphological structure and unavailability of resources, it is difficult to work on sentiment analysis in Urdu. The dataset was prepared using Twitter data throughout the pandemic to understand the pattern in people's sentiments. The author has given a general overview of the process of sentiment analysis of Roman-Urdu Tweets and has observed that TF_IDF with

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

Unigram and Bigram are the most used features used for the same. Currently, Hybrid approaches do not give better results since a lot more work is required in this field. Even when the dataset size is large, many studies have achieved good results using the Naive Bayes algorithm. In conclusion, the author concludes that the labeled dataset gives more accurate results than the unlabeled dataset. [71] Sentiment Analysis on Urdu Tweets Using Markov Chains: Compared to the work done in the field of NLP on the English language, minimal work has been conducted on the languages like Urdu, Bengali, Hindi, and other Asian languages. In this paper, the author focuses on developing a three-class sentiment analysis model for the Urdu language. The dataset for Urdu tweets was collected by using Twitter API. The author has proposed a methodology based on the Markov chain model to conduct sentiment analysis on the Urdu Tweets dataset. In conclusion, the author found that this methodology gave better results than the lexicon-based approach and other common machine-learning-based approaches to sentiment analysis. However, the author also observed that the model gives bad results for predicting positive Urdu Tweets because of the lack of positive tweets in the dataset.

[72] Even though almost 169 million people are familiar with the Urdu language and a large amount of data is generated on various internet platforms daily, very few efforts have been made to perform sentiment analysis on the Urdu language. The author has focused on evaluating various machine and deep learning algorithms based on two text representations: n-gram features and pre-trained word embeddings. In conclusion, the author mentions that the highest accuracy of the F1 Score of 82.05% was achieved using the LR with a combination of various features. SVM classifier gave the second best results for the sentiment analysis as compared to all the other machine learning classifiers.

III. SUMMARY OF FINDINGS

For sentiment analysis in the English language, there is an abundance of benchmark datasets from numerous platforms available on the internet. As a result, much research is already available on sentiment analysis for the English language. Nevertheless, significantly fewer datasets are available regarding regional languages other than English (Hindi, Tamil, Urdu). As there is a scarcity of benchmark datasets, not many previously done research results are available to compare with the current work.

1. The Lexicon (corpus) based approach and machine learning approach are commonly used for sentiment analysis.
2. We see that the majority of studies used SentiWordNets for sentiment scoring and classified the text into three classes: positive (+1), negative (-1), and neutral (0).
3. Various machine learning algorithms are used as classifiers in which SVM, RF, and NB are the most common.
4. Almost all the papers mentioned that the dataset required for analysis (either the lexicon/corpus or training dataset for the ML approach) was curated manually, either from scratch or expanding the WordNets using different techniques.
5. Many studies focused on bilingual, mixed-language, and transliterated text input practically used in day-to-day lives rather than monolingual texts.
6. Within the last couple of years, the study of deep learning methods for sentiment analysis has increased, with many researchers using neural networks alongside existing machine learning algorithms.

TABLE IX. SENTIMENT ANALYSIS OF DIFFERENT INDIAN REGIONAL LANGUAGES

Reference	Language	Dataset	Features and Algorithms	Results
[1]	Marathi	Text With Different Marathi Words	Lexicon	A User Interactive Webpage Classifying Input Sentence Was Created
[2]	Marathi	Text Files With Various Marathi Keywords And Their Meanings	Corpus (Lexicon), Marathi Sentiwordnet, English Sentiwordnet	Accuracy 60-70%
[3]	Marathi	Tweets	Cnn, Bert, Ulmfit, Bilstm, Indicbert, Various Word Embeddings	Accuracy 93.13%
[4]	Marathi	Users' Reviews	Lexicon, Marathi Wordnet, General Architecture For Text Engineering (Gate), A Nearly-New Information Extraction System (Annie),	Polarity Calculated Using Given Lexicon Approach
[5]	Marathi	Different Marathi Stories Belonging To Different Domains	Corpus, TF-IDF, K-Means, Fuzzy K-Means, Hierarchical Agglomerative Clustering, Word Cloud	Fuzzy K-Means Found To Have Better Accuracy.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

[6]	Marathi	MLT And Marathi Text Datasets	Machine Learning Approach, Random Forest, K-Nearest Neighbor, Naïve Bayes, Decision Tree, SVM, Logistic Regression Algorithm, Google Cloud Translator	Accuracy Random Forest - 65.41%, SVM - 64.16%
[7]	Marathi	Tweets	Machine Learning Approach, Bow, Tf-If, Unigram With Sentiwordnet, Nb, Svm, Rf	Sa Using Various Machine Learning Algorithms Is Done
[8]	Marathi	Web Scraping	Various Sentiment Classification Techniques	A Shaded-Based Approach Is Proposed Under Semantic-Corpus Based Sentiment Classification Problem.
[9]	Marathi	E-News	Machine Learning Approach, Lstm (Long Term Short Memory) Deep Learning Algorithm	Accuracy 72%
[10]	Hindi	Transliterated/ Bilingual Marathi And Hindi Texts, Hindi - English Transliteration Pairs Collected From Fire 2013	Language Identification, POS Tagging, Polarity Identification, Wordnets, SVM, Random Forest, And Naive Bayes	Accuracy Upto 95%
[11]	Hindi	Web Scraping	Lexicon And Machine Learning Based Approaches	Around 67% of Researchers Used A Machine Learning Approach And About 29% Used Lexicon.
[12]	Hindi	Travel Destination Reviews	Wordnets, Wordnet Senses, Corpus Of Synset Identifiers, Word Sense Disambiguation (IWSD) Algorithm, SVM	Accuracy 72%
[13]	Hindi	Web Scraping	Lexicon, Phonetic Algorithms Like Soundex And Dmetaphone, N-Gram	Accuracy 90.20%
[14]	Hindi	Reviews And Blogs	Subjective Lexicon, Wordnets, N-Gram, Weighed N-Gram, Svm, Naive Bayes	Accuracy 61.6%
[15]	Hindi	Reviews	Lexicon, Hindi Sentiwordnet	Accuracy 80.21%
[16]	Hindi	Movie Reviews	Semi-Supervised Learning, Deep Belief Network Model	Accuracy 64%
[17]	Hindi-English	Twitter	Pseudo Label With BERT & TF-IDF With SGD	0.731 F1 Score
[18]	Hindi-English	Youtube Comments	TF-IDF With NB Classifier	85% Accuracy
[19]	Hindi	Hindi Websites	Aspect Based Mbert	79.7% Accuracy
[20]	Hindi, Bengali	Sail-2015	Heterogeneous Ensemble Classifier	62.6% Accuracy
[21]	Hindi	Twitter	LSTM	84.5% Accuracy
[22]	Tamil	Mobile Product Reviews	Basic, Fuzzy, And Ensemble Classification Methods, Decision Tree, Naïve Bayes, Nbtree, Rough Set, And Svm, Bagging, Boosting, And Stacking Classification Techniques	Accuracy Basic - 77%, Fuzzy - 84%, Ensemble - 91%
[23]	Tamil	Web Scraping	Tf-Idf, Word2vec, Presence Of Words, Tf And Bow As Features, Tamil Sentiwordnet, N-Grams, Svm, Rnn	Accuracy Save - 75.96%, Rnn - 88.23%
[24]	Tamil	Tweets	Tamil Sentiwordnet, Tf-Idf, Python Nlp	Accuracy 77.89%

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

[25]	Tamil	Amazon & Movie Reviews	ESD-DA	70..5% Accuracy
[26]	Tamil-English	Youtube Comments	Levenshtein Distance Metric With Traditional Classifiers	81.5% Accuracy
[27]	Tamil, Malayalam	Twitter	LSTM	97% Accuracy
[28]	Tamil, Malayalam, Kannada	Youtube Comments	Members Subjected To Cross-Entropy Loss	75.3% Accuracy
[29]	Tamil-English	Dravidiancodemix FIRE 2020	3-Layer GCN On CMSA	0.75 F1 Score
[30]	Tamil-English	Youtube Comments	Random Forest	0.65 F1 Score
[31]	Telugu	E-News	Hybrid Query Selection Strategy, Active Learning, Svm, Extreme Gradient Boosting (Xgboost), Gradient Boosted Trees (Gbt)	Accuracy 79%
[32]	Telugu	E-News	Lexicon, Telugu Sentiwordnet	Accuracy : Subjectivity Classification- 74% Sentiment Classification- 81%
[33]	Telugu	Twitter, Youtube Comments	Hyperbolic Feature Based NB	94.14%
[34]	Telugu	Movie Reviews	Transliteration	96%
[35]	Telugu	Amazon Product Reviews	Feature-Level Rating With Xgboost Classifier	81% Accuracy
[36]	Telugu	Twitter	RNN And NB	80.5% Accuracy
[37]	Telugu	Amazon Product Reviews	Adabooster Classifier	80.5% Accuracy
[38]	Malayalam	Youtube Comments	1D Dimensional Convolution	0.63 F1 Score
[39]	Malayalam, Kannada, Tamil	Twitter	CNN With ELU Activation Function	98.1% Accuracy
[40]	Malayalam-English	Dravidiancodemix FIRE 2020	Ulmfit Framework With AWD-LSTM Classifier	0.6 F1 Score
[41]	Bengali	E-News	Multilingual Bert, Gru, Lstm, Cnn, Word2vec, And Fast text Word Embeddings	Accuracy : 2 Class - 71% 3 Class - 60%
[42]	Bengali	Sports Comments (Cricket) And Drama Reviews	Google Machine Translation Service, Synthetic Minority Over-Sampling Technique (SMOTE), LR, RR, SVM, RF, ET, And LSTM	Classifier Performs Better With Machine Translated Text. Unigram Model Has Better Accuracy Than N-Gram Model
[43]	Bengali	Tweets	Bengali Sentiwordnet, LBM, SLM, Naive Bayes, LRC, SGDC, Code-Mixed Index (CMI), Code-Mixed Factor (CF)	Kappa Values Language Tag - 0.83 Sentiment Tag - 0.94
[44]	Bengali	Facebook Comments	TF-IDF, SVM, RF, KNN, NB, NN	Accuracy : SVM - 62%, RF - 58%, KNN- 55%, NB - 52%, NN - 50%
[45]	Bengali	Generic	WSD With Knowledge-Based Approach	75% Accuracy

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

[46]	Bengali	Bad	Ensemble Approach With TF-IDF Features	82% Accuracy
[47]	Bengali	Manual Data Collection	EPA Scale	0.93 Correlation
[48]	Bengali	Sail 2015	Heterogeneous Classifier Ensemble Model With Majority Voting Combination Rule	62.6% Accuracy
[49]	Bengali	BBC Bangla And Prothom Alo	A-CNN	72% Accuracy
[50]	Gujarati	Movie Reviews	TF-IDF, Countvectorizer, KNN, MNB	Accuracy : MNB - 87.14% KNN - 81.43%
[51]	Gujarati	Tweets	SVM, POS Tagging, N-Gram	Accuracy 92%
[52]	Gujarati	Web Pages	GUJSTER, SVM, RF, KNN, MNB, MLR, GB	Accuracy 75% To 97%
[53]	Gujarati	Tweets	Hindi Sentiwordnet, Indo Wordnet, Unigram Presence, Simple Scoring	Accuracy: Unigram Presence - 52.72% Simple Scoring - 52.95%
[54]	Gujarati	Movie Review	MNB Classifier With TF-IDF	86.1% Accuracy
[55]	Gujarati	Twitter	Linear SVC With TF-IDF Feature	84% Accuracy
[56]	Gujarati	Twitter	SVM Classifier With POS Tagging Feature	92% Accuracy
[57]	Urdu	Movie And Product Reviews	Sentiunits, Shallow Parsing, Lexicons	Accuracy: Movie Reviews - 72% Product Reviews - 78%
[58]	Urdu	Reviews	Word2Vec, RNN, LSTM, Softmax, NB, ELM	F-Measure 0.849
[59]	Urdu	Sports (Cricket And Football) Tweets	TF-IDF, N-Gram, NB, RF, KNN	Successfully Created ABSA Dataset Containing Four Types Of Information
[60]	Urdu	Tweets	Twitter Search API, Data Preprocessing	Successfully Created An Urdu Dataset Having 1,140,825 Tweets
[61]	Urdu	Youtube Comments	Linear Regression, SVM, KNN, Multi-Layer Perceptron, And NB, BOW	Accuracy 64%
[62]	Urdu	Youtube Comments	SVM	64% Accuracy
[63]	Urdu	Movie Review	LSTM	95% Accuracy
[64]	Urdu	Twitter	NB Classifier	97% Accuracy
[65]	Urdu	Twitter	Markov Chain Model	0.857 F1 Score
[66]	Urdu	User Reviews From Internet	LSTM Vs. LR	0.77 & 0.82 F1 Scores

IV. CHALLENGES

After studying the research papers, it became evident that the biggest challenge faced by all researchers was the lack of resources available for sentiment analysis of the regional languages. The availability of gold-standard benchmark datasets in the case of regional languages (Hindi, Tamil, etc.) is deficient. Datasets already available have much noise and the problem of class imbalance. As there was a lack of satisfactory datasets, data often had to be manually collected and annotated. Most Indian languages have different grammar, syntax, and composition than a language like English. Hence the extensive work done for the English language (like English SentiWord, etc.) cannot guarantee a high level of accuracy if the same method is used for these local languages. Grammatically incorrect sentences, transliterated text, and mixed language can affect accuracy and efficiency. A comparatively good amount of research is done on Hindi sentiment analysis which can be helpful for other languages, but the limited scope of respective WordNets does not help either.

On top of that, code-mixed datasets add another layer of complexity. To solve the problem of code-mixed datasets, traditional classifiers had to be combined with various features to adapt to a completely different hybrid approach. Further, the challenge of more than two languages in a single sentence appeared in some code-mixed datasets.

V. CONCLUSION

After studying the literature, it can be concluded that a significant amount of research has been done in recent years regarding the Sentiment Analysis of Indian regional languages. From creating publicly available high-quality datasets/corpus to using various machine learning algorithms for sentiment scoring, many different techniques and approaches have been tried to improve the overall accuracy and performance of the sentiment models. The performance of a given classifier was found to be directly proportional to the dataset's quality. When gold-standard datasets were available for a given language, most of the approaches used to perform sentiment analysis gave better performance. Similarly, poor results were observed when the dataset for a given language was of poor quality (poorly annotated datasets, class imbalance, noise, etc.).

Further, it was observed that dataset cleaning and data preprocessing were performed for code-mixed datasets. Using ensemble techniques (combining best-performing traditional classifiers) gave significantly better performance than traditional classifiers. It was also observed that ensemble techniques and feature extraction methods significantly boosted performance.

VI. FUTURE SCOPE

1. As few benchmark datasets are available for low-resourced languages, newly created gold-standard datasets would give better results and help the research community to proliferate and learn more. The most important and needed is the availability of large datasets like WordNets, SentiWordNets, and similar lexicons for each regional language. Data augmentation in NLP helps give different instances of data samples that can simulate real-time data. This will help improve the performance of the system significantly.
2. The current work mainly considers single sentences as input so that the study can be done for more significant text inputs. As many machine learning algorithms have been explored for analysis, using the ones with the highest accuracy and fine-tuning them to increase performance will be helpful in case of such large documents given as input.
3. Web-based platforms are ridden with fake news (Ex. Fake product reviews). Considering this challenge, fake opinion detection, and filtering, along with sarcasm/irony detection, should be implemented to achieve better results.
4. A system should also detect idioms and proverbs to get accurate context and sentiment out of the data. It should be able to analyze slang, emoticons, and mixed language text more commonly used daily.
5. Considering the vast number of Twitter-based datasets, a system should be built to successfully evaluate bi-lingual and font-mixed tweets to enhance overall accuracy.
6. NLP systems often face the challenge of adequately identifying the words and determining the specific usage. In the surveyed research papers, word sense disambiguation for text processing was the least addressed issue. If solved, this will help achieve better results.

CONFLICTS OF INTEREST

The authors declares that there is no conflict of interest regarding the publication of this paper.”

REFERENCES

[1] Snehal Pawar, Swati Mali, "Sentiment Analysis in the Marathi Language," IJRITCC, vol. 5, no. 8, pp. 21-25, Aug. 2017. Sentiment Analysis in the Marathi Language | International Journal on Recent and Innovation Trends in Computing and Communication (ijritcc.org)

[2] Sujata Deshmukh, Nileema Patil, Surabhi Rotiwar, Jason Nunes, “Sentiment Analysis of Marathi Language, "IJRPET, vol. 3, no. 6, pp. 93-97, Jun. 2017. SENTIMENT ANALYSIS OF MARATHI LANGUAGE

[3] Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, Raviraj Joshi, “L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset”, arXiv:2103.11408v1 cs.CL], 21 Mar 2021. https://arxiv.org/abs/2103.11408v1

[4] Chitra Chaudhari, Ashwini Khaire, Rashmi Murtadak, Komal Sirsulla, "Sentiment Analysis in Marathi using Marathi WordNet," IJIR, vol. 3, no. 4, pp. 1253-1256, 2017. Sentiment Analysis in Marathi using Marathi WordNet

[5] Prafulla Bafna, Jatinderkumar Saini, “Marathi Text Analysis using Unsupervised Learning and Word Cloud”, IJEAT, vol. 9, no. 3, pp. 338-343, Feb 2020. International Journal of Recent Technology and Engineering (IJRTE)

[6] Harry Gavali, “Text Sentiment Analysis of Marathi Language in English And Devanagari Script”, Dublin Business School, Jan. 2020. https://esource.dbs.ie/bitstream/handle/10788/4216/msc_gavali_h_2020.pdf

[7] Renuka Naukarkar, Dr. A. N. Thakare, “A Review on Recognition of Sentiment Analysis of Marathi Tweets using

Machine Learning Concept", IJSRSET, vol. 8, no. 2, pp. 190-193, Mar. 2021. IJSRSET

[8] Monali Patil, Nandini Chaudhari, B.V. Pawar, Ram Bhavsar, "Exploring various emotion-shades for Marathi Sentiment Analysis", 2021 Asian Conference on Innovation in Technology (ASIANCON), pp. 1-5, 2021. https://ieeexplore.ieee.org/document/9544961

[9] Manisha Date, "Sentiment analysis of Marathi news using LSTM", IJIT, vol. 13, 2021. https://link.springer.com/article/10.1007%2Fs41870-021-00702-1

[10] Kale Sunil Digamberrao, Rajesh S. Prasad, Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi, Procedia Computer Science, Volume 132, 2018, Pages 1086-1101, https://doi.org/10.1016/j.procs.2018.05.024.

[11] Kale, Sunil Digambarrao and Rajesh S. Prasad. "Influence of Language-Specific Features for Author Identification on Indian Literature in Marathi." (2019).

[12] Kale, Sunil Digamberrao and Rajesh S. Prasad. "Author Identification on Imbalanced Class Dataset of Indian Literature in Marathi." *International Journal of Computer Sciences and Engineering* (2018).

[13] Kale, Sunil Digamberrao and Rajesh Shardanand Prasad. "A Systematic Review on Author Identification Methods." *Int. J. Rough Sets Data Anal.* 4 (2017): 81-91.

[14] Kale. Sunil Digamberrao and R. S. Prasad, "Author Identification on Literature in Different Languages: A Systematic Survey," *2018 International Conference On Advances in Communication and Computing Technology (ACCT)*, 2018, pp. 174-181, DOI: 10.1109/ICACCT.2018.8529635.

[15] Amidwar, Shubhesh et al. "Text Analysis for Author Identification using Machine Learning." *Journal of emerging technologies and innovative research* (2017):

[16] Mohammed Ansari, Sharvari Govilkar, "Sentiment Analysis of Transliterated Hindi and Marathi Script", Sixth International Conference on Computational Intelligence and Information Technology – CIIT , pp. 142-149, 2016. (PDF) Sentiment Analysis of Transliterated Hindi and Marathi Script

[17] Sonali Shah, Abhishek Kaushik, "Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining", Preprints, 2019110338, 2019. Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining

[18] Balamurali A R, Aditya Joshi, Pushpak Bhattacharyya, "Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets ", Proceedings of COLING 2012: Posters, pp. 73-82, Dec. 2012. (PDF) Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets

[19] Deepali Londhe, Aruna Kumari, Emmanuel M., "Language Identification for Multilingual Sentiment Examination", IJRTE, vol 8, no. 2S11, pp. 3571-3576, Sep. 2019. Language Identification for Multilingual Sentiment Examination

[20] Piyush Arora, "Sentiment Analysis For Hindi Language", International Institute of Information Technology Hyderabad - 500 032, April 2013. Sentiment Analysis For Hindi Language

[21] Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek, "Sentiment Analysis of Hindi Review based on Negation and Discourse Relation", IJCNLP, pp. 45-50, Oct 2013. Sentiment Analysis of Hindi Reviews based on Negation and Discourse Relation

[22] Naman Bansal, Umair Ahmed, Amitabha Mukherjee, "Sentiment Analysis in Hindi", Indian Institute of Technology Kanpur, Sentiment Analysis In Hindi

[23] Bao, Wei, et al. "Will_go at SemEval-2020 Task 9: An Accurate Approach for Sentiment Analysis on Hindi-English Tweets Based on Bert and Pseudo Label Strategy." Proceedings of the Fourteenth Workshop on Semantic Evaluation. 2020.

[24] Thakur, Varsha et al. "Current State of Hinglish Text Sentiment Analysis." Social Science Research Network (2020): n. pag.

[25] Pathak, Abhilash & Kumar, Sudhanshu & Roy, Partha & Kim, Byung-Gyu. (2021). Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models. Electronics. 10. 2641. 10.3390/electronics10212641.

[26] Sarkar, Kamal. (2020). Heterogeneous classifier ensemble for sentiment analysis of Bengali and Hindi tweets. Sādhanā. 45. 10.1007/s12046-020-01424-z.

[27] Das, Sourav et al. "Sentiment classification with GST tweet data on LSTM based on polarity-popularity model." Sādhanā 45 (2020): 1-17.

[28] A. Sharmista, Dr. M. Ramaswami, "Sentiment Analysis on Tamil Reviews as Products in Social Media Using Machine Learning Techniques: A Novel Study", Madurai Kamaraj University Madurai - 625 021, Feb 2020. Sentiment Analysis on Tamil Reviews as Products in Social Media Using Machine Learning Techniques: A Novel Study

[29] Sajeetha Thavareesan, Sinnathamby Mahesan, "Review On Sentiment Analysis In Tamil Texts", JSC EUSL, vol. 9, no. 2, pp. 1-19, 2018. Review on sentiment analysis in Tamil texts

[30] Vallikannu Ramanathan, T. Meyyappan, S.M. Thamarai, "Predicting Tamil Movies Sentimental Reviews Using Tamil Tweets", Journal of Computer Science, vol. 15, no. 11, pp. 1638-1647, 2019. Predicting Tamil Movies Sentimental Reviews Using Tamil Tweets | Journal of Computer Science

[31] Elango, Sivasankar & Krishnakumari, Kalyan & Palani, Balasubramanian. (2021). An enhanced sentiment dictionary for domain adaptation with multi-domain dataset in Tamil language (ESD-DA). Soft Computing. 25. 10.1007/s00500-020-05400-x.

[32] 32]Srinivasan, Ramakrishnan and C. N. Subalalitha. "Sentimental analysis from imbalanced code-mixed data using machine learning approaches." Distributed and Parallel Databases (2021): 1 - 16.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
Article Received: 03 May 2023 Revised: 30 June 2023 Accepted: 21 July 2023

[33] Gokula Krishnan et al, . "TWITTER SENTIMENT ANALYSIS USING ENSEMBLE CLASSIFIERS ON TAMIL AND MALAYALAM LANGUAGES." OSF, 23 Aug. 2021. Web.

[34] Hande, A deep, et al. "Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced Dravidian languages." arXiv preprint arXiv:2108.03867 (2021).

[35] Dowager, Suman, and Radhika Mamidi. "Graph convolutional networks with multi-headed attention for code-mixed sentiment analysis." Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. 2021.

[36] 36]Chakravarthi, Bharathi Raja, et al. "Corpus creation for sentiment analysis in code-mixed Tamil-English text." arXiv preprint arXiv:2006.00206 (2020).

[37] Sandeep Mukku, "Sentiment Analysis for Telugu Language", International Institute of Information Technology Hyderabad - 500 032, Dec. 2017. (PDF) Sentiment Analysis for Telugu Language (researchgate.net)

[38] Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu, Ramesh Kumar Mohapatra, "Sentiment Analysis using Telugu SentiWordNet", WiSPNET, March 2017. Sentiment analysis using Telugu SentiWordNet | IEEE Conference Publication

[39] Bharti, Santosh Kumar, Reddy Naidu, and Korra Sathya Babu. "Hyperbolic Feature-based Sarcasm Detection in Telugu Conversation Sentences." Journal of Intelligent Systems 30.1 (2021): 73-89.

[40] Badugi, Srinivasu. "Telugu Movie Review Sentiment Analysis Using Natural Language Processing Approach." Data Engineering and Communication Technology. Springer, Singapore, 2020. 685-695.

[41] Suryachandra, Palli, and P. Venkata Subba Reddy. "CLASSIFICATION OF THE FEATURE-LEVEL RATING SENTIMENTS FOR TELUGU LANGUAGE REVIEWS USING WEIGHTED XGBOOST CLASSIFIER." Technology 11.12 (2020): 373-383.

[42] Priya, G. Balakrishna, and M. Usha Rani. "A Framework for Sentiment Analysis of Telugu Tweets." International Journal of Engineering and Advanced Technology (IJEAT) 9.6 (2020).

[43] Suryachandra, Palli, and P. Venkata Subba Reddy. "CLASSIFICATION OF THE SENTIMENT VALUE OF NATURAL LANGUAGE PROCESSING IN TELUGU DATA USING ADABOOSTER CLASSIFIER."

[44] Chakravarthi, Bharathi Raja, et al. "A sentiment analysis dataset for code-mixed Malayalam-English." arXiv preprint arXiv:2006.00210 (2020).

[45] Kumar, S. Sachin, M. Anand Kumar, and K. P. Soman. "Identifying Sentiment of Malayalam Tweets Using Deep Learning." Digital Business. Springer, Cham, 2019. 391-408.

[46] Kalaivani, A., and D. Thenmozhi. "SSN_NLP_MLRG@ Dravidian-CodeMix-FIRE2020: Sentiment Code-Mixed Text Classification in Tamil and Malayalam using ULMFiT." FIRE (Working Notes). 2020.

[47] Saiful Islam, Ruhul Amin, Khondoker Islam, "Sentiment analysis in Bengali via transfer learning using multilingual BERT", ICCIT, vol. 23, Jan 2021. (PDF) Sentiment analysis in Bengali via transfer learning using multilingual BERT

[48] Salim Sazzed, Sampath Jayarathna, "A Sentiment Classification in Bengali and Machine Translated English Corpus", IEEE IRI, vol. 20, pp. 107-114, Aug 2019. A Sentiment Classification in Bengali and Machine Translated English Corpus

[49] Simmi Bagga, Anil Sharma. (2023). Transformation from CIM to PIM for Querying Multi-Paradigm Databases. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 354–359. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2717

[50] Soumil Mandal, Sainik Kumar Mahata, Dipankar Das, "Preparing Bengali-English Code-Mixed Corpus for Sentiment Analysis of Indian Languages", ALR collocated with LREC, vol.13, March 2018. 1803.04000. Preparing Bengali-English Code-Mixed Corpus for Sentiment Analysis of Indian Languages

[51] Serajus Khan, Sanjida Rafa, Al Ekram Abir, Amit Das, "Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity", JEA, vol. 2 no. 3, pp. 118-124, 2021. Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity | Journal of Engineering Advancements

[52] Dawn, Debapratim Das, Sohrab Hossain Shaikh, and Rajat Kumar Pal. "A comprehensive review of Bengali word sense disambiguation." Artificial Intelligence Review 53.6 (2020): 4183-4213.

[53] Mamun, Md, et al. "Classification of Textual Sentiment Using Ensemble Technique." SN Computer Science 3.1 (2022): 1-13.

[54] Mukherjee, Shibashis, and David R. Heise. "Affective meanings of 1,469 Bengali concepts." Behavior research methods 49.1 (2017): 184-197.

[55] Sarkar, Kamal. "Heterogeneous classifier ensemble for sentiment analysis of Bengali and Hindi tweets." Sādhanā 45.1 (2020): 1-17.

[56] Sharmin, Sadia, and Danial Chakma. "Attention-based convolutional neural network for Bangla sentiment analysis." AI & SOCIETY 36.1 (2021): 381-396.

[57] Parita Shah, Priya Swaminarayan, Maitri Patel, "Sentiment analysis on film review in Gujarati language using machine learning", IJECE, vol. 12, no. 1, pp. 1030-1039, Feb 2022. http://doi.org/10.11591/ijece.v12i1.pp1030-1039

[58] Vrunda Joshi, Vipul Vekariya, "An Approach to Sentiment Analysis on Gujarati Tweets", ACST, vol. 10, no. 5, pp. 1487-1493, 2017. An Approach to Sentiment Analysis on Gujarati Tweets

[59] Chandrakant Patel, Jayesh Patel, "Influence of Gujarati STEmmeR in Supervised Learning of Web Page Categorization", IJISA, vol. 13, no. 3, pp. 23-34, Jun 2021. https://doi.org/10.5815/ijisa.2021.03.03

[60] Lata Gohil, Dharmendra Patel, "A Sentiment Analysis of Gujarati Text using Gujarati Senti word Net", IJITEE, vol.

ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7401
***Article Received**: 03 May 2023 **Revised:** 30 June 2023 **Accepted:** 21 July 2023*

8, no. 9, pp. 2290-2293, Jul 2019. International Journal of Soft Computing and Engineering

[61] Shah, Parita, Priya Swaminarayan, and Maitri Patel. "Sentiment analysis on film review in Gujarati language using machine learning." International Journal of Electrical & Computer Engineering (2088-8708) 12.1 (2022).

[62] Gohil, Lata, and Dharmendra Patel. "Multilabel Classification for Emotion Analysis of Multilingual Tweets." Int. J. Innov. Technol. Explore. Eng 9.1 (2019): 4453-4457.

[63] Dhabliya, D. (2021). Feature Selection Intrusion Detection System for The Attack Classification with Data Summarization. Machine Learning Applications in Engineering Education and Management, 1(1), 20–25. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/8

[64] Joshi, Vrunda C., and Vipul M. Vekariya. "An approach to sentiment analysis on Gujarati tweets." Advances in Computational Sciences and Technology 10.5 (2017): 1487-1493.

[65] Afraz Syed, Aslam Muhammad, Ana Martinez-Enriquez, “Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits”, MICAI, pp. 32-43, 2010. Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits

[66] Sajadul Kumhar, Mudasir Kirmani, Jitendra Sheetlani, Mudasir Hassan, “Sentiment Analysis of Urdu Language on different Social Media Platforms using Word2vec and LSTM”, TURCOMAT , vol. 11, no. 3, pp. 1439-1447, 2020. View of Sentiment Analysis of Urdu Language on different Social Media Platforms using Word2vec and LSTM

[67] Sadaf Rani, Muhammad Anwar, “Resource Creation and Evaluation of Aspect Based Sentiment Analysis in Urdu”, ACL-IJCNLP, vol. 10, pp. 79-84, Dec 2020. Resource Creation and Evaluation of Aspect Based Sentiment Analysis in Urdu

[68] Rakhi Batra, Zemun Kastrati, Ali Imran, Sher Daudpota, Abdul Ghafoor, “A Large-Scale Tweet Dataset For Urdu Text Sentiment Analysis”, PREPRINT, March 2021. A Large-Scale Tweet Dataset for Urdu Text Sentiment Analysis

[69] Tooba Tehreem, Hira Tahir, “Sentiment Analysis for YouTube Comments in Roman Urdu”, Feb 2021. 2102.10075. Sentiment Analysis for YouTube Comments in Roman Urdu

[70] There, Tooba. "Sentiment Analysis for YouTube Comments in Roman Urdu." arXiv preprint arXiv:2102.10075 (2021).

[71] Khattak, Asad, et al. "A survey on sentiment analysis in Urdu: A resource-poor language." Egyptian Informatics Journal 22.1 (2021): 53-74.

[72] Shah, Syed Muhammad Waqas, Muhammad Nadeem, and Muzamil Mehboob. "Sentiment Analysis of Roman-Urdu Tweets about Covid-19 Using Machine Learning Approach: A Systematic Literature." International Journal 10.2 (2021).

[73] Nasim, Zarmeen, and Sayeed Ghani. "Sentiment Analysis on Urdu Tweets Using Markov Chains." SN Computer Science 1.5 (2020): 1-13.

[74] Khan, Lal, et al. "Urdu sentiment analysis with deep learning methods." IEEE Access 9 (2021): 97803-97812.