A Comprehensive Review of Sentiment Analysis on Indian Regional Languages

Dr. Sunil Kale
Associate Professor, Department of Computer Engineering, Pune Institute of Computer Technology

Harshada Sasturkar
Department of Computer Engineering, Pune Institute of Computer Technology

Sentiment Analysis is the process of understanding the emotion present in a text. It helps in identifying 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 for SA of local languages has also been widely recognized. This has led to considerable research in analysis of Indian regional languages. This paper gives an overview of work done on SA in the following major languages : Marathi, Hindi, Tamil, Telugu, Malayalam, Bengali, Gujarati and Urdu. Furthermore this paper presents challenges and future scope for enhancing results of accuracy.

Keywords: Sentiment Analysis, Machine learning, Indian Regional Languages

1. INTRODUCTION

Sentiment Analysis, also known as opinion mining, is used to study and identify the emotion or opinion given by a piece of literature like 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 due to the fact that 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 is beneficial for 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 the healthcare industry where opinions and responses from people are important. Even though on social media the majority of such texts are in English language, there are many texts which 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 becomes important [11]. For this survey the focus is 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 gives an overview of the overall trend, results received, challenges faced and the possible future scope in this topic.

1.1 Approaches in Sentiment Analysis

There are mainly two types of approaches used for sentiment analysis [1][67].

1. Machine Learning Approach:
   Machine learning algorithms are used to classify the text. Depending on whether the data is labelled or not, either supervised or un/semi-supervised algorithms are used. A model is trained using these algorithms on a training dataset usually containing the same category of texts as that of the test data. The accuracy is then calculated by standard measures like precision, F-score, etc.

2. Lexicon Based Approach:

In this approach, an already prepared lexicon is used as a reference to classify the text. A lexicon is a set of words or phrases with a specific polarity score (+1 for positive, -1 for negative and 0 for neutral) assigned to them. These individually assigned polarities are then aggregated to get the total polarity. This has further two categories - corpus based and dictionary based.

1.2 Components of Sentiment Analysis

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

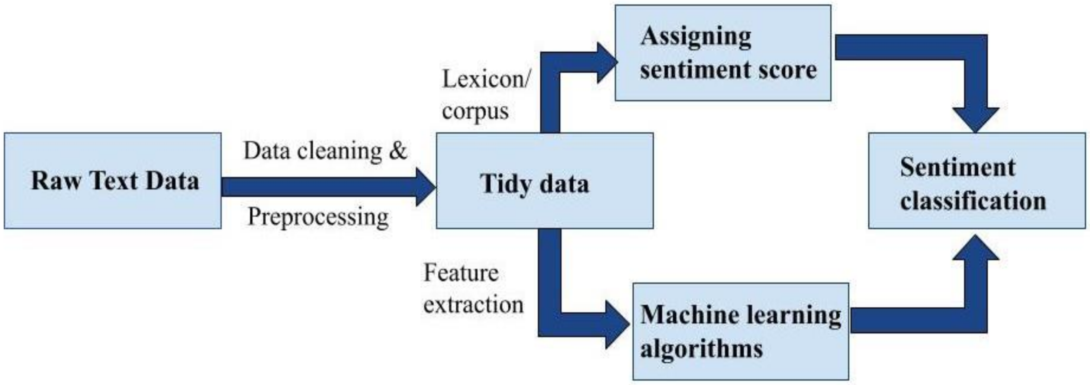

Figure 1: Components of Sentiment Analysis

The first step is to generate the raw input data needed for 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. For lexicon based approach a lexicon or corpus with predefined set of words along with their polarities is used. The data in the tidy dataset is compared with the lexicon and accordingly sentiment scores are assigned. Usually +1, -1 and 0 are used to indicate positive, negative and neutral polarities respectively. The cumulative sentiment score is calculated by adding the individual scores to determine overall polarity. For a machine learning based approach, the feature selection step is important where the data is first vectorized and then specific features (words) are selected which might contribute towards the sentiment. Various machine learning algorithms are used to create the classifiers which predict the sentiment category based on the selected features. After the final step of sentiment classification the accuracy of results is calculated along with the performance evaluation.

2. LANGUAGE WISE RESEARCH ON SENTIMENT ANALYSIS

2.1 Marathi

Marathi is the official language of the state of Maharashtra. It is the third most major language in India with around 99 million speakers. The script used in Marathi is called Devanagari.

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 1: Sentiment words in Marathi

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

Snehal Pawar et al. [1] proposed a Lexicon based approach where the system accurately classified sentiments by relying on a pre-defined lexicon. They concluded that to get efficient results one requires a richer database. Sujata Deshmukh et al. [2] used a Corpus based (Lexicon) approach and created a feasible corpus for 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 major dataset for Marathi Sentiment Analysis - L3CubeMahaSent which was publicly available was created by Atharva Kulkarni et al. [3] and had 16000 tweets. They used deep learning algorithms to perform SA with the dataset out of which the best accuracy was gained from IndicBERT and CNN with Indic fastText word embeddings. Chitra Chaudhari et al. [4] did Marathi SA with Marathi WordNet. They used tools like General Architecture for Text Engineering (GATE) for data processing and classification. It was concluded that to enhance the system more Marathi words need to be incorporated for which synset replacement algorithm is better. It was also found that an NLP based approach may perform poorly for grammatically incorrect text. Prafulla Bafna et al. [5] performed Sentiment Analysis of Marathi text using Unsupervised Learning and Visualization with Word Cloud. The system achieved the summarization of the clusters of Marathi corpus using unsupervised learning which is first of its kind. Among the algorithms, Fuzzy K-means clustering was found to have better accuracy. Harry Gavali [6] in his thesis performed SA on Mixed Language Text (MLT) containing Marathi+English and the same text written in Marathi Devanagari script. He concluded that if collected data is first translated and then classified then it shows better change compared to working on the original text. They observed that random Forest classifier had the best accuracy. They also created an open source dataset for MLT and Devanagari script. Renuka Naukarkar et al. [7] studied SA of Marathi tweets using various machine learning algorithms. It was concluded that in statistical machine learning approaches of sentiment analysis, the Bag-of-words (BOW) is the most commonly used method to model text. Monali Patil et al. [8] studied various sentiment classification techniques to perform a comparative study of Marathi sentiment analysis. They focused on the opportunities and challenges faced by researchers for Marathi SA and proposed a shaded based approach which is categorized under semantic-corpus based sentiment classification problem. Manisha Divate [9] used a Machine Learning approach to perform polarity-based sentiment analysis on Marathi using deep learning algorithm LSTM (Long Term Short Memory).

2.2 Hindi

Hindi is the official language of 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's also written in Devanagari script.

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 2: Sentiment words in Hindi

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

Mohammed Ansari et al. [10] used a Machine learning approach using WordNets and algorithms like SVM, etc. to perform SA on Hindi text. It built up on existing methods by integrating the best of them and concluded that results were better than individual methods. Sonali Shah et al. [11] studied the recent work that has been done in SA focusing on multilingual text containing indigenous language. The review mentioned the trends present in SA field, concluding that around 67% researchers used machine learning approach and about 29% used lexicon. For Indian languages, majority work has been done for Hindi compared to other languages. Balamurali A R et al. [12] presented an approach for cross-lingual SA for Marathi and Hindi that uses WordNet synset identifiers as features of a supervised classifier where SVM was used for classification. The study hopes to perform the same in a multilingual setup. Deepali Londhe et al. [13] wrote various approaches for language identification in multilingual texts for sentiment analysis task. Their survey concluded that for multilingual texts containing Marathi, Hindi and English, the N-Gram algorithm outperformed others. Piyush Arora [14] thesis focused on Independent Subjective Lexicon creation and sentiment analysis on Hindi using different machine learning algorithms. Namita Mittal et al. [15] created an annotated dataset for testing and devised rules for handling negation and discourse relation using HindiSentiWordNet for improved performance. All the inflected words of the existing root words were added in the HSWN increasing its coverage. Naman Bansal et al. [16] 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. [17] An Accurate Approach for

Sentiment Analysis on Hindi-English Tweets Based on Bert and pseudo Label Strategy : Mixing languages are widely used in social media, especially in multilingual societies like India. Detecting the emotions contained in these languages is of great significance to the development of society and political trends. In this paper, the authors propose an ensemble of pseudo-label based Bert (Bidirectional Encoder Representation from Transformer) model and TF-IDF (Term Frequency Inverse Document Frequency) based SGD Classifier model to identify the sentiments of Hindi -English (Hi-En) code-mixed data. The ensemble model combines the strengths of rich semantic information from the Bert model and word frequency information from the probabilistic n-gram model to predict the sentiment of a given tweet that is code-mixed. Finally, our team got an average F1 score of 0.686. This result, along with F1 scores of Bert with pseudo label and Ensemble being 0.725 and 0.731 respectively, showed that the model had a good effect on sentiment analysis for mixing languages. The future work for this project would be processing documents mixed in more languages, optimizing and extending the method to more aspects to learn more rich features in various languages. [18] Current State of Hinglish Text Sentiment Analysis : This paper reviews "sentiment analysis of hinglish text". Sentiment analysis, which is an extension of text mining, comprises research related to extracting sentiments, emotions from real world data. Traditional sentiment analysis of text, which is dependent on correct language syntax, can be useful for various decision-making processes. However, social media comments do not follow strict grammatical rules, with many occurrences of social media posts written in non-original scripts. In India, lots of people use Hinglish (combination of Hindi and English) as their colloquial language in their WhatsApp texts, Facebook posts, Instagram Reels, etc. In this paper, the recent developments on sentiment analysis from not only English text but also code mixed text and different problems related to the same have been presented. To overcome the challenge of aspect based sentiment analysis, selecting an appropriate classification model is vital. The validity of various classification procedures along with different types of features of emotion text data and extraction techniques have been analyzed in this paper. The conclusion from this paper was that TF-IDF is widely used in most of the papers for feature extraction because it performs better than the other techniques. The next vital component of sentiment analysis is the use of the Classifier. Naive Bayes is the machine learning algorithm that is used the most for solving Sentiment analysis issues. Ambiguity problems and aspect based sentiment analysis in hinglish text are some of the future scopes. Further research is expected to improve accuracy in sentiment analysis of code mixed language. [19] Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models : This paper focuses on the issue of most approaches of Sentiment Analysis to identify only the overall polarity of a sentence, instead of the polarity of each aspect mentioned in the sentence. It focuses on the Aspect-Based Sentiment Analysis (ABSA) approach, which identifies the aspects within the given sentence, and the sentiment that was expressed for each aspect. Recently, the use of pre-trained models such as BERT (Bidirectional Encoder Representation from Transformer) has yielded state-of-the-art results in the field of natural language processing. In this paper, the authors propose two ensemble models based on multilingual-BERT, namely, mBERT-E-MV and mBERT-E-AS. Using different methods, they construct an auxiliary sentence from this aspect and convert the ABSA problem to a sentence-pair classification task. Then, fine-tuning of different pre-trained BERT models is done, followed by creating an ensemble of them for a final prediction based on the proposed model; we achieve new,revolutionary results for datasets belonging to different domains in the Hindi language. The results indicated that overall, BERT-based models performed much better than the other models, which is possible because of the construction of auxiliary sentences from the aspect information, which is analogous to exponentially increasing the dataset. Also, the BERT model has an additional advantage in handling sentence pair classification tasks, which is evident from its impressive improvement on the QA (Question Answering) and NLI (Natural Language Interface) tasks. In future work, the system proposed in this paper can be applied to other NLP problems. The obtained results suggest that there is scope for augmenting the Hindi datasets for further improvements in performance. There is also scope for introducing a dataset for the TABSA task in Indian languages, as there is no dataset available for that very purpose. [20] Heterogeneous classifier ensemble for sentiment analysis of Bengali and Hindi tweets : Here authors presented polarity of sentiments on Hindi and Bengali tweets data. In the proposed method, three separate classifiers are constructed, with each having a unique feature set. Also, tweet words have been enhanced with sentiment polarity information retrieved from an external knowledge base called sentiment lexicon. Model performance was evaluated and compared using the 10-fold cross-validation performance metric. A dataset consisting of 1500 tweets is generated out of the SAIL 2015 Bengali training and test data, and the average accuracy of each model is computed over 10 folds. Similarly, SAIL 2015 Hindi training and test data were combined for a dataset containing 1760 tweets. Experiments have shown that ensembles of classifiers with a variety of features are effective and achieved 63.5% of accuracy. [21] Sentiment classification with GST tweet data on LSTM based on polarity-popularity model : The work presented in this paper was inspired by the large-scale reformation of opinion contrasts and conflicts regarding this new taxation system. The proposed approach is to analyze the reactions of public sentiment on Twitter based on popular words either directly or indirectly related to GST. The dataset comprises 200K tweets solely about

GST from June 2017 to December 2017 in two phases. In order to assure the relevance of the scraped tweets with respect to GST, we prepared a topic-sentiment relevance model. Furthermore, several state-of-the-art lexicons were employed for identifying sentiment words and assigned polarity ratings to each of the tweets. In order to extract the relevant words that are linked with GST implicitly, a new polarity-popularity framework was proposed and such popular words were also rated with sentiments. Next, an LSTM model was trained using both types of rated words for predicting sentiment on GST tweets and obtained an overall accuracy of 84.51%. For future work, taking the most occurring keyword in a given event and adding it to any such event to predict the course and trend of that event is one likely course. The authors would also like to develop a system to successfully evaluate bi-lingual font-mixed tweets to enhance the accuracy of our experiment.

2.3 Tamil

Tamil is one of the Dravidian languages which are spoken in many southern Indian states. The Tamil script is also called abugida script which is made up of units of consonants and vowels combined together. The total count of Tamil speakers is above 77 million making it the fifth most spoken language in India.

Sentimental features :

Positive: நல்ல(Good), வெற்றி(Success), அற்புதமான(Awesome),
கைத்தட்டல்(Applause), நன் றாக(Thank You)

Negative: வெறுப்பு(Dislike), சலிப்பு(Boring), குறைபாடு(Defect),
வெறுக்கிறேன்([I] Hate), குற்றம்(Offence)

Neutral: உரையாடல்(Conversation), காலநிலை(Climate),
சந்தித்தல்(Meeting), படம்(Picture), உணவு(Food)

Table 3: Sentiment words in Tamil

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

A. Sharmista et al. [22] 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. [23] studied the objectives, corpus, features, techniques and challenges along with the accuracy or F-measure presented in Tamil SA literature. It was found that the efficiency of SVM and RNN classifiers that used TF-IDF and Word2vec features of Tamil text was better than other classifiers. Vallikannu Ramanathan et al. [24] used enhanced Tamil SentiWordNet, TF-IDF feature to perform domain specific ontology and contextual semantic sentiment analysis. [25] An enhanced sentiment dictionary for domain adaptation with multi-domain dataset in Tamil language (ESD-DA) : The author has proposed an enhanced dictionary in the Tamil language which aims at building contextual relationships between the terms of multi-domain datasets that focuses on minimizing the feature mismatch problem. In short, the initial dictionary finds point wise mutual information to calculate contextual weights then the final dictionary calculates the rank score based on the significance of terms out of all the reviews. This work aims to classify reviews from multiple target domains in Tamil by using a unified dictionary with a substantial number of vocabulary. This extendible dictionary has significantly improved the accuracy of DA that is performed between various domains, giving out an accuracy of 70.5%, which is very high considering the multi-domain datasets in Tamil. [26] Sentimental analysis from imbalanced code-mixed data using machine learning approaches : Sentiment analysis is much more complex with class imbalance problems. One more major challenge in sentiment analysis is 'Code-Mixing' that involves the use of multiple languages in a text or sentence. The author proposes a solution to address both of these challenges by 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 pre-processing, SMOTE and ADASYN, these resampling techniques helped improve the F-1 score by 50%. Using the combination of sampling techniques along with the levenshtein distance metrics helped improve the code-mixing problem but there is a lot more work has to be done for improving the class imbalance problem. [27] Twitter sentiment analysis using Ensemble classifiers on Tamil and Malayalam languages : Organizations can get access to the real-time feelings of people towards a topic by performing sentiment on Twitter data. In this research, the author has proposed a solution to classify the general polarity of feelings expressed in Tamil and Malaya language over twitter in three categories: positive, negative and neutral. Dataset is prepared for each language and pre-processing is performed in order to label the data. The author has undertaken the deep learning approach to this classification problem for each dataset. In conclusion, among Recurrent NN, Recursive NN and LSTM, LSTM performed best on both the Tamil and Malaya languages with an average accuracy of 97%. [28] Benchmarking Multi-Task Learning for Sentiment Analysis and Offensive Language Identification in Under-Resourced Dravidian Languages : The author has conducted a survey to find out if training models using multi-task learning is beneficial as compared to the single-task learning approach in the case of sentiment analysis and offensive language identification. Dataset used in the survey is of code-mixed Youtube comments for Tamil, Malayalam and Kannada languages. Experiments have shown that the multi-task learning model performs significantly better compared to the single-task learning mode and also reduces the space and time constraints for the training of the individual models. For SA and OLI, the best model had a weighted F-1 score of (66.8% and 90.5%), (59% and 70%), and (62.1% and 75.3%) for Kannada, Malayalam, and Tamil respectively. [29] Graph Convolutional Networks with Multi-headed Attention for Code-Mixed Sentiment Analysis : A lot of work has been done on the code-mixed dataset sentiment analysis, but most of the 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 DravidianCodeMix 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 showed promising results by outperforming various traditional methods. In conclusion, the best results of a weighted F1 of 0.75 and an accuracy of 0.73 was obtained on the Malayalam-English CM dataset. [30] Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text : Sentiment analysis of code-mixed datasets is one difficulty, but the dimension of low-resourced languages add a whole new level of complexity to it. For this problem, the author created a standard Tamil-English code-mixed, sentiment annotated dataset consisting of 15,744 comments from various posts on YouTube. Out of many different classifiers, Random Forest, Logistic Regression and Decision Tree classifiers performed comparatively better on sentiment analysis while the SVM model actually had low performance on the same. Using deep learning methods also did not help in achieving better performance in terms of three automatic metrics. In conclusion, the author has achieved higher inter-annotator agreement in terms of Krippendroff alpha collected from the voluntary annotators through Google forms.

2.4 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.
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 4: Sentiment words in Telugu

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

Sandeep Mukku [31] created for Telugu text a manually annotated corpus and word embedding model. He suggested a hybrid approach of query selection strategies with active learning techniques. He found that the extreme gradient boosting (XGBoost) classifier in combination with Hybrid query selection approach gave the best accuracy. Reddy Naidu et al. [32] performed sentiment analysis with a two phased approach, subjectivity classification and sentiment classification using Telugu SentiWordNet. [33] Hyperbolic Feature-based Sarcasm Detection in Telugu Conversation Sentences : Detection of a sarcastic sentence is challenging as it contains only positive words conveying a negative sentiment. Existing systems for sarcasm detection are limited to the English language. To carry out the same in other less resourceful languages like Hindi and Telugu is challenging as there is a scarcity of useful datasets. The author has manually collected data from various resources and prepared a basic dataset for the sentiment analysis in sarcasm detection using Telugu language. The author has used algorithms based on hyperbolic features such as Intensifier, Interjection, Question mark and exclamation symbol and achieved an accuracy of 94%. [34] Telugu Movie Review Sentiment Analysis Using Natural Language Processing Approach : Opinion about a movie can be considered as a short description of the movie along with the review. It may be positive, negative or neutral. The author has proposed a framework to look through film surveys which are basically utilized transliteration plots. In short, Telugu to English into positive, negative and neutral for better classification. The system is built upon rule-based NLP and a machine learning approach. The author's main focus was to create a seed list of words based on category. With the proposed system, the author implemented an opinion extraction on Telugu movie review using NLP. Since the opinion extraction is based on verbs, adjacent and adverbs, the proposed model successfully provided good results as high as 96%. [35] Classification of the feature-level rating sentiments for Telugu language reviews using weighted XGBoost classifier : Sentiment analysis can also be used for opinion extraction. Since there is a huge amount of data that 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 data is collected and pre-processed to convert unstructured data into structured data. In the next step, the author used a weighted XGBoost classifier to classify the product reviews and ratings are generated concurrently. 5-star meaning excellent and 1-star rating meaning terrible review. For feature level rating method, the accuracy obtained was 81.02%, f-measure is of 87%, precision was of 78%, the recall was of 89% . [36] A Framework for Sentiment Analysis of Telugu Tweets : Even though sentiment analysis is one of the most researched fields in the NLP, most of the work is done considering the English language. In this paper, the author proposed a framework that is specifically designed to conduct sentiment analysis based on the Telugu language. Further, the framework was integrated with the word embedding model

Word2Vec, language translator and deep learning approaches like RNN and Naive Bayes algorithms. To construct the dataset, the data was collected manually and pre-processed before applying it to the framework. In conclusion, the accuracy was measured at 80.45%, specificity at 76.57% and precision at 82.46%. [37] Classification of the sentiment value of natural language processing in Telugu data using ADABooster classifier : Automatic text summarization under NLP can also be used to identify the sentiment behind the text. In this paper, the author completed the system in three steps: data pre-processing stage, classification stage and semantic analysis stage. The data pre-processing stage involved cleaning the manually obtained data, splitting the necessary data from the reviews. Then using the ADABooster classifier, semantic analysis is performed to find the sentiment behind the text and evaluate the compound polarity of each review. 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, 80.56 for the proposed Adabooster classifier.

2.5 Malayalam

Malayalam is a Dravidian language with Vatteluttu as its written script. It is spoken by over 35 million people mainly in the Kerala and the union territories of Lakshadweep and Puducherry. It is the tenth most spoken language in India.
Sentimental features :

Positive: പ്രസന്നമായ(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)

Table 5: Sentiment words in Malayalam

Category	Malayalam Word	English Representation
Positive	പ്രസന്നമായ അഭിനന്ദിക്കുക സ്വാദിഷ്ടമായ	prasannamāya abhinandikkuka svādiṣṭamāya
Negative	കഠിനമായ അപമര്യാദയായ വിരസത	kaṭhinamāya apamaryādayāya virasata
Neutral	സംസാരിക്കുക ഭക്ഷണം ജോലി	sansārikkuka bhakṣaṇaṁ jēāli

[38] A Sentiment Analysis Dataset for Code-Mixed Malayalam-English : In view of 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 it 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 neighbours (KNN), Contextualized DME (CDME), Bidirectional Encoder Representations for Transformers (BERT), 1D Dimensional Convolution (1DConv), Bidirectional Encoder Representations for Transformers (BERT). In conclusion,

1DConv shows a better score in recall, precision and F-1 score while BERT fails to identify some of the classes. DME and CDME are successful in identifying all of the classes. [39] Identifying Sentiment of Malayalam Tweets Using Deep Learning : As sentiment analysis lacks work in regional languages, the author has proposed an approach to perform sentiment analysis on the Malayalam tweets dataset by using deep learning methods. The author has used CNN and LSTM as primary classifiers and compared the results with traditional classifiers such as SVM. The output is classified into three categories 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 along with ReLU, ELU, SELU activation functions gave much better results. [40] Sentiment Code-Mixed Text Classification in Tamil and Malayalam using ULMFiT : To contribute towards 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 AWD-LSTM model along with 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 for both Malayalam-English and Tamil-English languages. The author also concludes that results can be further improved by handling data imbalances.

2.6 Bengali

In India Bengali language is mostly spoken in eastern region and is a native language of Bengal state. With a total over 100 million speakers, it is the second most spoken language in India. The written Bengali is an abugida script.

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 6: Sentiment words in Bengali

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

Saiful Islam et al. [41] worked with deep learning model for their sentiment classification task. Along with 2 class they also created a 3 class sentiment dataset and built a classifier for that 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. Salim Sazzed et al. [42] 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 work needs to be done so that the scope of the study could be extended from bilingual to multilingual classification. Soumil Mandal et al. [43] 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 own language tagging system with lexicon based module and supervised learning module, and a sentiment tagging system combining rule based and supervised methods. Out of all the supervised methods stochastic gradient descent had higher accuracy. Serajus Khan et al. [44] performed sentiment analysis using various 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 abusive and religious categories. It was concluded that the SVM algorithm gave the best accuracy. [45] A comprehensive review of Bengali word sense disambiguation : Word Sense Disambiguation (WSD), is the process of identifying the real meaning of a word based on what context it is used. Even though a lot of work is done in this field, there is a discontinuity when it comes to working on Bengali WSD. In this paper, the author has conducted a survey on various approaches to Bengali WSD and also surveyed 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 has surveyed two techniques and achieved an average accuracy of 58.5%. In knowledge-based approach, the author has surveyed two techniques and achieved an average accuracy of 75%. In the supervised approach, the author has surveyed three techniques and achieved an average accuracy of 71%. [46] Classification of Textual Sentiment Using Ensemble Technique : 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 (BSaD) 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 three best performing traditional classifiers which are: LR, RF, and SVM. In conclusion, the author has observed that the ensemble approach along with TF-IDF features performed significantly better than other traditional classifiers, giving the highest accuracy of 82% on BSaD. [47] Affective meanings of 1,469 Bengali concepts : 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 categorised into three stimuli: evaluation, potency, and activity. This study made use of pan-respondent component analysis to examine the respondents usage of the EPA scale. Interesting patterns were found when looked 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. [48] Heterogeneous classifier ensemble for sentiment analysis of Bengali and Hindi tweets : For sentiment analysis of Bengali and Hindi tweets, the author has used the SAIL 2015 dataset as a benchmark dataset. The proposed approach combines Multinomial NB classifier with character n-gram features, Multinomial NB classifier with word n-gram feature and SVM with unigram features in an ensemble. The author has also incorporated the sentiment lexicon in the model. The author has observed that ‘majority voting’ rule gave better results for sentiment analysis of Bengali tweets and ‘average of probabilities’ rule gave better results for sentiment analysis of Hindi tweets. In conclusion, out of all three base classifiers, Multinomial NB with character n-gram feature and sentiment lexicon performed the best. [49] Attention-based convolutional neural network for Bangla sentiment analysis : Due to the lack of benchmark datasets, the work done so far in the field of sentiment analysis in the Bangla language is low. Sentiment analysis is the technique of using the natural learning process along with text analysis to find out and extract subjective information from the text. Even though the attention mechanism in CNN (A-CNN) performed better than A-LSTM at 96.55%, it fails to achieve the overall performance in terms of accuracy (66.06%), precision (66.04%), recall (65.66%) and F-measure (66.02%). This inaccuracy was due to the noise present in the dataset. Further, the author used CGAN network to generate synthetic data, which resulted in a performance boost of A-CNN by almost 6%. In conclusion, training accuracy was around 99% and overall validation accuracy was around 72%.

2.7 Gujarati

Gujarati is the native language of the state of Gujarat. It has seventh rank 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.
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 7: Sentiment words in Gujarati

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

Parita Shah et al. [50] 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 along with TF-IDF had better accuracy than with CountVectorizer. The accuracy of KNN was the same in both cases. Vrunda Joshi et al. [51] combined POS tagging with SVM classifier to improve its performance and achieved 92% accuracy for senti analysis. They concluded that data pre-processing increases the performance of classifiers and other feature extraction techniques can be studied to test out their performance with such classifiers. Chandrakant Patel et al. [52] studied how much the stemming algorithms affect the categorization of Gujarati web pages. The focus was mainly on GUJarati STEmmeR (GUJSTER) algorithm along with syllable tokenizer and dynamic stop words identification as part of pre-processing. They concluded that stemmer algorithms have a considerable influence on supervised algorithms improving their accuracy specifically for problem statement of web page categorisation. Lata Gohil et al. [53] created a Gujarati SentiWordNet (G-SWN) with the help of Hindi SentiWordNet and IndoWordNet. Lexical approach was used to classify sentiments on a manually annotated corpora using the above G-SWN. They also stated that the accuracy can be improved in future using Word Sense Disambiguation. [54] Sentiment analysis on film review in Gujarati language using machine learning : For the purpose of sentiment analysis in the Gujarati language, the dataset is prepared by taking reviews of various products in the Gujarati language. In this paper, the author has implemented two machine learning classifiers on the collected dataset, K-nearest neighbours (KNN) and Multinomial Naive Bayes (MNB). The author has also considered TF-IDF and Word Level Count Vectorizer to achieve better results. In conclusion, the author has observed that MNB classifier with TF-IDF gave a better performance as compared to Word Count Vectorizer and KNN classifier had the same performance in case of TF-IDF as well as Count Vectorizer. [55] Multilabel Classification for Emotion Analysis of Multilingual Tweets : In this paper, the author has used sentiment analysis to perform the classification of tweets over eight categories as follows: joy, surprise, fear, trust, sadness, anticipation, anger, disgust. Two datasets are prepared by collecting tweets related to Indian politics and are annotated manually for each language, namely English, Gujarati and Hindi. The experiment is conducted for each language using basic machine learning classifiers and hybrid classifier with two feature generation algorithms namely SN and CS-SN which make use of SenticNet are utilized. In conclusion, it was observed that the machine learning approach gave better results as compared to the hybrid approach for the Gujarati language. The author has also mentioned various reasons for the low performance of the hybrid approach. [56] An Approach to Sentiment Analysis on Gujarati Tweets : Sentiment Analysis being one of the most popular topics in NLP, lacks work in the case of languages other than English. In this paper, the author has proposed a practical approach to perform sentiment analysis in the Gujarati language using data collected from Twitter. The author has focused upon classifying the data into positive and negative polarities. From the literature survey, the author came to an understanding 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 on SVM classifier. Results obtained by this approach gave an accuracy of 92%.

2.8 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.

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 8: Sentiment words in Urdu

Category	Urdu Word	English Representation
Positive	بہادر دلچسپ بہتر	Bahadur Dilchasp Behtar
Negative	مہنگا بحث مایوس	Mehnga Behas Mayoos

Neutral	ماحول پیشہ جگہ	Mahol Pesha Jagah

Afraz Syed et al. [57] focused on extracting the SentiUnits, sets of words/phrases which contribute towards 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 into the sentiment. It was concluded that work is needed for expanding the lexicon for better results. Sajadul Kumhar et al. [58] performed sentiment analysis using Word2Vec as text vectorizer and LSTM neural network model as a classifier. An activation function called SoftMax which assigns the polarity using probabilistic approach was included in LSTM. It was concluded that this approach had better accuracy. Sadaf Rani et al. [59] studied aspect based sentiment analysis which focuses 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. Rakhi Batra et al. [60] worked towards creating a dataset containing Urdu tweet texts. By adding emojis and their polarity the dataset is made useful for sentiment analysis purposes. Authors proposed that this large dataset can be used in fields like machine learning, NLP and information retrieval. Tooba Tehreem et al. [61] 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 in future various other feature extraction methods could be used to improve the accuracy. [62] Sentiment Analysis for YouTube Comments in Roman Urdu : Even though sentiment analysis is a vast area in machine learning, most of the 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 over 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%. [63] A survey on sentiment analysis in Urdu: A resource-poor language : Although the volume of studies on sentiment analysis is increasing rapidly, the primary language of concern is English. In this paper, the author mainly focused on describing three different dimensions used for Urdu sentiment analysis: text pre-processing, lexical resources and sentiment classification. In conclusions of the survey carried out for recognizing the progress and shortcomings of Urdu sentiment analysis, the author has proposed guidelines for future work to acknowledge with six critical points for rectification. [64] 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. For understanding the pattern in people’s sentiments, the dataset was prepared using Twitter data throughout the pandemic. The author has given a general overview of the process of sentiment analysis of Roman-Urdu Tweets and has observed that TF_IDF with Unigram and Bigram are the most used features used for the same. Currently, Hybrid approaches don’t give better results, since a lot more work is required to be done in this field. Even when the size of the dataset is large, many studies have achieved good results by using the Naive Bayes algorithm. In conclusion, the author concludes that the labeled dataset gives more accurate results as compared to the unlabeled dataset. [65] Sentiment Analysis on Urdu Tweets Using Markov Chains : As compared to the work done in the field of NLP on the English language, very little amount of 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 carry out a sentiment analysis on the Urdu Tweets dataset. In conclusion, the author found that this methodology gave better results as compared to the lexicon-based approach and other common machine-learning based approaches on sentiment analysis. However, the author also observed that because of the lack of positive tweets in the dataset, the model gives bad results for the prediction of positive Urdu Tweets. [66] Urdu Sentiment Analysis with Deep Learning Methods : Even though almost 169 million people are familiar with the Urdu language and a large amount of data is generated on various internet platforms every day, 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 namely n-gram features and pre-trained word embeddings. In conclusion, the author mentions that the highest accuracy of F1 score of 82.05% was achieved using the LR with a combination of various features. SVM classifier gave second best results for the sentiment analysis as compared to all the other machine learning classifiers.

Table 9: Sentiment Analysis on different Indian Regional Languages

Ref	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.
[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-IDF, 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% 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	Pseuto 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 SVM - 75.96%, RNN - 88.23%
[24]	Tamil	tweets	Tamil SentiWordNet, TF-IDF, python NLP	Accuracy 77.89%
[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	mBERT 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 fastText 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
[46]	Bengali	BSaD	Ensemble approach with TF-IDF features	82% accuracy
[47]	Bengali	Manual data collection	EPA scale	0.93 corelation
[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

3. 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, a lot of research is already available on sentiment analysis for the English language. But when it comes to regional languages or languages other than English (Hindi, Tamil, Urdu,), a very less amount of datasets are available. As there is a scarcity in benchmark datasets, there are not many previously done research results available to compare with the current work.

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

4. CHALLENGES

After studying the research papers it became evident that the biggest challenge faced by all of the researchers was lack of resources available for sentiment analysis of the regional languages. Availability of gold-standard benchmark datasets in the case of regional languages (Hindi, Tamil etc) is very low. Datasets that are already available have a lot of noise in them along with the problem of class imbalance. As there was a lack of satisfactory datasets, many of the times, data had to be manually collected and annotated. Most of the 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 the 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 it. To solve the problem for 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.

5. CONCLUSION

After Studying the literature it can be concluded that a significant amount of research has been done in recent years when it comes to Sentiment Analysis of Indian regional languages. From creating publicly available high quality datasets/corpus to using various machine learning algorithms for sentiment scoring, a lot of 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 quality of the dataset. When gold-standard datasets were available for a given language, most of the approaches used to perform sentiment analysis gave better performance. Similarly, when the dataset for a given language was of poor quality (poorly annotated datasets, class imbalance, noise etc.), poor results were observed. Further, it was observed that when dataset cleaning and data preprocessing was 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 along with feature extraction methods gave a significant performance boost.

6. Future Scope

1. As there are not many benchmark datasets available for low-resourced languages, newly creating gold-standard datasets would give better results as well as help the research community to grow rapidly and learn more. The availability of large datasets like WordNets, SentiWordNets and similar lexicons for each regional language is the most important and needed. Data augmentation in NLP helps give different instances of data samples that can simulate the real-time data. This will help improve the performance of the system significantly.
2. The current work mostly considers single sentences as input so study can be done for larger text inputs. As a lot of machine learning algorithms have been explored for the analysis, using the ones with 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 slangs, emoticons and mixed language text which is more commonly used on a daily basis.
5. Considering the huge amount of Twitter-based datasets, a system should be built to successfully evaluate bi-lingual and font-mixed tweets in order to enhance the overall accuracy.
6. NLP systems often face the challenge of properly 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.

7. References

[1] Snehal Pawar, Swati Mali, "Sentiment Analysis in Marathi Language", IJRITCC, vol. 5, no. 8, pp. 21-25, Aug. 2017. Sentiment Analysis in 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 Divate, “Sentiment analysis of Marathi news using LSTM”, IJIT, vol. 13, 2021. https://link.springer.com/article/10.1007%2Fs41870-021-00702-1

[10] 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

[11] 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

[12] 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

[13] 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

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

[15] 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

[16] Naman Bansal, Umair Ahmed, Amitabha Mukherjee, “Sentiment Analysis in Hindi”, Indian Institute of Technology Kanpur, Sentiment Analysis In Hindi

[17] 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 Pesudo Label Strategy." Proceedings of the Fourteenth Workshop on Semantic Evaluation. 2020.

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

[19] 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.

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

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

[22] 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

[23] 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

[24] 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

[25] 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.

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

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

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

[29] Dowlagar, 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.

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

[31] 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)

[32] 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

[33] 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.

[34] Badugu, Srinivasu. "Telugu Movie Review Sentiment Analysis Using Natural Language Processing Approach." Data Engineering and Communication Technology. Springer, Singapore, 2020. 685-695.
[35] 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.

[36] 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).

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

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

[39] 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.

[40] 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.

[41] Saiful Islam, Ruhul Amin, Khondoker Islam, “Sentiment analysis in Bengali via transfer learning using multi-lingual BERT”, ICCIT, vol. 23, Jan 2021. (PDF) Sentiment analysis in Bengali via transfer learning using multi-lingual BERT

[42] 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

[43] 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

[44] 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

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

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

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

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

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

[50] 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

[51] 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

[52] 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

[53] Lata Gohil, Dharmendra Patel, “A Sentiment Analysis of Gujarati Text using Gujarati Senti word Net”, IJITEE, vol. 8, no. 9, pp. 2290-2293, Jul 2019. International Journal of Soft Computing and Engineering

[54] 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).

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

[56] 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.

[57] 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

[58] 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

[59] 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

[60] Rakhi Batra, Zenun 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

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

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

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

[64] 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).

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

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

[67] Sentiment analysis

[68] List of languages by number of native speakers in India

[69] Google Translate