IRJSE International Research Journal of Science & Engineering
Int. Res. J. of Science & Engineering, Special Issue A14, June, 2024
ISSN: 2322-0015
SJIF Impact Factor 6.70
https://doi.org/10.5281/zenodo.12702222
RESEARCH ARTICLE OPEN ACCESS

A Framework for Opinion-Mining of Tweets

Krutik Mehta[1*], Guruanjan Singh[2], Aniket Kore[3] and Sridhar Iyer[4]

Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai, Maharashtra 400056
*Corresponding author Email: [1]krutikmehtaa@gmail.com
[2]singhguruanjan@gmail.com, [3]Aniket.kore@djsce.ac.in and [4]Sridhar.iyer@djsce.ac.in

Manuscript Details

Available online on https://www.irjse.in
ISSN: 2322-0015

Editor: Dr. Arvind Chavhan

Cite this article as:
Krutik Mehta, Guruanjan Singh, Aniket Kore and Sridhar Iyer. A Framework for Opinion-Mining of Tweets, *Int. Res. Journal of Science & Engineering*, 2024, Special Issue A14: 111-120. https://doi.org/10.5281/zenodo.12702222

Article published in Special issue of National Conference on Machine Learning and Data Science (NCMLDS-2024)" organized by College of Computer Science and Information Technology (COCSIT) Ambajogai Road, Latur, Maharashtra, India on date April 16th to 17th 2024

Abstract

Sentiment analysis, commonly referred to as Opinion Mining, represents by far the most prominent field of research in the domain of natural language processing. It can be described as the procedure of determining the polarities of a particular content to obtain access to the secret information contained in the text or piece of written material. The expansion of electronic data over the past two decades has led to substantial developments in several widely spoken languages in India, including Hindi, English, Marathi, and Tamil. Gujarati, the seventh most used dialect throughout India, has yet to be studied, however, because there aren't enough credible datasets available. By using the official Twitter API tweepy, we present the first significant dataset of Gujarati sentiment analysis in this study. Our dataset consists of approximately 5,500 unique tweets, categorized into three broad polarities: positive, negative, and neutral. We have annotated our dataset based on our customized framework formed by mapping synsets of existing wordnets - Hindi SentiWordNet (H-SWN) and IndoWordnet. In our final module, we present the dataset statistics and baseline classification results using BERT-based deep learning models - IndicBERT, mBERT, and XLM-roBERTa having accuracies of 72.97%, 65.81%, and 70.42% respectively.

Keywords: Gujarati Sentiment Analysis, Twitter Sentiment Dataset, Gujarati Text Classification, Deep Learning, BERT, Polarity

1. Introduction

The volume of people using the internet has dramatically increased during the past several years. With a record-breaking growth rate, there are already 4.62 billion registered active social media users globally in 2022. Due to unprecedented technological advancements, a dramatic increase in the amount of data was seen on the internet [1].

The way that people communicate has evolved in this period, which has been dramatically affected by technology and digital media. Fundamental means of communication, such as language, have undergone significant alteration. Information is now shared across cultures much more frequently as a result of a communication barrier that digitalization has created. A vast amount of such data is generated on online social networking platforms.

Twitter, amongst the most widely utilized social media and microblogging hubs in the world, is a very well-liked channel for expressing thoughts. How people get knowledge from other people and institutions that interest them has been shaped and transformed by it. Tweets are status update messages that users can post on Twitter to notify their followers about their thoughts, actions, or global events. Users can also communicate with one another by responding to or retweeting their tweets. Owing to its numerous uses and the ever-increasing volume of information that is readily accessible, the extraction of emotion orientations from Twitter posts has emerged as a popular study choice. For instance, several models have been created to offer political election plans by evaluating the sentiment polarization of Twitter users toward elections [2]. Businesses may quickly and efficiently track consumer sentiment regarding various brands and merchandise by using Twitter opinion mining [3].

The broad concept of sentiment, evaluation, appraisal, or attitude of a piece of information that reflects the writer's or speaker's viewpoint is encompassed by the phrase "opinion." Since these perspectives differ from user to user. It is essential to consider a diverse range of viewpoints to get a more realistic sense of how people feel about the subject. Classifying the overall sentiment of a Twitter tweet is the goal of opinion mining using Twitter data. Positive, negative, and neutral sentiment inclination or polarity are denoted. Opinion Mining for the English language expanded the field of research in recent years and its progress is quite high. For sentiment analysis, two methodologies, specifically Lexicon and Machine Learning, have been thoroughly investigated. The necessity for a Machine Learning approach is a large volume of annotated data. Because languages with resource constraints lack annotated corpora, the lexicon method is a good place to begin. For preparing lexical resources, many ways have been investigated, including approaches based on dictionaries, wordnets, and corpus. Despite having out-of-context polarity results, lexical elements have been demonstrated to provide a reliable basis. For the English language, several sentiment lexical resources have been established. Over 3 million phrases are included in SentiWordNet coupled with observations that are positive, negative, and objective. In OpinionFinder's Subjectivity Lexicon, includes words, POS tags, polarities, and strong or weak subjectivity. The Opinion Lexicon was constructed by selecting adjectives out of opinion statements in the annotated Twitter corpus. The words in AFINN-111 have been manually graded for valence [4-6].

For the majority of Indian languages, opinionated lexicon components are also being created. English SentiWordNet and English-Hindi WordNet Linking are used to create Hindi-SentiWordNet (H-SWN). With the help of a bilingual English-Bengali dictionary and SentiWordNet, a Hindi Subjective Lexicon was created utilizing original seed words that acquired relationships between synonyms and antonyms from Hindi WordNet [7]. The Odia lexicon resource was created by mapping a term's corresponding SynsetID of the Odia wordnet wherein the POS tag identifies either an adjective and perhaps an adverb to SentiWordNet of Indian languages & IndoWordNet to SentiWordNet for Indian languages [8]. SentiWordNet for the Tamil language was built using a translation technique [9]. For the objective of boosting lexicon dependability, four English-language repositories were utilized: Subjectivity Lexicon, AFINN-111, English SentiWordNet 3.0 [10], and Opinion Lexicon.

This has conclusively shown that a significant amount of work has been done on creating lexical resources for many regional Indian languages in the past decades. This has provided a boost in the field of Twitter Opinion Mining for these languages. However, the Gujarati language, which is the state's official primary language, where there are approximately 64.8 million people, has

not been explored due to the unavailability of dependable data sets. Additionally, most of Gujarat's population only converses in and writes in just this language. Hence, in this paper, we have created the first social media sentiment analysis dataset that can provide researchers to uncover the undiscovered Gujarati Language. This paper could serve as a baseline for further development of Information Extraction and NLP having numerous applications. We present approximately 5,500 unique and pre-processed Gujarati tweets which we have annotated into neutral, positive, and negative class categories. We also provide statistics for our dataset after experimentation. We then present the accuracy of pre-trained BERT models on our dataset.

The remainder of the manuscript is structured as outlined: we reviewed Related Work in Section II. The methods we used are described in Section III, in great detail, by explaining various phases such as Dataset Curation, Data Pre-processing, Dataset Annotation, and Dataset Statistics. Section IV goes on to present the outcomes of our methodology in detail. Section V goes over the future scope of this research, and concludes.

2. Related Work

Several methodologies have been introduced to conduct opinion mining to help with decision-making processes. There are numerous studies on the topic of analyzing user sentiments using language processing. Twitter Opining Mining, in particular, has now become a high-profile study topic among many researchers. Research papers relevant to our approach have been mentioned in this section.

Authors proposed an ML technique for performing Twitter Sentiment Analysis on tweets in Gujarati [11]. The pre-processing techniques employed in the actual model were stopword removal and stemming. The Support Vector Machine (SVM) classification technique and Parts of Speech (POS) tagging were applied in the system to extract features. SVM exhibited excellent performance and provided an accuracy of 92%. However, because the tool was only tested on 40 tweets, which is not regarded as a good dataset, the Gujarati language was briefly scratched in this study.

In Patel et al. [12], the authors particularly used a hybrid method to carry out opinion mining in the Gujarati language. An accuracy of 75% accuracy was achieved using the CNN approach, which also included N-gram and number features. Negation and Conjunction rules were also added to handle the scrutiny of words better in the sentence resulting in considerable accuracies of 90% and 70% respectively. However, their methodology concentrates on just two opinion polarities. Additionally, their SWN is limited to only the Education domain which restricts the versatility of the model to perform mining of opinions from different contexts.

The approach presented by Shah et al. [13] analyses sentiments in Gujarati movie reviews and can be roughly divided into five parts. Using the Python module beautiful soup, a dataset of 500 Gujarati movie reviews was compiled from the Gujarati Webidunia website. In the pre-handling step, stop words were eliminated and tokenization was completed. By dividing paragraphs into sentences and sentences into words, TF-IDF and CountVectorizer algorithms were used to separate tokenized features from clean data. Separated features were fed to two machine learning-based classifiers - KNN and MNB to assess the accuracy. Comparing the performance aspects of Accuracy, Recall, Precision, and F-score quality parameter, we found that the MNB model predicted opinions more accurately with TF-IDF features than with CountVectorizer features.

Because there aren't enough resources, opinion mining seems to be an interesting challenge when analyzing various attitudes for a language like Gujarati. Shah et al. [14] applied a sentiment analysis algorithm to Gujarati movie reviews. Five distinct datasets were produced by the authors to evaluate the effectiveness of the suggested algorithm. Four datasets were constructed by compiling movie reviews from four distinct websites, and one data set was personally created by obtaining reviews from various consumers. Pre-processing techniques Data Cleaning and Tokenization were

applied to the datasets. Multiple machine learning-based classifiers using unigram, bigram, and trigram characteristics are fed the feature vector created using the TF-IDF and Count Vectorizer routine as input. The performance of the classifiers is tested using this confusion matrix. Even though reviews were collected in English and translated into Gujarati, the authors' method of dataset preparation is particularly labor-intensive because one step involves manual preparation, and there is no mechanism to identify any ambiguities related to the transliterated text's language identification.

To enhance the segmentation of multilingual texts into various emotion categories for Emotion Analysis, authors in [15] proposed a new methodology. The dataset was generated by gathering and classifying tweets in Hindi, Gujarati & English according to the 8 emotional states on Plutchik's wheel. SenticNet resource-using feature-generating algorithms SN and CS-SN were presented for a hybrid approach. LinearSVC classifier's performance for Hindi and English was enhanced by their hybrid technique with the CS-SN feature generating technique. The efficiency of their hybrid technique, however, suffered for Gujarati because of its cultural relevance, colloquial tweets, and translation. The study also demonstrated that machine learning techniques performed superiorly in comparison.

The research in [16] established a novel viewpoint on sentiment acquisition of Gujarati Poems for expressing sentiments through Gujarati poems with a range of qualities. The authors created their dataset named ‘Kavan’ which has 300+ poems, written by well-known Gujarati literature poets. Poetry from "Kavan" was automatically input into the intended emotion detection system one at a time. Lines were processed to extract words consecutively. Zipf’s law was then employed to identify the number of occurrences of particular words statistically in Gujarati poetry that match predefined words in the ‘Rasa’ identification dataset. The final steps involved identifying the terms that appeared the most frequently and classifying the text as poetry based on the Indian concept ‘Navarasa’. Up to 87.62% accuracy was found in the findings for the emotion categorization task leveraging the Gujarati poetry corpus. However, their approach does not perform sentiment classification against standardized polarities and the problem associated with Zipf’s law related is an unambiguous formulation and lack of rigorous testing on many texts to prove its validity.

Utilizing an existent MT system to interpret phrase sections of different dialects is the easiest method for normalizing and translating code-mixed text. In 2020, authors presented the creation of a bilingual corpus for English and Gujarati [17]. Since there were fewer Gujarati terms in the vocabulary to handle unfamiliar phrases used by the Google API, they first manually created a lexicon of Gujarati terms with a suitable variety. To deal with word deviations and distinguish Gujarati languages when they were combined with any other language, they used the look-up approach in conjunction with Hidden Markov Model and Naïve Bayes Classifier. The work presented mainly concerns with transliteration for identifying the native Gujarati language used in Gujarati-English code-mixed text.

To consolidate text into numerical representations, feature extraction is done. The numerical data that has been processed serves as the input for data modeling processes. It is possible to extract features using a variety of methods. As input for the classifier, the textual data is transformed into a particular kind of vector using the most comprehensive count vectorizer technique. Every word sequence in the word is given a certain frequency when the TF-IDF with N-gram is used. Kazi et al. [18] used six machine learning classifiers to address the language identification dilemma with the romanized and non-romanized versions of Gujarati written in mixed-code script. According to the findings, the highest result was 92% for Support Vector Classifier (SVC) featuring RBF kernel and features associated with N-gram. However, TF-IDF features have a considerably smaller impact on efficiency than N-gram features. Deep learning techniques were presumed to have produced better outcomes in the field of language identification.

3. Methodology

3.1. Dataset Curation

Using the publicly broadcasted Twitter API and keywords to search for relevant tweets, we scraped Twitter data to construct the dataset. We have collected tweets from various Twitter accounts to increase the diversity of our dataset. We have only incorporated those tweets that have been primarily written in the Gujarati language. The tweets are stripped of any redundant textual elements such as hashtags, emoticons, mentions, numbers, special symbols, as well as infrequent English words. Multiple Python libraries and APIs are available which would aid us in scrapping the tweets. Tweepy, the official open-source library offered by Twitter, GetOldTweets, Twint, and Snscrape are just a few examples. We have used Tweepy for the creation of our dataset because it is more flexible and returns the maximum number of tweet results compared to other APIs.

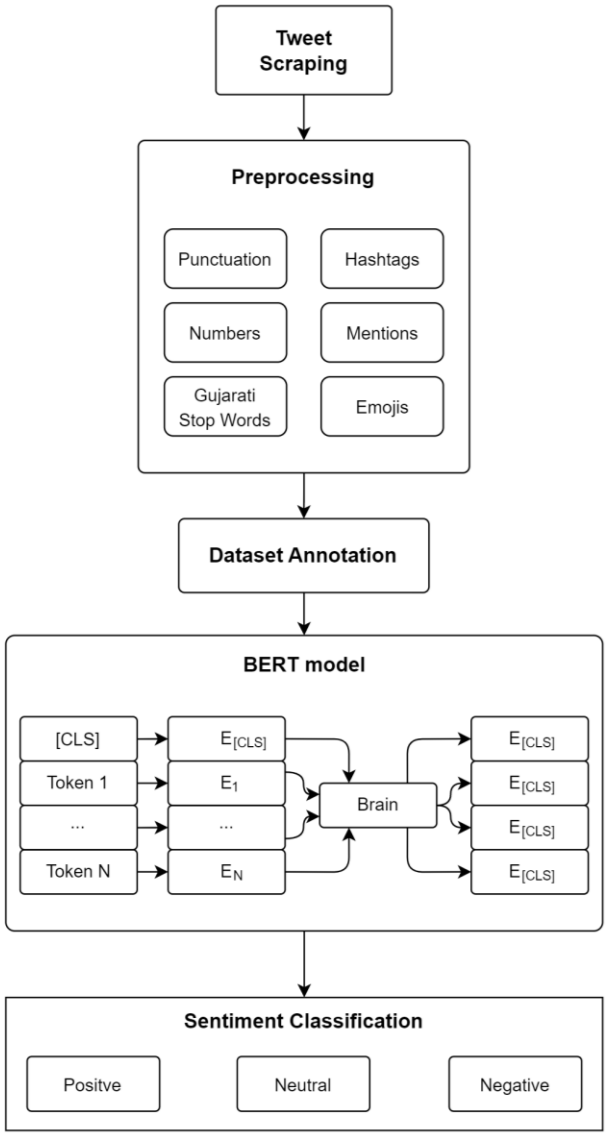

Figure 1. System Architecture

3.2. Data Pre-Processing

Because of how impulsively and intuitively people use digital platforms, the raw, unfiltered tweets obtained through Twitter's official API typically produces an extremely noisy and cryptic dataset. Cleaning of the dataset is necessary because we need only the tweet to analyze the sentiment and no other noisy sources. The official API returns 32 features of every tweet, and its unfiltered tweet consists of certain special features, such as emojis, hashtags, emoticons, email addresses, URLs, numerical data, and user mentions. We have removed all those words, and symbols from each tweet that do not add value to our classification's effectiveness in discerning sentiments. We divided pre-processing procedures into the following phases which are as follows:

3.2.1. Punctuations and Numbers

These are spaces and traditional symbols like “,”, “—”, “!”, “.”, “?” which are intended to assist readers to comprehend and read textual content correctly but are not necessary for assessing the sentiment. Additionally, since numbers have no significance in sentiment analysis, they are eliminated during pre-processing.

3.2.2. Gujarati Stop Word Removal

Sentiments are generally associated with adjectives and nouns, other natural language words which do not contribute to the sentiment such as pronouns, prepositions, and articles are categorized as stop words which we removed using IndoWordnet (IWN).

3.2.3. Hashtags and Mentions

Hashtags are terms that begin with the symbol # and are used to refer to well-known or well-liked topics or keywords. Hashtags act as the URL to a webpage that displays postings on the same subject. When speaking to or about another person on Twitter, mentions are keywords that are preceded by the symbol @ and include the username of that person. Tweets generally use a combination of @mentions and #hashtags so that readers can check out people involved or join the conversation about a topic. But none of them help us determine the sentiment of the tweet which makes them equivalent to noise.

3.2.4. Stemming

Stemming is a procedure in language processing where every derived word is transformed into its root or original form. We stemmed Gujarati terms in our suggested model to their root form so that it could be closely correlated with its lexicon utilizing - iNltk.

3.2.5. Negation List

To express words that alter the polarity of preceding words, we have prepared a list of popular Gujarati terms with negative polarity. When one of these words appears in the text, the next word's polarity is changed to be either positive or negative.

3.3. Dataset Annotation

We used our framework to annotate the entire dataset by mapping existing wordnet synsets. The labels "1", "-1" and "0" have been used to denote classes that are positive, negative, and neutral, respectively. The dataset was shared amongst the entire roster for parallel tagging. We established an annotation guideline to ensure consistency when tagging tweets. Positive feelings are defined as happiness, gratitude, respect, inspiration, and support. Hate, disrespect, sadness, insult, disagreement, and antagonism are all considered negative emotions. Neutral tweets do not evoke strong feelings, like simple assertions, statistics, or facts. Tweets using sarcasm or irony that overtly express a negative mood are assigned negative tags. Positive tweets include congratulatory and thank-you messages. A negative tweet criticizes something or someone or provides a fact about an undesirable event or reaction. Finally, tweets with mixed emotions are categorized according to the primary emotion conveyed.

A chain of wordnets from those in the Indo-Aryan, Dravidian, and Sino-Tibetan lineages of Indian languages makes up IWN. The Hindi wordnet, which was produced via the expansion method, serves as the base wordnet for all other IWN wordnets. By matching their synsets onto our dataset, we annotated it using the polarity scores of the two individuals mentioned above. The prior polarity of terms makes up the sentiment lexical resource. This polarity is unrelated to the situation. However, in a language with few resources,

lexical resources are an excellent place to start. Hindi SentiWordNet (H-SWN) and IWN were used to create our suggested lexicon resource. The following are the actions that were taken to create the lexical resource:

1) Repeat steps 2–5 for each H-SWN and SentiwordNet word (w).
2) Map w to the HindiWordNet (HWN) and SentiwordNet synset (s-id).
3) The synset of each synset lemma (s-id) is used to create a new synset lemma list (s-id)
4) H-SWN is used to project the polarity score of synset (s- id).

3.4. Dataset Statistics

We initially annotated 10,000 tweets in total, but we chose an equal number of tweets for every class at random to guarantee that the classifications are balanced. The result is that 5,487 tweets make up the complete set of our data. We divided the tweets into training and testing sets where 15% of the tweets, that is, 825 tweets, were set aside for testing purposes while the rest, 4662 tweets, were used to train the models - mBERT, XLM-RoBERTa, and IndicBERT. The remaining tweets aside from the ones involved in the final version of the data will also be published along with the dataset. These extra tweets have not been considered to measure the model performance.

4. Results

A sizable unlabeled corpus was used to train the deep bi-directional Transformer-based model BERT. Various BERT models, namely RoBERTa [20] and AlBERT [19], are considered as well within this study. Three primary paradigms for multilingualism can also be employed in the Gujarati language. These 3 models have been fine-tuned using a monolingual Gujarati dataset and distributed as part of this project. All models are trained using standard hyper-parameters and a masked language modeling aim for twenty epochs.

4.1. mBERT [21]

It's a vanilla BERT model that incorporates next-sentence prediction (NSP) and masked language modeling (MLM), and it has been pre-trained on 104 languages. Gujarati was among the languages operated during the pre-training.

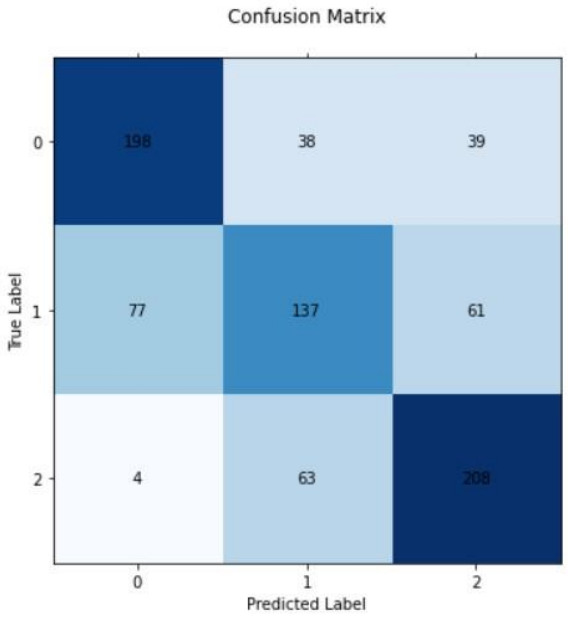

Figure 2. Confusion matrix for mBERT (3 class classification)

4.3. XLM-RoBERTa [22]

It is based on the RoBERTa model that's been well-trained in hundred languages with the goal of MLM. On a variety of tasks, the model outperforms mBERT. Gujarati is included as one of the pre-training languages in this model as well. The RoBERTa mostly alters the original BERT's hyper-parameters and eliminates the NSP task.

4.3. IndicBERT [23]

It's a multilingual AlBERT [19] model that's been specifically trained in twelve languages within India. AlBERT [19] is a simplified rendition of BERT. To lower the memory footprint, it employs parameter reduction techniques such as repeated layers. On the majority of Indic NLP tasks, the model has been shown to operate well.

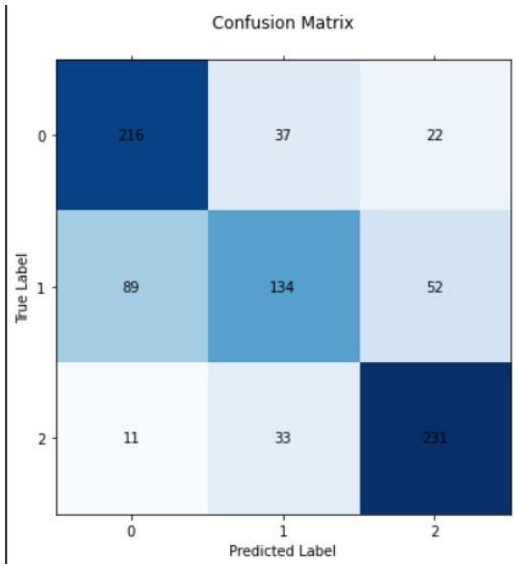

Figure 3. Confusion matrix for XLM-RoBERTa (3 class classification)

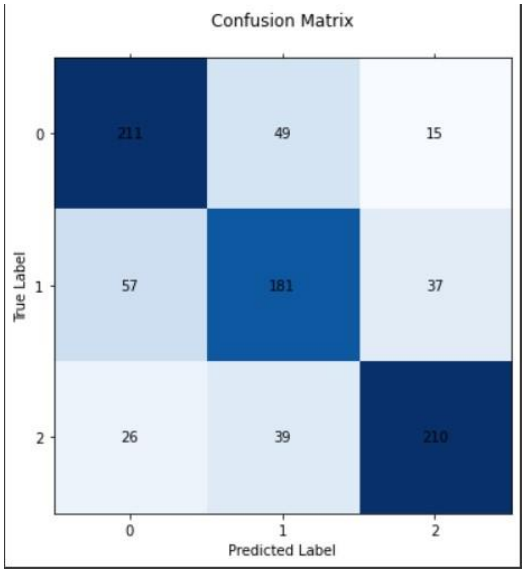

Figure 4. Confusion matrix for IndicBERT (3 class classification)

Table 1. Model Training Results

Model	*Training Data*	*Test Data*	*Accuracy*
mBERT [21]	4662	825	65.81%
IndicBERT [22]	4662	825	72.97%
XLM-RoBERTa [23]	4662	825	70.42%

5. Conclusion

When emphasizing Indian languages, opinion mining is an intriguing activity, but it becomes more difficult when attempting to interpret thoughts from a vernacular like Gujarati owing to a lack of adequate resources. In this research, we have outlined a framework to perform opinion mining in a low-resource Gujarati language.

We have harvested the first substantial Twitter dataset for Gujarati sentiment analysis, which consists of 5,487 tweets. After creating the dataset, we cleaned the unfiltered raw data. The data set was then annotated by building a lexical asset using IWN and H-SWN, mapping existing Wordnet synsets, and assigning sentiment polarity classes. The framework was trained on 4662 tweets using pre-trained BERT-based deep learning models - IndicBERT, mBERT, and XLM-roBERTa. Baseline classification results presented an accuracy of 72.97 percent on the IndicBERT model when testing all the models on 825 tweets. In the future, performance accuracies can be improved by incorporating a hybrid approach of feature extraction algorithms and machine learning models to identify culture-influenced patterns of Gujarati texts better.

References

1. Adke V Bakshi P and Askari M. Factors Impacting Adoption of Social Media Channels for Customer Service Management: A Review," 2022 17th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP), Corfu, Greece, 2022, pp. 1-5, doi: 10.1109/SMAP56125.2022.9942218.
2. Paul D, Li F, Teja MK, Yu X and Frost R. Compass: Spatio temporal sentiment analysis of US Election what Twitter says!, in Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Halifax, Canada, 2017, pp. 1585–1594.
3. Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: a review. Artif Intell Rev 53, 4335–4385 (2020). https://doi.org/10.1007/s10462-019-09794-5.
4. Esuli and Sebastiani F. ''SentiWordNet: A publicly available lexical resource for opinion mining,'' in Proc. 5th Conf. Lang. Resour. Eval. (LREC), Genova, Italy, 2006.
5. Wilson T, Hoffmann P, Somasundaran S, Kessler J, Wiebe J, Choi Y, Cardie C, Riloff E and Patwardhan S. OpinionFinder: A system for subjectivity analysis,'' in Proc. HLT/EMNLP Interact. Demonstrations, 2005, pp. 34–35, doi: 10.3115/1225733.1225751.
6. Nielsen FÅ. ''A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,'' in Proc. ESWC Workshop Making Sense Microposts, Big Things Come Small Packages, CEUR Workshop, vol. 718, May 2011.
7. Bakliwal, P. Arora, and V. Varma, "Hindi subjective lexicon: A lexical resource for hindi polarity classification," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC), pp. 1189–1196, 2012.
8. Mohanty, G., Kannan, A., and Mamidi, R. (2017). Building a sentiwordnet for odia. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017, pages 143–148.
9. Kannan, A., Mohanty, G., and Mamidi, R. (2016). Towards building a SentiWordNet for Tamil. In Proceedings of the 13th International Conference on Natural Language Processing, pages 30–35, Varanasi, India, December. NLP Association of India.
10. Baccianella S, A. Esuli, and F. Sebastiani, ''SentiWordNet 3.0: An enchanced lexical resource for sentiment analysis and opinion mining,'' in Proc. LREC, Eur. Lang. Resour. Assoc., vol. 10, 2010, pp. 2200–2204.

11. Joshi VC, Vekariya VM. An Approach to Sentiment Analysis on Gujarati Tweets," Advances in Computational Sciences and Technology, pp.1487-1493, 2017.
12. Himadri H. Patel, Bankim C. Patel, and Kalpesh B. Lad, 2022. Opinion Mining of Gujarati Language Text Using Hybrid Approach. United International Journal for Research & Technology (UIJRT), 3(4), pp.105-110.
13. Shah P, Swaminarayan P and Patel M. Sentiment analysis on film review in Gujarati language using machine learning," International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 1, pp. 1030–1039, 2022, doi: 10.11591/ijece.v12i1.pp1030-1039.
14. Parita Shah, Priya Swaminarayan, Maitri Patel, Nimisha Patel, "Sentiment Analysis on Movie Reviews in Regional Language Gujarati Using Machine Learning Algorithm," International Journal of Engineering Trends and Technology, vol. 70, no. 3, pp. 319-326, 2022. Crossref, https://doi.org/10.14445/22315381/IJETT-V70I1P236.
15. Gohil L and Patel D. Multilabel classification for emotion analysis of multilingual tweets," Int. J. Innov. Technol. Explor. Eng., vol. 9, no. 1, pp. 4453–4457, 2019.
16. Bhavin Mehta and Bhargav Rajyagor, "GUJARATI POETRY CLASSIFICATION BASED ON EMOTIONS USING DEEP LEARNING," International Journal of Engineering Applied Sciences and Technology, vol. 6, no. 1, pp. 358-362, 2021, doi: 10.33564/IJEAST.2021.v06i01.054.
17. Patel D and Parikh R. Language identification and translation of English and Gujarati code-mixed data,'' in Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng. (ic-ETITE), Vellore, India, Feb. 2020, doi: 10.1109/icETITE47903.2020.410.
18. Kazi M, Mehta H and Bharti S. Sentence level language identification in Gujarati-Hindi code-mixed scripts,'' in Proc. IEEE Int. Symp. Sustain. Energy, Signal Process. Cyber Secur. (iSSSC), Gunupur Odisha, India, Dec. 2020, pp. 1–6, doi: 10.1109/iSSSC50941.2020.9358837.
19. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In ICLR, 2020.
20. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
21. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
22. Conneau K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, ''Unsupervised cross-lingual representation learning at scale,'' 2019, arXiv:1911.02116.
23. Kakwani D, Kunchukuttan A, Golla S, Bhattacharyya A, Khapra MM and Kumar P. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages,'' in Proc. Findings Assoc. Comput. Linguistics (EMNLP), 2020, pp. 4948–4961.

Conflicts of interest: The authors stated that no conflicts of interest.

Publisher's Note
IJLSCI remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Correspondence and requests for materials should be addressed to Krutik Mehta.

Peer review information
IRJSE thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.