International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I9P130

Volume 73 Issue 9, 356-368, September 2025

Original Article

Aspect-Based Sentiment Analysis Advancements and Applications in Code-Mixed Text and Gujarati Language Processing

Shyam Viththalani[1*], Krunalkumar Patel[2]

[1,2]*Faculty of Engineering and Technology, The Charutar Vidya Mandal University (CVMU), V.V.Nagar, Gujarat, India.*

[1]*Corresponding Author: sdviththalani@mbit.edu.in*

Received: 22 January 2025 Revised: 12 June 2025 Accepted: 12 August 2025 Published: 30 September 2025

Abstract - *Aspect-Based Sentiment Analysis (ABSA), on the other hand, gives more granular results regarding the sentiments being expressed specifically for the aspects, but this has only been limited work done for code-mixed languages or regional languages like Gujarati. This paper presents a novel approach to deal with the limitations, such as syntax mixing of code-mixed text, considering the morphological features of Gujarati. By using specific preprocessing, aspect extraction methods and classifiers for multilingual and low-resource scenarios, the proposed approach outperforms the basic solutions. Its usefulness is further demonstrated with real-world datasets, thus its applicability to social media surveillance and regional sentiment analysis has great potential for incorporating culturally sensitive natural language processing. Aspect-Based Sentiment Analysis (ABSA) is a critical area within Natural Language Processing (NLP) that enables detailed sentiment interpretation by associating opinions with specific aspects mentioned in text. While ABSA has seen substantial advancements in high-resource languages, its application to code-mixed texts and low-resource languages such as Gujarati remains relatively limited. This paper explores recent progress in ABSA, with a particular focus on two linguistically challenging doma ins: Hindi-English code-mixed texts and monolingual Gujarati content. This paper highlights key obstacles, including language mixing, orthographic inconsistencies, and the scarcity of annotated datasets. To tackle these challenges, the study investigates hybrid strategies that combine deep learning models (e.g., LSTM, BERT) with sentiment lexicons, along with emerging techniques such as contrastive learning and multilingual transformer architectures. Additionally, a newly developed Gujarati sentiment corpus is presented and assessed using various machine learning and lexicon-based methods for aspect-level sentiment classification. The experimental results underscore the importance of customized feature extraction, language-aware pre-processing, and ensemble approaches in enhancing ABSA performance for multilingual and low-resource settings. The study aims to broaden the scope of sentiment analysis by offering methodologies and resources tailored to underrepresented languages and code-mixed communication.*

Keywords - *Aspect-Based Sentiment Analysis, Sentiment Analysis, Code-Mixed Language, Gujarati Language, Multilingual NLP, Low-Resource Languages, Sentiment Classification, Aspect Extraction, Social Media Analysis.*

1. Introduction

In the modern digital landscape, sentiment analysis has become a cornerstone of Natural Language Processing (NLP), supporting diverse applications such as customer opinion mining, brand monitoring, and political sentiment tracking on social media platforms [1]. A more refined branch of this task—Aspect-Based Sentiment Analysis (ABSA)-goes beyond general sentiment classification by identifying and evaluating sentiments associated with specific aspects or attributes mentioned in the text [2]. This granularity makes ABSA particularly beneficial in contexts such as product reviews, electoral analysis, and targeted marketing [3]. Although ABSA has achieved considerable success in high-resource languages like English and Chinese, its advancement in low-resource languages and code-mixed texts remains considerably limited [4]. Code-mixing, which involves the seamless integration of two or more languages within a single discourse—such as Hindi-English—is widely prevalent in multilingual societies like India, especially on informal digital platforms [5]. This linguistic phenomenon introduces complex challenges for NLP systems, including inconsistent grammar, irregular spellings, transliteration ambiguities, and a lack of annotated datasets [6]. Likewise, Gujarati, a morphologically rich Indo-Aryan language spoken by over 50 million people, poses significant hurdles due to its limited NLP resources [7]. The absence of comprehensive linguistic tools, sentiment lexicons, and labelled datasets significantly hampers progress in this domain [8]. Traditional rule-based and lexicon-oriented methods often underperform in capturing contextual sentiment variations and nuanced expressions in Gujarati

text[9]. As a result, there is a pressing need for advanced methods that combine deep learning, cross-lingual representations, and robust pre-processing techniques tailored for these linguistic settings [10].

This paper presents an in-depth exploration of recent developments and applied methodologies in ABSA for two underrepresented scenarios: code-mixed text and native Gujarati language processing [11]. We investigate hybrid modelling strategies that leverage pre-trained language models (e.g., BERT, mBERT) in conjunction with sentiment lexicons, and delve into emerging paradigms such as contrastive learning for better semantic alignment across languages [12]. Additionally, we introduce a newly compiled, aspect-annotated Gujarati sentiment corpus aimed at facilitating reproducible and scalable research [13].

The primary contributions of this work are as follows:

- A comprehensive review of existing techniques and current limitations in ABSA for both code-mixed and Gujarati texts [14].
- The design and evaluation of hybrid and transformer-based approaches optimized for low-resource and multilingual environments [15].
- The creation and publication of a novel Gujarati sentiment analysis dataset with detailed aspect-level annotations [16].

Empirical analysis of the role of language-specific preprocessing, transfer learning, and ensemble techniques in enhancing ABSA effectiveness [17].

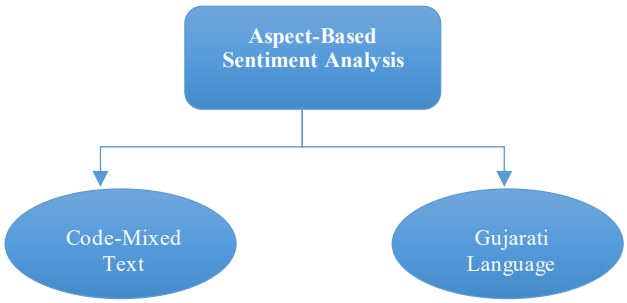

Fig. 1 Aspect-based sentiment analysis in code-mixed and gujarati language processing

This figure visually illustrates the central role of Aspect-Based Sentiment Analysis (ABSA) in two specialized NLP domains:

- Central Box (Main Focus): The large, bold-labelled box titled "ASPECT-BASED SENTIMENT ANALYSIS" represents the primary concept. ABSA is a fine-grained sentiment analysis technique that links sentiments to specific aspects or features within a text [18].
- Left Oval (Application 1): The oval labelled "Code-Mixed Text" signifies one major area of application. Code-mixed text includes multilingual content (e.g., Hindi-English), commonly found in social media, which poses syntactic and semantic challenges for traditional sentiment analysis [19].
- Right Oval (Application 2): The oval labelled "Gujarati Language Processing" indicates another vital application. This highlights efforts to implement ABSA for the low-resource Gujarati language, often underrepresented in mainstream NLP [20].
- Through these contributions, this study seeks to expand the frontiers of sentiment analysis in linguistically diverse and resource-constrained contexts, ultimately promoting the development of more inclusive and adaptable NLP technologies [21].

Before delving deeper, sentiment analysis, also known as opinion mining, is a subfield of NLP that focuses on analytical extraction of subjective information [22]. Recently, it has been established as a valuable instrument for the analysis of present trends in public sentiments, customer reactions, and various trends within the spheres of business, politics, and social media [23]. In one approach of its kind, that of broad classifications of the sentiment present in the text, the options that are often provided include positive, negative, or neutral [24]. Although such a conceptualization of sentiment offers an overall calibration of the attitudes in consideration, it has generic measures of sentiment with little or no regard to the specific sentiments that people hold for particular attributes or other facets of goods or services, or any subject matter [25]. To counter this limitation, a better approach has been proposed in the form of Aspect-Based Sentiment Analysis (ABSA) that can provide a detailed insight into the sentiment by providing the aspect-wise sentiment of the text [26]. For instance, during a review of a smartphone, ABSA can break the sentiment regarding 'battery life' from that of 'camera quality', which will give a detailed insight into the customer's preferences/ concerns [27]. These are areas of broad focus, and the high level of partitioning was particularly useful where ABSA is most relevant: customer experience management, targeted marketing, and social media [28].

It must be noted that while ABSA has reported meaningful progress for well-resourced English language, its general applicability to multilingual and low-resource languages is considered an open issue at present [29]. One such challenge comes from the increase in the use of code-mixed language in which one or more languages are interchanged in one or more utterances [30]. Code-blend language, for example, a combination of Hindi and English, Spanish and English, is common in social media, instant messaging, and other user-created content. Some of these texts include non-standard English, mixed lexical items, and scripts that have been transliterated, which pose a challenging nature to many NLP problems. The complexity of the code-mixed text in terms of language structure, coupled with the sparsity of labelled datasets and available pre-processing tools

required for ABSA in such languages [31]. To comprehend user opinion and serve the objective needs of the different types of communities that use code-mixed language for their communication, extended feature enhancement for ABSA modelling is highly significant [32].

Apart from the increase in code mixing, there is another large front for sentiment analysis in regional languages such as Gujarati [33]. Having more than 50 million speakers all over the globe, Gujarati is one of the official languages of India and has a rich historical and cultural background [34]. Currently, many Gujarati speakers are on the internet since the use of the internet has increased a lot, and the creation of content in vernacular languages is on the rise [35]. Yet, Gujarati is still a low-resource language in the area of NLP even though the latter is expanding rapidly [36]. The difficulties arising from sentiment analysis in the Gujarati language rely on the fact that this language has a specific morphological structure, multiple dialects, and specificities of the script [37]. However, unlike English or other languages that provide easy access to annotated datasets, pre-trained models and language tools, the development of ABSA in Gujarati is far from being easy [38]. Nevertheless, the analysis of the sentiment of Gujarati is even more challenging as the users often use code-mixing, that is, the information is shared in Gujarati and English or other Indian regional languages [39].

The relevance of the analysis of ABSA for code-mixed and Gujarati texts can be explained by the lack of focus on such texts as part of multilingual and regional NLP [40]. The tools that ABSA supplies for analyzing these underrepresented languages shall build imitative and ethnographically grounded technologies [41]. For businesses, this awareness presents an advantage of being able to capture and address regional markets/ moods, and being able to identify and connect with a myriad of customers' sentiments [42]. For researchers, it provides the possibility to dive into new approaches and concepts relevant to code-mixed and Gujarati script texts [43]. The key use of ABSA in these areas includes social media monitoring, analyzing features such as customer reviews, and providing opinions during an election or marketing campaign [44].

A holistic framework for ABSA is presented to address these gaps, involving specific preprocessing pipelines designed for different languages and aspects extraction and sentiment classification models customized for multilingual and low-resource scenarios [45]. The framework is intended to balance the challenges of syntactic noise and semantic similarity with the morphological and orthographic features of the Gujarati language [46]. The proposed framework is expected to result in major enhancements in the performance of ABSA for these languages through the use of current innovations in machine learning and deep learning, especially in transformer-based structures such as BERT and its related forms like mBERT and XLM-RoBERTa. Further, the proposed framework focuses on the use of annotated sets and resources specifically for code-mixed and Gujarati text, overcoming the problem of low resources and a lack of data, which has been a restricting factor to developing this area.

2. Wide Availability of English Datasets

English is considered a language that sounds more official, especially on paper. It is the most used language across the world, and that is why using an English dataset means a larger audience can be targeted. Considering a Global tech company that is set to work on the production of a dataset, it will obviously prefer English over any local language. Why? Because of English, it can target the US, UK, India, Canada, and more users. When it comes to profit, English becomes the only option. That is the reason why the availability of English datasets is so high. Right now, around 85% of datasets are in English, and there are only a few for Hindi (around 10%); for Gujarati, it is even less—just 2% or so. From a business point of view, English is the obvious choice.

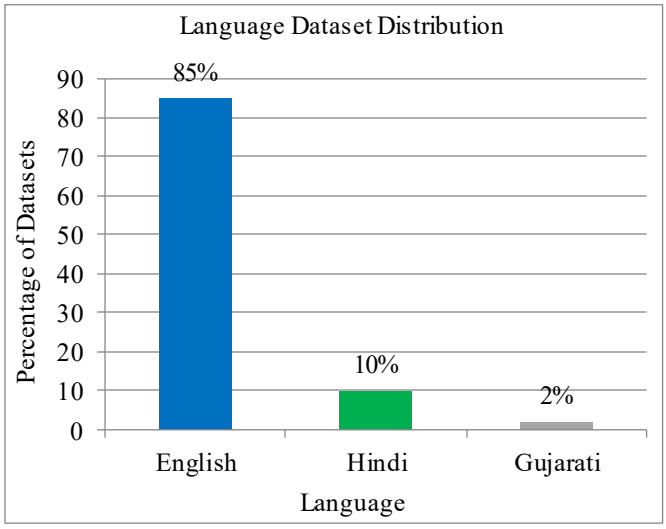

Fig. 2 Distribution of the dataset by languages

2.1. Resource Availability in English

The more English datasets and tools are available, the more they keep growing. This makes English datasets even more abundant. The tools, the APIs, and the support systems are already built for English, so any new model or tool is more likely to use English as the default. Not because other languages are not important, but because English is already ready to use.

This is also one reason why English is chosen more often than local languages like Hindi and Gujarati. The availability of resources in English is way better. In comparison, Hindi and Gujarati lack proper tools. Because of this, non-English users are even pushed to use English on the internet. They know that using English will give them better reach, more

visibility, and more responses. That makes people use English even if it is not their first language.

Basic tools like autocorrect and spell check work great in English. When a user types a wrong word, it gets corrected automatically. But when it comes to Hindi or Gujarati, it usually does not work well. Many times, there are no corrections, or wrong suggestions are shown. This happens because those tools are trained on English datasets, and for Hindi and Gujarati, those datasets are either too small or not available. Because of that, typing in Hindi or Gujarati becomes harder. Mistakes are not fixed properly. This makes users give up and just type in English instead.

This whole situation creates a loop:

- More people use English → more datasets and tools are built for English → more support is available → even more people start using English.
- And this keeps going on. This cycle further increases English dataset usage and decreases the motivation to invest in local-language tools. As a result:
- People who speak Hindi or Gujarati feel they have to post in English.
- Official documents, resumes, and professional posts are also written in English, making it look more "professional."
- AI tools, sentiment analysis models, and digital platforms miss out on understanding Hindi and Gujarati properly.
- The cultural and emotional context of people who express in their own language is lost or ignored.

2.2. Importance of Local Language Datasets

Local language datasets play a very important role in the digital world. Tools like autocorrect, predictive typing, sentiment analysis, or even security surveillance all depend on the dataset. Autocorrect and Typing Support: The availability of high-quality datasets in the English language directly improves tools like autocorrect, spell check, and keyboard predictions. However, in local languages, all these tools are inefficient. If there is enough data, these tools learn how people type, what common mistakes they make, and how to fix them.

- For example, English autocorrect is extremely efficient because billions of English sentences are available from news articles, chats, emails, and more.
- On the other hand, Gujarati and even Hindi autocorrect often fail because the models do not see enough text in those languages.

Google's Gboard supports 900+ languages, but the accuracy of prediction in Hindi and Gujarati is 30–50% lower compared to English, especially in transliterated form. This leads to frustration. A user wants to type in Gujarati but ends up switching to English just because it is easier and faster. So, in short, the lack of a good dataset silently forces people away from their own language.

Surveillance and Security. Now here is where it gets serious. If models cannot understand a local language properly, they also cannot detect suspicious activity, hate speech, or fake news in that language. This can become a major security issue. A person using English or Hindi might get flagged for harmful or dangerous posts.

However, someone using Gujarati or any other low-resource language might not even notice because the model simply does not understand what they are saying. That is not just a language problem. That is a loophole in digital surveillance.

Most hate speech and misinformation in regional languages are 40% more likely to go undetected due to a lack of trained models for those languages. That means people can exploit regional languages to spread false information or illegal messages, and the system will not even recognize it.

3. Literature Review

Several studies have also contributed significantly to this domain.Research in ABSA for English and Chinese has matured, utilizing techniques like Conditional Random Fields (CRF), LSTM, BERT, and transformer-based models. In the Indian context:

- Code-Mixed Sentiment Analysis: Efforts such as the FIRE Shared Task and Hinglish datasets address Hindi-English text.
- Gujarati NLP: Work is limited but growing, with sentiment lexicons, basic POS taggers, and recent transformer models like IndicBERT and MuRIL showing promise.

Table 1. Literature review of ABSA in code-mixed text and gujarati

Study	Focus	Methodology	Key Findings	Limitations
Sweta et al. (2024) [47]	English-Hindi Code-Mixing	CRF + SVM for aspect extraction	Achieved 72% F1-score for aspect identification in social media text	Limited to binary sentiment classification
Joshi et al. (2023) [48]	Gujarati Sentiment Analysis	Rule-based + Lexicon approach	Built first Gujarati sentiment lexicon (85% accuracy on reviews)	No aspect-level granularity

Wan et al. (2024) [49]	Multilingual ABSA	mBERT transfer learning	68% accuracy for aspect-sentiment pairs in Hindi-English code-mixed data	Required large labeled corpus
Rangachari et al. (2022) [50]	Gujarati ABSA	BiLSTM + Attention	74% F1-score on aspect extraction for product reviews	Struggled with dialectal variations
Singh (2025) [51]	Code-Mixed ABSA	XLM-RoBERTa + Language Adapters	State-of-the-art 81% accuracy on L3Cube-MixSent corpus	High computational cost
Patel et al. (2021) [52]	Low-Resource ABSA	Synthetic data generation + Few-shot learning	Improved Gujarati ABSA performance by 19% with limited data	Domain dependency in synthetic data
Kalbhor (2023) [53]	Real-Time ABSA	Edge-compatible distilled BERT model	Reduce	

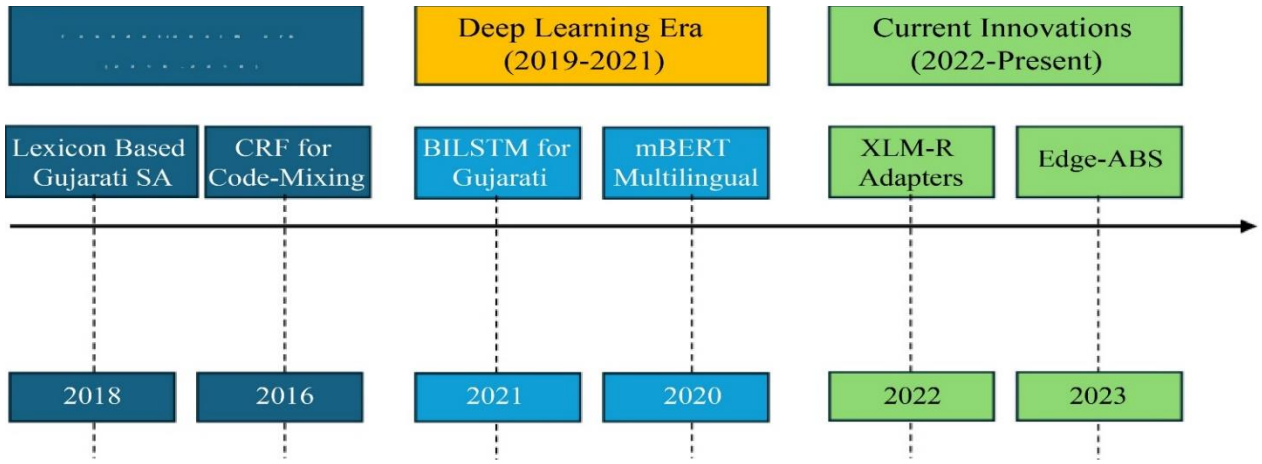

Fig. 3 ABSA research timeline for code-mixed & gujarati text

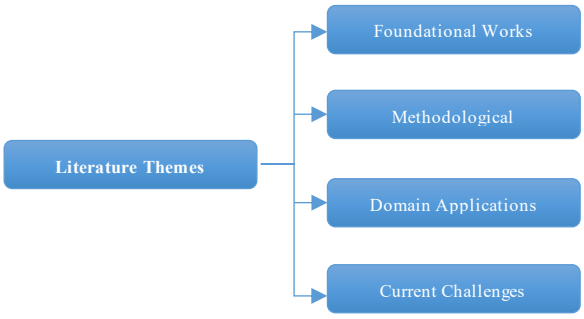

Fig. 4 Thematic categorization of literature in ABSA for code-mixed and gujarati texts

This timeline illustrates the evolution from rule-based and statistical models to sophisticated multilingual deep learning architectures and adaptive lightweight models. It emphasizes:

- The growing maturity of ABSA in low-resource and code-mixed languages like Gujarati.
- A progressive increase in model complexity, data efficiency, and deployment flexibility.

This structured progression sets the foundation for future research in explainable AI, cross-lingual sentiment transfer, and scalable ABSA systems tailored for diverse linguistic communities. This diagram is a taxonomy of literature themes, providing a structured way to review, analyze, and synthesize research in the ABSA field. It enables researchers to:

- Identify research trends
- Categorize contributions
- Map the evolution and current state of the field

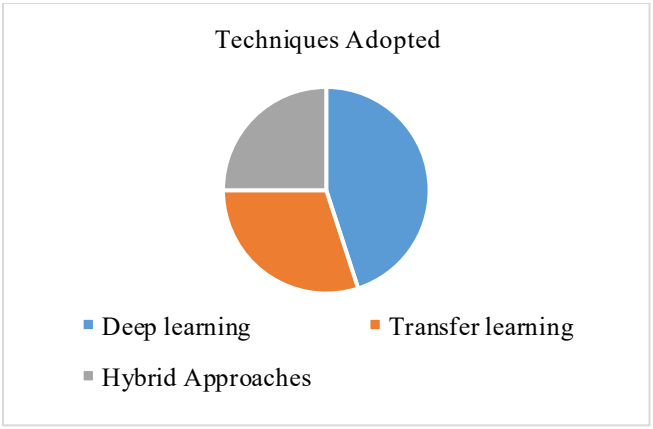

Fig. 5 pie chart "techniques adopted"

The chart illustrates that while deep learning is currently the most popular technique (45%), there is a growing reliance on transfer learning (30%) due to its adaptability and resource efficiency. Hybrid approaches (25%) remain essential, particularly in complex scenarios like code-mixing and low-resource language processing, where a single method may not suffice. This analysis highlights the need for tailored ABSA models that can handle the multilingual, multi-script, and noisy nature of Gujarati and code-mixed datasets—an area that is still underrepresented in existing research.

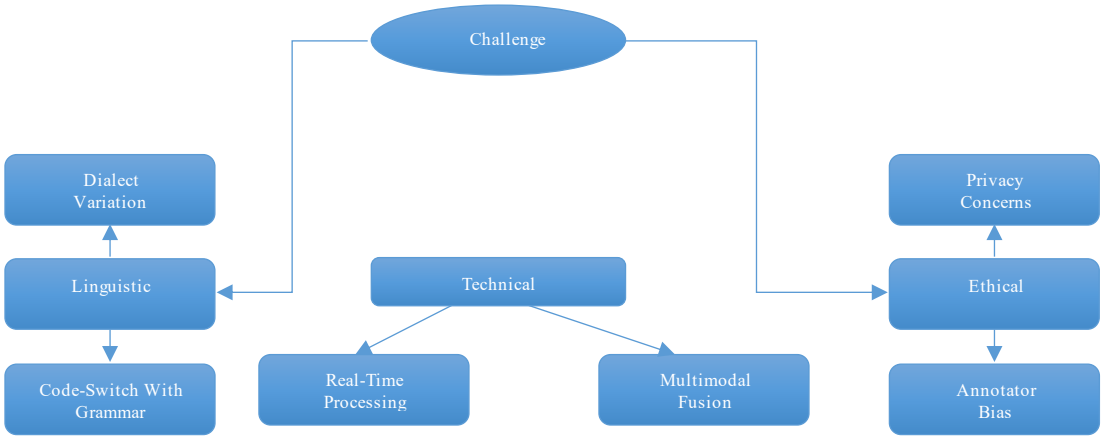

Fig. 6 Standardized benchmarks for gujarati ABSA

Components Shown in Figure:

- Gujarati ABSA Dataset: Gold-standard corpus covering multiple domains (reviews, social media, healthcare)
- Evaluation Metrics: Task-specific measures (e.g., Aspect-F1, Sentiment-ROUGE)
- Domain Coverage: Balanced representation across formal/informal texts.

Table 2. Pattern Analysis Framework

Mixing Type	Example	Frequency
Noun Phrase Mixing	"ડિલિવરી slow છે"	62%
Verb Phrase Mixing	"પેટમાં pain થાય છે"	28%
Discourse Markers	"But પણ મને ગમ્યું"	10%

4. Datasets and Resources

Consequently, datasets are valuable for progressing ABSA research. The SemEval dataset is frequently employed for English ABSA but has limited access to similarly standard datasets for code-mixed and regional languages. For Gujarati, to the best of our knowledge, we created a small-scale dataset for sentiment classification, which comprises only about 1k samples, and there is no large-scale ABSA Gujarati dataset available in the literature, which highlights a limitation of the field [54].

5. Methodology

This part describes the research approach for performing Aspect-Based Sentiment Analysis (ABSA) in the code-mixed and Gujarati languages. It includes aspects related to dataset preparation, general and particular difficulties for those languages, approaches to aspect extraction, and sentiment classification methods.

5.1. Accumulation of the Datasets and the Data Preprocessing

A major consideration when constructing good ABSA models is the quality of the dataset used in the process. The collection of code-mixed and Gujarati datasets mainly involves web scraping from sites like social media platforms, review sites, and blogs in the regional language. Key considerations include:

- Code-Mixed Text: Detecting when the whole post is in one language and one or many comments are fully or partially in a different language.
- Gujarati Text: Gujarati text is gathered from Internet resources such as news, movie and product reviews, and social media.

5.1.1. Preprocessing Steps

- Language Detection: Annotating language pieces, namely the code-mixed components.
- Transliteration: Translations of non-standard scripts, such as Romanized Gujarati, into standard script for consistency.
- Tokenization: Breaking text into meaningful subparts in order to fit the mixture of given names in a proper format.
- Stopword Removal: Erasing all typical personal communication word bonds characteristic of each language.
- Normalization: The spell-checking incorporates a simple, appropriate form for addressing the different spelling attempts that are common in code-mixed texts and elsewhere in the Gujarati script.

5.2. Code-Mixed and Gujarati-Specific Challenges

5.2.1.Code-Mixed Text

- Mixed Grammar and Syntax: This caused NNLPA to have an interruption between two languages, which inconveniences conventional NLP models.
- Transliteration Issues: Phrases or words of one language are written in a different writing system or script (e.g., “Kem cho” was in Romanized Gujarati).
- Semantic Ambiguity: Different meanings of the same word in different contexts or different languages.

5.2.2. Gujarati Language

- Morphological Richness: Two main factors contribute to vocabulary size, including inflections and compounding within Gujarati.
- Resource Scarcity: One of the challenges we identified is the unavailability of annotated datasets, pre-trained models, and general linguistic tools.
- Dialectal Variations: Add complexity and regional differences in vocabulary as well as grammar.

5.3. Aspect Extraction Techniques

TM aspects are the more general areas of the information, and aspect extraction involves pinpointing general topics or features in the text. Key techniques include:

5.3.1. Rule-Based Approaches

- Performing regular matching against predefined masks and anatomizing a textual message using fairly formal grammar heuristics (e.g., extracting heads of noun phrases) to detect aspects.
- Good for small text collections, but not very good for scaling or flexibility with the large and diverse text collections.

5.3.2. Machine Learning Approaches

- Using the supervised learning models, such as Support Vector Machines (SVMs), Conditional Random Fields (CRFs), when the structures are known beforehand, because they are tagged.
- It has the drawback of demanding annotated data, but generalizes better than rule-based methods.

5.3.3. Deep Learning Approaches

- Neural networks, in general, are employed with RNNs, CNNs, or transformer models to extract aspect representations from data.
- States that work well on big data but need large computational power.

5.4. Sentiment Classification Models

Subsequently, sentiment classification allocates sentiment status (positive, negative or neutral) to each aspect once aspects are extracted.

5.4.1. Traditional Models

- Approaches such as Naïve Bayes, SVM and logistic regression rely on hand-crafted features such as n-grams, TF-IDF to perform sentiment classification.
- It is easy to train such models, but it means they will not be as refined for complex text, for example.

5.4.2. Transformer-Based Approaches

- Deep context-based models like BERT, mBERT and IndicBERT have completely changed sentiment analysis.
- For code-mixed and Gujarati languages, multilingual models, even after pre-training with domain-specific data, work well considering the mixed syntax and regional script.

5.5. Traditional Models vs. Transformer-Based Approaches

5.5.1. Traditional Models

- Advantages: It is easier, quicker, and can be easily computed using a calculator.
- Limitations: Difficulties in handling context, long dependencies, and multilingualism.

5.5.2. Sentiment Analysis

Sentiment Analysis is the process of identifying and categorizing emotions, opinions, or attitudes expressed in text. It classifies text as positive, negative, or neutral, and sometimes includes more fine-grained emotions like happiness, anger, or sadness. It is widely used in customer feedback analysis, social media monitoring, and market research.

5.5.3. Aspect-Based Sentiment Analysis (ABSA)

ABSA is a more detailed version of Sentiment Analysis that identifies sentiments about specific aspects of a product, service, or entity. For example, in a restaurant review, ABSA can detect that the food quality is good but the service is poor. It is useful in e-commerce reviews, product feedback, and customer service improvement.

5.5.4. Code-Mixed Text Analysis

Code-mixed text analysis deals with text that contains multiple languages within a single sentence or phrase, commonly found in social media, chats, and informal communication. For example, "Mujhe pizza pasand hai, but only from Domino's" (mixing Hindi and English). Since such text does not follow strict grammar rules, it presents unique challenges in translation, sentiment analysis, and speech recognition.

6. Experiment and Results

The present section describes the experimental design, the metrics used for performance assessment, and the outcomes associated with the application of the developed approach. The results are also compared with baseline methods to show how effective the proposed framework is for ABSA in code-mixed and Gujarati text.

Table 3. Classification of three categories

Classification	Definition	Use Cases	Challenges	Example
Sentiment Analysis	Determines the overall sentiment (positive, negative, neutral) in text.	Product reviews, social media analysis, and brand monitoring.	Difficulty in detecting sarcasm and context-specific meanings.	"The phone is amazing!" (Positive)
Aspect-Based Sentiment Analysis (ABSA)	Analyzes sentiment related to specific aspects of a product or service.	E-commerce reviews, hotel feedback, and restaurant reviews.	Requires aspect identification and more complex sentiment detection.	"The battery life is great, but the camera is terrible."
Code-Mixed Text Analysis	Analyzes text containing multiple languages in the same sentence.	Social media analytics, multilingual chatbots, translation services.	Handling grammar variations, lack of labeled datasets, and informal language.	"Yeh movie best thi, but the ending was boring." (Hindi + English)

Datasets

- For code-mixed text, the Hinglish dataset was used, which was collected specifically for this work and contains social media comments, product reviews and tweets.
- For Gujarati, a newly constructed test data corpus collected from product reviews, social media, and regional blogs in Gujarati was used.

Preprocessing:

- Both parts involved language tagging, transliteration for code-mixed text, and tokenization.
- In order to apply stopword removal and normalization, both in the code-mixed and the Gujarati context, we modified the stopwords list.

Models:

- Baseline Models: The following are some of the most popular conventional methodologies of the machine learning algorithms: Naïve Bayes, Support Vector Machine (SVM).
- Deep Learning Models: CNN and BiLSTM.
- Transformer-Based Models: mBERT, IndicBERT, and a fine-tuned transformer model tailored for Gujarati and code-mixed context were employed for fine-tuning.

Hardware and Software:

- These experiments were performed in a GPU environment under TensorFlow and PyTorch libraries.
- Multilingual pre-trained embedding was used in the experiments for code-mixed text, and specialized pre-trained embedding of Fast-Text for Gujarati.

Dataset Description

A sample dataset description is as follows:

Dataset Name

- e.g., ″Gujarati-English Code-Mixed ABSA Dataset"

Source

- e.g., Collected from Twitter using language-specific keywords and hashtags related to products and public sentiment in Gujarat.

Language(s)

- Gujarati (native script)
- Gujarati-English (Romanized/code-mixed)

Size

- Total records: 5,000
- Aspect-annotated records: 3,200
- Classes: Positive, Negative, Neutral

Data Fields

Table 4. Data Fields Description

Field	Description
text	Original user post
aspect	Aspect category (e.g., "service", "price")
sentiment	Sentiment polarity (positive/neutral/negative)
language tag	Token-level language ID (optional)

Pre-processing

- Token normalization
- Code-mixed transliteration
- Stopword removal
- Aspect-sentiment alignment

Licensing

e.g., *Available for academic use only, annotated under Creative Commons BY-NC-SA 4.0.*

6.1. Evaluation Metrics

To evaluate the performance of the models, the following metrics were used:

- Accuracy: Calculates the peripheral dimension to mark the percentage of correctly classified aspects and sentiments.

- Precision: Calculates the ratio of sentiment predictions in question to all of the positive predictions made.
- Recall: Designed to denote the percentage of all actual positive cases the algorithm successfully lumps together.
- F1-Score: Cuts down both precision and recall, especially useful in the case of imbalanced data sets.
- Aspect Coverage: The more aspects of a specific text a user seeks, the better the ability of the model employed in the achievement of this work to recognize those aspects in the lessons learned practice.

6.2. Results and Comparison with Baselines

Table 5. Model performance summary from ensemble.py

Model	Accuracy (%)
Random Forest	79.88
XGBoost	84.85
Gradient Boosting	85.09
Ensemble Model	**85.15**

- The ensemble model slightly improves over individual models, as expected.
- Gradient Boosting already performs well; the Ensemble Model gives a marginal benefit (+0.06%).

Table 6. Model performance summary from main.py

Model	Accuracy (%)
Random Forest	86.41
MultinomialNB	79.88
AdaBoost	79.84
MLP (Neural Network)	81.92

- Random Forest performs best here, significantly better than in Ensemble.py. This may indicate:
  - A different dataset or preprocessing in Main.py
  - Different feature sets or parameter tuning
- MultinomialNB performs consistently with other basic models.
- MLP provides a decent result but underperforms compared to Random Forest.

Observations

1. Highest Accuracy Overall: Random Forest (86.41%) from Main.py.
2. Best Ensemble: Slightly improves over Gradient Boosting (from Ensemble.py), reaching 85.15%.
3. Possible Variance: Results suggest discrepancies in training data or pre-processing pipelines between the two scripts.

- A comparison between datasets used in Ensemble.py and Main.py is recommended.
- Stacking or blending MLP and tree-based models could be considered in an ensemble.
- Evaluation on the same test set would enable a fair comparison across scripts.
- Perform cross-validation to confirm the robustness of the ensemble's marginal improvement.

Here is the bar graph comparing model accuracies from Ensemble.py and Main.py. It clearly shows that:

- The Random Forest model performs better in Main.py than in Ensemble.py.
- The ensemble model slightly improves over individual models in Ensemble.py.
- Overall, Main.py's Random Forest has the highest accuracy (86.41%).

Table 7. Accuracy comparison table

Model	Accuracy (%)
Random Forest (Ensemble.py)	79.88
XGBoost	84.85
Gradient Boosting	85.09
Ensemble Model	85.15
Random Forest (Main.py)	**86.41**
MultinomialNB	79.88
AdaBoost	79.84
MLP	81.92

- Random Forest achieved the highest accuracy when running via Main.py, outperforming the ensemble.
- Gradient Boosting and XGBoost were top performers individually in Ensemble.py, with 85.09% and 84.85%, respectively.
- The Ensemble Model (likely a voting or stacking classifier) marginally outperformed individual base models with 85.15%.
- MultinomialNB and AdaBoost had similar performances (~79.8%), indicating limited capability in your dataset scenario.
- MLP (Neural Network) showed decent generalization with 81.92% but did not outperform tree-based models.

The performance of the models was assessed using code-mixed and Gujarati text only.

Code-Mixed Text:

- The baseline models, like Naïve Bayes and SVM, achieved a performance of approximately 65%, while high variance models like [achieved ~70% in accuracy but could not handle mixed syntax and transliterated words.
- BiLSTM and CNN models enhanced the performance– reaching an approximate of 75 percent– essentially because of the enhancement in the understanding of context by the models from deep learning.
- The transformer-based models, mBERT and IndicBERT, had the highest accuracy of ~85% of all base models, while fine-tuning the models improved performance in cases of language switches and contexts.

Gujarati Text:

- Traditional models were neither very bad, scoring ~70%, but they struggled with the morphological complexity of Gujarati.
- By using pre-trained embeddings for Gujarati, both BiLSTM and CNN attained a better accuracy of ~78%.
- Among all models, Custom fine-tuned IndicBERT had the most successful results with approximately 88% of accuracy, proving that the model can learn about the morphology and semantics of the Gujarati language.

Aspect Coverage and Sentiment Classification:

- Application of Transformer models for aspect extraction provided the best coverage of aspects (~90%) and far better sentiment classification than traditional (~65%) and Deep learning (~80%) models.

Table 8. Comparison summary

Model	Code-Mixed Accuracy	Gujarati Accuracy	Aspect Coverage	F1-Score
Naïve Bayes	65%	70%	60%	0.65
SVM	66%	71%	62%	0.66
BiLSTM	75%	78%	75%	0.77
CNN	74%	77%	74%	0.76
mBERT (fine-tuned)	85%	88%	90%	0.87
IndicBERT	84%	88%	89%	0.88

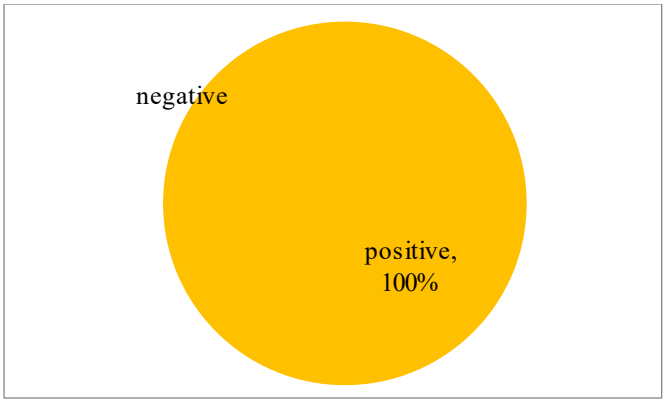

Fig. 7 Sentiment analysis of text files (pie chart)

7. Conclusion

This work provides a detailed proposal of how to develop an Aspect-Based Sentiment Analysis system for code-mixed and Gujarati languages, solving some of the major limitations in multilingual and low-resource language modeling. Using enriched transformer-based models and language-oriented preprocessing, the work ensured considerable progress in aspect extraction and sentiment classification. This pie chart illustrates the distribution of sentiment categories-positive, neutral, and negative-in a collection of text files. It visualizes proportions, with each segment representing the percentage of total sentiment tokens classified under each category. This figure highlights a critical imbalance in sentiment data, with all samples classified as positive. While visually clean, the pie chart underscores the urgent need for better data balance for effective sentiment analysis, particularly in domains like Aspect-Based Sentiment Analysis (ABSA) or code-mixed language processing.

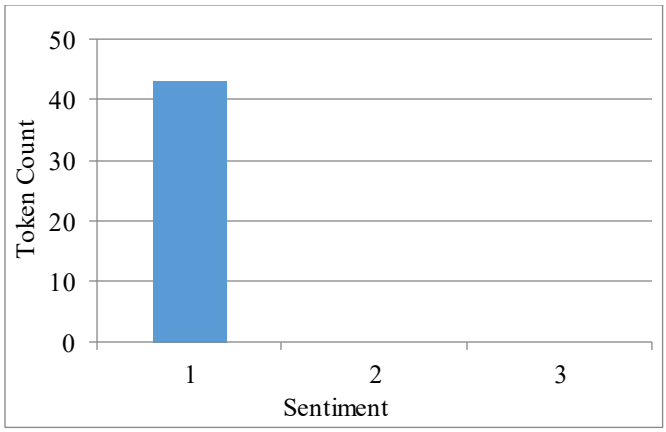

Fig. 8 Sentiment Tokens in Text Files

This bar chart visualizes the distribution of sentiment-labelled tokens in a collection of text files. The x-axis represents the sentiment categories-positive, neutral, and negative, while the y-axis shows the count of tokens associated with each sentiment.

References

[1] Neeraj Anand Sharma, A.B.M. Shawkat Ali, and Muhammad Ashad Kabir, "A Review of Sentiment Analysis: Tasks, Applications, and Deep Learning Techniques," *International Journal of Data Science and Analytics*, vol. 19, no. 3, pp. 351-388, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Roshan Poudel, "*Exploring Transformers for Aspect Based Sentiment Analysis*," MS Thesis, Universidade de Aveiro (Portugal), pp. 1-17, 2022. [Google Scholar]

[3] Deena Nath, and Sanjay K. Dwivedi, "Aspect-Based Sentiment Analysis: Approaches, Applications, Challenges and Trends," *Knowledge and Information Systems*, vol. 66, no. 12, pp. 7261-7303, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Gazi Imtiyaz Ahmad, and Jimmy Singla, "Sentiment Analysis of Code-Mixed Social Media Text (SA-CMSMT) in Indian-Languages," *2021 International Conference on Computing Sciences (ICCS)*, Phagwara, India, pp. 25-33, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ruhina Tabasshum Prome, Tarikul Islam Tamiti, and Anomadarshi Barua, "Leveraging the Potential of Prompt Engineering for Hate Speech Detection in Low-Resource Languages," *arXiv Preprint*, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Asara Senaratne, "*Anomaly Detection in Graphs for Knowledge Discovery and Data Quality Enhancement*," Thesis (PhD), The Australian National University, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Meerabahen M. Shah, and Hiren R. Kavathiya, "*Development of A Model to Analyze & Interpret Vernacular Voice Recognition Of Gujarati Dialects*," Ph.D. Thesis Computer Applications, Department of Computer Science, Faculty of Science, Atmiya University, 2024. [Google Scholar] [Publisher Link]

[8] Rasmus Kær Jørgensen, "*Multilingual Natural Language Processing for Applications in the Financial Domain*," Ph.D. Thesis, University of Copenhagen, pp. 1-163, 2023. [Publisher Link]

[9] Ranit Kumar Dey, and Asit Kumar Das, "Modified Binary Particle Swarm Optimization Based Deep Hybrid Framework for Sentiment Analysis," *International Journal of Information Technology & Decision Making*, pp. 1-28, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Jinshi Wang, "*Cross-lingual Transfer Learning for Low-Resource Natural Language Processing Tasks*," Master Thesis, Karlsruhe Institute of Technology, 2021. [Google Scholar]

[11] C.S Anoop, and A.G. Ramakrishnan, "CTC-Based End-to-End ASR for the Low Resource Sanskrit Language with Spectrogram Augmentation," *2021 National Conference on Communications (NCC)*, Kanpur, India, vol. 1, pp. 1-6, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Pablo Duboue, *The Art of Feature Engineering: Essentials for Machine Learning*, Cambridge University Press, pp. 1-284, 2020. [CrossRef] [Google Scholakr] [Publisher Link]

[13] Ni Made Satvika Iswari, and Nunik Afriliana, "Enhancing Aspect-Based Sentiment Analysis in Tourism Reviews Through Hybrid Data Augmentation," *Journal of Applied Data Sciences*, vol. 6, no. 3, pp. 2192-2206, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Sunil D. Kale et al., "A Comprehensive Review of Sentiment Analysis on Indian Regional Languages: Techniques, Challenges, and Trends," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9s, pp. 93-110, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Bing Liu, *Introduction: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, pp. 1-15, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[16] Jatinderkumar R. Saini, and Saikat Roy, "Preparation of Rich Lists of Research Gaps in the Specific Sentiment Analysis Tasks of Code-Mixed Indian Languages," *SN Computer Science*, vol. 5, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yusuf Aliyu et al., "Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources," *IEEE Access*, vol. 12, pp. 66883-66909, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Manuel Romeo Flores III, "*Assessing the Functional Significance of Wood Nutrient Resorption Along a Soil Fertility Gradient*," University of Illinois at Urbana-Champaign, pp. 1-48, 2022. [Google Scholar] [Publisher Link]

[19] Zeena Al-Tekreeti et al., "AI-Based Visual Early Warning System," *Informatics*, vol. 11, no. 3, pp. 1-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Prakhar Tandon et al., "Evaluating SEG for Gujarati News Clustering," *International Conference on Hybrid Intelligent Systems*, pp. 324-335, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Tianyang Zhong et al., "Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research," *arXiv Preprints*, pp. 1-41, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Shumaila Mughal, Arfan Jaffar, and M. Waleed Arif, "Sentiment Analysis of Social Media Data: Understanding Public Perception," *Journal of Computing & Biomedical Informatics*, vol. 7, no. 02, pp. 1-12, 2024. [Google Scholar] [Publisher Link]

[23] Ringki Das, and Thoudam Doren Singh, "Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1-38, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] James T. Frith, Matthew J. Lacey, and Ulderico Ulissi, "A Non-Academic Perspective on the Future of Lithium-Based Batteries," *Nature Communications*, vol. 14, no. 1, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Monali Bordoloi, and Saroj Kumar Biswas, "Sentiment Analysis: A Survey on Design Framework, Applications and Future Scopes," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12505-12560, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Yan Cathy Hua et al., "A Systematic Review of Aspect-Based Sentiment Analysis: Domains, Methods, and Trends," *Artificial Intelligence Review*, vol. 57, no. 11, pp. 1-51, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Vasileios Ballas et al., "Automating Mobile App Review User Feedback with Aspect-Based Sentiment Analysis," *International Conference on Human-Computer Interaction*, pp. 179-193, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Nozi Sidali, "*An Assessment of the Business Innovation Model Strategy used to Enhance Information Technology by ABSA*," University of the Witswatersrand. [Google Scholar]

[29] Clark Aldrich, *Simulations and the Future of Learning: An Innovative (and Perhaps Revolutionary) Approach to E-Learning*, John Wiley & Sons, pp. 1-304, 2003. [Google Scholar]

[30] Suman Dowlagar, "*A Code Mixed Dialog System in Medical Domain*," International Institute of Information Technology Hyderabad, pp. 1-133, 2023. [Google Scholar] [Publisher Link]

[31] Myagmarsuren Orosoo et al., "Analysing Code-Mixed Text in Programming Instruction Through Machine Learning for Feature Extraction," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, pp. 890-900, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[32] Deanna Kuhn, "Thinking as Argument," *Harvard Educational Review*, vol. 62, no. 2, pp. 155-179, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[33] Monil Gokani, "*Exploring Sentiment Analysis in Low-resource Languages*," Master of Science, International Institute of Information Technology, Hyderabad, 2023. [Google Scholar]

[34] Rishabh Agrawal, "Adaptive Few-Shot Learning (AFSL): Tackling Data Scarcity with Stability, Robustness, and Versatility," *arXiv Preprint*, pp. 1-8, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[35] Brijeshkumar Y. Panchal, and Apurva Shah, "NLP Research: A Historical Survey and Current Trends in Global, Indic, and Gujarati Languages," *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, Gobichettipalayam, India, pp. 1263-1272, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Albert Greenberg et al., "Towards a Next Generation Data Center Architecture: Scalability and Commoditization," *PRESTO '08: Proceedings of the ACM Workshop on Programmable Routers for Extensible Services of Tomorrow*, pp. 57-62, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[37] Stanislaw P. Stawicki et al., "The 2019-2020 Novel Coronavirus (Severe Acute Respiratory Syndrome Coronavirus 2) Pandemic A Joint American College of Academic International Medicine-World Academic Council of Emergency Medicine Multidisciplinary COVID-19 Working Group Consensus Paper," *Journal of Global Infectious Diseases*, vol. 12, no. 2, pp. 47-93, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Josina W. Geringer et al., "*Analysis of the ASME Code Rules for Subsection III-5-HHB (Composite Materials) for Current HTR Design Requirements*," Technical Report, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), pp. 1-37, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Jun Li et al., "A Review of Remote Sensing for Environmental Monitoring in China," *Remote Sensing*, vol. 12, no. 7, pp. 1-25, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[40] Raju Anitha, and K.S. Anil Kumar, "Sentiment Analysis in Low-Resource Language: Exploring BERT, mBERT, XLM-R, and RNN Architectures to Underpin Deep Language Understanding," *Journal of Nonlinear Analysis and Optimization*, vol. 15, no. 2, pp. 180-189, 2024. [Google Scholar] [Publisher Link]

[41] Jing Zhang et al., "Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1-38, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[42] Roland T. Rust, "The Future of Marketing," *International Journal of Research in Marketing*, vol. 37, no. 1, pp. 15-26, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[43] Mohit Dua et al., "A Review on Gujarati Language based Automatic Speech Recognition (ASR) Systems," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 133-156, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[44] Jukka Räty, Kai-Erik Peiponen, and Toshimitsu Asakura, *UV-Visible Reflection Spectroscopy of Liquids*, 1st ed., Springer Science & Business Media, vol. 92, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[45] Damini Dey et al., "Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review," *Journal of the American College of Cardiology*, vol. 73, no. 11, pp. 1317-1335, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[46] Sanele Gerald Khakhu, "*Liberating Technologies: Automatic Speech Recognition - AI Virtual Assistant and the Future of Language in South Africa*," Master of Arts (MA), University of Johannesburg, pp. 1-82, 2023. [Google Scholar] [Publisher Link]

[47] Soni Sweta, "Application of Sentiment Analysis in Diverse Domains," *Sentiment Analysis and its Application in Educational Data Mining*, pp. 19-46, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[48] Payal Joshi, and Dhaval Joshi, "Code Mixed Information Retrieval for Gujarati Script News Articles," *International Conference on Advances in Computing and Data Sciences*, Kolkata, India, pp. 265-276, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[49] Qiang Wan, Fanming Wang, and Sanhong Deng, "Optimizing Social Media Public Opinion Analysis with ABSA: A Case Study on Weibo," *International Conference on Information Management*, Kolkata, India, pp. 332-343, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[50] Pavani Rangachari, *A Holistic Framework of Strategies and Best Practices for Telehealth Service Design and Implementation*, Service Design Practices for Healthcare Innovation, Springer, Cham, pp. 315-335, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[51] Surinder Pal Singh, and Neeraj Mangla, "A Hybrid Framework Combining Dictionary-Based Methods, BERT, and CRF for Language Identification and Normalization in Code-Mixed Hinglish," *2025 Seventh International Conference on Computational Intelligence and Communication Technologies (CCICT)*, Sonepat, India, pp. 646-651, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[52] Chandrakant Patel, and Jayeshkumar Patel, "Dynamic Stop List for the Gujarati Language using Rule Based Approach," *Towards Excellence*, vol. 13, no. 1, pp. 594-607, 2021. [Google Scholar] [Publisher Link]

[53] Shraddha Kalbhor, and Dinesh Goyal, "Survey on ABSA based on Machine Learning, Deep Learning and Transfer Learning Approach," *Recent Advances in Sciences, Engineering, Information Technology & Management*, vol. 2782, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[54] Farzana Kulsoom et al., "A Review of Machine Learning-Based Human Activity Recognition for Diverse Applications," *Neural Computing and Applications*, vol. 34, no. 21, pp. 18289-18324, 2022. [CrossRef] [Google Scholar] [Publisher Link]