Research Article

Aibi revista de investigación, administración e ingeniería. Volumen 13, Número 1, Pág 234-249, Enero - Abril de 2025, ISSN 2346-030X

A survey on Gujarati NLP research work.

Brijeshkumar Y. Panchal[1], Apurva Shah[2]
[1,2]*The Maharaja Sayajirao University of Baroda, Gujarat - India*
[1]*Gujarat Technological University, Gujarat - India*

Received: august 22, 2024.
Accepted: december 05, 2024.
Publicado: january 01, 2025.

***Abstract*—** The intriguing field of Natural Language Processing (NLP) is a captivating Artificial Intelligence (AI) segment that explores text, speech, and translation intricacies. The number of internet users communicating in regional languages is soaring each day. Consequently, NLP enthusiasts have turned their attention to local dialects; among these, Gujarati stands out. Gujarati is an Indo-Aryan tongue native to the vibrant Indian state of Gujarat and is passionately spoken by its people. Approximately 62 million individuals around the globe articulate in Gujarati, positioning it as the 26th most widely spoken language in the world. Yet, Gujarati is considered the youngest and least-resourced Indian language within the NLP landscape. However, few groundbreaking advancements have emerged in the Gujarati NLP (GNLP) field. E.g., WordNet, Morphological, Stemmer, optical character recognition (OCR), Speech Recognition, Parts of Speech, Machine Translation, etc. Many researchers have been working with a rule-based approach for GNLP. After that, only a few researchers have attempted the machine learning, deep learning, and reinforcement learning approaches. This paper focuses on a critical survey of existing GNLP research, covering research papers from 1999 to August 2024 in this study. This survey predicts GNLP study until 2030 with the use of available data. This study explores gaps in present studies and offers suggestions for the newly active research field of GNLP. Through this paper, one can develop a deep Learning-based GNLP system to achieve greater accuracy.

Keywords: natural language processing (NLP), Gujarati language, artificial intelligence (AI), Gujarati, machine learning (ML), deep learning (DL).

*Corresponding author.
Email: panchalbrijesh02@gmail.com (Brijeshkumar Y. Panchal).
Peer reviewing is a responsability of the Universidad de Santander.

How to cite this article: B. Y. Panchal and A. Shah, "A survey on Gujarati NLP research work", *Aibi research, management and engineering journal*, vol. 13, no. 1, pp. 234-249 2025, doi: 10.15649/2346030X.4445

I. INTRODUCTION

Natural language is a dialect that has developed naturally in people through use over time. Individuals interact with one another using natural language to share knowledge, thoughts, and sentiments. Different native languages, respectively, with their alphabet, signs, and grammar, are spoken in diverse areas across the globe. India is the country wherein the oldest and most diverse local languages in terms of morphology could be discovered. Compared to other natural languages, the information conveyed in English by the commonly used code for ASCII is relatively simple for machines to handle. However, developing a machine's capacity to comprehend other languages naturally is complex and is done in various ways.

a. History of the Gujarati Language

The melodic Indo-Aryan tongue of Gujarati harmonizes beautifully with the vast Indo-Iranian lineage of the grand Indo-European language family. The Indian constitution formally recognizes Gujarati among 23 official languages spoken by over 46 million individuals. While substantial diaspora communities flourish worldwide, especially in the United Kingdom and the United States of America, many of these individuals call the Indian state of Gujarat their home [1].

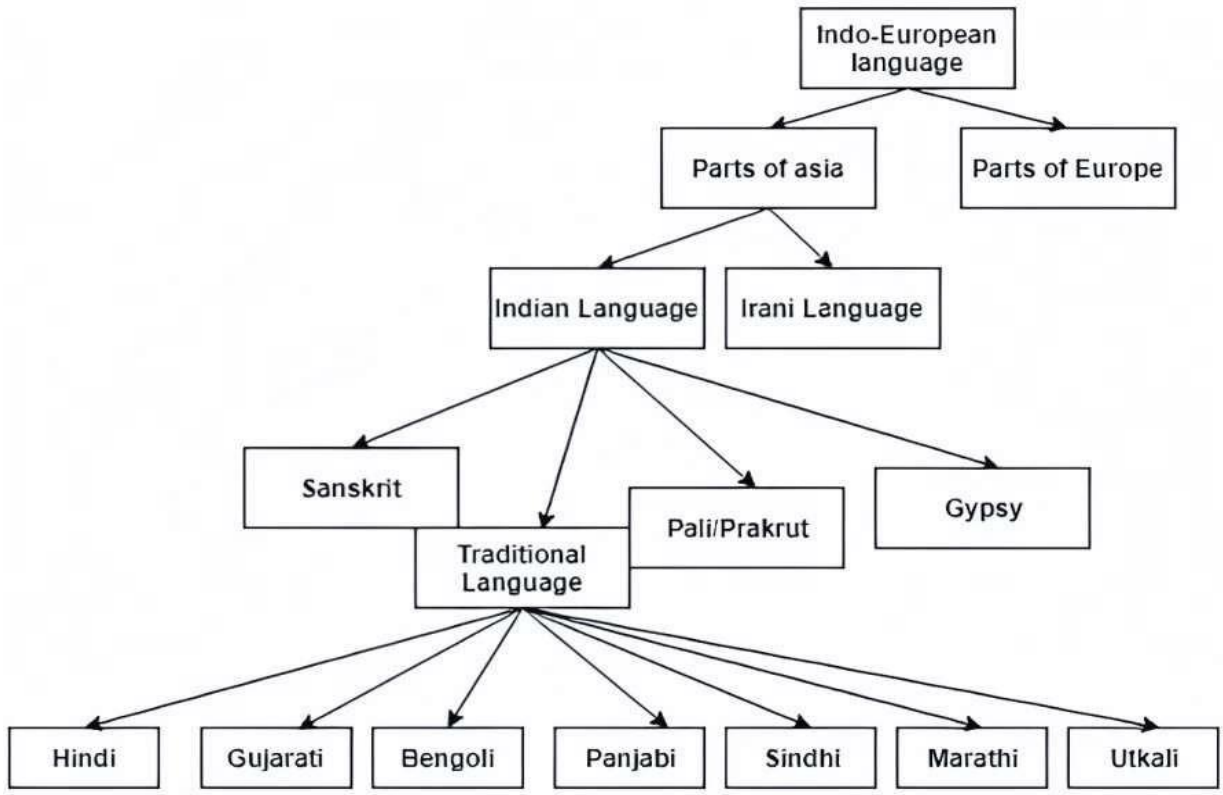

Figure 1: Origin of Gujarati Language [2].
Source: Own elaboration.

The origin of the Gujarati language dates back to the 12th century. Sanskrit was taught to the Parsis then, and many of their sacred books were converted from their Middle Persian originals. When Islam began to dominate Gujarat, Arabic and Persian were learned/taught. Parsis quickly mastered Persian, but interest in Sanskrit academics decreased. Gujarati translations of Avestan and Pahlavi writings were made using the existing Sanskrit versions. Persian and Urdu, which took over as the court tongue, significantly impacted Gujarati at this time. The Parsis spoke Gujarati in the villages nearby Surat, which was freely borrowed from Persian, Pahlavi, and Zand. Gujarati underwent westernization, and literature incorporated British Romantic traditions and aesthetics. Some people attempted to utilize Sanskrit but were seen as pedants [3].

Table 1: Gujarati consonants, vowels, and numbers. Tense, Vachans, Sentence Structure [1].

Components	No.	Details
Consonants	34	ક, ખ, ગ, ઘ, ચ, છ, જ, ઝ,ટ, ઠ, ડ, ઢ, ણ, ત, થ, દ, ધ, ન, પ, ફ, બ, ભ, મ, ય, ર, લ, વ, શ, ષ, સ, હ, ળ, ક્ષ, જ્ઞ
Vowels	12	અ, આ, ઇ, ઈ, ઉ, ઊ, ઋ, ઌ, ૠ, ૡ, એ, અઃ
Digit	10	૦, ૧, ૨, ૩, ૪, ૫, ૬, ૭, ૮, ૯
Tenses	3	Present, Past, Future
Vachans	2	Plural, Singular
Sentence Structure	3	Object, Verb, Subject

Source: Own elaboration.

The Devanagari script was modified to write the Gujarati language. The Gujarati script first appeared in print in a commercial in 1797, and the initial piece of writing was in the form of a document in 1592. Its primary use till the 19th century was maintaining records and writing letters, while literary and scholarly works were written in the Devanagari script [4].

b. Background of INDIC NLP

The time of the English language's dominance over India may be nearing an end, claims a study conducted by Google in 2017 in association/collaboration with KPMG (Klynveld Peat Marwick Goerdeler). Based on the study entitled "Indian Languages: Defining India's Internet," there are more than 234 million online handlers of Indian languages in India, compared to 175 million users of English. Between 2011 and 2016, 234 million more internet users spoke an Indian language, growing at a CAGR of 41%. As a result of this remarkable development, Indian language Internet users have exceeded their English-speaking peers. Of 100% of Indic Languages internet users, 6% belong to Gujarati language speakers. It is the 4th most extensively spoken language in India. There are 62 million Gujarati speakers globally. It is the 26th most broadly spoken language internationally. Numerous studies have been carried out to brand it informal so that individuals can collaborate with and intermingle with computers in their local natural languages. Kannada, Hindi, Tamil, Bengali, Marathi, Telugu, Punjabi, Malayalam, and Gujarati are just a few of the Indian Regional Languages (IRL) that Google offers transliteration for in addition to its 13 other language search options [5, 6].

c. Application of GNLP

The number of NLP applications attempted in GNLP, like WordNet, Morphological, Stemmer, optical character recognition (OCR), Speech Recognition, Part of Speech, and Machine Translation. Let us see a few short discerptions of each one.

1. WordNet

WordNet is a lexical library that ties words into semantic relationships between them, such as synonyms, hyponyms, and meronyms. The synonyms are compiled into synsets that include examples and succinct explanations. As a result, it may be supposed to be a synthesis and expansion of a dictionary and thesaurus. Although humans may access it using a web browser, its main uses are in automated text analysis and artificial intelligence. The English WordNet database and software tools were initially developed in the English language, and they are freely downloadable from that WordNet website and published under a BSD-style license. WordNets are now available in more than 200 languages. Dharmsinh Desai University, Nadiad, developed Gujarati languages WordNet, Total 24,896 Nouns, 5828 Adjectives, 2805 Verbs, and 445 Adverbs [7].

2. Morphological

The analysis of words, how they are constructed, and how they relate to additional words in similar language is recognized as morphology in linguistics. It inspects the arrangement of words and their basic stems, roots, prefixes, and suffixes. Morphology inspects how the setting moves word diction and sense and the parts of speech, intonation, and stress [8].

3. Stemmer

By peeling back the layers of word modulation to unveil their foundational essence, stemming emerges as a dazzling NLP technique that navigates through the vast ocean of text, words, and documents for text normalization. Inflection is a transformative process where a word morphs to express a kaleidoscope of grammatical categories, including tense, voice, case, person, aspect, gender, number, and mood. Thus, even though a single word may possess many altered versions, the NLP procedure takes on an added layer of intricacy when multiple inflected forms coexist within the same textual tapestry. To distill words down to their elemental essence, this may or may not resonate as a practical term within the language [9].

4. Optical Character Recognition (OCR)

Character and Digit recognition are additional names for OCR. Data is removed and recycled from documents using an OCR application. The unique material may be retrieved and corrected by OCR, which separates letters on the picture, turns them into words, and then goes the words into sentences. Also, it remains absent due to the necessity for human data input [10].

5. Speech Recognition

A program’s dimensions to change the vocal language into printed language are speech recognition, automated speech recognition (ASR), and computer voice recognition. Speech recognition emphasizes changing speech from a verbal to a written arrangement, whereas voice recognition aims to differentiate the voice of a convinced individual [11].

6. Part of Speech (POS)

A segment of language, once dubbed a word category or grammatical faction, represents a collection of terms with similar grammatical attributes. It is also known as POS or PoS in the realm of grammar. When words are categorized under the same language segment, they frequently display analogous morphological traits, where they undergo variations for related characteristics, akin syntactic patterns, and even

comparable semantic functions. Nouns, adjectives, verbs, adverbs, prepositions, pronouns, conjunctions, numerals, interjections, articles, and determiners are some of the language components that are commonly acknowledged in English [12].

7. Machine Translation

Machine translation, often abbreviated as MT, is a fascinating branch of computational linguistics that harnesses the power of computer algorithms to transform text or spoken words from one language into another [13].

II. RELATED WORK

Now, researchers present a few related research works of GNLP. From these many papers, one can predict how much work has been done till now in GNLP and what kind of research work will be improved and discovered.

Antani and Agnihotri [14] proposed a technique for a group of Gujarati characters with similar appearances, which was selected and classified using several classifiers. The imageries for the fonts were gathered from digital images accessible on cyberspace. Innovative scholars were classified employing the k-Nearest Neighbor and Euclidean Minimum Distance classifiers, utilizing moments that remained constant and unchanging. In addition, the Hamming Distance classifier also played a role in categorizing the characters within the realm of binary feature space. The report gives the recognition rates for various classifiers. An achievement rate was made. Selecting an appropriate array of traits for classification emerged as the most challenging aspect of the experiment. This highlights the crucial need for creating an expansive ground-truthed array of traits across various resolutions, enabling further exploration for identifying languages originating from the Indian subcontinent. The traits in the trial were also encapsulated within a defined bounding rectangle of fixed dimensions. Emphasis is placed on the reality that basic shape-based recognition may not consistently yield favorable results, especially when the system encounters ligatures, although this might not always hold true. Various font families show The same character differently, and the relationship between related characters differs from font to font.

Patel and colleagues [15] ingeniously crafted a hybrid lightweight stemmer tailored for the Gujarati language. To elevate the caliber of the stems and suffixes acquired during the training phase, the scholars harnessed linguistic proficiency by incorporating a meticulously curated Gujarati suffix list instead of relying solely on an unsupervised technique. To assess and validate the stemmer's performance, the researchers utilized the EMILLE corpus alongside Goldsmith's method. This way, the accuracy of the stemmer was increased by roughly 17% because of the inclusion of hand-crafted suffixes, helping them reach an accuracy of 67.86%. The authors also discovered that the system's performance is diminished by setting the minimum stem size and giving the suffix and stem different weights. In rolling out this, the stemmer was compact and only eliminated inflectional ends since it was designed to be used with an IR system. Derivational suffixes for full-fledged stemming, which may be needed in applications like presenting words in a user interface, may be added to the list of hand-crafted suffixes.

Desai [16] envisioned an innovative OCR for deciphering handwritten Gujarati numerals. This initiative introduces a sophisticated neural network dedicated to recognizing Gujarati handwritten digits. A multi-tiered feed-forward neural network has been suggested for precisely categorizing numbers. Four distinct profiles of digits encapsulate the essence of Gujarati numerals. Handwritten numerals are pre-processed by skew correction and thinning, which has resulted in an 82% achievement rate for Gujarati handwritten number identification. Many techniques are used in the pre-processing stage before implementing classification. Although the network's overall performance is 81.66%, it still falls short of its goal. The act of any classification model depends mainly on its feature abstraction. This prototype may need to be improved with the features abstraction or pre-processing techniques to improve its performance. This model has a success rate overall, but there are still areas for improvement.

Bhensdadia et al. [17] unveiled the latest addition to the vibrant tapestry of IndoWordnet: Gujarati. The Gujarati Wordnet is being crafted from the rich linguistic foundation of the Hindi Wordnet as a centerpiece of the IndoWordnet initiative, employing an extension technique. This research delves into the evolution of the Gujarati Wordnet. It explores the core traits of Gujarati and evaluates the suitability of Hindi as its foundational language.

Furthermore, the current landscape of the work and any burgeoning challenges are brought to the forefront. The development of the Gujarati Wordnet benefited significantly from the existence of the Hindi Wordnet and the linguistic affinities shared between Hindi and Gujarati. Moreover, the Bhagavad-Go-Mandal and the Gujarati Lexicon emerged as invaluable assets in creating synsets. The proliferation of wordnets across various Indian languages through an expansion technique promises a treasure trove of lexical resources indispensable for applications in machine translation and natural language processing [18].

Desai [19] made an effort /tried to extract words from lines of handwritten Gujarati language. Word extraction is a crucial step in the optical character recognition (OCR) procedure since most Indian scripts are quite cursive. It is challenging to separate characters and extract modifiers because of the cursive style of the writing. Word mining is one of the crucial stages of OCR, and it directly impacts how accurate OCR is. A mix of tried-and-true techniques, including projection profiles and morphological operations, is applied to improve the word extraction's accuracy. The procedure is discovered to be both clear and precise enough. Special characters like full stops, commas, and hyphens, among others, are to blame for most errors. This cannot be avoided since researchers are thinking about handwritten text forms, and the size of such symbols is unknown. Therefore, the existing technology cannot provide a universal fix. The best thing to do is to save these symbols as words, which can be recognized later according to their category and associated with the appropriate term. The accuracy rate for cases with the same writer may be taught into the system since it is greater than for cases with different authors. Five hundred distinct text lines have been used to test the algorithm. The lines are divided into two groups: those written by individuals of various ages and genders and those where the same author may have written more than one line. When comparing text lines written by the same author to those written by various authors, it is shown that the correctness rate for word extraction is more significant for the former.

Patel and Apurva [20] introduced an innovative OCR system. Nonetheless, numerous challenges remain unresolved for the research community in the realm of NLP. The study proposes three groundbreaking elements to represent handwritten Gujarati characters. These features encompass structures derived from zone pattern matching, structural decomposition, and normalized cross-correlation. The methods of Support Vector Machine (SVM) and Naive Bayes (NB) classifiers have been examined to categorize Gujarati characters defined by the proposed features. A compilation of 20,500 manually crafted Gujarati characters has formed the basis of these experiments. When classifiers were trained using structural decomposition-based features, the results revealed a significant advancement over existing benchmarks. Handwriting presents a challenge in 79.84 and 62.69% of instances, respectively. The mean accuracy achieved by employing features grounded in structural decomposition is an impressive 98.77%. Additionally, feature vectors have been constructed utilizing zone pattern matching, normalized cross-correlation, and structural decomposition-based features.

Kartik et al. [21] unveiled two remarkable stemmers tailored for Gujarati light inflection, one crafted through a hybrid model and the other a more intricate derivational stemmer stemming from a rule-based framework. Beyond a module designed to promote the unsupervised gathering of stems and suffixes for light stemming, the researchers incorporated a POS stemming module utilizing POS and a module grounded in substitution rules to elevate both their quality and suffixes. Integrating these modules enhanced the inflectional stemmer's precision by 9.6 percent and 12.7 percent, respectively, enabling us to achieve an impressive accuracy of 90.7 percent. The pinnacle of index compression attainable through the inflectional stemmer reached approximately 95 percent. Conversely, the derivational stemmer operates entirely on rules, which has led the researchers to attain a precision of 70.7 percent by employing suffix-stripping and substitution rules. Both methodologies were devised to support applications such as Information Retrieval, corpus compression, and dictionary search, serving as pre-processing units for various NLP challenges.

Juhi et al. [22] proposed the implementation of a Gujarati rule-based stemmer. The researchers have shown the development of stemming rules and Gujarati's rich morphology. Authors have shown the conception and application of a Gujarati rule-based stemmer. The stemmer may retrieve the majority of the morphological variations. Researchers tested their systems to validate our claim, and the results showed a 91.5% accuracy rate. To give a thorough error analysis report, the researchers would want to carry out a more thorough error analysis as an extension of this effort. Additionally, because of the greater likelihood of over- stemming faults with our technique, the researchers want to eliminate them in the future, further increasing accuracy.

Sheth and Patel [23] developed an approach for Gujarati Stemmer that emphasizes old-style techniques and joins non-traditional methods to produce an efficient and precise stemmer. By utilizing a range of new and relevant developments, better Gujarati stemmers could be created and, consequently, be useful in other computer science in linguistics. The stemmers defined for Gujarati have not examined over-stemming or under-stemming. Additionally, the inclusion of Named Entity Recognition is missing. Parts that make up the Stemmer GUJ Stem-suffix frequency tables (GSFT) record the stem-suffix pairing frequency based on the learning conducted using a Gujarati Corpus. This can be useful in determining the different inflections that can be made using the stem. In Gujarati, the Gujarati Language Rule Set (GLRS) contains the rules for stripping suffixes.

The rules are drawn up by studying the distinctive features of the Gujarati language. For instance, words like કર (to do) are a suffix that includes. The same suffixes can be used for the words like પડ (to be fall). Therefore, one can devise an algorithm to strip these suffixes and numerous others of the same type. Rules of GLRS are applied by the modules that strip suffixes to determine the beginning place of the suffix within the word in question and split the stem from the suffix. It also updates how often the stem-suffix pairing is in the GUJ Stem-suffix frequency tables (GSFT). Researchers propose a function called GUJ_STRIP based on an artificial neural network, smoothing, and probability. If the rules set are ineffective in aiding or recognizing stems, the program will assist in determining the stem. This function uses the GUJ Stem-suffix frequency tables to identify the probabilities of the suffixes that are known to date. In addition, it uses flattening methods to control the likelihood of a suffix that was not seen until this point. Entity recognition methods are incorporated into the model to determine words that are people's names or locations. If using this function, an entirely new suffix is discovered, and the stem-suffix pairing is delivered to the Gujarati expert in linguistics for validation. Once it is approved, it is included in the GUJ Stem-Suffix frequency table [23].

Patel and Desai [24] proposed identifying a word's zone borders and using zone boundary information. It is possible to identify a word's constituent parts using connected component labeling, which may then be further broken down if necessary to find more parts. It is the first effort to break down handwritten Gujarati words into their constituent parts. It has been discovered that characters with several constituent parts and characters in any combination are the main culprits in excess segmentation, which leads to inaccurate word segmentation. For identification, it is thus required to unite the excessively divided characters at the letter level into a single character. In literature, a method based on neural networks is used to distinguish original words from the text line of handwritten English text and several languages of Indian origins, such as Assamese. The trained ANN is then fed each of the segments that are thusly produced. A segmentation border for the character is believed to exist at the point of segmentation when the ANN detects a segment or a collection of segments to be similar to a handwritten character, and segmentation is then carried out. The segmentation border is then validated by comparing the segmented character to the best match. A similar strategy may be devised for Gujarati words with such excessively split characters.

Sheth et al. [9] introduce DHIYA, a remarkable stemmer inspired by the nuances of the Gujarati language. This innovative stemmer is intricately built upon the rich morphology of Gujarati. Common inflections in Gujarati literature were meticulously identified to enrich the stemmer's capabilities. This foundational work led to the formulation of a comprehensive set of rules. The EMILLE corpus is vital for training and evaluating the stemmer’s performance. The accuracy of this stemmer stands impressively at 92.41%. By expanding the rule set, the stemmer can be fine-tuned to reduce instances of over- and under-stemming. Moreover, a hybrid strategy, which combines rule-based techniques with statistical methods, can further elevate the effectiveness of stemming.

Patel and Goswami [25] proposed an innovative technique for rectifying words in Gujarati documents through a probabilistic lens. To forge a rich and varied corpus of Gujarati documents, collections are amassed from the vast expanse of the internet, particularly from Gujarati newspapers, utilizing the prowess of web crawlers. The statistical analyses yield data derived from bigram, unigram, and trigram techniques, subsequently informing the construction of a probabilistic model. This ingenious system undergoes evaluation using a synthetic dataset that deliberately infuses random errors at the word level of authentic text. This research unveils a method for rectifying words in Gujarati documents by applying a probabilistic strategy. To craft a rich and varied Gujarati document corpus, documents are sourced from the digital realm,

particularly from Gujarati newspapers, with the assistance of web crawlers. The statistical analyses will generate data rooted in unigram, bigram, and trigram methodologies, which will be the foundation for constructing a probabilistic model. This model will be rigorously tested using synthetic data that introduces random word-level errors into genuine documents.

Patel et al. [26] have improved to a level suitable for international languages like English. In this research, researchers developed a model to translate a speaker's spoken Gujarati number into machine-editable numeral text. Mel Frequency Cepstral Coefficients (MFCC) served as the vibrant feature palette in the envisioned system, while K-Nearest Neighbor (K-NN) took on the role of the astute classifier. The performance of the proposed model in recognizing spoken numerals in Gujarati reached an impressive average success rate of 78.13%. The suggested model in this study receives the spoken Gujarati number through the speaker's microphone before being translated into word format using MATLAB. The proposed recognition method was tested using voice samples from datasets of various sizes. The total system presentation of the planned model is 78.13% using speech samples from a diverse range of speakers for the train and test datasets, which have a combined size of 300. The experimental outcome demonstrates that using more train speech samples improves the presentation of the planned system for all Gujarati numerals, i.e., the more train speech samples utilized, the higher the likelihood that Gujarati numerals would be correctly identified. For a homogenous group, the accuracy rate is 64.48% for a speaker in the 5–15 age range with a dataset size of 105 and 60.45% for a speaker in the 16–40 age range with a dataset size of 155. Therefore, researchers can state that the system's effectiveness increases with more speech patterns of younger speakers. The majority of Gujarati numerals had the best right rates. However, the poorest correct rates were found for the numerals four and seven for both heterogeneous and homogeneous groups. Only isolated Gujarati numerals may be used with the suggested technique. In the future, researchers may adapt this algorithm to recognize spoken Gujarati words and uninterrupted and solitary numerals.

Varghese [27] proposed a Text-to-Speech (TTS) synthesizer, which is a tool that turns a given text into the sound that corresponds to it. Speech synthesis, also known as text-to-speech creation, is discussed in this research. The synthetic voice is produced artificially by generating human speech from Gujarati text. The PRAAT software records the necessary database for the character-to-sound conversion, which is then stored in a directory in a.wav format. The required words are produced by concatenating the shortest part of the recorded voice. The primary goal of using this approach is its simplicity. The concatenation method and MATLAB software are used. This study covers the planning and execution of a Gujarati text-to-speech system for concatenation-based TTS. This article only includes C, CV, V, and VC patterns, C is the consonant and V is the vowel. Since comparing to other ways, this one is really simple and effective to use. This holds true for several languages that have a higher naturalist population. Work on Word documents will be possible in the future, and with an advanced version, work on PDFs, scanned data, etc., with a few tweaks.

Jatayu Baxi [28] introduced The Morphological Analyzer as an ingenious instrument that delves into the syntactic framework of a word and uncovers the foundational essence of an inflected word form provided as input. Many NLP applications use morph analyzer as a pre-processing tool. There has been significant research in this field for several Indian languages, but nothing has been reported for Gujarati. The researchers provide a morph analyzer for Gujarati. Statistical, knowledge-based, and paradigm- based approaches are used in the development of the Morph analyzer. Researchers provide a thorough analysis of various strategies. With the knowledge-based hybrid approach and the statistical hybrid method, researchers show a considerable increase in general correctness and attain 92.34% and 82.84% correctness, respectively. The existing system's flaw is that it cannot handle derivational morphology.

Desai [29] embarked on a quest to unravel the complexities of OCR for the intricately handwritten Gujarati alphabets in this endeavor. A collection of forty handwritten characters was meticulously amassed from approximately 189 authors for this undertaking. Quality, alphabet length, and image segmentation techniques were utilized as feature dimensions, while the support vector machine (SVM), tasked with classification, achieved a remarkable accuracy of 86.66%. Alongside SVM, the k-nearest neighbors (kNN) algorithm was also applied for classification, with the results being meticulously compared. A support vector machine is also described in the study. Gujarati alphabet identification is difficult due to the large quantity and variety of alphabets. An algorithm for recognizing the handwritten Gujarati script is presented in this paper. This study demonstrates that a hybrid feature set is more useful for Gujarati handwritten alphabet recognition than a straightforward structural feature set. Additionally, compared to other classifiers like KNN and SVM with Gaussian kernel, SVM with the polynomial kernel (c = 2) provides the greatest accuracy for identifying Gujarati handwritten alphabets. Even though this work's recognition accuracy was 86.66%, further improvement is still required.

Kapadia and Desai [30] proposed morphological analysis is the foundation of language study. The authors explain several techniques, including rule-based and machine learning, for morphological analysis. It is necessary to authenticate words against morphological rules and a lexicon before doing an analysis using any method. This research unveils the intricate morphological guidelines for the classes of the Gujarati language alongside a comprehensive lexicon database. A Gujarati vocabulary with more than 15000 words has been constructed using the UNICODE coding scheme. All words are divided into several grammatical classes according to their affixes, and inflection and derivation rules are described. In this remarkable endeavor, researchers have assembled an invaluable toolkit encompassing intricate morphological grammar rules, a treasure trove of test data, a comprehensive dictionary, and an array of APIs. These rules have been meticulously integrated into a database to facilitate advanced processing and to forge a morphological analyzer tailored for the Gujarati language.

Jariwala and Patel [31] envisioned a remarkable way to convert Gujarati text into Braille, seamlessly integrating English and Hindi text and storing the results in a data file ready for instant printing on an embosser. The methodology unfolds, focusing on the traditional Braille system, accompanied by a thoughtfully crafted transformation table. The results yielded by the system were impressive. The planned method would improve the quantity of Braille material that Gujarat's visually impaired population can access, as there is now a relatively small amount of it accessible. This essay also advances fresh ideas for helping Braille learners and those with visual impairments. The approach discussed in the article is an effort to provide low-cost technology to help persons with vision impairments. The process will translate text from Gujarati, Hindi, and English into Braille, which may be printed directly on a Braille printer. The suggested work has undergone extensive testing and produced positive outcomes.

Kapadia and Desai [32] unveiled a morphological analyzer grounded in intricate rules. A Lexical Dictionary brimming with Root Words has been meticulously crafted. Ingeniously devised regulations inspired by the expertise of linguists have been established. The analyzer tool takes a Gujarati sentence as its muse. It meticulously unfolds its grammatical classification, gender, number, and tense, alongside enlightening details about the person and the root words. It works with both derivational and inflectional morphemes. Researchers have measured the accuracy

of 87.48 % when evaluating texts from short stories and essays. In addition to the high accuracy level, a rule-based system has disadvantages. First, developing a comprehensive rule- based system for all natural languages is difficult. Furthermore, the system will not be able to produce any results when a suffix does not meet any rules. Finally, because rules heavily rely on the authors' language, it is not easy to brand them as self-governing of language.

Sheth et al. [33] Patel proposed that the main goal of creating a text summarizer is to have features that improve the choice of a particular piece of information from the wealth of Gujarati-language content that suits their needs. The Gujarati text summarizer provides a concise summary of the Gujarati text, allowing readers to quickly choose the information they need without having to read the whole text. Gujarati text summarizing is complicated since, when people summarize a text, researchers often read it completely to get a knowledge of its sentence structure, vocabulary, nouns, gender, and grammar before writing a review that emphasizes the essential ideas. Gujarati text summary is extremely challenging and time-consuming since computers lack human understanding and linguistic skills, particularly in Indian languages like Gujarati. This study offers a technique for summarizing Gujarati text. When linguistics is included in text categorization, its performance increases. Although producing summaries by humans is challenging, utilizing linguistic tools like the Stemmer and String similarity scale makes it possible to produce summaries with a high recall. Using machine learning techniques and other NLP technologies may improve the summary.

Naik and Desai [34] introduced an innovative real-time method for recognizing Gujarati handwritten characters. They employed a blend of sophisticated features to classify strokes, utilizing SVM with linear, polynomial, and RBF kernels, KNN with diverse k values, and a multi-layer perceptron (MLP). A dataset comprising 3000 samples was harnessed to train this advanced system, which 100 unique authors subsequently assessed. With the SVM-RBF kernel, the researchers achieved an impressive accuracy of 91.63%, alongside the lowest correctness rate of 86.72%. Utilizing the SVM linear kernel, the authors managed to attain a swift average processing time of 0.056 seconds per stroke, while the MLP recorded a longer processing time of 1.062 seconds per stroke. The researchers proposed a hybrid feature-centric methodology for recognizing Gujarati characters written by hand online. They scrutinized the performance of SVM, MLP, and k-NN classifiers under various parameter configurations. The suggested system's flaw is that it offers poor accuracy for characters with high similarity levels and confusion. For characters that are difficult to recognize, a two-layer classifier technique may be utilized to increase recognition rates.

Patel et al. [35] presented 'The Stemmatizer,' a novel integration of a stemmer and lemmatizer intended for the Gujarati language. It possesses the remarkable ability to identify and incorporate new vocabulary. A total of 2197 words were employed to evaluate the proposed solution, yielding impressive outcomes. Stemming serves as a pivotal element in the realm of information extraction (IR). The stem produced by the stemmer isn’t required to be a valid term found in a dictionary. In contrast, the lemma generated by a lemmatizer, as mandated by numerous IR systems, is consistently recognized as a legitimate dictionary term. Resources for Indian languages are scarce. Significantly, although the Gujarati language is equipped with a stemmer, it is conspicuously devoid of a lemmatizer.

Addressing the Gujarati-English news translation challenge in the WMT19 setting, Goyal et al. [36] contribute to the discourse, and this article sheds light on the Neural MT system at IIIT-Hyderabad. Constructed upon an encoder-decoder framework featuring an attention mechanism, the study examined Multilingual Neural MT models. This research illustrates that for low-resource language pairs such as Gujarati-English, leveraging Multilingual Neural Machine Translation with parallel data from related languages can enhance BLEU scores by as much as 11.5. The researchers advocate that NMT is a promising technique for machine translation in languages with limited resources. Nevertheless, to combat the challenge of data scarcity, the researchers emphasize the necessity for an array of innovative strategies. Considerable research has been done in transfer learning and multilingual machine translation to address this issue. Researchers will develop efficient one-to-many Multilingual NMT systems in the future [36].

Sen et al. [37] introduced constrained systems, implying that only the provided data for this language pairing is utilized. By employing a synthetic parallel corpus fashioned through the back-translation of monolingual data and the authentic parallel corpus, the researchers develop Transformer-based, subword-level neural machine translation (NMT) systems. The Gujarati-English and English-Gujarati pairs' leading systems achieve impressive BLEU scores of 10.4 and 8.1, respectively. The BLEU score for this language duo significantly increases when monolingual data from back translations is integrated, especially in contrast to the baseline NMT and SMT systems. The authors proudly presented their contributions to the collaborative Gujarati-English news translation initiative for the WMT 2019 conference in this study. This marks the inaugural instance of the Gujarati language participating in a WMT challenge. The researchers deliver transformer-based NMT systems tailored for both English and Gujarati. Given the scarcity of parallel sentences within the training dataset and many of these sentences consisting of only two to three tokens, the BLEU scores for English-Gujarati pairs relying solely on the parallel corpus remain notably low.

Patel and Desai [38] compare the performance of wavelet-based methods for extracting Gujarati text from images using the wavelet filters Haar, Symlets, Daubechies, Bior, and Coiflets. Using the discrete wavelet transform (DWT) technique, this system recognizes Gujarati text from photos. A set of rules and geometric attributes have been developed on the linked component to localize the actual text portions. Finally, the Gujarati text was localized using bounding box information. The study on many photos demonstrates that the Haar wavelet filter can accurately detect text sections. The researchers suggested a 2D DWT wavelet-based technique for localizing Gujarati text in this work. In this case, wavelet functions, including Haar, Daube, Symlets, Coiflets, and Biorthogonal, have been used by researchers to localize Gujarati text. With text characteristics collected using the Haar wavelet, this approach has the highest success rate for identifying Gujarati text.

Tailor and Patel [39] introduced an innovative framework for statistical unsupervised machine learning alongside a method for rule-based sentence tokenization tailored for Gujarati text. In the realm of continuous Gujarati text, punctuation marks such as the dot ('.'), exclamation ('!') and question mark ('?') serve as vital indicators of sentence termination; leveraging Punkt unsupervised learning followed by meticulously crafted rules proves highly effective in sentence tokenization. This hybrid approach shines particularly bright due to the lack of POS tags and other resources like a compilation of abbreviations or a verb list. Tested on 140 distinct articles, this strategy proved to be fairly reliable. Researchers think they may get even better outcomes by including other features specifically designed to address repeated punctuation and abbreviated words. Authors are attempting to construct a list of abbreviated terms.

Deepang Raval et al. [40] proposed an innovative approach for elevating the performance of the Gujarat language End-to-End voice recognition system. This methodology centers on deep learning, incorporating Connectionist Temporal Classification (CTC) for the loss function, augmented by layers of Bi-directional Long Short-Term Memory (BiLSTM), dense layers, and Convolutional Neural Networks

(CNN). The researchers introduce a hybrid language model (WLM and CLM) based prefix decoding technique and a post-processing strategy rooted in bidirectional encoder representations from transformers (BERT) to amplify the system's output while navigating the limited dataset size. Researchers unveiled various analytical methods to unearth insights from our Automatic Speech Recognition (ASR) system. These discoveries facilitate a deeper comprehension of the ASR system grounded in the Gujarati language and may steer ASR systems toward enhancing their performance for lesser-known languages. The word error rate (WER) has diminished by 5.11% when juxtaposed with the base-model WER following the model's training on the Microsoft Speech Corpus. This research has yielded a comprehensive End-to-End Gujarati voice recognition system. The researchers present a prefix decoding technique that propels the system's efficiency by leveraging dual language models. In a quest to further accelerate the performance of the ASR system, which is anchored in the Gujarati language, these insights may also serve to fine-tune ASR systems.

Mehta and Mitra [41] proposed research that investigates different approaches to the Gujarati language's embedding and classification. The collection consists of multiple-classified Gujarati news headlines. The headlines are sorted into various categories through various classifiers and distinct embedding techniques tailored to the Gujarati language. The Gujarati language possesses a scarcity of resources. Diving into this language is quite an uncommon endeavor. This investigation tackles one of the crucial challenges in Natural Language Processing: classification. Furthermore, readers are encouraged to delve into an array of embedding techniques for Gujarati, which are essential in the feature extraction process during classification. This research utilizes robust classifiers that have been previously established to categorize the embedded data, following an initial embedding phase that crafts a valid representation of the textual information. This investigation also clarifies how several NLP operations can be managed in a language that lacks sufficient resources, like Gujarati. To determine which combination produces the most advantageous results, the research wraps up with a comparative evaluation of the performances of multiple existing embedding and classification strategies.

A hybrid chunker tailored for Gujarati, as proposed by [42], is unveiled in this study. The concluding two Unicode characters of the word and the part-of-speech (POS) are contextual elements in the chunker's evolution, leveraging a machine-learning approach. The best suitable method for Chunking Gujarati text has been determined using four distinct statistical approaches: SVM, Naive Bayes CRF, and HMM. Additionally, linguistic guidelines have been developed to enhance performance. The final result was 98.21% attained accuracy with 96.42% precision, 95.62% recall, and 96.02% F1 score, respectively.

Patel et al. [43] proposed Jodani, a typical spell-checking program that checks the entered string to a vocabulary of acceptable terms. However, since it relies on string matching, it cannot handle inflected words. Therefore, a spell checker that gets around this restriction is necessary. The Gujarati spell checker tool "Jodani" is provided in this study. It employs string similarity measurements to detect misspelled words, tries to auto-correct the term, or recommends other phrases that are syntactically appropriate. Jodani determines if a Gujarati word is spelled correctly and then provides a list of alternatives. Using Levenshtein edit distance, the list of recommendations and the spelling accuracy are produced, and rules are constructed for handling inflected words and organizing the proposals. The accuracy of 91.56% demonstrates the effectiveness of Jodani's method, which outperforms the presently offered Gujarati spell checker. More improvements may be made by challenging the presumption that the initial character of the entered word is valid. Neural networks and other machine learning techniques may also enhance the ideas' order.

Patel and Patel [44] proposed an action plan for Gujarati's "sandhi" grammar idea. The majority of South Asian languages, including Devnagri, Sanskrit, Hindi, and Gujarati, as well as Chinese and Thai, use a word segmentation method known as Sandhi. The sounds at the end of a word combine to create a unified chunk of the character sequence due to Sandhi, which causes phonetic alteration at the word borders of a written chunk. The primary focus is on "sandhi's" rule-based implementation. Like other languages written in Indian script, Gujarati grammar has its own rules for combining consonants, vowels, and modifiers. The researchers have established specific guidelines by which authors carry out the actual application of "sandhi." The Gujarati grammatical system has multiple sandhi rules, each designating a particular set of phonetic alterations. The question words undergo neither syntactic nor semantic modification by the Sandhi. A discretionary procedure called Sandhi only relies on the writer's attentiveness.

Raulji et al. [45] introduced a mesmerizing MT framework that employs a technique of grammatical transmission to translate the elegant written Sanskrit language into the expressive Gujarati tongue. Given that both languages boast a lavish morphological tapestry, grasping the intricacies of each component's morphology and significance proves to be a daunting yet vital endeavor for seamless integration into the system's execution. A careful study of the complicated methods used to develop Nouns, Verbs, Pronouns, and Indeclinables, including their links, has been carried out to refine the accuracy of execution and the clarity of translation. The cornerstones of this endeavor include lemmatization, tokenization, and a thorough morphological analysis of the Sanskrit-Gujarati bilingual dictionary, which is founded on a synthesis of synonyms and linguistic artistry. The outcomes of this implementation were scrutinized across 1,000 sentences utilizing an automated Bilingual Evaluation Understudy (BLEU) scale, culminating in a commendable score of 58.04.

Moreover, the structure was assessed using the ALPAC scale, yielding an Intelligence score of 69.16 and a Fidelity score of 68.11. The results were heartening and illuminated the tenacity and reliability of the system devised for real-world applications that surround us. Designing a rule-based system is perpetually a formidable task owing to the complex nature and sheer volume of rules involved. This difficulty is particularly magnified for linguistically rich languages such as Gujarati and Sanskrit. The challenge lies in encompassing every grammatical nuance. The suggested implementation will guarantee dependable outcomes by integrating every grammatical component. The initial MT framework from Sanskrit to Gujarati achieved commendable results when evaluated through ALPAC's manual fidelity and intelligibility parameter scale. It also demonstrated remarkable outcomes when assessed with the BLEU scale. As the architectural implementation could enhance the accuracy of the results by embracing a broader spectrum of vocabulary and addressing any grammatical constraints, the researchers are also keen to explore these additional elements. Furthermore, should a substantial bilingual corpus emerge in the future, machine and deep-learning frameworks could be incorporated to elevate the precision of the system and its overall efficacy.

Audichya et al. [46] conducted this research to provide a method for resolving WordNet inclusion difficulties, including loanwords, missing words, and freshly created terms. One of the most well-known NLP toolkits is WordNet. Any language, such as WordNet, may be improved using this method. Given that these are the languages with the most widespread dialects, the authors decided to focus on Hindi and Gujarati for

this study task to provide higher-quality research results. Instead of a prose-based data corpus, the study endeavor used more than 5000 Hindi verses.

Consequently, over 14000 Hindi words, or 13.23% of the current word count of 105,000+, were missing from the well-known Hindi IndoWordNet. Idiomatic use was a special technique for the Gujarati language. A maximum of 3500 idioms were employed, and 900 Gujarati terms not in the IndoWordNet were found, making up little under 1.4 percent of the 64000+ Gujarati words in the IndoWordNet. Additionally, it will add around 900 Gujarati terms and almost 14000 Hindi words to the IndoWordNet project. Ongoing work is always needed to improve WordNets. As a result, nearly any language's WordNet may be strengthened and improved by implementing these techniques and procedures. Belong to some neighbouring languages for future improvement. Additional investigations may be embarked upon to scrutinize errant spellings and unsuitable terms, including collaborative linguistic inquiries or studies associated with language identification [46].

Jatayu Bax and Bhatt [47] embarked on the fascinating journey of crafting an annotated morphological dataset for the Gujarati Indo-Aryan language, known as GujMORPH. Their exploration delved deep into the intricate rules of word formation, the nuances of language grammar, and the delicate art of suffix attachments to bring this dataset into existence. Within this treasure trove lies a wealth of information on the morphological segmentation and grammatical feature tagging of 16,527 unique inflected words. In the realm of the Gujarati language, this dataset is a pioneering creation and holds the potential to fuel the development of morphological analyzer and generator models. The dataset underwent a meticulous analysis utilizing a reference system, meticulously annotated in alignment with the esteemed Unimorph schema. The researchers also shed light on the innovative software that facilitated data annotation in the specified format. In addition to the library, this invaluable dataset is generously made accessible to the public. Any machine learning model may be trained using the data that can be retrieved to use this package. In earlier work, researchers constructed the baseline system for the Gujarati morphological analyzer utilizing the requested dataset using a Bi-LSTM-based model. The system's accuracy for the morpheme boundary detection test is 89.05%, while for predicting grammatical features, the F1 scores for the POS categories for nouns, verbs, and adjectives are 0.68, 0.12, and 0.68. The Unimorph 4.0 release will include gathering corpora, preparing, and annotating low-resource language data such as the dataset. Researchers manually check the GujMORPH dataset to ensure appropriate annotation, which is a costly but necessary task to maintain the dataset's gold standard. The suggested dataset is available to the public and tested using the standard system. The dataset will eventually be expanded by adding new instances and other speech types.

Parikh and Desai [48] introduced an innovative approach for pinpointing offline handwritten Gujarati conjuncts, revealing a fascinating array of 767 distinct varieties of commonly utilized conjuncts within this study. The identification of handwritten Gujarati conjuncts employs the sophisticated architectures of convolutional neural networks (CNN) - namely AlexNet, GoogLeNet, Inception V3, and ResNet50. A meticulous evaluation of these CNN architectures' performance is scrutinized. Before being acknowledged, Gujarati conjuncts undergo segmentation from their designated locations. Conjuncts are classified into preceding and following elements before they are recognized. This mechanism enables the system to identify each of these earlier and subsequent components independently. The proposed research endeavor harnesses a vast collection of 28050 exemplar images of the following components of conjuncts and 19694 sample visuals of their preceding counterparts. Through the utilization of Google's neural network, the pinnacle of accuracy achieved for the preceding components and the utmost precision for the subsequent components both reached an impressive 93.41%.

Utkarsh Kapadia and Apurva Desai [49] proposed grammatical inference, information retrieval, and machine translation, which had to include part of speech tagging. The challenge with tagging is determining the best suitable tag for each word in a phrase based on its lexical and contextual characteristics. The two basic approach categories are monitored and unsupervised. The unsupervised technique does not rely on labeled text or any pre-established rules. Here, researchers have a hybrid Gujarati Part-of-Speech (POS) tagger. Rules are developed in collaboration with linguists and native speakers. With a lexicon of 30,050 words and 12,637 phrases, the authors assessed performance using 30 distinct standard part-of-speech tags for Gujarati. The system is evaluated using text from several Gujarati domains.

News, essays, and short stories are all included in these areas. The method has an 82.52% accuracy rate. Part of the Speech tagger with custom suffix replacement rules has been explored. Tokenization comes first, after which words are looked up in the database; if they can't be discovered, the proper rules are used. When researchers apply rules, the system sometimes tags words with incorrect POS tags. If more than half of the words in a phrase are unknown, the algorithm will not categorize them. The reasoning for this is that word tagging resolution is reliant on word context. Thus, the system must determine which rules to apply first. As a result, sometimes it is challenging to identify tags when numerous tags are unknown. By expanding the database, part of speech tagger accuracy may be improved. There is a potential that more than one rule will be relevant in the same circumstance, but if it produces a different tag for the same word, the system may fail. In this case, researchers may change the rules' priority to choose the most helpful tag for the word [49].

Limbachiya et al. [50] introduced an innovative system that harnesses the power of transfer learning and CNN to recognize handwritten Gujarati alphanumeric characters. The realm of pattern recognition, often referred to as offline handwriting recognition, has captivated the interest of scholars owing to its significance. Transforming handwritten content into machine-readable textual data is a complex endeavor that involves uncovering hidden patterns and deciphering the words within the documents. India boasts 22 officially designated scheduled languages, and Gujarati proudly stands among them. The Gujarati language presents a multitude of challenges in the realm of optical character recognition (OCR), especially when it comes to discerning consistent patterns and peculiarities within handwritten Gujarati text. One significant concern regarding handwritten Gujarati script is the absence of a comprehensive standard dataset. The matter was recognized, and a dataset including 75,600 photos encompassing 54 distinct classes of Gujarati characters was constructed. Despite the dataset being of a considerable size, it remains insufficient for training deep neural networks from the ground up, mostly due to overfitting issues. To tackle this issue, we have incorporated transfer learning into a CNN to recognize handwritten characters in Gujarati. We have employed five unique pre-trained models and attained an approximate accuracy rate of 97% while evaluating photos belonging to 54 different categories.

Parth Goel and Amit Ganatra [51] applied the deep transfer learning technique to categorizing handwritten Gujarati digits using ten pre-trained networks. When training the intended model, researchers tackled the key transfer learning challenge of how deeply to tweak the previously learned convolutional neural network. To tackle this conundrum, the authors embraced three innovative strategies of transfer learning to illustrate the impact of diverse fine-tuning methods on the model's efficacy. Initially, pre-trained models were leveraged as feature extractors with the aid of linear SVM and softmax classifiers, revealing that this technique yielded the least favorable outcomes due to the weight

parameters remaining unrefined during the target task's training. In the second approach, two newly adjusted fully connected layers were appended following the last convolutional layer to enhance the pre-trained models further. This strategy allowed us to grasp the nuanced high-level features through retraining the final layers, culminating in the pinnacle of performance. By fine-tuning the models starting from the midpoint of the network in the third method, we noted median effectiveness while experiencing the most prolonged training duration. Ultimately, an experiment was conducted using a self-generated handwritten Gujarati digit dataset, employing a variety of performance evaluation metrics. Researchers got the best accuracy using the second method instead of starting from scratch using the EfficientNetV2S model. Researchers discovered that the efficient transfer learning fine-tuning technique enhances the model's performance and greatly shortens the training period. Additionally, it can be concluded that if accuracy is the primary criterion for selecting the top performer, EfficientNetV2S may be the best model. In contrast, LeNet may be the greatest option amongst all models if memory is the primary consideration for selecting a lightweight model for resource-constrained devices.

Abhinav Sharma et al. [52] unveiled the innovative CNN-driven EfficientNet B3 model alongside the YOLO v4 model, tailored explicitly for recognizing Gujarati text. Both models underwent a thorough examination. The EfficientNet B3 model, celebrated for its heightened accuracy and efficiency, played a pivotal role in crafting the system. A visual featuring optical Gujarati text serves as the system's entry point, and the computer deftly generates an editable document reflecting the text extracted from the image. This ingenious technique has proven effective in establishing a digital repository of articles from Gujarati newspapers. At the core of the proposed method lies a character recognition model. Upon receiving an input image, it meticulously dissects the content into lines, words, and individual characters. The prediction model scrutinizes each character independently before synergizing them to yield the final representation of the image. Consequently, segmenting text into characters and predicting Gujarati letters significantly influence the system's overall accuracy. The strategy of breaking down input text into discrete characters by leveraging pixel values across each row and column of the image has emerged as a dependable approach. EfficientNet B3 and CNN have attained outstanding results in character prediction. This remarkable initiative represents a significant stride towards the cultural and linguistic preservation of the Gujarati language. It can enhance by integrating automatic correction and completion techniques.

In this research endeavor, Uttam Chauhan [53] introduced a DFA-inspired approach to lemmatization tailored specifically for the Gujarati language. The innovative technique for transforming Gujarati text into its base forms encompasses an impressive array of over 59 meticulously crafted rules. The authors achieved remarkable success, lemmatizing an impressive 83% of the vocabulary. To evaluate the impact of the proposed lemmatization technique on textual analysis, researchers employed Latent Dirichlet Allocation to extract underlying themes from a text corpus. Their findings illustrated that while lemmatization did not diminish the total number of tokens, it led to a noteworthy reduction in the overall vocabulary size. The experimental data revealed that lemmatizing the corpus significantly enhanced the understanding of the themes. The subjects became more concentrated and meaningful. This conclusion was further validated by the Hellinger distance and Jaccard distance metrics.

Moreover, the measures of semantic coherence underscored an increase in topic quality. All three methodologies normalized Pointwise Mutual Information, Pointwise Mutual Information, and Log Conditional Probability, confirmed an uptick in coherence values. Additionally, it was found that the themes derived from the lemmatized corpus exhibited greater specialization. Furthermore, the terms associated with the themes' semantic connections experienced significant enhancement. A universal technique may be devised for any text corpus in the future. For example, a compilation of rules could facilitate lemmatization across various text domains, including news articles, forums, books, and social media, among others [53].

Monil Gokani and Radhika Mamidi [54] introduce the Gujarati Sentiment Analysis Corpus (GSAC), comprising more than 6500 tweets that have been carefully annotated. To the best of our current understanding, this corpus represents a notable and accessible resource for the assignment at hand in the Gujarati language. Researchers introduce our annotation system and perform comprehensive testing to establish baseline results for this novel dataset. It has been observed that language models that have undergone pre-training or fine-tuning, including Gujarati as one of the languages, exhibit superior performance on this dataset compared to other models. Among these models, IndicBERT achieves the highest weighted and macro F1 scores. As a component of our future research, Researchers intend to investigate approaches for augmenting this dataset in an automated manner. It involves utilizing the current database to label supplementary data, employing techniques like bootstrapping. Additionally, Researchers want to investigate alternative means of data acquisition, such as leveraging machine translation to convert existing datasets in languages like English or Hindi [54].

Jatayu Baxi and Brijesh Bhatt [55] introduced the innovative neural morph analyzer tailored for the Gujarati dialect, showcasing remarkable effectiveness across all parts of speech categories while triumphantly surmounting the limitations of the foundational system. The dataset crafted by the researchers stands as a valuable resource for academics, enabling them to delve deeper into studies related to computational morphological tools within the Gujarati language. The linguistic analysis of the gathered data has illuminated the complex morphological tapestry of the Gujarati language. To enhance the prowess of the current system, the researchers set their sights on expanding the horizons of the training data in their forthcoming ventures. In addition, there exists an aspiration to collect feedback at the level of sentences, breaking away from the prevailing focus on individual words. This transformation will unlock the door to a deeper investigation of sentence-level connections within the realm of morphological analysis. Moreover, the researchers' ambition includes the concurrent prediction of both part-of-speech (POS) classifications and the morphological characteristics of words. This endeavor aspires to elevate the precision of the current system while examining how POS tagging shapes the prediction of morphological traits. Looking toward the horizon, the morph analyzer model holds the promise of enhancing the accuracy of subsequent NLP tasks, especially for the Gujarati language.

Nitesh Patel and Dhiren Patel [56] explore the captivating influence of rule-based "sandhi" splitting and joining techniques on the Gujarati language, a revered Indian tongue known for its rich tapestry of expression. "Sandhi" serves as a vital thread in the intricate fabric of grammar, involving modifications to words during their amalgamation, thus affecting both written and spoken communication. This investigation meticulously examines conventional rule-based methods for dissecting and uniting Gujarati compound words, which constitute the core of the language's morphology. We evaluate the efficacy of various techniques in improving the processing and understanding of text, particularly within the domain of NLP applications. This research focuses on the complex linguistic phenomena surrounding Gujarati sandhi and endeavors to create rule-based algorithms that can accurately separate and merge words according to established grammatical rules. We delve into the utilization of these techniques across a variety of natural language processing tasks, including text analysis, machine translation, and information retrieval. The intricate art of sandhi splitting, known as vicched, poses a greater challenge than Sandhi joining due to its inherent complexity

and dependence on context. Furthermore, we assess the efficacy and suitability of rule-based sandhi techniques by comparing them with alternative methods, such as machine learning-based approaches, to evaluate their performance and adaptability in various situations. Additionally, we ponder potential advancements and future directions for the incorporation of these rule-based techniques into contemporary computer science and computational language processing.

Table 2: Research Papers Methods and Limitations.

Paper	Year	Approach /Method	Accuracy	Limitation
[14]	1999	k-Nearest Neighbour and Euclidean Minimum Distance classifiers	67 %	Simple shape-based recognition won't continuously work well
[15]	2010	Goldsmith Method	67.86 %	Researcher used hand-crafted Gujarati suffix list
[16]	2010	Multi-Layered Feed-Forward Neural Network	81.66 %	Few work available for the comparison purpose for Indic Languages
[17-18]	2010/ 2016	Using Hindi WordNet- Synsets	Not Applicable	Limited WordNet
[19]	2017	Pre-process technique global threshold, erosion and dilation, skew correction - Structural and Statistical approach, KNN	96.99 %	In classification other approaches can be used, for comparison purpose
[20]	2011	Support Vector Machine (SVM) and Naive Bayes (NB)	98.77 %	Only 20500 manually produced dataset used, which is small amount to know accuracy of system compare to other system
[21]	2011	Rules-based Method -– Hybrid	70.7 %	Only use of suffix-stripping and substitution rules
[22]	2012	Rules-based Method	91.5 %	System inclined to make over-stemming mistakes and rigorous error analysis needed
[23]	2013	Entity Recognition Methods	Not Mentioned	POS tagger should be improved accuracy of system
[24]	2013	Connected Component Labeling and Neural Network Method	87.4 %	Reason for poor performance include having too tiny character sizes and improper inter character alignment.
[9]	2014	Rules Based statistical methods	92.41 %	System tested by small amount of Dataset
[25]	2014	Probabilistic Approach- Unigram, Bigram and Trigram/ Naïve Bayes and Hidden Markov Model	Not Mentioned	System tested on synthetic dataset
[26]	2014	MelFrequency Cepstral Coefficients and KNN	78.13 %	System needed speech patterns of younger speakers
[27]	2014	Concatenation Technique	Not Mentioned	There are no any depth clarification for how to develop this system and future scope of this research paper.
[28]	2015	Paradigm based approach with knowledge based and Statistical Approach	92.34 % knowledge based / 82.84 % statistical method.	Cannot Handle Derivational Morphology
[29]	2015	SVM and KNN	86.66 %	'Th- થ' Character is confused for identification by the system
[30]	2015	Rule-Based and Machine Learning for morphological analyzer	Not Applicable	Not Mentioned
[31]	2016	Braille system American Grade 0- Nemeth Code, Literary Code	Satisfactory Result	Not Mentioned
[32]	2017	Rule Based	87.48 %	No Result when suffix does not meet any rules
[33]	2017	Stemmer and String similarity scale	Not Mentioned	Machine learning and different NLP tools can be used
[34]	2019	Multi-layer Classification Approach	94.13 %	When the second layer classifier is unable to provide correct findings, the first layer classifier fails.
[35]	2019	Lemmatization	98.33 %	The real inflection cannot be located and eliminated.
[36]	2019	Multilingual Neural Machine Translation model - RNN	Not Mentioned	This system not using one to many Multilingual NMT systems
[37]	2019	Statistical Machine Translation and Neural Machine Translation	BLEU scores rose from 8.1 to 10.4.	There are relatively few parallel phrases in the training set, and most of the sentences are just two or three tokens long.
[38]	2019	2D DWT wavelet features	Haar orthogonal wavelet- 92 %	Proposed algorithm is tested on 800 images only
[39]	2019	Unsupervised Learning	99.34 %	Many punctuation marks and truncated word missing in dataset
[40]	2020	Convolutional Neural Network (CNN), Bi-directional Long Short Term Memory (BiLSTM) layers, Dense layers, and Connectionist Temporal Classification (CTC)	Reduces the overall Word Error Rate by 5.11 %	Half conjugate erroneous words
[41]	2020	Embedding Techniques	Not Applicable	Only for Gujarati News Paper Headlines, other text has low accuracy
[42]	2021	SVM, HMM, Naïve Bayes, CRF	98.21 %	Linguistic Rules should be updated
[43]	2021	Levenshtein edit distance	91.56 %	Neural networks and other machine learning techniques missing, work only on string similarity measurements
[44]	2021	Rule Based	Not Mentioned	Sandhi is an optional procedure that only relies on the writer's attentiveness.

[45]	2022	Bilingual Evaluation Understudy and Automated Language Processing Committee	Intelligibility score was 69.16 % and Fidelity score 68.11 %	Huge bilingual corpus not available.
[46]	2022	Data Collection, Data Pre-processing, Data Filtering, Data Logging, Data Rechecking	Not Applicable	Still Gujarati WordNet rich WorddNet.
[47]	2022	Bi-LSTM-based model	Not Applicable	Unavailability of few part of the speech categories
[48]	2022	CNN, AlexNet, GoogLeNet, Inception V3, and ResNet50	93.41 % for the preceding components and 89.01 % using GoogLeNet	Few comparison are mission in related work or survey work.
[49]	2022	Hybrid Part-of-Speech (POS) tagger- Rule Based	82.52%	Few tagging issues identified by researchers
[50]	2022	Transfer learning with CNN	97 %	5 distinct pre-trained models use, novelty is low.
[51]	2023	Deep transfer learning technique- linear SVM and softmax classifiers	EfficientNetV2S - 97.32%	Researcher should be used multiple OCR with available Dataset
[52]	2023	CNN based EfficientNet B3 and YOLO v4	CNN- 79.78 % / EfficientNet B3 98.92%	System can be improved by using autocorrect and auto complete mechanisms.
[53]	2023	DFA-Based Lemmatized	83 %	Unavailability of rules, which can achieve lemmatization for news articles, discussion forums, textbooks, novels, social media text domains.
[54]	2023	Bag of Words, TF-IDF, Pretrained LMs - GujaratiBERT and IndicBERT	IndicBERT - 66 %	Unavailability of labeled additional data
[55]	2024	Deep Learning, Bi-LSTM - Mor	91.57%	Sentence-level dependencies not available
[56]	2024	Rule Based - Sandhi	-	Only rule based approach available

Source: Own elaboration.

As per research need, the above table has 56 papers covering GNLP research work till August 2024. There are several limiters for each paper. After examining each limitation, we found that the most popular limitation is the unavailability of a good dataset. There is no accurate data set to test the system. The foremost requirement of GNLP research work is to create a quality dataset of Gujarati language. One can see that the average accuracy of the GNLP system is very low. It should be enhanced with the help of current technology. This survey proves that a good Rules-based, machine-learning, and deep-learning hybrid approach is needed for accurate GNLP systems

III. GNLP ONLINE RESEARCH STUDIES

Figure 1 demonstrates the availability of several landmark online GNLP studies. Please note that only unique studies are counted in this survey. As shown in the figure, only one research paper was found in 1999, 2004, and 2007. With slow progress, three papers were found in 2009, while in 2010 and 2011, the number of papers doubled, i.e. 6. In 2012, 2013, and 2015 the number of papers fluctuated. However, GNLP again picked up the speed of research in 2016 and 2017, with 10 and 11 papers, respectively. However, in 2018 a research gap is observed in GNLP with only five papers. Then, 10 and 13 research papers will be available in 2021 and 2022, respectively, except for a few in 2020. Unique and quality-based research work done in 2023 & 2024, respectively 4 and 2 papers. This chart indicates that the GNLP will also gain useful research in the coming years.

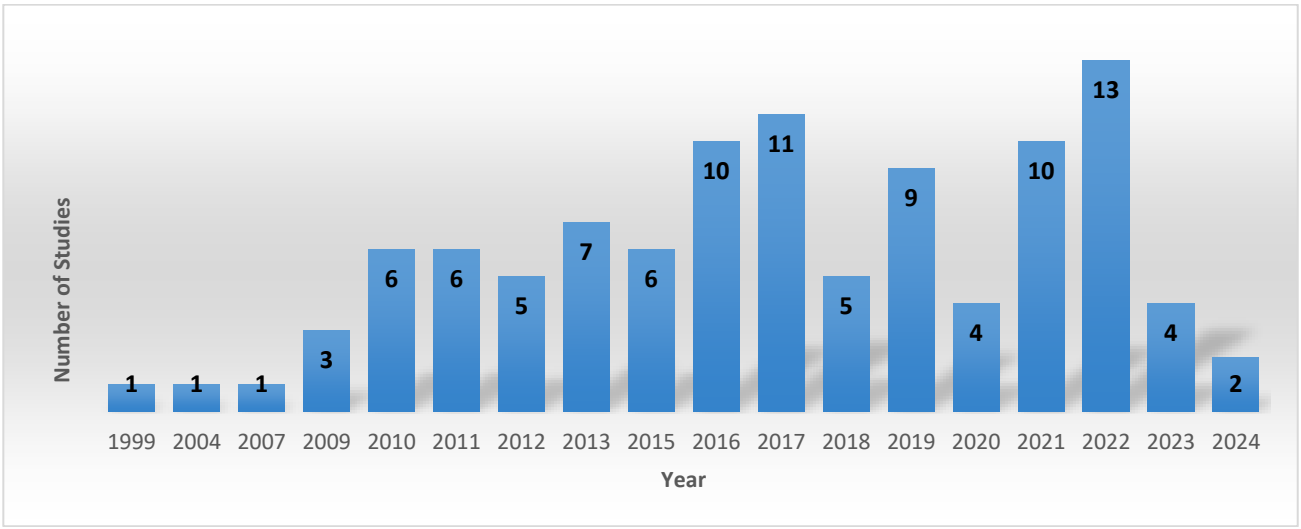

Figure 1: GNLP Online Research Studies.
Source: Own elaboration.

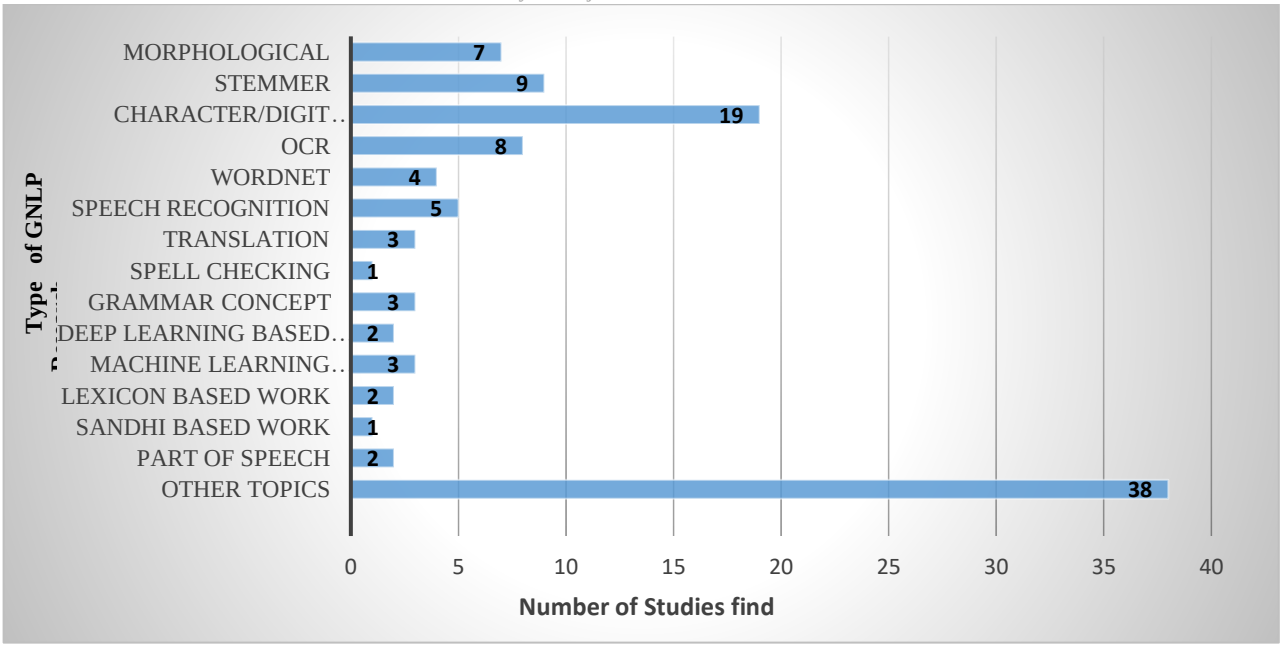

Figure 2: GNLP Resources and Tools.
Source: Own elaboration.

As shown in Figure 2, Gujarati Character and Digit Recognition have the highest number of studies. Afterward, stemmer and morphological have 9 and 7 research paper, respectively. Spell Checking, Grammar Checking, and Translation-based GNLP research have the lowest numbers. Figure 2 explains that only a few studies were done on GNLP; the big task remains. So, GNLP wants more focus from Indic NLP researchers.

Figure 3 is an output of a Microsoft Excel function called prediction. The researcher has put a GNLP Forecasting for Research Studies from 2023 to 2030. Please note that this figure is an imaginary part of GNLP research that uses only available data. It is predicted that, on average, ten papers will be available as part of GNLP research.

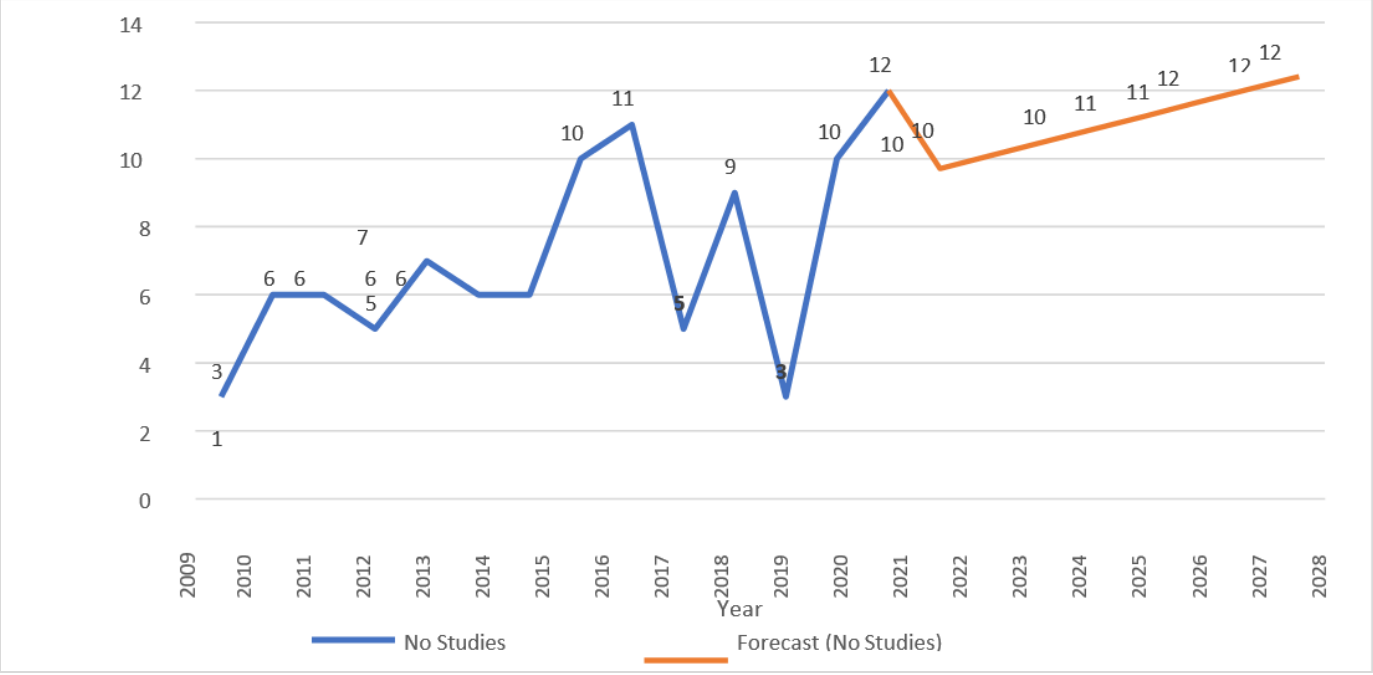

Figure 3: GNLP Forecasting for Research Studies.
Source: Own elaboration.

IV.RESEARCH GAP AND SOLUTION

Many parts and assets have been developed and studied in GNLP but have not yet been made available to the public. There is not enough annotated data available in Gujarati for various purposes and sectors, so alternative languages are actively sought. A tree bank for Gujarati must be constructed to facilitate the development of the parsers and, eventually, the machine translation of Gujarati into source-rich languages like English or Hindi. This might facilitate the dissemination of knowledge. The language translator, anaphora resolver, word sense disambiguates, and parser all need much research and development in Gujarati. Although much research has gone into creating the morphological analysis and lexical analyzer components for NLP's initial phases, the latter stages' elements still need more in-depth development to meet the format

requirements of languages with a wealth of resources, like English. Human experts should be assisted in developing answers based on rules. Then, learning should be automated using machine learning, which can later be made accessible to other academics.

Moreover, Gujarati data needed for various NLP activities can be created. The intricate bond shared by Gujarati and its Indo-Aryan counterparts, such as Marathi, Hindi, and Konkani, is widely acknowledged. With commendable dedication, NLP innovations are emerging for both Marathi and Hindi. For Hindi, invaluable co-referenced dependency-tagged data is provided. Functionalities like POS tagging and morphological analysis tools are also accessible for the corpora of Hindi and Marathi, along with search engine capabilities. Additional components and characteristics may also be built, especially from the relevant sources provided in the linked languages, such as the extension method used to develop the Gujarati wordnet. For text summarization in the medical field, models should be created for Gujarati NLP utilizing transfer learning and the language models currently available for resource-rich languages like English.

There is no equivalent of a spell-checking or grammar checker for the Gujarati language. The Chandaria Foundation's Gujarati Lexicon website has Saras Spellchecker [57], but is inoperable. The efforts Rajesh Mashruwala and colleagues (USA) undertook were unsuccessful. The "Jodani" technology in the GNLP spell checker area has cast a glimmer of hope. An essential component of writing is proofreading. Gujarati NLP researchers must focus on product-based systems rather than an academic effort to address spelling and grammar issues. All those who deal with the Gujarati language will find these techniques to be of great use. Market demand-oriented Gujarati Spell checkers are a good place for GNLP researchers to start, after which they may go on to grammar checkers. There is no single accurate grammar checker or spell checker in GNLP market. Researchers of GNLP should start their efforts for Gujarati languages' Spell Checker and Grammar Checker [58-60].

As per Figure 2, much of the research on GNLP is Rules-based work and machine learning only. Researchers of GNLP should try Deep Learning based GNLP systems. There are more chances for accurate results in deep learning-based work. Researchers can use RNN, LSTM, Bi-LSTM, GRU and BERT to clarify research work. A good number of GNLP datasets are still unavailable after so many decades of Indic NLP research. The techniques that have already failed in past research work should not be tried repeatedly. Rather than the old one, researchers should use a new combination for GNLP development.

V. CONCLUSIONS

This paper demonstrates a survey of existing research work on GNLP. Initially, the authors describe the history of the Gujarati language, including its consonants, vowels, and numbers tense, Vachans, and sentence structure. Afterward, a brief note on the background of Indic NLP is explained, and one can find an overview of Indic language internet users. Before discussing the related work of GNLP, researchers discussed the number of GNLP applications like WordNet, Morphological, Stemmer, optical character recognition (OCR), Speech Recognition, Part of Speech, and Machine Translation. In related work on GNLP, the authors write a short review of several papers. Each survey has pros and cons. Through each paper, researchers get a critical summary of GNLP's online research studies as well as its recourses and tools present in graphical form. Finally, the authors critically analyze a research gap and a solution for GNLP. Overall, this paper gives a historical to present scenario of GNLP. From this paper, academicians or industry people get an idea regarding remaining and completed research work in GNLP. Lastly, Through the whole paper, researchers get an idea that the Deep Learning-based solution of GNLP is more beneficial.

VI. REFERENCES

[1] B. Waghmar, Gujarati, in Concise Encyclopedia of the Languages of the World, K. Brown and S. Ogilvie, Eds. Oxford: Elsevier, pp. 468–469, 2009.

[2] T. Vyas and A. Ganatra, "Gujarati language: Research issues, resources and proposed method on word sense disambiguation," Int. J. Recent Technol. Eng. (IJRTE) ISSN, vol. 8, no. 2, suppl. 11, p. 2277-3878, pp-3745- 3749, 2019.

[3] C. Masica, The Indo-Aryan Languages. Cambridge: Cambridge University Press, ISBN 978-0-521-29944-2, 1991.

[4] P. J. Mistry, "International encyclopedia of linguistics," in Gujarati, 2nd ed, vol. 2, W. Frawley, Ed. Oxford: Oxford University Press, 2003.

[5] Census of India. India: LANGUAGE Atlas, 2011.

[6] A Study by KPMG in India and Google- "Indian Languages – Defining India's Internet", Apr. 2017.

[7] P. Bhattacharyya, "IndoWordNet," in Proc. Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA), 2010.

[8] J. Baxi, P. Patel, and B. Bhatt, "Morphological Analyzer for Gujarati using Paradigm based approach with knowledge based and Statistical Methods," in Proc. 12th International Conference on Natural Language Processing, pp. 178–182, 2015.

[9] J. Sheth and B. Patel, "Dhiya: A stemmer for morphological level analysis of Gujarati language," in international conference on issues and challenges in intelligent computing techniques (ICICT). IEEE, pp. 151–154, 2014.

[10] H. Patel, "Gujarati OCR: Compound character recognition using Zernike moment feature extractor," Int. J. Comput. Sci. Trends Technol. (IJCST), vol. 8, no. 5, pp-45–50, 2020.

[11] J. H. Tailor and D. B. Shah, "Speech recognition system architecture for Gujarati language," Int. J. Comput. Appl., vol. 138, no. 12, 2016.

[12] A. A. Desai Kapadia, "U.N.,Paradigm based Part of Speech Tagging with priorities: Implantation for Gujarati Script," Int. J. Comput. Sci. Trends Technol., vol. 10, no. 1, pp. 104–112.

[13] V. Goyal and D. M. Sharma, "The IIIT-H Gujarati-English machine translation system for WMT19," in Proc. Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 191–195, 2019.

[14] S. Antani and L. Agnihotri, "Gujarati character recognition," Proc. Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (cat. No. PR00318). Bangalore, India, pp. 418–421, 1999.

[15] P. Patel, K. Popat, and P. Bhattacharyya, "Hybrid stemmer for Gujarati," in Proc. 1st Workshop on South and Southeast Asian Natural Language Processing, pp. 51–55, 2010.

[16] A. A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network," Pattern Recognit., vol. 43, no. 7, pp-2582-2589, 2010.

[17] B. S. Bhatt, C. K. Bhensdadia, P. Bhattacharyya, D. Chauhan, and K. Patel, "Gujarati WordNet: A profile of the IndoWordNet," in The WordNet in Indian Languages, N. Dash, P. Bhattacharyya, and J. Pawar, Eds. Singapore: Springer, 2017.
[18] C. K. Bhensdadia, B. Bhatt, and P. Bhattacharyya, "Introduction to Gujarati wordnet," in Third national workshop on indowordnet Proceedings, vol. 494, 2010.
[19] A. A. Desai, "Handwritten Gujarati numeral optical character recognition using hybrid feature extraction technique," in IPCV 2010, Proc. 2010 International Conference on Image Processing, Computer Vision, & Pattern Recognition, Las Vegas, NV, Jul. 12–15, pp. 733–739, 2010.
[20] C. Patel and A. Desai, "Zone identification for Gujarati handwritten word," in Second International Conference on Emerging Applications of Information Technology. IEEE, pp. 194–197, 2011.
[21] K. Suba, D. Jiandani, and P. Bhattacharyya, "Hybrid inflectional stemmer and rule-based derivational stemmer for Gujarati," in Proc. 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP), pp. 1–8, 2011.
[22] J. Ameta, N. Joshi, and I. Mathur, "A lightweight stemmer for Gujarati. ArXiv:1210.5486". Available at: https://arxiv.org/abs/1210.5486, 2012.
[23] J. R. Sheth and B. C. Patel [Article], "Stemming techniques and naive approach for Gujarati stemmer" IJCA Proceedings on International Conference on Recent Trends in Information Technology and Computer Science 2012 ICRTITCS (2), vols. 9–11, Feb., 2013.
[24] C. Patel and A. Desai, "Extraction of characters and modifiers from handwritten Gujarati words," Int. J. Comput. Appl., vol. 73, no. 3, 7–12, 2013.
[25] D. B. Patel and M. M. Goswami, "Word level correction in Gujarati document using probabilistic approach," International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE), Coimbatore, India, pp. 1–5. 2014.
[26] B. C. Patel and A. A. Desai, "Recognition of spoken Gujarati numeral and its conversion into electronic form," Int. J. Eng. Res. Technol. (IJERT), Vol., no. 9, Sept., p. 3, pp-474–480, 2014.
[27] J. M. Varghese and S. S. Hande. "text-to-speech System for Gujarati Language.", Int. J. Adv. Comput. Electron. Technol. (IJACET), vol. 2, no. 4, pp-78–81, 2015.
[28] J. Baxi, P. Patel, and B. Bhatt, "Morphological Analyzer for Gujarati using Paradigm based approach with Knowledge based and Statistical Methods," in Proc. 12th International Conference on Natural Language Processing, pp. 178–182, 2015.
[29] A. A. Desai, "Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space," CSIT, vol. 2, no. 4, 235–241, 2015.
[30] U. Kapadia and A. Desai, "Morphological rule set and lexicon of Gujarati grammar: A linguistics approach," VNSGU J. Sci. Technol., vol. 4, no. 1, Pp-127–133, 2015.
[31] N. B. Jariwala and Dr. B. Patel, "Transliteration of digital Gujarati mathematical text into Braille for visually impaired people," Int. J. Latest Trends Eng. Technol., Issue (3), vol. 7, no. 3, pp. 217–229, 2016.
[32] U. Kapadia and A. Desai, "Rule based Gujarati morphological analyzer," Int. J. Comput., vol. 14, no. 2, pp. 30–35, 2017.
[33] J. Sheth and B. Patel, "Saaraansh: Gujarati text summarization system," Int. J. Comput. Sci. Inf. Technol. Sec., vol. 7, no. 3, pp. 46–53, 2017.
[34] V. A. Naik and A. A. Desai, "Multi-layer classification approach for online handwritten Gujarati character recognition," in Adv. Intell. Syst. Comput., vol 799, N. Verma and A. Ghosh, Eds. Computational Intelligence: Theories, Applications and Future Directions. Singapore: Springer, vol. II, 2019.
[35] H. Patel and B. Patel, "Stemmatizer-Stemmer-based lemmatizer for Gujarati text," in Emerging Trends in Expert Applications and Security. Advances in Intelligent Systems and Computing, vol. 841, V. Rathore, M. Worring, D. Mishra, A. Joshi, and S. Maheshwari, Eds. Singapore: Springer, 667–674. 2019.
[36] V. Goyal and D. M. Sharma, "The IIIT-H Gujarati-English machine translation system for WMT19," in Proc. Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 191–195, 2019. doi: 10.18653/v1/W19-5316.
[37] S. Sen, K. K. Gupta, A. Ekbal, and P. Bhattacharyya, "IITP-MT system for Gujarati-English news translation task at WMT 2019," in Proc. Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 407–411, 2019. doi: 10.18653/v1/W19-5346.
[38] M. Patel and A. Apurva Desai, Performance analysis of various wavelet filters for Gujarati text localization in images, IJRAR-International Journal of Research and Analytical Reviews ,, "Jagin," vol. 6, no. 2, pp. 96–100, 2019.
[39] C. Tailor and B. Patel, "Sentence tokenization using statistical unsupervised machine learning and rule-based approach for running text in Gujarati language," in Emerging Trends in Expert Applications and Security. Advances in Intelligent Systems and Computing, vol. 841, V. Rathore, M. Worring, D. Mishra, A. Joshi, and S. Maheshwari, Eds. Singapore: Springer, 319–326, 2019.
[40] D. Raval, V. Pathak, M. Patel, and B. Bhatt, "End-to-end automatic speech recognition for Gujarati," in Proc. 17th International Conference on Natural Language Processing (ICON). Patna, India: Indian Institute of Technology Patna. NLP Association of India (NLPAI), pp. 409–419, 2020.
[41] S. Mehta and S. K. Mitra, "Text classification of Gujarati newspaper headlines," Int. J. As. Lang. Proc., vol. 30, 2020.
[42] C. Tailor and B. Patel, "Chunker for Gujarati language using hybrid approach," in Rising Threats in Expert Applications and Solutions. Advances in Intelligent Systems and Computing, vol. 1187, V. S. Rathore, N. Dey, V. Piuri, R. Babo, Z. Polkowski, and J. M. R. S. Tavares, Eds. Singapore: Springer, 77–84. 2021. doi: 10.1007/978-981-15-6014-9_10.
[43] H. Patel, B. Patel, and K. Lad, "Jodani: A spell checking and suggesting tool for Gujarati language," Data Sci. Eng. (Confluence), (Noida, India) 11th International Conference on Cloud Computing, vol. 2021, pp. 94–99. 2021.
[44] N. Patel and D. Patel, "'Implementation Approach of Indian Language Gujarati Grammar's Concept "sandhi' using the Concepts of Rule-based NLP," 8th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, vol. 2021, pp. 481–485, 2021.
[45] J. K. Raulji, J. R. Saini, K. Pal, and K. Kotecha, "A novel framework for Sanskrit-Gujarati symbolic machine translation system," IJACSA, vol. 13, 2022. doi: 10.14569/IJACSA.2022.0130444, Pp-374–380.
[46] M. K. Audichya, J. R. Saini, and J. C. Modh, "Towards a richer IndoWordNet with new additions for Hindi and Gujarati languages," IJACSA, vol. 13, Pp-832–842. 2022.
[47] J. Baxi and B. Bhatt, "GujMORPH-A dataset for creating Gujarati morphological analyzer," in Proc. Thirteenth Language Resources and Evaluation Conference, pp. 7088–7095, 2022,

[48] M. Parikh and A. Desai, “Recognition of handwritten Gujarati conjuncts using the convolutional neural network architectures: AlexNet, GoogLeNet, inception V3, and ResNet50,” in Advances in Computing and Data Sciences, Revised Selected Papers: 6th International Conference, ICACDS 2022, Kurnool, India, April 22–23, part II, p. 291–303, 2022.
[49] 303. Cham: Springer International Publishing, 2022.
[50] N. Kapadia Utkarsh and A. Deasi Apurva, “Paradigm based Part of Speech Tagging with priorities: Implantation for Gujarati Script,” Int. J. Comput. Sci. Trends Technol. (IJCST), vol. 10, no. 1, pp. 104–112, 2022.
[51] K. Limbachiya, A. Sharma, P. Thakkar, D. Adhyaru, “Identification of handwritten Gujarati alphanumeric script by integrating transfer learning and convolutional neural networks” Sādhanā, vol. 47, no. 2, p. 102, 2022.
[52] P. Goel and A. Ganatra, “Handwritten Gujarati numerals classification based on deep convolution neural networks using transfer learning scenarios,” IEEE Access, vol. 11, pp. 20202–20215, 2023.
[53] A. Sharma et al., “Gujarati script recognition,” Procedia Comput. Sci., vol. 218, pp. 2287–2298, 2023.
[54] U. Chauhan et al., “Modeling topics in DFA-based lemmatized Gujarati text,” Sensors (Basel), vol. 23, no. 5, p. 2708, 2023. doi: 10.3390/s23052708.
[55] M. Gokani and G. S. A. C. Radhika Mamidi, “A Gujarati sentiment analysis corpus from Twitter,”, in Proc. 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. Association for Computational Linguistics, Jul. 14, 2023, pp. 129–137.
[56] J. Baxi and B. Bhatt, “A bidirectional LSTM-based morphological analyzer for Gujarati,” Nat. lang. processing. Cambridge University Press, pp. 1–17, 2024. doi: 10.1017/nlp.2024.14.
[57] N. G. Patel and D. B. Patel, “NLP-based processing of Gujarati compound word sandhi’s generation and segmentation,” in International Conference on Universal Threats in Expert Applications and Solutions. Singapore: Springer Nature Singapore, pp. 263–271, Jan., 2024.
[58] B. Y. Panchal and A. Shah, “Spell checker using Norvig algorithm for Gujarati language,” in. ICSMDI 2024. Algorithms for Intelligent Systems, R. Asokan, D. P. Ruiz, and S. Piramuthu, Eds. Smart Data Intelligence. Singapore: Springer, 2024. doi: 10.1007/978-981-97-3191-6_21.
[59] Available at: https://www.gujaratilexicon.com/saras-spellchecker/.
[60] Y. Gondaliya, P. Kalariya, B. Y. Panchal, and A. Nayak, “A rule-based grammar and spell checking,” SAMRIDDHI A J. Phys. Sci. Eng. Technol., vol. 14, no. 1, pp. 48–54, 2022. doi: 10.18090/samriddhi.v14i01.8.