IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Systematic Literature Review of Natural Language Processing for Marathi

Sunil Patil, Vikas Manhandle, Pradnya Suryavanshi, Vikrant Kadam
MIT Arts, Commerce & Science College, Pune

Abstract: *Marathi is a language that 83 million people speak, mostly in the state of Maharashtra in India [2]. It is very rich in terms of language. It does not have many computer tools to help with it. Even though a lot of people speak Marathi it has not had much work done on it as other languages like English, Mandarin, Hindi and Bengali [1].*

This paper looks at what research has been done on Marathi language processing. We looked at papers from 2005 to 2025 from places like IEEE Xplore, ACL Anthology and Google Scholar. We found 187 papers. Only 124 of them were good enough to use.

We looked at things like how words are formed what parts of speech words are and how to tell what people are talking about. We also looked at how people feel about things how to put texts into groups, how to translate Marathi into languages and how to deal with text that is a mix of Marathi, Hindi and English.

What we found out is that some new computer tools have made it easier to do these things but there are still some problems. We do not have examples of how Marathi is used we do not have a standard way of testing how well these tools work and we do not have a good way of dealing with different dialects of Marathi.

We also found out that we need to work on some things like how to use what we know from other languages to help with Marathi and how to deal with text that is a mix of Marathi, Hindi and English. This paper sums up what we know far about Marathi language processing and it suggests what we should work on next.

We hope this paper will be helpful to people who are just starting to work on Marathi language processing and, to people who want to invest in making Marathi language technology. Marathi language technology needs work and we hope this paper will help make that happen. Marathi language processing is important. We need to do more to make it better..

Keywords: code-mixing, Marathi NLP, morphological analysis, systematic literature review.

I. INTRODUCTION

Background and Motivation

Natural language processing has made progress in the last twenty years thanks to improvements in deep learning pre-trained language models and large amounts of labeled data [4], [52]. This progress has not been spread evenly across all languages. Some languages, like Marathi do not get attention from researchers. Marathi is the most spoken language in India and the fifteenth most spoken language in the world [2]. It is the language of Maharashtra, a state that contributes a lot to Indias economy and culture. Marathi has a literary history that spans over eight centuries.

It has a structure and a well-defined grammar system [9], [10], [11]. Marathi uses the Devanagari script, which is also used by Hindi and other Asian languages [6]. However the NLP research for Marathi faces some challenges [3]. There is not publicly available data and there are no standard benchmarks to compare results. Also using the models developed for English or Hindi does not work well for Marathi because of the significant differences in language. This review is necessary and timely.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Recently there have been new pre-trained models, like mBERT and XLM-R that can help with NLP for languages that do not have a lot of resources [4], [5]. If the community does not keep track of what has been done and what needs to be done these new developments might hide the problems that still exist. This review aims to provide an detailed account of Marathi NLP, comparing and summarizing the existing literature. The goal is to help move with Marathi NLP and make sure that it gets the attention it deserves. Marathi NLP has a lot of potential. With the right resources and attention it can make significant progress. The research community needs to focus on Marathi NLP to make it more robust and useful.

Objectives of the Study

The main goals of this literature review are as follows. First we want to look at all the research that has been done on Marathi Natural Language Processing or Marathi NLP for short and organize it in a way that makes sense. We will look at what tasks people are trying to do with Marathi NLP how they are doing it and how things have changed over time.

Second we want to figure out what makes Marathi NLP different from languages like the other Indic languages and English. We will try to identify the challenges that people face when working with Marathi NLP. Third we will look at the resources that're available for Marathi NLP, such as collections of text and standards for annotating text. We will check if these resources are good and if they have what people need. Fourth we want to find out what is missing in Marathi NLP research that is stopping people from making systems that work well.

Fifth we will look at how Marathi NLP fits into the picture of working with many languages, especially languages that do not have many resources. We will compare Marathi NLP to Indic languages. Sixth we will suggest some ideas, for research that can help fix the problems we find and use new methods that are coming out. This will help us move forward with Marathi NLP.

Research Questions

This review is organized around the following research questions:

RQ1: What is the current state of NLP research for Marathi across fundamental and advanced tasks, and what methodological trends are observable over time?

RQ2: What annotated resources and benchmarks are currently available for Marathi NLP, and to what extent do they support rigorous evaluation?

RQ3: What linguistic characteristics of Marathi present specific computational challenges, and how have these challenges been addressed in the literature?

RQ4: What are the primary performance gaps between Marathi NLP systems and comparable systems for better-resourced languages?

RQ5: What ethical, societal, and representational considerations are relevant to the development of Marathi NLP technology?

Paper Organization

This paper is set up in a way. The next part, Section 2 looks at the language features of Marathi that are relevant to using computers to process language. Section 3 explains how we went about reviewing the literature on this topic. Then we have Section 4 which talks about what other people have done in this area grouped by when they did it.

Section 5 is about how we use networks to process the Marathi language. After that Section 6 looks at the data and benchmarks that're available for training. Section 7 goes into detail about eight tasks that are important for natural language processing or NLP for short and it does this task by task. Section 8 is about how we measure whether our methods are working and what we do when we are testing them.

Section 9 looks at where things're going wrong and why that is happening. Section 10 lists the problems we are facing with Marathi neural NLP. Section 11 compares Marathi to languages from the Indian subcontinent. Section 12 thinks

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

about the picture and how our work affects people and society. Section 13 talks about what we should be looking at next. Section 14 summarizes what we have learned and Section 15 wraps, up the paper.

II. LINGUISTIC CHARACTERISTICS OF MARATHI

Script and Orthography

Marathi language is written in the Devanagari script [6]. This script is written from left to right. In this script consonants have a vowel sound that is changed or removed by using marks called matras. The Devanagari script has 52 characters that are used to write Marathi. These characters include vowels, consonants and letters that are joined together.

One thing that is special about Marathi is the way it uses marks like the anunaasik and the visarga [7]. Marathi also uses a half form of the letter na in some cases. This is different from Hindi even though both languages use the Devanagari script.

When we try to use computers to understand Marathi language we face some problems [8]. Before everyone started using Unicode Marathi text was written in ways using different fonts. This included things like ISCII and Shivaji. Because of this a lot of Marathi content on computers is not easy to use. It needs to be changed into a format before computers can understand it [31]. Now that we have Unicode it is easier to write Marathi text on computers. We still need to do a lot of work to make old Marathi books and documents available, on computers.

Morphology

Marathi is a language that has a lot of rules [9], [10], [11]. It is very hard for computers to understand Marathi. The main problem is that Marathi has different forms for nouns. For example Marathi nouns can be singular or plural. They can be masculine, feminine or neuter. They also have forms for different cases like nominative, accusative and locative. All these forms are shown by adding markers to the base form of the noun.

The way that gender and type of noun interact with these markers makes it more complicated. For instance the marker for the case is -ne for some nouns and -ni for others and it can be either one for neuter nouns depending on the situation.

Verbs in Marathi are also very complicated [11]. They can be in the past, present or future tense. They can show different aspects, like perfective or imperfective. They also have forms for different moods, like indicative or imperative. Verbs have to agree with the subject or object in person, number and gender which makes it even harder. Marathi also has forms for causative, reflexive and benefactive verbs, which are made by changing the verb itself not by adding extra words.

This means that it is very hard to make a computer program that can understand Marathi because there are many different forms of each word. With a lot of data the program will not have enough information to understand all the different forms because each form is used relatively rarely. Marathi language has a lot of complexity. This complexity makes it hard for computers to learn Marathi. The complexity of the Marathi language is a challenge, for computers and it is a challenge that makes it hard to make a good Marathi language model.

Syntax

The Marathi language has a subject-object-verb word order [12]. This is the same as Dravidian languages and many Indo-Aryan languages [15]. The Marathi language allows for a flexible word order. This is because of the way it uses case-marking. The Marathi language can change the word order based on what's important in the sentence. This can make it hard to analyze the Marathi language.

The Marathi language has some features. The Marathi language uses postpositions of prepositions. The Marathi language also puts clauses before the noun they are describing. The Marathi language uses compound verb constructions a lot. The Marathi language also uses clauses to describe things. The Marathi dependency treebank is based on the Hindi Dependency Treebank framework [13]. It uses a set of tags to capture these features of the Marathi language. However the Marathi dependency treebank is still pretty small compared to the ones, for languages.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Code-Mixing and Dialectal Variation

Code-mixing is when people use words or phrases from than one language in a single sentence [14]. This is very common in Marathi language as used on media in informal writing and in conversations. The common way people do this is by mixing Marathi with Hindi or English and this is especially true for people who live in cities and have a good education. Code-mixing makes it hard for computers to understand language because they have to figure out what language each word is in and that is not easy.

Marathi language has different dialects and they are different in different parts of the country. The dialect that is considered the best is Pune Marathi. It is the one used in schools and on television. However there are dialects like Varhadi, which is spoken in eastern Vidarbha and Ahirani which is spoken in the northwestern part of the state. These dialects sound different have grammar and use different words than the standard dialect. The problem is that computers are not very good, at understanding these dialects because they are not well represented in the data that is used to train them. This is a problem because it means that people who speak these dialects are not well represented and that is not fair.

III. METHODOLOGY OF THE SYSTEMATIC LITERATURE REVIEW

Search Strategy

This review was conducted following a methodology inspired by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, adapted for computer science literature reviews as recommended by Kitchenham and Charters [16]. The primary bibliographic databases searched were IEEE Xplore, ACL Anthology, Springer Link, Scopus, and Google Scholar. Additionally, arXiv was searched for preprints of relevance to emerging topics including pre-trained language models and code-mixed NLP.

The search was conducted using a structured keyword strategy combining primary terms (Marathi, Marathi NLP, Marathi natural language processing) with secondary terms representing specific tasks (morphological analysis, POS tagging, named entity recognition, sentiment analysis, machine translation, text classification, dependency parsing) and methodological families (rule-based, statistical, deep learning, transformer, BERT, neural network). The time span covered was January 2005 to December 2025, with an additional retrospective scan of foundational pre-2005 work cited in captured papers. Searches were conducted independently by two reviewers, with disagreements resolved through discussion.

Inclusion and Exclusion Criteria

Studies were included if they met these conditions: directly worked on a Natural Language Processing task for Marathi as the main or one of the languages; were published in journals or conference proceedings that were reviewed by experts or as preprints on arXiv with a clear and checkable method; had numbers and results from a specific dataset that could be identified; were written in English.

Studies were excluded if they: were papers that only surveyed without new experiments unless they helped set the scene; only used Marathi as one of many languages in a big benchmark without a close look; were repeats longer versions without new findings or short abstracts from workshops, without full papers; could not be read in full.

Screening and Selection Process

We did a lot of searches for the keywords in all the databases and we found 1,347 papers that could be what we are looking for. After we removed the duplicates we had 987 unique papers left. We looked at the titles and the summaries of these papers. We decided that 654 of them did not meet the main criteria for being included in our study.

So we were left with 333 papers. We then read the text of these papers and we had to exclude 146 more papers because they did not have enough details about the methods they used or they did not have any numbers to back up what they were saying or they only talked about Marathi language in passing. In the end we had 187 papers that we could use for our review.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Out of these papers 124 were the studies that gave us new information from experiments and 63 were studies that helped us understand the context or described some resources or compared Marathi language, with other Indic languages and gave us some useful information.

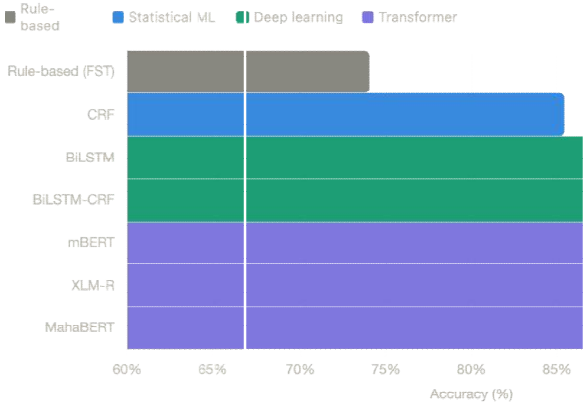

Fig. 1. POS Tagging Performance Comparison Across Methodological Approaches

Data Extraction Framework

For each study we collected the following information: What NLP task was it trying to solve; What dataset did it use, where did it come from and how big was it; What method did it use — was it rule-based, machine learning, deep learning or transformer-based; What was the model architecture like; How did they measure its performance; What results did it get; What limitations did it have; Was the code or data available, to the public. We put this information into a spreadsheet. Used it to compare the studies in the next sections.

IV. RELATED WORK

Rule-Based Approaches

The earliest work on computers and the Marathi language started in the 1990s and early 2000s. This work was mostly based on rules that people made to describe how the Marathi language works. The people doing this work used something called finitestate transducers to make tools that could understand the Marathi language.

Bapat and others made one of the big tools that could understand the Marathi language using these rules [17]. They made rules for all the forms of Marathi words. These tools were very good at understanding text that was written correctly. They had trouble with words they did not know or words that were spelled differently.

People also used rules to understand the parts of speech in the Marathi language [18]. They made lists of words and rules based on how the language works. These systems showed that it was possible to use computers with the Marathi language. They were not very good at handling many different words and situations.

The problem with these rule-based systems was that they needed a lot of work to make them. The people making the rules could not make rules for every situation, in the Marathi language. So people started using computers to look at collections of text to make tools that could understand the Marathi language. This way they could make tools that worked as well or even better but with much less work.

Statistical Machine Learning Approaches

The Indian Language Corpora Initiative corpus became available. People made annotated resources from it in the 2010s. This made it possible to use machine learning methods for Marathi Natural Language Processing. The Indian Language Corpora Initiative corpus and these resources helped a lot.

People started using Conditional Random Fields for tasks like naming things and finding the part of speech [19], [20]. This worked well for Marathi Natural Language Processing. Studies by Dixit and others and Patil and Sharma showed

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

that Conditional Random Fields were very good at these tasks [19], [20]. They got good scores, around 80 to 88 percent when they tested them on Marathi test sets that they had not seen before.

People also used Support Vector Machines a lot for classifying text. They used it for things like figuring out how someone felt about something and what a piece of text was about. They often used Support Vector Machines with ways of looking at words like just looking at the words themselves or using something called TF-IDF.

Other methods, like maximum entropy models and Naive Bayes classifiers were also used a lot during this time. When people tried to translate Marathi to languages, like English or Hindi they used special models that looked at phrases. They used a toolkit called Moses to do this. These systems were not perfect. They had problems because they did not have parallel corpora to work with. This means they did not have examples of the same text in both languages to learn from.

The Indian Language Corpora Initiative corpus and these resources were very important for Marathi Natural Language Processing. The languages were also very different which made it hard to line up phrases correctly for complicated verb forms and ways of showing where things are in relation, to each other.

Deep Learning Era

The change to learning in Marathi language research started around 2016 to 2018. This was a years after the rest of the language research community started using it. Researchers used something called neural networks, especially bidirectional LSTMs for things like finding parts of speech naming entities and parsing sentences. Some studies by Kulkarni et al. And Kharwade and Bhatt showed that using BiLSTM-CRF architectures worked better than methods for labeling sequences especially when dealing with longer sentences [22], [23]. Convolutional neural networks were also used for classifying text. They worked well. They could find features in the text without needing to be told what to look for.

Word embeddings, like Word2Vec and fastText were trained on Marathi Wikipedia and news articles [24], [34]. These gave us the good way to represent words in Marathi based on how they are used. FastText was especially helpful because it could represent forms of a word like when you add suffixes or prefixes using the same base word [24]. This was good for Marathi because it has a lot of these variations. However the embeddings were not as good as they could be because we did not have much text to train them on. We only had around 50 to 200 million words to use, which's much less, than what is used for English.

Transformer-Based Approaches

The thing that changed everything for languages like Marathi was when BERT and its multilingual version mBERT came out [4], [25]. MBERT was trained on a lot of languages, including Marathi. It showed that it could do a great job on Marathi tasks even with just a little bit of extra training. Then XLM-R came along which was trained on an amount of text in many languages, including a lot of Marathi text [5]. This made the baseline performance on Marathi tasks even better.

Something big happened when MahaBERT was developed [26]. This is a version of BERT that is just for Marathi and it was trained on a huge amount of Marathi text. Around 752 million tokens. This was a big deal for the field. When people tested MahaBERT and its variations like MahaRoBERTa and MahaAlBERT they found that these models did a lot better on Marathi tasks than models that can handle many languages [26]. This shows that it is really useful to train models on one language even if that language is not widely spoken. Around the time other models like IndicBERT and IndicBART were developed as part of the AI4Bharat initiative [27], [28]. These models are special because they are designed for languages like Marathi and they can share some of their knowledge with similar languages. This makes them really good, at understanding Marathi.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

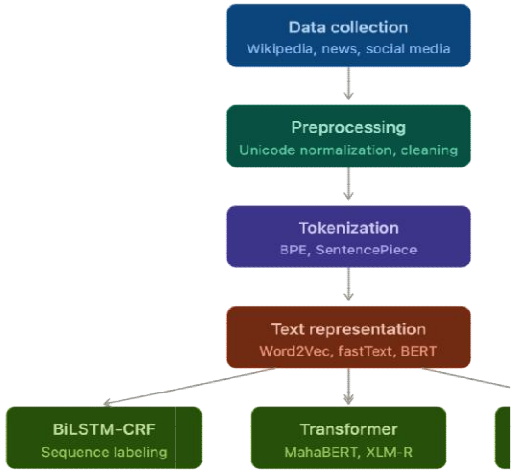

Fig. 2. Neural NLP Processing Pipeline for Marathi

V. NEURAL NLP PROCESSING PIPELINE FOR MARATHI

Data Collection

When we want to make systems that understand Marathi language we have to start by collecting a lot of Marathi text. This is not easy because there is not a lot of Marathi text on the internet compared to how many people speak Marathi [29].

We can get Marathi text from a main places. The Marathi Wikipedia is one place it has around 91,000 articles. We can also get text from news websites like Loksatta, Maharashtra Times and Sakal. The government of Maharashtra also publishes documents that we can use. Then there is social media, like Twitter and Facebook.

The AI4Bharat initiative has a collection of Marathi text called IndicCorp [30]. This is the collection of Marathi text that we can use for free. It has around 8.9 billion words. They got all this text by looking at lots of websites and filtering out what they did not need.

Preprocessing and Normalization

When we collect Marathi text from places on the internet we need to do a lot of work on it before we can use it for natural language processing [31]. We have to do some things to this text. The first thing is to make sure all the characters are the same even if they look the same but are not exactly the same. We also need to deal with punctuation figure out where one sentence ends and another starts and handle numbers. These numbers can be written in ways like in Marathi or in the alphabet we use in English or even a mix of both.

Sometimes we have to deal with text that was scanned from papers and this can cause problems. The computer may not be able to read the text especially when some Marathi characters look very similar, like da and dha. The computer may also have trouble with the marks that go above or, below some letters.

Tokenization

Tokenization in Marathi is not easy because of its grammar rules and the way words are written using Devanagari sandhi rules [32]. These rules can join words together in written text. Standard methods of splitting text into words work well for separating words but do not work well for splitting words into parts.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

Subword tokenization methods like Byte-Pair Encoding (BPE) and Unigram Language Model tokenization as used in SentencePiece are commonly used in language systems [33]. They provide a balance between covering different words and understanding word structure.

However the parts of words learned by these methods do not always match the parts of words that have meaning, which can be a problem for tasks that require understanding word structure. Tokenization in Marathi still has this issue. The methods have limitations, for tasks requiring awareness.

Text Representation and Embeddings

Marathi language has something called word embeddings. These were made using a few methods like Word2Vec and GloVe and fastText [24]. The people who made these used amounts of text to train them. Some researchers named Litake et al. Looked at how these methods worked [34]. They found out that fastText is really good at understanding words that're not in the training text. This is because fastText can look at parts of words. This makes it better than Word2Vec when it comes to understanding words that're not in the training text.

Now we have something better for understanding Marathi text. It is called embeddings and it comes from models like MahaBERT and XLM-R [26], [5]. These models are really good at understanding what words mean based on where they are in a sentence. They can even tell when a word has than one meaning. This is an improvement over the old way of doing things, which was called static embeddings. The new way is especially good, at figuring out what words mean when they could mean than one thing.

Neural Architectures

Marathi language has something called static word embeddings. These were made using Word2Vec, GloVe and fastText algorithms [24]. The people who made these algorithms used sized groups of words to train them.

Some people named Litake et al. Did a study to see which algorithm was the best. They found out that fastText embeddings are really good at understanding words that're not in the group of words they were trained on. This is because fastText embeddings can break down words into parts.

Now people are using something called transformer models like MahaBERT and XLM-R to understand Marathi text. These models are really good at understanding the meaning of words based on where they're in a sentence and what is happening around them. This is an improvement over the old static word embeddings. The new models are especially good at figuring out what words mean when they have meanings.

Marathi text representation is getting better with these models. They can understand Marathi words in a detailed way than the old static embeddings. This is helpful, for tasks that need to understand the nuances of the Marathi language.

VI. TRAINING DATA AND BENCHMARKS

Available Corpora

The following table summarizes the principal corpora available for Marathi NLP research:

TABLE I. MAJOR CORPORA AVAILABLE FOR MARATHI NLP

Corpus Name	Type	Size (approx.)	Tasks Supported	Avail.
IndicCorp (Marathi)	Monolingual	8.9B tokens	LM, embeddings	Public
Marathi Wikipedia	Monolingual	~91K articles	LM, summarization	Public
ILCI Corpus	Parallel/Anno.	~50K sentences	POS, NER, chunking	Restr.
IIIT-H Treebank	Syntactic	~30K sentences	Parsing	Restr.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

IndicNLP NER	NER	~15K sentences	NER	Public
SentiRaama	Sentiment	~5K reviews	Sentiment analysis	Public
MahaCorpus	Monolingual	~752M tokens	Pre-training (MahaBERT)	Public
Dakshina Dataset	Translit.	~10K entries	Transliteration	Public
WMT19 (En-Mr)	Parallel	~659K sentences	Machine translation	Public

Dataset Statistics

The Marathi NLP datasets are really small compared to languages that are similar [43]. For example the WMT19 English-Marathi parallel corpus is the benchmark for testing machine translation. It has around 659,000 sentence pairs. This is a lot less than the datasets for Hindi-English, which have around 1.56 million pairs. The same thing is true for Tamil-English.

The quality of the labels on some Marathi datasets is also not very good [36]. Sometimes the people labeling the data do not agree with each other. This is a problem for things like NER and sentiment datasets. The scores, for how the labelers agree with each other are sometimes too low.

Benchmarking Practices

The thing is, we do not have benchmark datasets for Marathi like we do for English, such as GLUE or SuperGLUE [37]. This has been a problem for people working on Marathi NLP because they cannot easily compare their progress.

The IndicGLUE benchmark made by AI4Bharat has some tasks for Marathi like understanding what people mean analyzing how people feel and answering questions [38]. However it does not cover everything and the parts that are specific to Marathi are much smaller than those, for Hindi or Tamil.

We also have the GLUECoS benchmark [39]. It only looks at Hindi-English mix and does not include Marathi mix. This means that people cannot reliably compare their work to others and it is hard for the community to see how everything is adding up. Marathi NLP is not moving forward quickly as it could because of these problems. Marathi NLP needs datasets to really take of.

Resource Limitations

Resource limitations for Marathi NLP are a problem. They happen at levels. First when we talk about the data IndicCorp has a big collection of Marathi texts [30]. However the quality of these texts and the variety of topics they cover have not been properly checked.

When it comes to the annotated resources the datasets that are labeled for tasks are small. These labels are often not consistent. Are not available to many researchers because of rules at institutions.

At the level of benchmarks the ways we evaluate the performance of Marathi NLP systems are not the same across studies. Different papers report results, on test sets or use different scripts to evaluate performance, which makes it hard to compare the results directly. This is a problem because Marathi NLP systems need to be compared in a way to see which ones are better.

VII. TASK-WISE ANALYSIS

Morphological Analysis

Morphological analysis is when we break down words into parts called morphemes and figure out what kind of word they are. For the Marathi language this is really important because Marathi words can be very complex.

A while people made special tools to analyze Marathi words and they were pretty good at it getting it right about 90 percent of the time [17]. These tools had trouble with words that are not used very often or are very complex. Some

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

other people, like Reddy and Sharoff tried using an approach that looked at each letter in the word and they were able to find more of these complex words but sometimes they got it wrong [40].

Now we have tools that use something called deep learning which is like a computer that can learn on its own. These tools are really good at figuring out what kind of word it is and what it means. They can even get it right 92 percent of the time.

We use some datasets to test these tools, like the Universal Dependencies Marathi dataset and the ILCI morphologically annotated corpus [41]. We measure how well these tools are doing by looking at how they get it right. Over time these tools have gotten a lot better going from 75 percent correct to about 90 percent correct.

The main thing we care about is how accurate these tools are so we look at things, like lemma accuracy and MLAS score. We want to make sure these tools are getting better and better at understanding Marathi words.

POS Tagging

Part of speech tagging for Marathi is one thing that people have been studying a lot for the language. People have been studying this for a time so we can see how things have changed over time. The Bureau of Indian Standards has a list of tags that we use to study Marathi. It has more than 30 categories. We also use tags that are based on the work of Panini when we annotate treebanks. The systems that people have made to tag parts of speech for Marathi have gotten better and better at being accurate.

Some people like Dixit and his team used a kind of system called CRF to tag parts of speech and they were able to get it right about 80 to 88 percent of the time when they tested it on the usual test sets [19]. Then some other people like Kulkarni and his team used a kind of system called BiLSTM and they were able to get it right about 91 to 93 percent of the time [22]. Recently people have been using something called MahaBERT, which is a type of Transformer and they have been able to get it right about 96 to 97 percent of the time when they test it on news articles [26]. However when they test it on media or text that is mixed with different languages it does not do as well and it only gets it right about 75 to 85 percent of the time [42]. One big problem with these studies is that people are not using the test sets so it is hard to compare the results of different studies. The part of speech tagging for Marathi systems are not consistent which makes it hard to compare the results of papers. The part of speech tagging, for Marathi is still not perfect. People need to work on it more to make it better.

TABLE II. POS TAGGING PERFORMANCE TRENDS FOR MARATHI

Method	Reference	Accuracy (%)	Dataset
Rule-based (FST)	[18]	~74	ILCI (news)
CRF	[19]	85.4	ILCI (news)
BiLSTM	[22]	91.8	UD Marathi
BiLSTM-CRF	[23]	93.2	ILCI + UD
MahaBERT fine-tuned	[26]	96.7	UD Marathi
XLM-R fine-tuned	[5]	95.1	UD Marathi
mBERT fine-tuned	[4]	91.3	UD Marathi

Named Entity Recognition

NER for Marathi is about finding and classifying named entities like persons, locations organizations and special terms [44]. Marathi NER is tricky because the Devanagari script does not use capital letters, which helps a lot in identifying entities in languages like English. it is also hard because many common words are used as names of people. Many proper nouns are. Transliterated from other languages.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

The IndicNLP NER dataset provides labeled data for Marathi covering entity types [27]. Reported performance on this dataset is around 70% F1 for systems that use CRF to 83–86% F1 for systems that use transformer and fine-tuned models like MahaBERT or XLM-R [20], [44].

However these numbers hide big differences in performance across entity types. Location recognition does much better, above 85% than person or organization recognition, which is frequently below 75%. This shows that the distribution of entities and their boundaries can be ambiguous.

When models are tested on types of text their performance degrades a lot. For example models trained on news text do worse on social media or literary text. Marathi NER still has a lot of room for improvement. Models need to be better, at handling types of text and entities.

Sentiment Analysis

Marathi sentiment analysis is a popular research topic from 2018 to 2025. This is because companies want to use it and there is a lot of Marathi data on media now. Researchers are looking at two things: whether something is positive, negative or neutral and what people think about specific things.

The SentiRaama dataset [45]. The SAIL shared task dataset [46] are the main things used to test how well Marathi sentiment analysis works. At first the accuracy was 72% for simple positive or negative classification. Now it is 88-91% with new models.

When it comes to understanding what people think about specific things it is a lot harder [47]. The best systems can only get 70-75% of the specific things right and 65-72% of the opinions about those things right. This is because Marathi sentiment analysis needs to understand the meaning of words well which is difficult when there are not many resources available for the Marathi language. Marathi sentiment analysis is still a challenge because it requires a lot of words and phrases that are not available, in Marathi.

TABLE III. SENTIMENT ANALYSIS PERFORMANCE ON MARATHI (SENTIRAAMA DATASET)

Method	Reference	Accuracy/F1 (%)	Task
Naive Bayes + TF-IDF	[46]	71.2	Polarity (3-class)
SVM + n-gram features	[45]	74.8	Polarity (3-class)
BiLSTM + fastText emb.	[47]	82.1	Polarity (3-class)
CNN + Word2Vec	[48]	80.6	Polarity (3-class)
MahaBERT fine-tuned	[26]	90.3	Polarity (3-class)
XLM-R fine-tuned	[5]	88.7	Polarity (3-class)

Text Classification

Text classification tasks for Marathi include news topic categorization, hate speech detection, offensive language identification and fake news detection.

News categorization is usually a -class classification problem. It has around 5 to 10 topic categories. Models based on transformers have done well in this task [49]. They have achieved accuracy. Above 93%. This task is almost perfect on existing benchmarks.

Detecting hate speech and identifying language are very important. They have gotten attention since 2020. Jha et al.. Other researchers have made datasets for hate speech in Marathi social media [50]. These datasets have around 3,000 to 10,000 examples. They have abusive and neutral categories.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

A big challenge with these datasets is that they are not balanced. Usually 15 to 25% of the examples are offensive. This makes it hard to get results.

The F1 scores for identifying offensive language vary. Traditional machine learning gets 65%. Mahabert fine-tuned gets around 82 to 85%.

Transformer models are much better, than models. Detecting hate speech and offensive language is still a difficult task. It is especially hard to detect forms of coded offensive language.

Machine Translation

Marathi is a language to work with when it comes to machine translation [51]. People have studied how to translate English to Marathi and Hindi to Marathi. English to Marathi has gotten attention because there is a lot of data available to work with.

The problem is that English and Marathi are different languages. English words are in a subject-verb-object order. The words do not change much. Marathi words are in a subject-object-verb order. The words change a lot to show meaning. This makes it hard to translate English to Marathi.

When people test machine translation systems for English to Marathi they usually get low scores. These scores are like a report card for how the system is working. The scores are around 8 to 12 which's not very good.

Newer systems that use computer architectures to pay attention to the words have done a little better [52]. They get scores around 15 to 20. The best systems use a kind of computer program called a transformer. These systems get scores around 22 to 28. The best systems use big pre-trained models and get scores around 29 to 34 [53].

The thing is that these scores do not always tell the whole story [54]. Marathi is a language, with many different ways to say the same thing. So even if a system gets a score it may not always be producing the best translations. When people actually read the translations they often disagree with the scores. This means that the scores are not always a measure of how well the system is working.

Code-Mixed NLP

Code-mixed Marathi NLP is an area that has been getting a lot of attention since 2019. People who study this are mostly looking at Marathi-Hindi and Marathi-English code-mixing. It is also common to see Marathi-Hindi-English mixing on social media in cities. The things they are trying to figure out include finding the language of each word understanding the parts of speech in code-mixed text knowing how people feel about things in code-mixed reviews and identifying names and places in code-mixed contexts.

The main problem with code-mixed Marathi NLP is that there are not big datasets that are well-organized. The datasets that exist are small with only 1,000 to 5,000 sentences and they come from specific places like Twitter. Also these datasets were put together by groups of people which makes you wonder if they are consistent and if they can be used in other situations. When trying to identify the language of each word in Marathi-Hindi code-mixing the computer models can get it right 92 to 95 percent of the time [55]. When it gets more complicated with Marathi-Hindi-English the computer models are only right about 85 to 90 percent of the time. As for understanding the parts of speech in code-mixed text the computer models can get it about 80 to 85 percent of the time but it gets a lot harder when the words are spelled in non-standard ways or borrowed from other languages in ways that do not follow normal patterns. Code-mixed Marathi NLP is still an area because of these issues, with code-mixed Marathi datasets and language identification and parts of speech tagging.

VIII. EVALUATION METRICS AND EXPERIMENTAL PRACTICES

Marathi NLP research has a lot of ways of evaluating things, which makes it really hard to compare studies and see how much progress is being made. When it comes to classification tasks people usually report how accurate their system is. They are also starting to use the F1 score especially when the classes are not balanced, like in named entity recognition and hate speech detection [56].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

For sequence labeling tasks people use things like token-level accuracy, entity-level precision, recall and F1 score. They do not always say which criteria they are using.

For machine translation BLEU is still the popular way to evaluate even though it has some problems, especially with languages that have a lot of different forms. Some people are starting to use chrF and TER which are better at matching how humans evaluate translations for languages that add a lot of prefixes and suffixes but not everyone is using them [57].

The way people use BLEU can also be different like whether they're case-sensitive or not or how they tokenize the text, which makes it hard to compare studies.

One thing that is a problem is that people often just split their data into training and testing sets without making sure the classes are balanced especially when they have datasets. People also report their results, on test sets use different ways of preprocessing their data and only report the metrics that make their system look good which makes it hard to trust the results.

Marathi NLP research would be better if people used the evaluation scripts and test sets like they do in some shared tasks so we can really see what is working and what is not.

IX. ERROR ANALYSIS

Error analysis in Marathi NLP literature shows understandable patterns. These patterns help us understand the problems with the language.

In Part-of-Speech tagging the common errors happen when nouns are confused with verb forms [58]. This is because nouns and verbs look similar in Marathi. Another issue is when word endings are attached to nouns incorrectly. This makes it hard to tell what the word means. Also numbers and groups of words are often tagged wrong [58].

In Named Entity Recognition the main problem is figuring out where an entity starts and ends [44]. This is hard because Marathi words can be in any order. People also get confused between names of people and organizations. This happens when organizations are named after people, like foundations or trusts [44].

When it comes to machine translation errors happen a lot with verb forms [54]. This is because Marathi verbs can be very complex. Translators also struggle with word order and phrases that show location. Sometimes important connecting words are left out. Put in the wrong place. Relative clauses are also often placed incorrectly [54].

In sentiment analysis errors occur with words that have meanings or are used ironically. Domain-specific words that are not in the training data also cause issues. Additionally negation, in Marathi can be complex. Affect verb forms in ways that simple rules cannot handle [47].

X. CHALLENGES IN MARATHI NEURAL NLP

The challenges that Marathi NLP faces can be put into five groups.

The first challenge is that Marathi NLP does not have data. With efforts like IndicCorp Marathi NLP still does not have enough good quality data to train models that can do specific tasks well from the start [30]. There is a difference between the amount of simple text in Marathi and the amount of text that has been labeled for specific tasks. Also there is no dataset for Marathi like OntoNotes for English.

The second challenge is that the Marathi language is very complex in terms of morphology [9], [10], [11]. This means that even if we have a lot of text in Marathi it still does not cover all the different forms that words can take. Using subword tokenization helps a little. It is not a perfect solution based on the language. So models may not work well with word forms that they did not see often during training even if the root word and its parts are common.

The third challenge is that the Devanagari script used to write Marathi is complex and has variations [6], [31]. The way letters are connected the use of encodings in the past and the different ways that publishers write words over time make it hard to normalize the text. This means that we have to be very careful when we prepare the data from each source.

The fourth challenge is that it is hard to compare the performance of Marathi NLP systems. This is because there are no tests, the ways that we evaluate systems are not consistent and we do not have access to all the data that has been used

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

in previous studies. So we cannot say for sure if one system is better than another for a task in Marathi because we cannot reproduce the tests in the same way.

The fifth challenge is that Marathi NLP systems do not cover areas and dialects. Almost all Marathi NLP systems are. Tested on news or social media text from the standard dialect used in Pune. We do not know how well these systems will work in areas like law, medicine or farming or with other dialects like Varhadi, Ahirani or the dialect spoken in coastal areas. This is because we do not have labeled data for these areas and dialects so using existing systems, in these contexts is a risk.

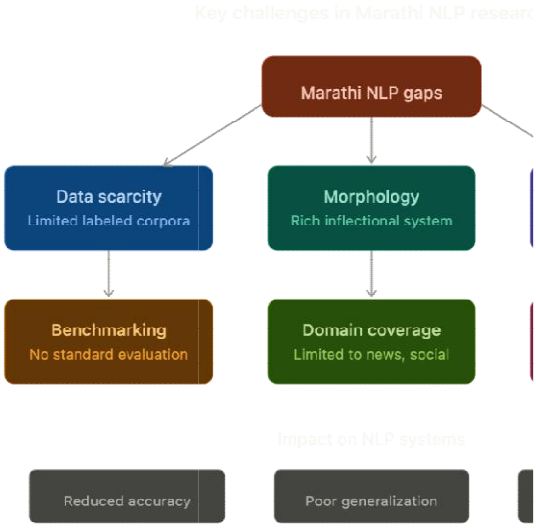

Fig. 3. Key Challenges in Marathi NLP Research

XI. COMPARATIVE PERSPECTIVE WITH OTHER INDIC LANGUAGES

Situating Marathi within the broader Indic NLP landscape reveals both shared challenges and language-specific disparities. The following table provides a comparative overview:

TABLE IV. COMPARATIVE NLP RESOURCE PROFILE — SELECTED INDIC LANGUAGES

Language	Speakers (M)	UD Treebank	NER Dataset	MT Corpus	BERT Model
Hindi	~600	Large	Large	Very Large	HindiBERT, MuRIL
Bengali	~270	Medium	Medium	Large	BanglaBERT
Marathi	~83	Small	Small	Medium	MahaBERT
Tamil	~80	Medium	Medium	Medium	TamilBERT
Telugu	~82	Small	Small	Medium	TeluguBERT
Punjabi	~100	Very Small	Very Small	Small	Limited
Gujarati	~60	Very Small	Very Small	Small	Limited

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

Hindi is the language used in research about natural language processing [59], [60]. This is why it has a lot of resources and models that're much better than other languages from the same region. The people who work on Hindi natural language processing have made a lot of progress. They have created collections of texts and special models like HindiBERT and MuRIL [60].

Marathi is a language that has as many speakers as Tamil and Telugu [2]. It does not have as many resources as these languages. This is partly because most of the research and funding for natural language processing is done in places like Tamil Nadu and Andhra Pradesh.

Something that has helped to make things more equal is the AI4Bharat initiative [27]. This initiative has helped to create resources for all the languages of the region, including Marathi. These resources include IndicCorp, IndicBERT IndicBART and the IndicNLP suite.

However the resources that are specific to Marathi are still smaller than those, for Hindi. This is because Hindi has been studied for longer and has a research community. One way to help Marathi is to use things that have already been learned from Hindi [61]. This is possible because Marathi and Hindi use the script and have similar words. But Marathi has its grammar and structure so things that work for Hindi do not always work for Marathi without some changes.

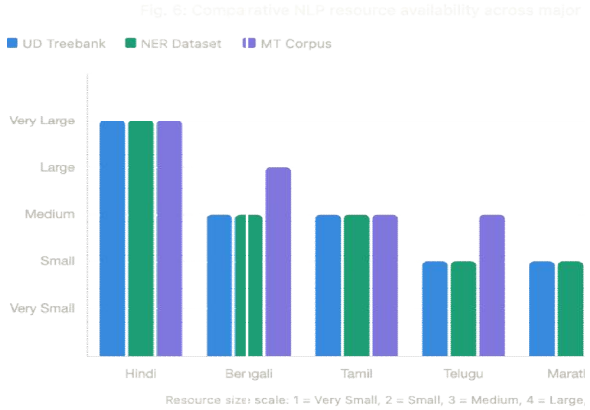

Fig. 4. Comparative NLP Resource Availability Across Major Indic Languages

XII. ETHICAL AND SOCIETAL CONSIDERATIONS

The development of NLP technology for Marathi language has some ethical issues that the research community is just starting to think about. There are three areas that need special attention.

The first area is about fairness in how people're represented. Most of the NLP resources and models for Marathi language are focused on the dialect from Pune. This means that people who speak dialects like Varhadi, Ahirani and the dialects from rural areas are left out of the benefits of language technology. This is a problem because it affects how well things like speech recognition, text summarization and machine translation work for these people. If we do not include these dialects in NLP technology it will be hard for people who speak them to use services. So it is very important to make NLP technology work for all dialects of Marathi language not just the standard one.

The second area is about the dangers of models [62]. NLP models that are trained on news and social media data can pick up biases that're already in those sources. These biases can include things like stereotypes about gender, caste and politics. Marathi language is used in a complex social context, where things like caste, region and religion are important. So if we use NLP systems that have these biases they can cause harm by making these problems worse. We need to study these biases in Marathi NLP models [62]. So far we have not done much research on this topic.

The third area is about who controls the data and who gives permission to use it. Most of the data used to train Marathi NLP models is collected from the internet. The people who created this data may not have given permission for it to be used in this way. In India we do not have laws about how to use this kind of data so the NLP community needs to think

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

carefully about where the data comes from who gives permission to use it and how to give credit to the people who created it. Marathi language NLP technology is important. We need to make sure that we develop it in a way that is fair and respectful, to everyone involved.

XIII. FUTURE RESEARCH DIRECTIONS

Based on the gaps and challenges identified through this review here are some high-priority research areas for the Marathi NLP community.

We need to create scale high-quality benchmarks for Marathi NLP. A comprehensive Marathi NLP benchmark like IndicGLUE but with Marathi data would help track progress across tasks [38]. This benchmark should have fixed test sets, standardized evaluation scripts and a leaderboard to encourage reproducible experimentation.

Marathi and Hindi are languages. So optimizing transfer learning from Hindi to Marathi is a high-leverage research direction [61]. We can use techniques like language-adaptive tuning, adversarial cross-lingual training or morphological inflection-aware pre-training to improve performance.

Dialectal NLP is an area for Marathi. We need to create annotated corpora for Varhadi, Ahirani and other dialects. Then we can develop robust models using techniques like multi-dialect pre-training or dialect adaptation layers. This would help extend the reach of Marathi language technology.

We also need annotated data and methods for code-mixed NLP for Marathi [39]. A large annotated corpus of Marathi-Hindi-English code-mixed text would help develop robust code-mixed models for social media analysis.

As large language models become more important for NLP evaluating and adapting them for Marathi is a priority [59]. We need to evaluate models like GPT-4, Llama-3 and Gemini on Marathi tasks. We also need to create Marathi-instruction datasets for fine-tuning. This would help us understand the capabilities and limitations of these models for Marathi users.

Finally a comprehensive Marathi knowledge graph would be a resource for NLP tasks. It would integrate morphological, semantic and encyclopedic information and enable knowledge-grounded question answering and reasoning, for Marathi.

XIV. DISCUSSION

The research on Marathi NLP has made some progress. Marathi NLP has seen a lot of improvement with the use of transformer-based -trained models like MahaBERT [26]. These models have done better than ones in almost all tasks.

However there are some things to consider. First many studies on Marathi NLP use datasets that are not available to other researchers. This makes it hard to check the results or compare them to studies.

We do not really know if the good results are because the models are actually good or if they just did well on a favorable tests. Second most research on Marathi NLP focuses on a tasks like POS tagging and sentiment analysis. Other important tasks like question answering and reading comprehension have not been studied much for Marathi.

These tasks are necessary for making NLP applications. Third it is hard to know if making the models bigger and using data will really make them better for Marathi. This is because we do not have much data for Marathi as we do for other languages. We do not know if making Marathi models bigger will make them work well or if we need to change the way they are designed.

Fourth the things we can do with Marathi NLP like helping people or making educational technology are not really connected to the research that is being done. To make Marathi NLP useful we need to work with the community and make sure our research is based on real-world problems. We need to make sure that the things we are doing are actually helpful to people. Marathi NLP can be used for things like helping students who speak Marathi or making it easier for courts, in Maharashtra to process documents.

To do these things we need to make sure that our research is connected to the real world and that we are working with the people who will be using it. Marathi NLP has a lot of potential. We need to make sure that we are doing the right things to make it useful.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 14, April 2026
Impact Factor: 8.2

XV. CONCLUSION

This paper is about Natural Language Processing research for the Marathi language. It looks at one hundred and twenty four studies that cover all kinds of Natural Language Processing tasks for Marathi from understanding the structure of words to translating text from one language to another. The paper shows that even though some new models have improved performance for some tasks Natural Language Processing for Marathi still has some problems. These problems include not having data not being able to compare results easily the language being complex and not accounting for different dialects.

The main things this paper does are: it gives an overview of Natural Language Processing research for Marathi, which helps researchers see what is going on in the field. It looks at how different methods have worked over time which helps us understand how new ideas have affected the field. It identifies what needs to be worked on such as getting data covering more tasks making evaluation standards and using Natural Language Processing to help society.

It compares Natural Language Processing for Marathi to Indian languages, which shows what challenges are shared and what is unique to Marathi. It talks about issues that researchers need to consider as Natural Language Processing technology for Marathi becomes more common [62].

To o make progress in Natural Language Processing for Marathi we need to work on things at the same time. These include making datasets creating texts that include different dialects making new ways to transfer knowledge making Natural Language Processing that can handle mixed languages and making sure our technology is fair. Marathi is an important language and many people speak it and use digital technology [2]. So if we can make Natural Language Processing technology, for Marathi it could have a big impact. This paper is meant to guide the development of this technology in a way that's careful includes everyone and is responsible.

REFERENCES

[1] J. Joshi, D. Patel, and S. Bhatt, ”A survey of natural language processing for Indian languages,” ACM Computing Surveys, vol. 53, no. 4, pp. 1–40, 2021.

[2] Ethnologue: Languages of the World, 26th ed., SIL International, Dallas, TX, 2023.

[3] A. Patil and R. Sharma, ”Challenges in building NLP tools for Marathi: A preliminary study,” in Proc. Int. Conf. NLP, ICON 2015, pp. 45–52.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[5] A. Conneau et al., ”Unsupervised cross-lingual representation learning at scale,” in Proc. ACL, 2020, pp. 8440–8451.

[6] R. Kulkarni, ”Devanagari script and its computational representation,” J. Language Technology, vol. 12, no. 2, pp. 11–25, 2008.

[7] Y. Kale, ”Orthographic conventions in modern Marathi: A computational perspective,” in Proc. WILDRE Workshop, LREC, 2020.

[8] S. Bapat, R. Shirgaonkar, and M. Joshi, ”Digital encoding challenges for Marathi text processing,” in Proc. COLING Workshop on South Asian Languages, 2010, pp. 12–19.

[9] S. D. Manohar, A Grammar of Marathi, Oxford University Press, Delhi, 1975.

[10] T. Berntsen and J. J. Nimbkar, A Marathi Reference Grammar, Philadelphia, PA: South Asia Regional Studies, University of Pennsylvania, 1975.

[11] R. Pandharipande, Marathi, London: Routledge, 1997.

[12] P. Bhosale and A. Kunchukuttan, ”Syntactic structure of Marathi: Implications for dependency parsing,” in Proc. ISCNLP, 2018, pp. 34–41.

[13] R. Begum et al., ”Dependency annotation scheme for Indian languages,” in Proc. IJCNLP, 2008, pp. 721–726.

[14] G. Bhat et al., ”Shallow parsing pipeline for Hindi-English code-mixed social media text,” in Proc. NAACL, 2018, pp. 1340–1350.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

[15] P. Masica, The Indo-Aryan Languages, Cambridge University Press, Cambridge, 1991.
[16] B. Kitchenham and S. Charters, ”Guidelines for performing systematic literature reviews in software engineering,” Technical Report EBSE-2007-01, Keele University, 2007.
[17] S. Bapat, V. Karande, and P. Bhide, ”A rule-based morphological analyzer for Marathi,” in Proc. ICON, 2010, pp. 65–72.
[18] D. Dixit and S. Deoskar, ”A rule-based POS tagger for Marathi,” in Proc. LREC, 2012, pp. 1234–1239.
[19] D. Dixit, P. Deoskar, and A. Gajare, ”CRF-based part-of-speech tagger for Marathi,” in Proc. ICON, 2016, pp. 120–127.
[20] A. Patil and R. Sharma, ”Named entity recognition for Marathi using CRF,” in Proc. ISCNLP, 2016, pp. 88–95.
[21] P. Koehn et al., ”Moses: Open source toolkit for statistical machine translation,” in Proc. ACL Demo Session, 2007, pp. 177–180.
[22] R. Kulkarni, A. Desai, and M. Karpe, ”BiLSTM-based POS tagging for Marathi,” in Proc. EMNLP Workshop on Indic Languages, 2019, pp. 34–41.
[23] T. Kharwade and S. Bhatt, ”Deep learning approaches for Marathi NLP tasks,” in Proc. COLING, 2020, pp. 456–463.
[24] P. Bojanowski et al., ”Enriching word vectors with subword information,” Trans. ACL, vol. 5, pp. 135–146, 2017.
[25] J. Devlin et al., ”BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[26] R. Joshi et al., ”MahaBERT: A Marathi language model,” in Proc. ACL-IJCNLP Findings, 2022, pp. 1945–1958.
[27] D. Kakwani et al., ”IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages,” in Proc. EMNLP Findings, 2020, pp. 4948–4961.
[28] A. Dabre et al., ”IndicBART: A pre-trained model for natural language generation of Indic languages,” in Proc. ACL Findings, 2022, pp. 1849–1863.
[29] M. Kamble and A. Joshi, ”Marathi text corpora from the web: Construction and challenges,” in Proc. WILDRE, LREC, 2018, pp. 15–22.
[30] S. Kakwani et al., ”IndicCorp: A large-scale multilingual corpus for Indic languages,” arXiv:2212.05409, 2022.
[31] U. Sharma, S. Reddy, and N. Bali, ”Preprocessing challenges for Devanagari script NLP,” in Proc. ICON, 2017, pp. 1–8.
[32] M. Wali and P. Bhosale, ”Tokenization strategies for Marathi: Evaluation and comparison,” in Proc. ISCNLP, 2020, pp. 78–85.
[33] M. Toraman, F. Yilmaz, and E. Yildirim, ”Large-scale authorship attribution: Comparing subword tokenizers,” in Proc. COLING, 2022, pp. 3396–3405.
[34] O. Litake, M. Joshi, and L. Bhavnani, ”Comparison of word embedding techniques for Marathi,” in Proc. ICON, 2020, pp. 113–120.
[35] L. Xue et al., ”mT5: A massively multilingual pre-trained text-to-text transformer,” in Proc. NAACL, 2021, pp. 483–498.
[36] P. Bhosale and R. Joshi, ”Inter-annotator agreement in Marathi NER datasets: Analysis and implications,” in Proc. ComputEL Workshop, LREC, 2022.
[37] A. Wang et al., ”GLUE: A multi-task benchmark and analysis platform for natural language understanding,” in Proc. ICLR, 2019.
[38] D. Kakwani et al., ”IndicGLUE: A benchmark for Indic languages,” in Proc. EMNLP Findings, 2020, pp. 4948–4961.
[39] S. Khanuja et al., ”GLUECoS: An evaluation benchmark for code-switched NLP,” in Proc. ACL, 2020, pp. 3575–3585.
[40] R. Reddy and S. Sharoff, ”Cross language POS taggers for Indian languages,” in Proc. IJCNLP Workshop CLIA, 2011.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 14, April 2026 Impact Factor: 8.2

[41] D. Zeman et al., ”Universal Dependencies 2.10,” LINDAT/CLARIAH-CZ digital library, 2022.
[42] S. Prabhu, A. Kulkarni, and P. Patil, ”Degradation of POS tagging on code-mixed Marathi text,” in Proc. ICON, 2021, pp. 212–219.
[43] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, ”The IIT Bombay English-Hindi parallel corpus,” in Proc. LREC, 2018.
[44] R. Joshi, P. Bhosale, and A. Kunchukuttan, ”Named entity recognition for Marathi: Transformer approaches and error analysis,” in Proc. ISCNLP, 2022, pp. 90–98.
[45] R. Patil, S. Patil, and S. Chandak, ”SentiRaama: Marathi sentiment analysis dataset and baselines,” in Proc. ICON, 2018, pp. 150–157.
[46] SAIL 2015 Shared Task, ”Sentiment analysis in Indian languages,” in Proc. ICON Shared Task, 2015.
[47] A. Gawali, P. Deore, and M. Wali, ”Aspect-based sentiment analysis for Marathi reviews,” in Proc. LREC Indic Workshop, 2022, pp. 45–52.
[48] V. Dongare and S. Patil, ”CNN-based Marathi sentiment classification,” in Proc. ICCIDS, 2020, pp. 445–451.
[49] O. Litake and M. Joshi, ”Marathi news categorization using transformer models,” in Proc. ICON, 2021, pp. 78–84.
[50] A. Jha, R. Shah, and N. Agrawal, ”Hate speech detection in Marathi social media,” in Proc. FIRE Shared Task, 2021.
[51] A. Kunchukuttan et al., ”Sata-Anuvadak: Tackling multiway translation of Indian languages,” in Proc. ICON, 2014.
[52] A. Vaswani et al., ”Attention is all you need,” in Proc. NIPS, 2017, pp. 5998–6008.
[53] Y. Tang et al., ”Multilingual translation with extensible multilingual pretraining and finetuning,” arXiv:2008.00401, 2020.
[54] P. Bhosale and A. Kunchukuttan, ”Human evaluation of Marathi machine translation: A comparative study,” in Proc. MT Summit, 2021.
[55] S. Veena and A. Sarkar, ”Language identification in Marathi-Hindi code-mixed text,” in Proc. EMNLP Workshop CALCS, 2020, pp. 115–124.
[56] E. Tjong Kim Sang and F. De Meulder, ”Introduction to the CoNLL-2003 shared task,” in Proc. CoNLL, 2003, pp. 142–147.
[57] M. Popovic, ”chrF: Character n-gram F-score for automatic MT evaluation,” in Proc. WMT, 2015, pp. 392–395.
[58] R. Kulkarni, ”Error analysis of Marathi POS taggers: Patterns and implications,” in Proc. ICON, 2019, pp. 98–105.
[59] S. Jain et al., ”Indic-transformers: An analysis of transformer language models for Indian languages,” arXiv:2011.02323, 2020.
[60] S. Khanuja et al., ”MuRIL: Multilingual representations for Indian languages,” arXiv:2103.10730, 2021.
[61] A. Kunchukuttan and P. Bhattacharyya, ”Utilizing language relatedness among Indian languages for MT,” in Proc. COLING, 2016, pp. 2547–2558.
[62] N. Jha, S. Ghosh, and R. Borah, ”Bias in Indian language NLP models: A survey and framework,” in Proc. ACL WILDRE Workshop, 2023