Fuzzy Systems and Soft Computing
ISSN : 1819-4362

MARATHI NATURAL LANGUAGE PROCESSING: A REVIEW

Dr A A Jinturka, Assistant Professor, Department of Computer Science, SIES (Nerul) College of Arts, Science and Commerce (Autonomous), Navi Mumbai :jinturkaraditya@gmail.com

ABSTRACT
Data is being seen as an "Oil" of the 21st century fuelling the growth and innovation. With the exponential use of Internet, Social media platforms; the size of data is increasing humongous. With Computing devices supporting multiple languages, this data is now not only in English, but regional languages also. Processing of this data has given rise to new domains in Artificial Intelligence such as Data Science (DS), Information Retrieval (IR), Natural Language Processing (NLP). This Paper reviews the various phases of NLP, challenges associated with these phases and work done in case of Marathi Language.
Keywords: Natural Language Processing, Marathi, Artificial Intelligence.

INTRODUCTION:
With growing use of digital devices, humongous data is being available in digital format and in many languages. (Borisova, Karashtranova, & Atanasova, Vol. 15, No. 2, April 2025) To process this data and get insights from this data; Researchers are focusing on Natural Language Processing domain. In recent years; advancements in computing tools have been instrumental in development of sophisticated NLP tools. This progress resulted into GenAI chatbots like ChatGPT. Tools like ChatGPT are based on Large Language Models (LLM). They require large amount of data which gets processedand new text / data is generated.
(Shinde & Joshi) Most of work done in NLP research is in 20 out of world's 7000 languages. Therefore, there is scope for research in these underexplored languages. This paper focuses on review of work done in the area of NLP for Marathi Language. Paper is organized as follows. Next section discusses NLP definition and generic pipeline of NLP tasks. Followed by discussion on some of the common NLP techniques such Topic modelling, Text analysis, Named Entity recognition, etc. Then Literature review of work done in Marathi Language. Followed by conclusion of paper.

NATURAL LANGUAGE PROCESSING:
(Natural Language Processing (NLP) – Overview - GeeksforGeeks, 2025)Natural Language is a branch of Artificial Intelligence which enables Computing machines to understand and generate Natural language which is having some potential use and at the same time the generated text is meaningful also. NLP primarily focuses on two aspects: Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU is more like a listener whose job is to understand and define meaningful representation whereas NLG focuses on algorithms which gives or generates the meaningful text in Natural Language. Figure 1 shows the basic pipeline for NLP tasks.

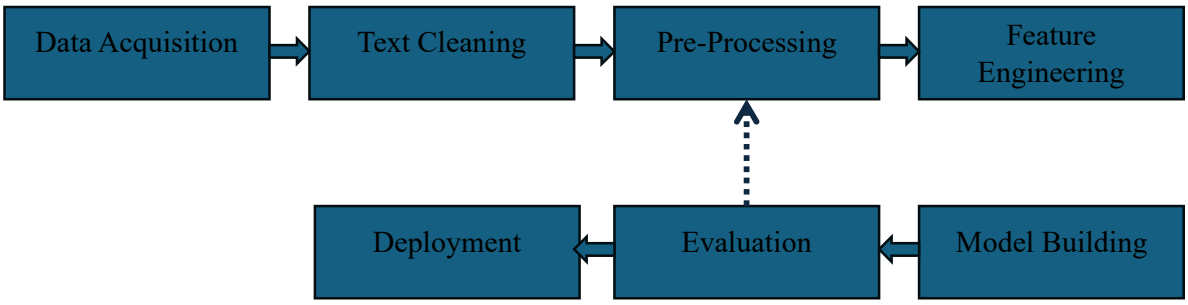

Figure 1: Natural Language Processing Pipeline

Processing starts with Data acquisition in which data written in natural language is collected. Such data can be found in Websites, social media, Public databases, etc. This data needs to be collected as per requirements of Task. The data which is collected from these sources may not be useful in its original form because; it may contain special characters, emojis, Spelling mistakes, formatting tags. In order to proceed further, we need to clean this data using techniques such as Unicode normalization, Spelling correction.

Once the input data is cleaned; this data is then subjected to undergo pre-processing tasks. Objective behind this is that since NLP algorithms work mainly at sentence level; so the words are also required at their minimum level. These preprocessing tasks mainly include Tokenization, Stemming, stop word removal, etc. Feature engineering deals with the techniques which convers this pre-processed data into vectors so that Machine can understand and learn from this text using mathematical / statistical models.

Once the machine has features to learn; a model is required to facilitate the learning process. Model can be Heuristic, Machine or Deep learning model. Once model is built; it is then tested using evaluation metrics.

MARATHI NATURAL LANGUAGE PROCESSING:

(Marathi language - Wikipedia, 2025)Marathi is an Indo-aryan language spoken by over 83 million people of Maharashtra and other states. Recently, Government of India recognised and awarded Marathi as a “Classical” language. It is also Official language of Maharashtra. The language is based on Devnagri script which makes it morphologically complex. Like English, it is not case-sensitive language. Therefore, a different approach is needed for Natural language processing in Marathi.

REVIEW OF WORK DONE IN MARATHI NATURAL LANGUAGE PROCESSING:

NLP has been an active area of research from last few years. Therefore, one can expect advancements in Marathi NLP like English language. However, diverse nature of Marathi, scarcity of publicly available benchmark databases poses the significant problems in the research of Marathi NLP. In this section; a review of literature has been done of work done in Marathi NLP.

(Date, Deshmukh , Boyd, Ashokkumar , & Pennebaker, 2024)In this paper Author has presented a novel framework for the design and development of a Marathi translation of the LIWC dictionary. In this paper; a Linguistic Inquiry and Word Count (LIWC) tool has been proposed for Marathi Language. LIWC is a tool originally developed for English language which is used for linguistic analysis of text. Experimental results show that Marathi version of LIWC performs at par with its English counterpart. There is scope to expand the LIWC with more lexicons and can be applied in various domains.Author has observed that the performance of Marathi Version of LIWC captures 72% of the dictionary words

(Dani & Sathe) Paper discusses the progress of NLP research in Indian languages including Marathi. Paper also puts focus on various tools and techniques available for research. Author has also discussed the models like BERT, BART, IndicNLP etc. used to analyse sentiments. Author has observed that there was a improvement in the outcomes in medium resource languages like Marathi using LLMs. The various Tools and resources have been useful to increase the accuracy in results of sentiment analysis like XLM-R, NLLB, IndicTrans, mahaNLPNMT, NER, etc. During last few years, experimental results have been improved for task such as Sentiment analysis, Named Entity Recognition. However, there is still scope for improvement in task like Text-to-speech.

(Shinde & Joshi) In this paper; Topic modelling techniques have been proposed for Marathi. BERT and non-BERT models have been examined for Multilingual as well as monolingual BERT models. The author has trained BERT model with three different Parameters that are LDC, LPC and SHC. The author has also analysed the performance of various models which gives insights for Marathi language topic modelling. The Coherence metric shows the accuracy of the models. Experimental results shows that BERTopic models outperform LDA in terms of Coherence score.

(Chavan , Madle, Patil, & Joshi, 2024) This paper deals with stop words in Marathi Language using TF-IDF approach. A list of 400 words have been curated and incorporated into the database. The Author has observed that the IndicBERT model gives highest recognition accuracy 94.24 % for recognition of stopwords as compared to other models. Experimental results shows that this curation of stop words has significantly improved the performance of NLP tasks in Marathi language.
(Deshmukh , et al., 2024)Paper focuses on need of Named Entity Recognition (NER) methods for Marathi. With growing digital content in Marathi, there is need for an efficient NER technique. Paper analyses existing NER techniques and suggests an approach on how to adopt these techniques for Marathi. Paper mentions need for additional research in this area using various databases. Also, Research can be extended to other similar languages.
(Vivek A. Manwar) Paper discusses the problem of "words having multiple meanings", which is a common problem across the languages. Word Sense Disambiguation is NLP task which deals with this problem has been proposed for Marathi Language. The Author has used the unsupervised learning method for disambiguation of words in Marathi language.
(Sarode & Sultanova, Vol: 06 Issue: 06)In this paper,hate speech detection in Marathi Language has been discussed. The author has used two datasets short text dataset and long text dataset for the experiment. The Analysis of various pre-processing methods and their effect on hate speech detection has been performed. In this study, three models have been used that are ROBERT, IndicBERT and MahaBERT. The Transfer learning approach was proposed to detect hate speech and how results vary depending upon size of data was discussed. Need of further research for efficient hate speech detection model has been identified.

CONCLUSION AND FUTURE SCOPE:

Over the years, Marathi NLP is evolving as many researchers are contributing to find out efficient solutions to different NLP problems. However, diverse nature of Marathi, Non-availability of benchmark databases, morphological nature are some of the hurdles in this field. Computing advances in hardware and evolution of tools are helping researchers to get improved results.

BIBLIOGRAPHY:

1. Borisova, N., Karashtranova, E., & Atanasova, I. (Vol. 15, No. 2, April 2025). The advances in natural language processing technology and its impact on modern society. *International Journal of Electrical and Computer Engineering (IJECE) ISSN: 2088-8708*, 2325-2333.
2. Chavan , R., Madle, V., Patil, G., & Joshi, R. (2024). Curating Stopwords in Marathi: A TF-IDF Approach for Improved Text Analysis and Information Retrieval. *arXiv:2406.11029v1 [cs.CL] 16*.
3. Dani, A., & Sathe, S. R. (n.d.). A Review of the Marathi Natural Language Processing.
4. Date, S., Deshmukh , S., Boyd, R., Ashokkumar , A., & Pennebaker, J. W. (2024). Designing of a Novel Framework for Marathi Natural Language Processing: MR-LIWC2015. *International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING ISSN:2147-6799*, 1-14.
5. Deshmukh , P., Kulkarni , N., Kulkarni, S., Manghani, K., Kale, G., & Joshi, R. (2024). Long Range Named Entity Recognition for Marathi Documents. *arXiv:2410.09192v1 [cs.CL]*.
6. Kadam Vaishali P, N. M. (Vol-6-3-May-2024). A Named Entity Recognition System for the Marathi Language. *JOURNAL OF ADVANCED APPLIED SCIENTIFIC RESEARCH-ISSN(O): 2454-3225*, 229 - 243.
7. *Marathi language - Wikipedia*. (2025, March 12). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Marathi_language
8. *Natural Language Processing (NLP) – Overview - GeeksforGeeks*. (2025, March 11). Retrieved from https://www.geeksforgeeks.org/: https://www.geeksforgeeks.org/natural-language-processing-overview/

9. Oyewole, A. T., Adeoye, O. B., Addy, W. A., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Automating financial reporting with natural language processing: A review and case analysis. *World Journal of Advanced Research and Reviews eISSN: 2581-9615*.
10. Sarode, A., & Sultanova, N. (Vol: 06 Issue: 06). Detection of Hate Speech in Marathi Using Language Specific Pre-Processing. *International Journal of Data Science and Advanced Analytics (ISSN: 2563-4429)*, 297-301.
11. Shinde, S., & Joshi, R. (n.d.). Topic Modeling in Marathi. *arxiv*.
12. TEXT CLASSIFICATION IN MARATHI LANGUAGE. (Volume: 08 Issue: 11 | Nov - 2024). *International Journal of Scientific Research in Engineering and Management (IJSREM) ISSN: 2582-3930* , 1-7.
13. Vivek A. Manwar, R. L. (n.d.). Word Sense Disambiguation for Marathi Language in Cross Language. *Proceeding of National Conference on "Recent Advancements in Science & Technology"*.