



A SURVEY OF PART OF SPEECH TAGGING APPROACHES IN HINDI

Vijay Kumar Soni ¹, Dr. Smita Selot²

^{1,2}Department of CSE, SSTC, Chhattisgarh

¹vijaysoni81@gmail.com, ² smitaselot504@gmail.com

Abstract

Various forms of speech Tagging is the procedure of giving individual words in a text a tag, which indicates the category of grammar to which that word belongs. It is a basic step that is needed in many natural language applications, from translation systems to speech recognition. Because different languages have different grammatical structures and morphologies, the ways to tag in different languages range from rule-based to statistical. Processing English and European languages takes a lot of work, but processing Indian languages takes less work because there isn't a lot of annotated corpora. This paper is a summary of the research that different institutions and organizations have done on how to process the Hindi language. We talk about the different ways language is processed, the available corpora, and how their results compare. A possible model for making a POS tagger is given based on comparisons and the study.

Keywords: NLP, HMM, HINDI, POS.

DOI Number: 10.48047/nq.2021.19.11.NQ21310

NeuroQuantology 2021;19(11):952-962

1. INTRODUCTION

The term "natural language processing" (NLP) refers to the process of analyzing human language (natural language) using a computer and then converting the results of that analysis into a form of representation that might be helpful. The primary focus of the discipline of NLP is on developing methods through which computers can interpret human languages. Developing applications like translation systems and query-based systems, among other things, is a subfield that falls within the umbrella of the study of natural language processing (NLP), which makes use of NLP tools like POS taggers and semantic analyzers, among other things. The Natural Language Processing (NLP) group's objective is to create software that is capable of analyzing, comprehending, and generating the languages that humans use in their everyday lives. Eventually, the group hopes that people will be able to communicate with computers in the same manner that they would communicate with another person. Dealing with the syntactic and semantic ambiguity of the language is the primary obstacle that must be overcome in the processing of language.

In POS tagging, each word that makes up a phrase is given a part of speech or lexical class identifier. This can also be completed by designating a part of speech to each word. The terms "noun," "verb," "preposition," "pronoun," "adverb," and "adjective" are all examples of different components of speech. It is a crucial part of a wide range of natural language processing (NLP) applications, such as word meaning disambiguation, text-to-text translation systems, speech recognition, and information abstraction. Rigid-word-order languages have a predetermined order for its words, whereas the more common SVO (subject-verb-object) pattern and free-order-word languages do not have such a structure. Example of former is English and that of later is Sanskrit. Free-order-word languages with rich morphological structure require robust methodology for tagging words [1].

The tagging of text is a difficult operation since, depending on the context, a single word may have numerous tags at any given time. The word for this kind of occurrence is called lexical ambiguity. For example, the two word like 'आम', and 'सूरज' have different meanings in different contexts. The first word has two different meanings; a common noun as a meaning of common people and a second meaning as a name of fruit. Similarly, the second



word has two different meanings; first, the name of a person, and second, the name of the sun in Hindi.

The solution to this issue lies in analyzing the word-and-tag combinations of the terms that surround the confusing word in question (the word that has multiple tags). Despite the recent surge in interest in POS tagging in such languages, the lack of available annotated corpus resources hinders research efforts in Hindi languages, in addition to other disambiguation and language-specific difficulties. Standardization is another key issue that needs to be addressed because various research academics employ a variety of tag sets. In this survey study, many ways of tagging words and phrases in Hindi are examined. Section 2 provides a summary of the fieldwork conducted. Section 3 provides several POS tagging approaches, whereas Section 4 explains the accessible corpus.

2. RELATED WORK

The statistical method has been used in several different implementations of part of speech taggers, most often for morphologically dense languages like Hindi. Numerical methods are simple to use and require only a small amount of linguistic knowledge, whereas rule-based methods require neither a corpus nor a large amount of linguistic expertise, but they do require an extensive understanding of the language. Table 1 shows the different types of POS tagging approach with accuracy.

Table 1. Comparison of POS Tagging Approach

Year	Author Name	Methodology	Dataset Size/data types	Accuracy
2006	Aniket Dalal	Maximum Entropy Markov Model	15562 words 27 POS Tags	94.81%
2006	Smriti Singh	Decision tree-based learning algorithm	15,562 words	93.45%
2006	Himanshu Aggarwal	Conditional Random Fields	21000 words 27 POS Tags	82.67%
2006	Pranjal Awasthi	HMM and error-driven learning: 1. Statistical tagging. 2. HMM with transformation rules.	26 POS Tags	80.74%
2006	A. Ekbal , S. Mondal, S. Bandyopadhyay	HMM	21470 words	82.05%
2008	Manish Shrivastava	HMM, Stemmer as pre-processor	66900 words 18 POS Tags	93.12%
2011	Shachi Mall	Rule Based Approach	35340 words	88.40%
2012	N. Garg, V. Goyal, and Suman Preet	Rule Based Approach	26149 words	87.55%
2013	N. Joshi, H. Darbari, and I. Mathur	HMM	3,58,288 words	92.13 %
2014	Ravi Narayan	Quantum neural network	11500 words	99.13%
2015	Deepa Modi	Rule-Based Approach	66900 words	91%
2016	Tang et al.	Neural network	Document-level	81%
2017	Vilares et al.	Sentiment classifier	Sentence level	90%
2018	Pham and Le	Neural network	Aspect Level	89%
2019	Go et al	NB, MaxEnt,SVM	Unigrams, bigrams, POS	83%
2020	Malhar and Ram	NB, SVM, MaxEnt,ANN classifiers.	Unigrams, bigrams, hybrids (unigrams+ bigrams)	92%
2021	Anton and Andrey	NB and SVM classifiers	Unigrams, bigrams, hybrids (unigrams+ bigrams)	81%
2021	Pak and Paroubek	Multinomial NB and SVM	Unigrams, bigrams, trigrams	91%

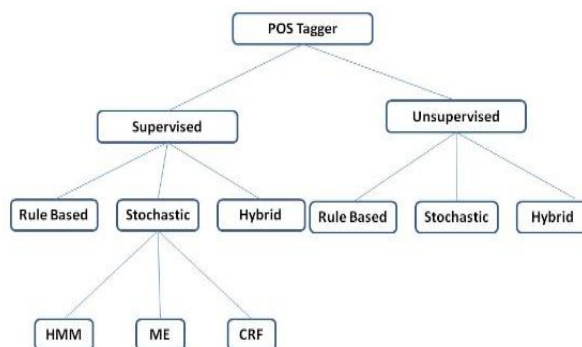


Fig. 1. Classification of POS Tagging

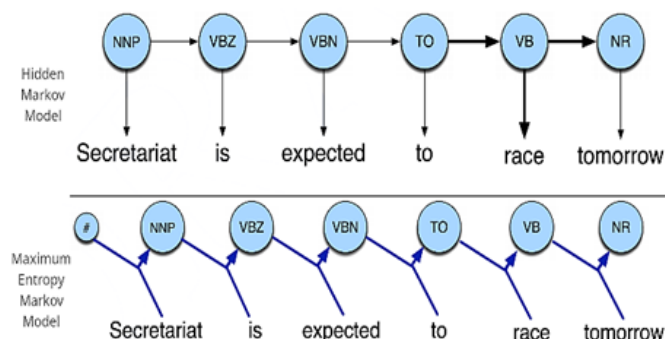


Fig. 2. The HMM and MEMM model

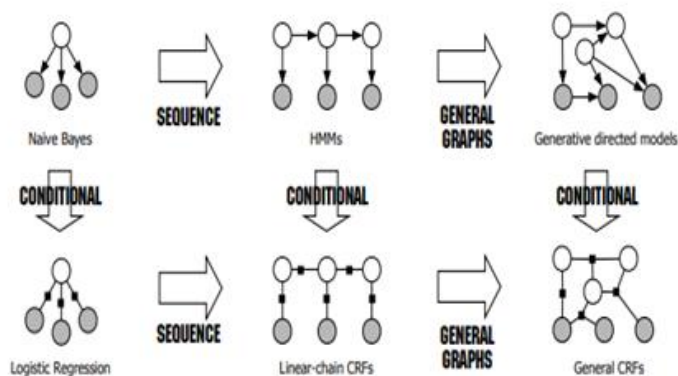


Fig. 3. Relationship between different classifiers

3. TAGGING METHODOLOGY

Approaches to POS tagging may generally be broken down into three categories: Methods that are based on rules, statistical approaches, and hybrid methods are all included. The rule-based tagging approach necessitates an extensive language knowledge base, whereas the stochastic

model calls for a massive annotated corpus but requires far less linguistic expertise. The probabilities and rates of occurrences of words for a specific tag are used as the foundation for the numerical Part of Speech tagger. In most cases, a Part-of-speech tagging using a mix approach will combine rule-based and statistical methods [2]. The classification of point-of-sale

(POS) taggers according to various methodologies is depicted in Figure 1.

3.1 RULE BASED POS TAGGING

Rule-based tagging representations are used to label words with POS labels, which incorporate information from their context in addition to a set of rules that have been hand-crafted. The rules in question are commonly referred to as context frame rules. The early algorithms for automatically allocating parts of speech were based on architecture that consisted of two stages. In the initial step of the process, possible parts of speech were identified for each word by consulting a dictionary. In the second step, vast lists of hand-written disambiguation rules were utilised in order to narrow down this list such that only one part of speech was assigned to each term. [3].

Most tokens may obtain POS tags using the rule-based data preparation, which entails three phases. The initial step is to create each token's potential POS tags. The pruning process comes next, when the context is taken into account to lower the number of potential POS tags for each coin. Masking POS tags that are ready for model training or inference is the final stage. The construction of the system makes use of specific language-based knowledge expressed in the form of rules. One example of a rule that may be found in a context frame is the following: "Tag an ambiguous or unknown word as an adjective if it comes before a determiner and is followed by a noun.": *det - X - n = X/adj*.

Algorithms:

Input: The rules, all the words in a phrase, and their potential POS tagsets are all listed here.

```
reformed = TRUE
while reformed == TRUE
  ensign = FALSE
  for j = 1 to rules1.Length
    rule1 = rules1[j]
    ensign = (ensign || ApplyRule (rule1, words, sets))
  reformed = ensign
```

3.2 STATISTICAL BASED POS TAGGING

The statistical method determines the lexical and contextual frequencies of individual words as well as the frequencies of word-tag pairs. The system searches through the annotated training data to determine which tag applies to a certain word the most frequently and then utilises this knowledge to allocate a tag to the word in the simple text without any notes or explanations. A statistical method requires a corpus that is a sufficiently enough size. Once this need has been met, the method may then calculate the frequency, possibility, or statistics of each and every word that is contained within the corpus. Several of the most widely-used ML models for POS tagging are as follows:

3.2.1. Hidden Markov Model: -

The HMM (Hidden Markov Model) may be thought of as a generative model. A rough estimate of the overall probability is provided by the model for every combined remark and label sequence. After that, the restrictions are educated so that the maximum possible probability is obtained from the combined efforts of the training sets [4].

It is helpful due to the fact that its fundamental idea is sophisticated yet simple to grasp. As a result, both its implementation and analysis are simplified. It utilises only positive data, which allows for easy scaling of the results. It has a limited number of drawbacks. Before the HMM can calculate the joint probability across the observation series and the label sequence, it first has to enumerate all of the possible sequences of observations that may occur. As a consequence of this, it makes a variety of assumptions about the data, such as the Markovian statement, which claims that the present label relies only on the label that came before it. This is one of the many assumptions that it makes. Representing many characteristics that overlap one another and dependencies that last for a long time is also impractical. There are a staggering number of parameters to be examined. Therefore, it requires a substantial data collection in order to train. Before we could move further with the development of an HMM-based tagger, we started by having to tag a corpus according to some criteria. [5].

By evaluating the forward and backward probabilities of tags in addition to the input sequence, HMM-based POS taggers choose the optimal tag to apply to a word. The equation that is presented here provides an explanation for this phenomenon.

$$P(t_i|w_i) = P(t_i|t_{i-1}) \cdot P(t_{i+1}|t_i) \cdot P(w_i|t_i)$$

The current tag's probability, $P(t_i|t_{i-1})$, is determined by the previous tag, whereas the next tag's probability, $P(t_{i+1}|t_i)$, is determined by the current tag. This encapsulates the change that takes place among the tags. These likelihoods are arrived at by applying an equation to the data in question.

$$P(w_i|t_i) = \frac{\text{freq}(t_i, w_i)}{\text{freq}(t_i)}$$

In addition to this, we determined the word likelihood probabilities by applying the formula $P(w_i|t_i)$, which stands for the possibility of the word given the current tag. This probability is determined by the use of an equation. [10].

3.2.2. Maximum Entropy Markov Model: -

The MEM models a probabilistic sequence model with conditions. It is capable of managing dependence over an extended length of time and may also express several facets of a word at the same time. The maximization of entropy is based on the maximization of entropy principle, which argues that the model with the least amount of bias is the one that takes into account all of the facts that are previously known. The maximization of entropy is reliant on this maximization principle. The observation feature is fed into an exponential model for each source state, which in turn provides a distribution across the space of possible following states. The models are connected by an "exponent" variable. There is an association between the states and the output labels. [6]

This model provides a solution to the issue of large dependency that was present in HMM. In comparison to HMM, it possesses a higher recall rate as well as greater precision. The problem of

label bias is one of the drawbacks of pursuing this approach. The probabilities of moving on from a certain condition have to add up to one in order to make sense. MEMM gives preference to those states that are subject to a lower total number of transitions. [7]

3.2.3. Conditional Random Field Model: -

The abbreviation "CRF" refers to the term "Conditional Random Field." It belongs to the class of probabilistic models known as discriminative models. It possesses all of the benefits of MEMMs but does away with the concern of label bias. CRFs are a type of undirected graphical model that are sometimes referred to as random fields. They are used in the process of determining the likelihood that a certain value would be assigned to an output node given the values allocated to other input nodes. Specifically, they are utilised in the process of calculating the restricted possibility of values on allocated output nodes.

Conditional Random Fields are a popular tool for automatically assigning tags to input sequences. To do this, we first have to figure out the restricted probability of standards on output nodes assumed values on input nodes. The term "undirected graphical models" refers to the CRFs, which are conditional random fields. The provisional probability of a state/ tag sequence $Y = y_1, y_2, \dots, y_n$ given an remark sequence $X = x_1, x_2, \dots, x_n$ is calculated as:

The feature function $f_j(y_{i-1}, y_i, x, l)$ and the weight j are both learned during training [11].

3.3 HYBRID POS TAGGING

Combining rules-based and statistical models to form a hybrid model is essentially what hybrid models are. In a hybrid system, the strategy employs a combination of both rule-based and ML methodology. It then creates new ways by drawing on the aspects of each method that are considered to be its strongest feature. It is an effective algorithm because it combines the characteristics of statistically based machine learning algorithms with language rule systems.



$$P(y|x) = 1/Z(x) \left[\exp \sum_{i=1}^n \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i) \right]$$

This system was built with a hybrid-based methodology, and it makes use of seven distinct standard parts of speech tags. The structure operates primarily through two stages: first, the arguments that are input are looked up in the database, and if they are there, the entry is tagged. Second, if it is not there, then a number of other rules or the HMM model will be used. [8]

4. CORPUS

A huge collection of texts is referred to as a corpus. A linguistic analysis will be based on a corpus, which might be a collection of written or spoken information. Corpora is the correct plural version of the word "corpus." The information obtained from corpus analysis includes lexical information, information on morph syntactic structures, semantic information, and pragmatic information. Corpora are essential to the progress of natural language processing (NLP) techniques. Examples of applications that make use of this technology include automated spellchecking, automated abstracting and indexing, information retrieval, machine translation, and speech recognition are all examples of AI applications. New dictionaries and grammars for language students can be derived using corpora. There are many different corpora available, but among the most well-known ones are the British National Corpus (BNC), the COBUILD/Birmingham Corpus, and the IBM/Lancaster Spoken English Corpus. [9]. Hindi corpus by the center for Indian Language Technology, IIT, Bombay [12], IIT Hyderabad provides a Hindi corpus with a size of 30 lakh words. [13]. Hindi Dataset FIRE 2013 with 30,823 transliterated Hindi words (Roman script) [14].

Mobile and laptop reviews are available Hindi news websites were the sites, we collected data from for this study. We collect the positive, negative, and neutral remarks given by the customer. Some of the standard mobile reviews are available online in English language, we have to convert thesis remarks into Hindi language

using google Translate. More than 100000 datasets were created by us.

5. POS Tagging of Hindi Language

Hindi Language Structure Since there are several strategies and public library accessible for English text but not frequently for Hindi text, tagging is still a topic of research. [1] Manish and Pushpak conducted research on Hindi POS utilising an easy, 93.12% accurate HMM-based POS tagger. [2] Nisheeth Joshi, Hemant Darbari, and Iiti Mathur conducted research on Hindi POS by dividing the occurrence total of two tags seen collected in the corpus by the occurrence amount of the prior tag seen self-sufficiently in the corpus. This was done in order to get a better considerate of the relationship between the tags. [3] For the Hindi POS Tag, S Phani Kumar Gadde and Meher Vijay Yeleti employed the Brants TnT (Brants, 2000) HMM-based tagger and CRF-based tagger, achieving an accuracy of 94.21%.

Smriti Singh's first attempt consisted of the presentation of a POS tagging system that was adaptable enough to be used by languages that had limited assets [2]. The POS tagger relies on manually generated morphological instructions rather than any kind of learning or disambiguation method. Its foundation is in morphological rules. A locally annotated corpus of 15,562 words, thorough morphological study supported by a rich lexicon, and a decision-tree-based learning method are all utilised by the system (CN2). Lexicon lookup is used by the system to determine whether POS categories are still missing. Through a four-fold cross validation of the corpora, the system's performance was determined to be 93.45% precise.

Aniket Dalal, Kumar Nagaraj, Uma Sawant, and Sandeep Shelke [2] suggested a POS tagger created on Maximum Entropy (ME). When building a POS tagger using the ME technique, it is important to extract feature functions from a training corpus in order to complete the process. A feature function is often a Boolean function that captures some component of the language that is relevant to the sequence labelling task. This component of the language could be anything from a word to a phrase. The findings of the experiment revealed that the effectiveness of the



system is directly proportional to the amount of data used for the training corpus. There is a growth in performance up until it influences 75% of the exercise corpus; beyond that, there is a loss in precision due to the trained model being overfit to the training corpus. This is the case even though there was an initial increase in performance. After conducting 10 separate tests, the POS tagging correctness of the system was originate to variety from a low of 87.04% to a high of 89.34%, with an average accuracy of 88.4%.

The third POS tagger was developed by Himashu Agarwal and Aniruddha Amni in 2006 using Provisional Random Fields [2]. This technique kind's use of a Hindi morph analyzer for preparation purposes, as well as for determining the origin word and probable POS tag for each word contained in the corpus. The preparation is carried out with the help of CRF++, and the preparation data includes supplementary data such suffixes, word length indicators, and special characters. For training and testing purposes, a 1.5 million-word corpus was utilised, and the system's accuracy was 82.67%.

The HMM-based technique was developed with the intention of making use of the morphological variety offered by the languages without the need to resort to complex and time-consuming research [2]. The essential idea behind this tactic was to "explode" the input in instruction to increase its total length and reduce the total number of exclusive types that were encountered during the learning process. Because of this approach, the chance score of making the right choice goes up while at the same time the doubt of the choices available at each level goes down. In addition, the appearance of new morphological forms for well-known root words helps to lessen the sparsity of data. A corpus containing 81751 tokens that had been exploded and divided into sections of 80% and 20%, respectively, was used for training and testing purposes.

An enhanced Hindi POS tagger was created by using a naïve stemmer, which matches the longest suffix, as a pre-processor for an HMM-based tagger [3]. This led to the invention of a naive stemmer. This technique does not need any linguistic resources other than a list of candidate suffixes, which may be easily generated by

making use of the various machine learning techniques that are already in existence. The performance of the system was 93.12%, according to the records.

In 2011, Nidhi and Amit Mishra made a suggestion on the Part of Speech Tagging for the Hindi Corpus [4]. The structure is responsible for scanning the Hindi corpus and extracting the phrases and words from the assumed corpus when the proposed method is used. In addition, the system searches the database for the tag pattern and displays the tag for each Hindi word. This includes the tag for the noun, the tag for the adjective, the tag for the number, the tag for the verb, and so on.

Using lexical sequence limitations as the foundation, Pradipta Ranjan Ray, Harish V., Sudeshna Sarkar, and Anupam Basu [5] created a POS tagger technique for the Hindi language. The projected procedure acts as the initial level of a part-of-speech tagger that makes use of constraint propagation. It does this by making use of information derived from ontological analysis, morphological investigation, and lexical rules. Even however the presentation of the POS tagger has not been statistically evaluated due to a lack of lexical resources, it covers an extensive range of linguistic phenomena and accurately catches the four most important local dependences in Hindi. The lack of lexical resources prevents statistical evaluation of the POS tagger's performance.

In a POS tagging, we consider following text in Hindi like: text = "इराक के विदेश मंत्री ने अमरीका के उस प्रस्ताव का मजाक उड़ाया है"

After the apply POS tagging, we got the following tags:

[('इराक', 'NNP'), ('के', 'PREP'), ('विदेश', 'NNC'), ('मंत्री', 'NN'), ('ने', 'PREP'), ('अमरीका', 'NNP'), ('के', 'PREP'), ('उस', 'PRP'), ('प्रस्ताव', 'NN'), ('का', 'PREP'), ('मजाक', 'NVB'), ('उड़ाया', 'VFM'), ('है', 'VAUX')]

A significant amount of study has been conducted on point-of-sale (POS) tagging over the years. All of the efforts can be broken down into three

distinct categories on a broad scale. They are as follows: the rule-based approach, in which we make use of both mathematical formulations and rule-based approaches to tag words; the statistical approach, in which we apply mathematical formulations and rule-based methods to tag words; and the hybrid approach, in which we make use of both rule-based methods and numerical methods. Machine learning is commonly utilised in the generation of point-of-sale (POS) taggers in the setting of European languages; however, in the case of Indian languages, we do not yet have a strategy that is both clear and effective. In this study, we cover the process of developing a POS tagger for the Hindi language using the Hidden Markov Model (HMM).

6. ANALYSIS

As observed in the Table1, statistical method gives improved results than rule-based method. It also shows that neural network performs better, even under sparse data. So, a POS Tagger with hybrid approach is expected to show better results. POS tagger can be designed using HMM method for basic level classification and NN can be applied for solving the ambiguous cases. Combination of two techniques at two stages will improve the overall accuracy.

Verb and VAUX accuracy both significantly improve. This is caused by Hindi's extremely inflective verb morphology. HMM often mislabelled several major verbs (VMs) as VAUX or vice versa in the Verb Group. When dealing with copula verb forms (h, TA, etc.), HMM often makes the mistake of labelling them as VAUX. This is because VAUX instances of these kinds are more common than VM instances. It doesn't help because there are typically three or more forms (TA, TF, and T). This form is stemmed down to (T), which more equally distributes the likelihood of (T) forms among VM and VAUX. Additionally, stemming aids in locating verbs in inflected forms that were not existing in the exercise set. As verbs inflect for Gender, Number, Person, Tense, Aspect, or Mood, this is a typical occurrence. This indicates that a verb or verb auxiliary may appear in several forms.

We relied on a statistical approach that was based on HMM in order to train our POS tagger for the Hindi language. We were able to successfully

disambiguate the appropriate word-tag pairings by making use of the contextual information that was presented within the text. On the basis of the examination data, we were able to achieve a correctness of 92.13%. As part of this project, improving the tagger's accuracy is something that could be focused on in the future. It is feasible to achieve this by refining the tag set and adding extra tags, which will enable the tagger to categorise the content in a manner that is clearer to the reader. We can also use this tagger to produce some plausible 8–10 tags for a word. These tags can then be translated into basic trees, and we can produce super tags by combining them with origin trees. This will be of great assistance to us in the training of a super tagger, this can then be applied to the development of a Tree Adjoining Grammar (TAG)-based parser for the Hindi language that is fully operational.

7. CONCLUSION

In this study, we have described many distinct strategies for tagging parts of speech, Numerous NLP-based apps rely on it as their foundation. Accuracy in NLP programs is proportional to how well a POS tagger performs. A strategy based on statistics may involve frequency analysis and probability calculations. This method is difficult because it may provide sequences of tags for sentences that deviate from standard grammatical practices in a given language. In other words, it can produce grammatically incorrect phrases. The problem can be solved using a hybrid approach, which is a mix of two or more different methodologies. These kinds of hybrid approaches can also lead to improvements in the system's degree of precision. One of the most effective methods for gaining knowledge from a limited amount of data is the use of neural networks. Most of the previous approaches used for POS tagging of Hindi were unable to capture this, so a combination of statistical and NN is proposed for developing a tagger.

REFERENCES

- [1] Merin Francis, A Comprehensive Survey On Parts Of Speech Tagging Approaches In Dravidian Languages , 09th IRF International



- Conference, 27 th July-2014, Bengaluru, India, ISBN: 978-93-84209-40-7.
- [2] Navneet Garg, Vishal Goya, Suman Preet, Rule Based Hindi Part of Speech Tagger, Proceedings of COLING 2012: Demonstration Papers, pages 163–174, COLING 2012, Mumbai, December 2012.
- [3] Deepika Kumawat, Vinesh Jain, POS Tagging Approaches: A Comparison, International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 6, May 2015.
- [4] Sanjeev Kumar Sharma and Gurpreet Singh Lehal. (2011). Using Hidden Markov Model to Improve the Accuracy of Punjabi POS Tagger, Computer Science and Automation Engineering(CSAE), 2011 IEEE International Conference on June 2011, pp. 697-701.
- [5] Dinesh Kumar, Gurpreet Singh Josan, Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey, International Journal of Computer Applications (0975 – 8887) Volume 6– No.5, September 2010.
- [6] Andrew Borthwick. 1999. “Maximum Entropy Approach to Named Entity Recognition” Ph.D. thesis, New York University.
- [7] Nidhi Mishra and Amit Mishra. (2011). Part of Speech Tagging for Hindi Corpus, In the proceedings of 2011 International Conference on Communication systems and Network Technologies, pp.554-558.
- [8] Kanak Mohnot, Neha Bansal, Shashi Pal Singh, Ajai Kumar, Hybrid approach for Part of Speech Tagger for Hindi language, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 4, Issue 1, February 2014.
- [9] Shubhangi Rathod, Sharvari Govilkar, Survey of various POS tagging techniques for Indian regional languages, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2525-2529.
- [10] Nisheeth Joshi, Hemant Darbari and Iti Mathur, HMM Based Pos Tagger For Hindi , CCSIT, SIPP, AISC, PDCTA - 2013pp. 341–349, 2013. © CS & IT-CSCP 2013.
- [11] Modi Deepa, Nehra Maninder, Nain Neeta and Ahmed Mushtaq, A Survey of Techniques for Two Level Corpus Annotation for Hindi, International Bulletin of Mathematical Research Volume 02, Issue 1, 2015, Pages 194-202, ISSN: 2394-7802.
- [12] V. K. Soni and S. Selot, "A Comprehensive Study for the Hindi Language to Implement Supervised Text Classification Techniques," 2021 6th International Conference on Signal Processing, Computing and Control (ISPPCC), 2021, pp. 539-544, doi: 10.1109/ISPPCC53510.2021.9609401.
- [13] V. K. Soni and S. Selot, Classification Technique Approach in Aspect Based Sentiment Analysis-A Survey Report, AICTE Sponsored National E - Conference On DATA SCIENCE AND ITS APPLICATIONS, July 2021
- [14] V. K. Soni and S. Selot, Big Data for Natural Language Processing: A Survey Report, All India Conference on “DISRUPTIVE TECHNOLOGIES, February 2020.
- [15] Mohammad Al-Smadi, Bashar Talafh, Mahmoud Al-Ayyoub Yaser Jararweh: “Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews” ,2018, International Journal of Machine Learning and Cybernetics.
- [16] Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yaser Jararweh, Omar Qawasmeh: “Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels’ reviews using morphological, syntactic and semantic features”, 0306-4573, 2018 Elsevier.
- [17] Aitor Garc Pablos, Montse Cuadros, German Rigau, “Almost Unsupervised System for Aspect Based Sentiment Analysis”, arXiv:1705.07687v2 [cs.CL] 18 Jul 2017.
- [18] Mauro Dragoni, Marco Federici, Andi Rexha: “An unsupervised aspect extraction strategy for monitoring realtime reviews stream”, 2015, @ Elsevier.
- [19] M. Rathan Vishwanath R. Hulipalled K.R. Venugopal: “Consumer Insight Mining: Aspect Based Twitter Opinion Mining of Mobile Phone Reviews”, <http://dx.doi.org/doi:10.1016/j.asoc.2017.07.056>.
- [20] Erik Cambria, Soujanya Poria, Devamanyu Hazarika: “Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings”, 2018, Association for the Advancement of Artificial Intelligence.
- [21] Fu Xianghua , Liu Guo, Guo Yanyan, Wang Zhiqiang, “Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon”,

- 2012Elsevie,<http://dx.doi.org/10.1016/j.knosys.2012.08.003>[29] Zhiyuan Chen Arjun Mukherjee Bing Liu: "Aspect Extraction with Automated Prior Knowledge Learning", 2014 Association for Computational Linguistics.
- [22]S. Poria, E. Cambria, L.-W. Ku, C. Gui, A. Gelbukh: "A rule-based approach to aspect extraction from product reviews", Proceedings of the Second Workshop on Natural Language Processing for Social Media , 2014.
- [23]Duc-Hong Pham, Anh-Cuong Le: "Learning multiple layers of knowledge representation for aspect based sentiment analysis", 2017 Elsevier B.V, <http://dx.doi.org/10.1016/j.datak.2017.06.001>.
- [24]Qian Liu, Zhiqiang Gao, Bing Liu and Yuanlin Zhang: "Automated Rule Selection for Aspect Extraction in Opinion Mining", Twenty - Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015).
- [25]Y. Ma, H. Peng, E. Cambria, "Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM", Thirty- Second AAAI Conference on Artificial Intelligence, 2018.
- [26]Subhabrata Mukherjee and Pushpak Bhattacharyya: "Feature Specific Sentiment Analysis for Product Reviews", LNCS 7181, 2012, Springer-Verlag Berlin Heidelberg.
- [27]Neha Nandal, Jyoti Pruthi, Amit Choudhary , "Aspect Based Sentiment Analysis Approaches with Mining of Reviews: A Comparative Study", IJRTE, ISSN: 2277-3878, Volume-7, Issue-6, March 2019
- [28]Mr.Ishaan Tamhankar¹, Dr.Ashysh Chaturvedi², "Classification Of Spam Categorization On Hindi Documents Using Bayesian Classifier", IOSR Journal Of Computer Engineering, Nov - Dec 2018.
- [29]Sonali Rajesh Shah, Abhishek Kaushik, "Sentiment Analysis On Indian Indigenous Languages: A Review On Multilingual Opinion Mining ", 27 November 2019.
- [30]Shalini Puri And Satya Prakash Singh, "An Efficient Hindi Text Classification Model Using Svm", Springer Nature Singapore, S.-L. Peng Et Al. (Eds.), Computing And Network Sustainability, Lecture Notes In Networks And Systems,2019.
- [31]Mukesh Yadav, Varunakshi Bhojane, "Semi-Supervised Mix-Hindi Sentiment Analysis Using Neural Network", 2019 IEEE.
- [32]Rahul Patel, Omprakash Yadav, Yash Shah, Saneesha Talim, " Sentiment Analysis On Hindi News Articles ", International Research Journal Of Engineering And Technology (IRJET), May 2020.
- [33]Ramchandra Joshi¹, Purvi Goel², And Raviraj Joshi, "Deep Learning For Hindi Text Classification: A Comparison", Arxiv:2001.10340v1 [Cs.Ir],19 Jan 2020.
- [34]Raktim Kumar Dey, Debabrata Sarddar, Indranil Sarkar, Rajesh Bose, "A Literature Survey On Sentiment Analysis Techniques Involving Social Media And Online Platforms", International Journal Of Scientific & Technology Research Volume 9, Issue 05, May 2020.
- [35]Leila Moudjari, Karima Akli-Astouati, " An Experimental Study On Sentiment Classification Of Algerian Dialect Texts ", International Conference On Knowledge-Based and Intelligent Information & Engineering Systems,2020.
- [36][7] Sandeep Singh Sikarwar, Dr.Nirupma Tiwari, " Analysis The Sentiments Of Amazon Reviews Dataset By Using Linear Svc And Voting Classifier ", International Journal Of Scientific & Technology Research, June 2020.
- [37]Nikita Kolambe, Yashashree Belkhede, Nikhil Wagh, " A Review On Sentiment Analysis On Hindi Language Using Neural Network ", The International Journal Of Analytical And Experimental Modal Analysis, September-2020.
- [38]Regatte Yashwanth Reddy, Gangula Rama Rohit Reddy, Radhika Mamidi, "Dataset Creation And Evaluation Of Aspect Based Sentiment Analysis In Telugu", Language Resources And Evaluation (Lrec 2020), Pages 5017–5024 11–16 May 2020.
- [39]Xiaoyu Luo, "Efficient English Text Classification Using Selected Machine Learning Techniques ", Alexandria Engineering Journal -2020.
- [40]Rajesh Kumar Chakravarti, Jayshri Bansal & Paritosh Bansal, "Machine Translation Model For Effective Translation Of Hindi Poetries Into English ", Journal Of Experimental &

- Theoretical Artificial Intelligence, 26 Nov 2020.
- [41]Nurul Husna Mahadzir, Mohd Faizal Omar, Mohd Nasrun Mohd Nawi, Anas A. Salameh, " Sentiment Analysis Of Code-Mixed Text: A Review ", Turkish Journal Of Computer And Mathematics Education,2021.
- [42]Daler Ali, Malik Muhammad Saad Missen, Mujtaba Husnain, " Multiclass Event Classification From Text ", Hindawi Scientific Programming Volume 2021.
- [43]Srinivasan, R., Subalalitha, C.N. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. *Distrib Parallel Databases* (2021). <https://doi.org/10.1007/s10619-021-07331-4>
- [44]P. C. Shilpa, R. Shereen, S. Jacob and P. Vinod, "Sentiment Analysis Using Deep Learning," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 930-937, doi: 10.1109/ICICV50876.2021.9388382.
- [45]M. Anusha and R. Leelavathi, "Analysis on Sentiment Analytics Using Deep Learning Techniques," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, pp. 542-547, doi: 10.1109/I-SMAC52330.2021.9640790.
- [46]"Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge" by Manish and Pushpak <https://www.cse.iitb.ac.in/~pb/papers/icon08-hindi-pos-tagger.pdf>
- [47]"HMM BASED POS TAGGER FOR HINDI" by Nisheeth Joshi, Hemant Darbari and Iti Mathur. <https://airccj.org/CSCP/vol3/csit3639.pdf>
- [48][3] "Improving statistical POS tagging using linguistic features for Hindi and Telugu" by S Phani Kumar Gadde, Meher Vijay Yeleti. <https://researchweb.iiit.ac.in/~mehervijay.yeleti/papers/icon08-pos.pdf>
- [49]Selvam, M., Natarajan, A.M., (2009) "Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques". *International Journal of Computers*, 3(4).
- [50]Dhanalakshmi, V., Kumar, A., Shivapratap, G, Soman, K.P., Rajendran, S, (2009) "Tamil POS Tagging using Linear Programming". *International Journal of Recent Trends in Engineering*, 1(2).
- [51]Dhanalakshmi V, Anand kumar M, Rajendran S, Soman K P., (2009) "POS Tagger and Chunker for Tamil Language". *Proceedings of Tamil Internet Conference 2009*.
- [52]Singh, J., Joshi, N., Mathur I., (2013) "Part of Speech Tagging of Marathi Text Using Trigram Method", *International Journal of Advanced Information Technology*, pp 35-41, Vol 3. No. 2.
- [53]Bharati, A., Sharma, D.M., Bai, L., Sangal, R., (2006) "AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages",<http://ltrc.iiit.ac.in/tr031/posguidelines>.
- [54]Nidhi Mishra Amit Mishra (2011), "Part of Speech Tagging for Hindi Corpus", *International Conference on Communication Systems and Network Technologies*.
- [55]Pradipta Ranjan Ray, Harish V., Sudeshna Sarkar and Anupam Basu, "Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi",Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur, INDIA 721302. www.mla.iitkgp.ernet.in/papers/hindipostagging.pdf.
- [56]Debasri Chakrabarti (2011), "Layered Parts of Speech Tagging for Bangla", *Language in India* www.languageinindia.com, May 2011, Special Volume:Problems of Parsing in Indian Languages.
- [57]Vijayalaxmi .F. Patil (2010), "Designing POS Tagset for Kannada, Linguistic Data Consortium for Indian Languages (LDC-IL), Organized by Central Institute of Indian Languages, Department of Higher Education Ministry of Human Resource Development, Government of India, March 2010..
- [58]Hammad Ali (2010), "An Unsupervised Parts-of-Speech Tagger for the Bangla language", Department of Computer Science, University of British Columbia. 2010.