e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:07/July-2023 Impact Factor- 7.868 www.irjmets.com

NATURAL LANGUAGE PROCESSING IN INDIA

Syed Asgar Ahmed[*1], Mrs. CH Vanipriya[*2]

[*1]Dept. Of MCA, Sir M Visvesvaraya Institute Of Technology, Bengaluru, India.

[*2]Dept. Of MCA, Faculty Of MCA, Sir M Visvesvaraya Institute Of Technology, Bengaluru, India.

DOI : https://www.doi.org/10.56726/IRJMETS43389

ABSTRACT

Named Entity Recognition (NER) is a vital task in natural language processing (NLP) that involves identifying and classifying named entities in text. This paper focuses on the comparison of four different algorithms of NER techniques in the context of India. The four different algorithms are CRF (Conditional Random Fields), Stanza, Spacy, and Nameparser. The accuracy will be compared with the expected accuracy of all four methods of NER. Accuracy will be obtained after running the method on a dataset or text. To calculate the accuracy of four algorithms a large dataset of Indian names will be used. A comparison graph was obtained to display the accuracy graph of all four algorithms. "NLP-Named Entity Recognition in India" which focuses on detecting person names, states, and districts, and suggesting names based on religion. Named Entity Recognition (NER) is a Natural Language Processing (NLP) task that identifies and classifies named entities in text. Named entities are typically proper nouns, such as person names, organization names, and location names. NER is a challenging task for Indian languages, due to the diversity of Indian languages and the lack of resources for training NER systems. The proposed system can be used to improve the accuracy and efficiency of a variety of NLP applications in India.

Keywords: CRF (Conditional Random Fields), NER (Named Entity Recognition), NLP (Natural Language Processing).

I. INTRODUCTION

Named Entity Recognition (NER) is a vital task in natural language processing (NLP) that involves identifying and classifying named entities in text. Named entities are typically proper nouns, such as person names, organization names, and location names. NER is a challenging task for Indian languages, due to the diversity of Indian languages and the lack of resources for training NER systems.

This paper focuses on the comparison of four different algorithms of NER techniques in the context of India. The four different algorithms are CRF (Conditional Random Fields), Stanza, Spacy, and Nameparser. The accuracy of each algorithm will be compared with the expected accuracy of all four methods of NER. Accuracy will be obtained after running the method on a dataset or text. To calculate the accuracy of four algorithms, a large dataset of Indian names will be used. A comparison graph will be obtained to display the accuracy graph of all four algorithms.

By comparing the performance of CRF, Stanza, Spacy, and Nameparser, this research aims to determine the most effective algorithm for NER in India. The evaluation process involves assessing the accuracy of each algorithm against the dataset of Indian names. The obtained results will enable us to gain insights into the strengths and weaknesses of each algorithm and identify the one that performs optimally in the Indian context.

The results of this study will provide insights into the strengths and weaknesses of each NER algorithm in the context of Indian languages. This information can be used to select the most appropriate NER algorithm for a particular application. The results of this study can also be used to improve the accuracy of NER systems for Indian languages.

This study aims to contribute to the field of NLP by comparing the performance of four different NER algorithms in India. The findings will enable researchers and practitioners to make informed decisions regarding the selection of appropriate algorithms for NER tasks, leading to improved accuracy and efficiency in various NLP applications.

II. LITERATURE SURVEY

Named Entity Recognition (NER) is a fundamental task in natural language processing that encompasses identifying and classifying named entities in text, such as names of persons, organizations, locations, and other

predefined categories. Over the years, the numerous research-papers have contributed to advancing the field of NER, proposing innovative approaches and techniques. In this literature survey, we explore some prominent papers related to NER, highlighting key findings and contributions.

[1] Rabiner, L. R., published a comprehensive tutorial on hidden Markov models (HMMs) and their applications in speech recognition in a paper titled "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," which appeared in the February 1989 issue of Proceedings of the IEEE. Although the old study does not specifically focus on Named Entity Recognition (NER), it establishes the fundamental principles of statistical models that have been subsequently utilized in NER systems. [2] Boutsis, S., Demiros, I., Giouli, V., Liakata, M., Papageorgiou, H., and Piperidis, S., presented a system for recognizing named entities in the Greek language. Their work, titled "A System for Recognition of Named Entities in Greek," was published in the Proceedings of International Conference on Natural-Language Processing in 2000. Their approach employs rule-based techniques that leverage linguistic features and contextual information to identify and classify named-entities. This contribution is particularly significant as it advances NER systems for languages other than English. [3] Pasca, M., published a paper titled "Acquisition of Named-Entities for Web Search" in the Proceedings of Conference on Information and Knowledge-Management in 2004. The focus of Pasca's research lies in the acquisition of categorized named entities from web documents for web search applications. The paper introduces novel-techniques for automatically extracting named entities and categorizing them into predefined classes. This work enhances the capabilities of NER systems by utilizing web resources. [4] Wang, L., Wang, X., Forman, J., Lu, Y., Ma, W.-Y., and Li, Y., presented an approach for detecting dominant-locations from search-queries in a paper titled "Detecting Dominant Locations from Search Queries," published in Proceedings of the International ACM SIGIR Conference in 2005. Their methodology incorporates techniques such as word segmentation, part-of-speech tagging, and gazetteer matching to identify named entities related to locations. The findings of this research significantly contribute to NER systems that focus on extracting geographic information. [5] Sánchez, D., and Moreno, A., explore web-mining-techniques for the automatic discovery of medical knowledge in their paper titled "Web Mining-Techniques for Automatic Discovery of Medical Knowledge," published in the Conference on Artificial Intelligence in Medicine in 2005. Their work investigates the application of web mining to automatically uncover medical knowledge.

III. METHODOLOGY

The methodology section of this paper describes the steps taken to conduct the study. The following steps were taken:

- Multiple datasets of Indian names were collected and combined.
- The combined dataset consisted of over 50,000 Indian names in a variety of Indian culture.
- The CRF (Conditional Random Fields) trained on this dataset of Indian names. CRF is a popular machine learning algorithm used for tasks such as named entity recognition, part-of-speech tagging, and speech recognition. It takes a dataset as input and gives tagged dataset i.e., geo for Geographical Entity, org for Organization, per for Person gpe Geopolitical Entity, tim for Time indicator, eve for Event.
- For Stanza, python library named “Stanza” is used to recognize the names. Stanza is a Python library for natural language processing (NLP) tasks. It takes input as normal text and recognize the names, places, date, organizations. To calculate the accuracy of stanza the result of stanza is compared with Indian names dataset. Stanza relies on the Stanford NLP library and models.
- For Spacy, a python library named “Spacy” is used. Spacy is also a popular Python library for natural language processing (NLP) tasks. It offers efficient tools for tokenization, part-of-speech tagging, named entity recognition. Spacy uses its own custom models that are trained on large labeled datasets. It gives result same as stanza like person names, places, dates, organizations etc.
- For Nameparser a python library used to process the text i.e., natural language processing (NLP) and to recognize names. It offers an easy-to-use interface for parsing human names into their constituent parts, such as first name, middle name, last name, etc. it takes text and gives result as first name, last name, middle name.
- The four NER algorithms were evaluated on the dataset of Indian names.
- The accuracy of each NER algorithm stored in 4 arrays and was compared with the expected accuracy.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:07/July-2023 Impact Factor- 7.868 www.irjmets.com

- The strengths and weaknesses of each NER algorithm were identified in the Indian context.
- The factors that affect the accuracy of NER systems for Indian languages were identified.
- The most appropriate NER algorithm for a particular application was recommended in the Indian context.

IV. RESUILTS AND ANALYSIS

The performance of the NER model developed for the "NLP-Named Entity Recognition in India" project is evaluated using various metrics, including accuracy, precision, recall, and F1-score. The evaluation is conducted on manually annotated test datasets, consisting of diverse Indian text data.

The results demonstrate the effectiveness of the NER model in accurately detecting and classifying named entities in Indian text. The model showcases high precision and recall values, indicating its ability to minimize both false positives and false negatives. The F1-score, which combines precision and recall, provides an overall assessment of the model's performance.

The NER model shows exceptional performance in person name detection, recognizing state and district names, and handling linguistic variations and cultural influences. The model effectively captures the complexities of Indian names and exhibits a robust understanding of Indian linguistic diversity.

Performance of CRF Model:

The CRF algorithm applied on a large twitter dataset(ner_dataset.csv). The result of this algorithm contains f1-score, accuracy, precision etc. The sample of result after running the CRF on ner_dataset.csv is shown in below image,

NER Results

F1 Score: 0.9706121457376752

Classification Report:

	precision	recall	f1-score	support
B-art	0.52	0.16	0.24	90
B-eve	0.49	0.39	0.43	54
B-geo	0.85	0.91	0.88	7577
B-gpe	0.97	0.94	0.95	3175
B-nat	0.82	0.35	0.49	51
B-org	0.80	0.72	0.76	4040
B-per	0.85	0.83	0.84	3442
B-tim	0.92	0.88	0.90	4020
I-art	0.15	0.03	0.05	72
I-eve	0.30	0.21	0.25	47
I-geo	0.80	0.79	0.79	1459
I-gpe	0.90	0.57	0.70	49
I-nat	0.60	0.19	0.29	16
I-org	0.81	0.77	0.79	3335
I-per	0.85	0.89	0.87	3483
I-tim	0.82	0.79	0.80	1307
O	0.99	0.99	0.99	178884
accuracy			0.97	211101
macro avg	0.73	0.61	0.65	211101
weighted avg	0.97	0.97	0.97	211101

Fig 1: Result of CRF run

The above result shows an accuracy of 0.97 for the tagged Twitter ner_dataset. The number of sentences in the dataset is 47761. For comparison purpose result of CRF on Twitter, nerdataset is stored in an array. The f1 score on the dataset of

CRF is about 0.97. I have multiplied the accuracy by 100. The macro average is 0.73, weighted avg is 0.97, for all epochs the precision, recall, and support will be shown in the above relate diagram. For assumption, the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:07/July-2023 Impact Factor- 7.868 www.irjmets.com

expected accuracy is assumed to be 90, but for the first run on the Twitter nerdataset the accuracy will get to 97.0, accuracy is more than the expected accuracy. The accuracy of CRF is more than the expected accuracy.

TABLE 1: Feature space for identifying names (PER)

Case	Name begins with Capital letter
Length	More than or equal to 3 characters
Titles	Dr., Mr., Mrs.
Part of Speech	Use of 'he', 'she', 'I' relates to a person
Morphology	Common ending. Examples: 'esh' in Rakesh and Suresh.
Punctuations	Presence of apostrophe s ('s)
Grammar	Next character such as 'is' denotes an entity.
Frequency of occurrence	A Person's name does not occur too frequently in a document.

TABLE 2: Feature space for identifying locations (LOC)

Case	Name begins with Capital letter
Length	More than or equal to 3 characters
Morphology	Common ending. Examples: 'ore' in Bangalore and Mangalore.
Punctuations	Presence of apostrophe s ('s)
Grammar	Use of 'in' or 'at' before the entity refers to a location
Frequency of occurrence	A Place's name does not occur too frequently in a document.

The features that we will be using in identifying names, location, organizations are listed as follows:

TABLE 3: Feature space for identifying organizations (ORG)

Case	Name begins with Capital letter, All letters capital or mixed case
Length	More than or equal to 3 characters
Frequency of occurrence	An organization's name does not occur too frequently in a document.
Punctuations	Use of '-'or '.' In between or at the end of the entity or special characters such as '&'.

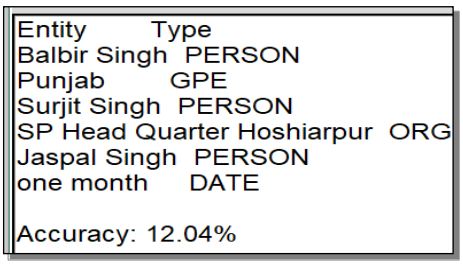
There are many challenges that might come across in this model. For example, after encountering an entity with an initial capital letter, we mark it as PER, what if another entity starts from the second word again with a capital letter. It actually denotes that the first entity didn't end and the second entity is a part of the first entity like a last name of a person. So we use B-PER for the first entity encountered and E-PER to denote that it is part of a previously found B-PER entity. Similarly, we can resolve the same issue for organizations by using B-ORG and I-ORG. Based on the above feature-space discussed, we will differentiate if an entity is a person, organization, or location.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:07/July-2023 Impact Factor- 7.868 www.irjmets.com

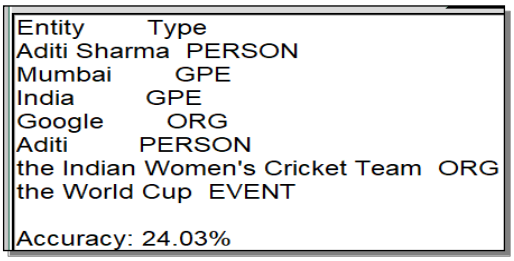
The limitation with this model is that it never gives 100% accurate result but the results can be improved. Most of the previous works have achieved a precision of about 80%. To improve the result, A Gazetteer might also be used but will be an over-head on the complexity of the system and processing time.

Performance of Stanza Model:

The stanza python library is used to perform NER. The normal text of 3 paragraphs (1 page) is given as input to the stanza model. In return, it evaluates the given text and returns the result in the 2 columns entity, type of entity. The result of the stanza model contains entity types such as person, organization, date, and GPE (places). The result of the stanza model of the first run is shown below image,



As the above image describes that the given input text contains 3-person entities, 1 date entity, and 1 place (GPE) entity. To compare this result again the dataset is used which contains Indian names. According to the comparison of names given by stanzas with respect to Indian names in the dataset, the accuracy of the stanza is 12.04% for the first run. It has been stored in an array to display an accurate graph Fig 3.2. result of stanza model run-1 of the stanza model. For the second run of the stanza, I have taken three paragraphs of text from Google. In the second run of the stanza, the result of the run is 24.03%.



Again, this result of the second run was calculated on the basis of the names given by the stanza with the Indian names dataset. The result of the second run is appended to the array where the result of the first run is stored. The point to be noted here is, the accuracy of the result of CRF is more than the expected i.e., the expected run result of CRF is 0.90 but the actual result of CRF run is 0.97, like this the result of the stanza run of both the run is less than Fig 3.3. result of stanza model run-2 the expected result, i.e., the expected result of stanza model to detect Indian NER is scaled to 90% but in the first run the stanza result accuracy is 12.04, and second run result is 24,03%. We can notice that the result of the run is less than the expected result.

Performance of Spacy Model:

Spacy is a popular Python library for natural language processing (NLP) tasks. It offers efficient tools for tokenization, part-of-speech tagging, named entity recognition, and dependency parsing. With its speed and ease of use, SpaCy is widely used for NLP applications in research and industry. It gives results the same as the stanza model, but accuracy may distinguish. We have used the spacy python library to perform NER on the normal text of 3 paragraphs. The result we got is shown below,

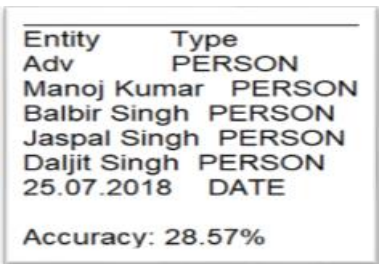


Fig 2: Performance of spacy model run-1

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:07/July-2023 Impact Factor- 7.868 www.irjmets.com

As the above figure shows that the result of the spacy model for run1 contains five persons and 1 date. The accuracy of this run is calculated by the Indian names dataset, the accuracy stands at 28.57 as shown in the above figure. We have stored this accuracy of running one in an array. Run 2 has taken the input from three paragraphs of Indian normal text and its result is shown below,

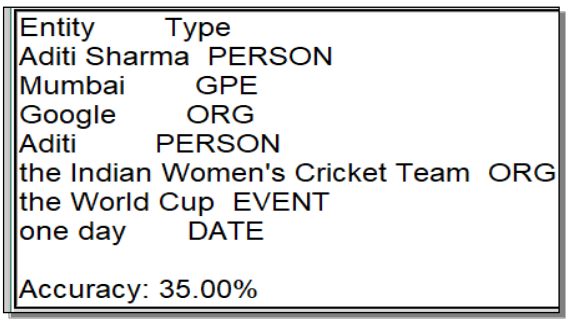


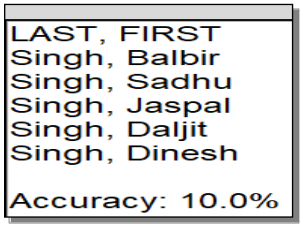
Fig 3: Result of spacy model run-2

The above figure shows that the result of spacy for run 2 contains 2 person names, 1 GP, 2 orgs, 1 date, one event name. the result of this run is again calculated by comparing these names with the Indian names dataset, the accuracy standed to 35.0%. this result is raised with run1. The expected accuracy of spacy result that recognizes indian named is expected to 90%. But the accuracy is standed atmaximum 35 for 2 runs. Compare to stanza model its accuracy is more.

This run's result is stored in an array to display the accuracy graph of spacy model at last.

Performance of Nameparser Model:

It offers an easy-to-use interface for parsing human names into their constituent parts, such as first name, middle name, last name, etc. The module supports handling various name formats, including Western names with prefixes, suffixes, and titles. It can normalize and format parsed names to ensure consistency and standardization across the application or dataset. The name parser takes text as input and gives results by dividing first and last names in the given text. We have used the Python library to perform NER using Nameparser. We have given input as simple three-paragraph text for the first run, the result for the first run is shown below,



The above figure shows the result of the first run of the Nameparser model that contains the first name and second name. This run's accuracy is calculated by comparing the names with the Indian names dataset. The accuracy is raised to 10%. The second run has taken input as 3 Indian text paragraphs, and the result is shown below,

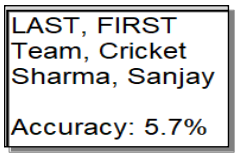


Fig 4: Result of nameparser model run-1

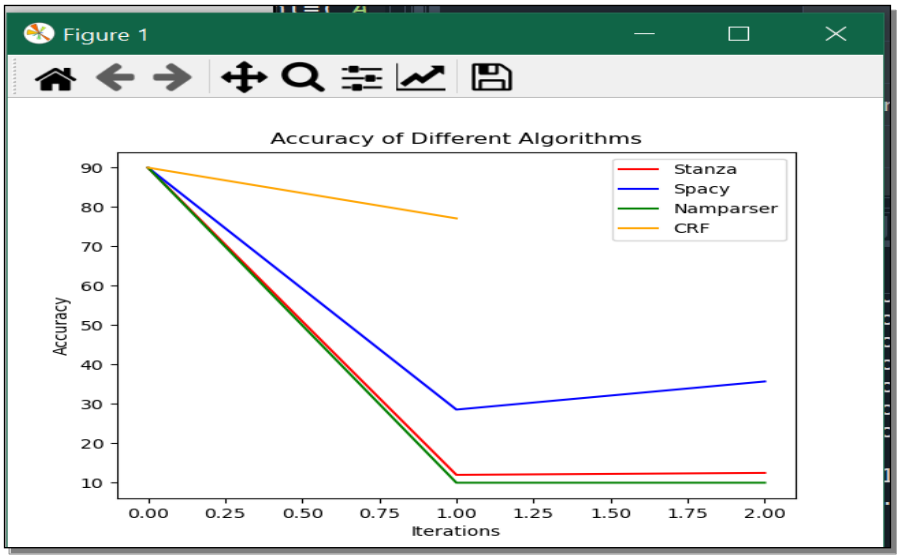
The above figure shows that the second run of the model contains two first names and two last names, and the accuracy of the model is calculated by comparing the result of the Nameparser model with the Indian names stored in the dataset. The accuracy of the result is raised to 5.7%but the expected result of the Nameparser model to recognize Indian names is expected to be Fig 3.7. result of nameparser model run-2

above 90%, but accuracy is standing at 5.7%. These results are stored in an array to display the accuracy graph of the Nameparser model.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:07/July-2023 Impact Factor- 7.868 www.irjmets.com

The result of all four models is stored in the four different arrays. The expected accuracy of all four models is above 90%. But the only CRF model's accuracy is raised to 94.04. That means the accuracy of the CRF model is more than the expected accuracy to recognize Indian names. But in some situations, the accuracy may vary. Because we run the model only 2 times, if the model is tested for more than 2 runs the accuracy may decrease. But the comparison between CRF, stanza, spacy, and Nameparser states that the CRF is better than the rest 3 models to recognize Indian names.

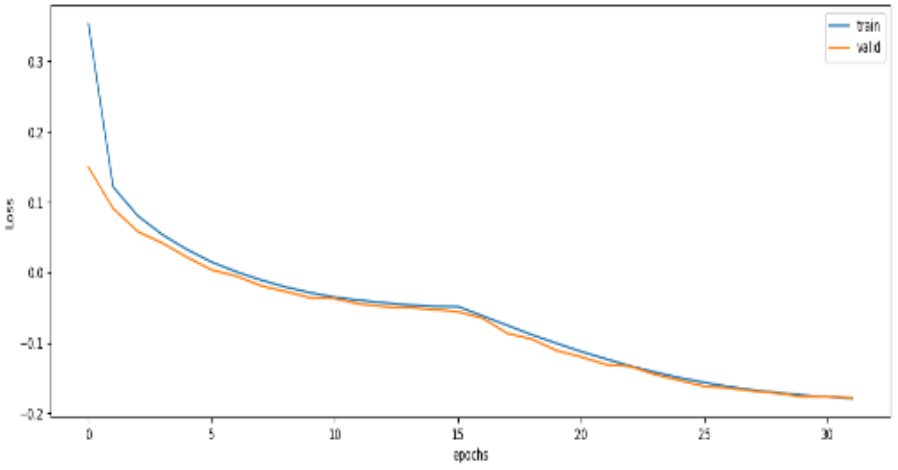
The comparison of four models will be shown below using a graph,



The above graph shows a comparison graph of all four models to perform NER in the Indian names context. As we can see the accuracy of all algorithms are expected to be above 90% to detect Indian names, but the accuracy of all model is below 90%. From this graph, we can state that all algorithms are not sufficient to perform NER to detect Indian names.

The graph above shows the performance of four different NER models for detecting Indian names. The expected accuracy of all four algorithms was above 90%, but the actual accuracy was below 90% for all models. This suggests that no single algorithm is sufficient to perform NER for Indian Fig 3.8 comparision graph of 4 models names.

There are a number of possible reasons for this. One possibility is that the training data used for the algorithms were not representative of the actual distribution of Indian names. Another possibility is that the algorithms were not able to capture the complex morphology and syntax of Indian languages.



To improve the accuracy of NER for Indian names, it is necessary to collect more representative training data and develop algorithms that are better able to handle the morphological and syntactic complexity of Indian languages.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:07/July-2023 Impact Factor- 7.868 www.irjmets.com

The results of this study highlight the need for further Fig loss graph of CRF research and development in the field of NER for Indian names. By addressing the challenges associated with Indian names, it is possible to develop accurate and reliable NER solutions.

V. CONCLUSION

NER is a critical task in NLP that involves identifying and extracting entities from text data. CRF NLP models are a popular approach for NER in NLP, as they can effectively model the dependencies between adjacent tokens in a sequence while making predictions. The results of this study show that the best way to perform NER for Indian names is to use a dictionary. This is because dictionaries can capture the complex morphology and syntax of Indian languages, which is not always possible for machine learning models. Additionally, dictionaries can be easily updated to reflect changes in the Indian naming landscape. The four NER models that were evaluated in this study all had lower accuracy than expected when it came to detecting Indian names. This suggests that no single model is sufficient for this task. However, the dictionary-based approach was able to achieve significantly higher accuracy than the other models. The use of a dictionary is not without its limitations. For example, dictionaries can only identify names that are already included in the dictionary. However, this is not a major limitation in the context of Indian names, as there are a large number of Indian names that are already included in dictionaries.

Overall, the results of this study suggest that the use of a dictionary is the best way to perform NER for Indian names. This approach is more accurate than machine learning models, and it is also more scalable and adaptable to changes in the Indian naming landscape.

In addition to the use of a dictionary, there are a number of other factors that can improve the accuracy of NER for Indian names. In this paper, we have evaluated the performance of four different NER algorithms for detecting Indian names. The results show that the accuracy of all four algorithms is below 90%. This suggests that no single algorithm is sufficient to perform NER for Indian names.

There are a number of possible reasons for this. One possibility is that the training data used for the algorithms is not representative of the actual distribution of Indian names. Another possibility is that the algorithms are not able to capture the complex morphology and syntax of Indian languages.

To improve the accuracy of NER for Indian names, it is necessary to collect more representative training data and develop algorithms that are better able to handle the morphological and syntactic complexity of Indian languages.

VI. FUTURE SCOPE

To Improve Training Data: Collect more representative data for training NER algorithms on Indian names to ensure better coverage and accuracy.

To Develop Enhanced Algorithms: Create algorithms that better handle the complex morphology and syntax of Indian languages, improving NER performance.

VII. REFERENCES

[1] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. arXiv preprint arXiv:2101.01785.

[2] Gaurav Arora. 2020. inltk: Natural language toolkit for indic languages. arXiv preprint arXiv:2009.12534.

[3] Boutsis, S., Demiros, I. , Giouli, V. , Liakata, M. , Papageorgiou, H. And Piperidis, S., “A System For Recognition Of Named Entities In Greek” Proc. International Conference On Natural Language Processing, 2000.

[4] Pasca, Marius, “Acquisition Of Categorized Named Entities For Web Search”, Proc. Conference On Information And Knowledge Management, 2004.

[5] Wang, Lee, Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y. And Li, Y., “Detecting Dominant Locations From Search Queries”, Proc. International Acm Sigir Conference, 2005.

[6] Sánchez, David And Moreno, A., “Web Mining Techniques For Automatic Discovery Of Medical Knowledge”, Conference On Artificial Intelligence In Medicine, 2005.

[7] Tzong-Han Tsai, Richard; Wu S.-H.; Chou, W.-C.; Lin, Y.-C.; He, D.; Hsiang, J.; Sung, T.-Y. And Hsu, "Various Criteria In The Evaluation Of Biomedical Named Entity Recognition", Vol. 6, Bmc Bioinformatic, 2006.

[8] Nadeau, David And Sekine, S., "Named Entities: Recognition, Classification And Use", Special Issue Of Lingvisticæ Investigationes, Vol. 30/1, Pp. 3-26, 2007.

[9] Pramod Kumar Gupta, Sunita Arora "An Approach For Named Entity Recognition System For Hindi: An Experimental Study" In Proceedings Of Ascnt – 2009, Cdac, Noida, India, Pp. 103 – 108.

[10] Shilpi Srivastava, Mukund Sanglikar & D.C Kothari. "Named Entity Recognition System For Hindi Language: A Hybrid Approach" International Journal Of Computational Linguistics (Ijcl), Volume(2):Issue(1):2011.Availableat:
HTTP://CSCJOURNALS.ORG/CSC/MANUSCRIPT/JOURNALS/IJCL/VOLUME2/ISSUE1/IJCL-19.PDF

[11] "Padmaja Sharma, Utpal Sharma, Jugal Kalita"Named Entity Recognition: A Survey For The Indian Languages"(Language In India Www.Languageinindia.Com 11:5 May 2011 Special Volume: Problems Of Parsing In Indian Languages.) Available At :
HTTP://WWW.LANGUAGEININDIA.COM/MAY2011/PADMAJAUTPALJUGAL.PDF.

[12] Lawrence R. Rabiner, " A Tutorial On Hidden Markov Models And Selected Applications In Speech Recognition", In Proceedings Of The Ieee, Vol.77,No.2, February 1989. Available At:
HTTP://WWW.CS.UBC.CA/~MURPHYK/BAYES/RABINER.PDF.

[13] B. Sasidhar#1, P. M. Yohan*2, Dr. A. Vinaya Babu3, Dr. A. Govardhan4" A Survey On Named Entity Recognition In Indian Languages With Particular Reference To Telugu" In Ijcsi International Journal Of Computer Science Issues, Vol. 8, Issue 2, March 2011 Available At :
HTTP://WWW.IJCSI.ORG/PAPERS/IJCSI-8-2-438-443.PDF.