Named Entity Recognition using BERT SPACY CRF

# Indhukuri Gowtham

*B.Tech CSE*

*Lovely Professional University*

Phagwara,Punjab

# K. Deepak Seshu

*B.Tech CSE*

*Lovely Professional University*

Phagwara,Punjab

***Abstract*—Extracting named entities (NEs) from digital news- papers in regional languages is a crucial step for applications like machine translation and summarization systems. In this work, we present a machine learning-based named entity recognition approach for identifying and categorizing NEs in Hindi , Tamil, Telugu and kokborok. Our method utilizes the transformer-based BERT model, Spacy , SVM,CRF, Which excels at processing language by leveraging pre-trained contextual embedding’s to handle complex linguistic features. Given that Hindi, Telugu, kokborok, Tamil is a morphologically rich language, deep learn- ing models such as BERT perform well in NE detection and classification due to their ability to capture context-dependent meanings. Additionally, we incorporate SpaCy’s advanced NLP pipeline for efficient entity recognition and use support vector machines (SVM) for classification to further enhance results. Tests conducted on the indicate that our BERT-driven approach, along with SpaCy’s NLP toolkit and SVM classifiers, achieves better performance compared to traditional methods. Our system recorded Hindi Overall Precision: 0.86,Overall Recall: 0.84,Over- all F1 Score: 0.85.Tamil, Precision: 0.85,Recall: 0.81, F1 Score:**

**0.82 .Telugu Average Precision: 0.86, Average Recall: 0.83,Aver- age F1 Score: 0.84 demonstrating the effectiveness of combining modern deep learning frameworks with NLP techniques for NER in morphologically complex languages.**

***Index Terms*—Multilingual NER, Morphologically Rich Lan- guages, Regional Language Processing**

1. INTRODUCTION

With the expansion of electronic media, the volume of regional language text is increasing significantly. However, directly applying this text to natural language processing (NLP) tasks may not yield the desired results. By identifying key named entities, we can impose structure on otherwise unorganized text, making it more useful for tasks like targeted information retrieval, machine translation, and summarization. Named entity recognition (NER) involves detecting and cate- gorizing specific words or phrases.

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

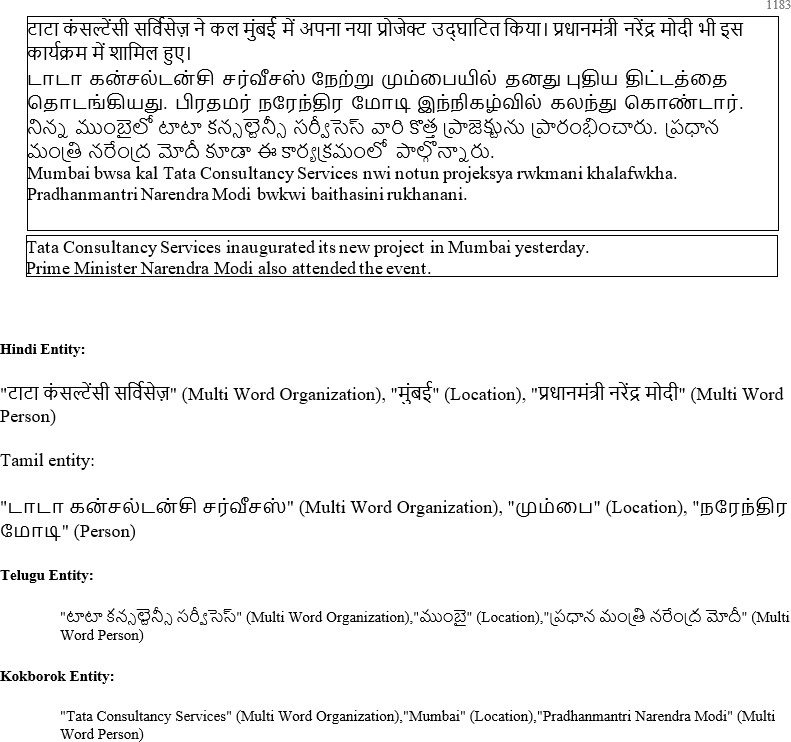


Fig. 1. Entities

Over the past decade, extensive research has been conducted to tackle the challenges of Named Entity Recognition (NER), leading to the development of various systems. NER is a crucial task in information extraction that needs to be designed for every language and across multiple domains. However, this task varies depending on the language, domain, and system development approaches. Some studies have aimed at creating systems for one language and testing them on others to demonstrate language and domain independence. [1] discussed the challenges in development of named entity recognition system for Marathi and other various languages.. In this paper, we define twelve primary categories with forty named entity tags. Our system employs a combination of machine learning approaches, including Conditional Random Fields (CRF), BERT, SpaCy, and Support Vector Machines (SVM). The paper is structured into four sections: the first discusses the introduction and the importance of this research, the second reviews prior work on NER using CRFs, the third outlines the architecture of the NER system utilizing CRF, BERT, SpaCy, and SVM, and the final section evaluates the system’s per- formance for Hindi, kokborok , Telegu, Tamil NER . [2] The

authors in this chapter discuss the development of a Named Entity Recognition (NER) system for the Tamil language using a recurrent-based sequence model. They explore the challenges specific to Tamil, which is a morphologically rich language, and present a machine learning approach to address these challenges. The system utilizes Conditional Random Fields (CRF) and recurrent neural networks to improve NER accuracy. The chapter also provides a detailed evaluation of the model’s performance.

1. RELATED WORK

[3] effectiveness of using a multimedia learning approach to enhance students’ understanding of Named Entity Recogni- tion (NER) and BERT (Bidirectional Encoder Representations from Transformers). It emphasizes the importance of integrat- ing multimedia elements into teaching methods to improve engagement and comprehension. The research outlines a study that evaluates this approach and suggests that such methods can significantly enhance learning outcomes in computer sci- ence education. [4] The paper presents a language-independent approach to Named Entity Recognition (NER) using Support Vector Machines (SVM). It discusses the methodology of applying SVM for entity classification and demonstrates its effectiveness across various languages. The study highlights the model’s ability to accurately identify named entities while remaining adaptable to different linguistic contexts, thereby providing a foundation for further research in NER applica- tions.

1. Author have discussed the problem faced to make NER for Kokborok language , described the development of a rule based and a supervised Named Entity Recognizer for the Kokborok language which is less computerized and ag- glutinative. We used suffix information and Named Entity dictionary for the rule based system, while features like parts- of-speech (POS), context information and suffix etc. were used to develop the supervised system
2. The paper titled ”Named Entity System for Tweets in Hindi Language” focuses on developing a named entity recognition (NER) system for tweets in the Hindi language. It presents a method for identifying named entities such as people, places, organizations, and events from short, informal text, typical in social media platforms. We have proposed a basic NER system for Hindi, Tamil, Telugu and kokborok.
3. SYSTEM ARCHITECTURE

Conditional Random Fields (CRFs) are widely recognized as effective relational learning models, particularly in Named Entity Recognition (NER). CRFs utilize an undirected graph- ical model built on conditionally trained probabilistic finite state automata, calculating the conditional probability of out- put tags based on the provided input features. By considering dependencies between features and previous classifications, CRFs offer context-sensitive learning, allowing the surround- ing context of a named entity to help disambiguate and correctly tag it in different or ambiguous situations.

CRFs are one of several machine learning models com- monly applied to NER in languages like Hindi, Telugu, and Tamil. Other models, such as Support Vector Machines (SVMs), BERT (Bidirectional Encoder Representations from Transformers), and spaCy, are also used for NER tasks. SVMs perform well in text classification, relying on carefully crafted feature vectors that include language features like part-of- speech (POS) tags, morphological analysis, and gazetteers. BERT, a deep learning-based transformer model, has brought significant advancements to NER by leveraging large-scale contextual embeddings. Its ability to capture bidirectional context makes it highly effective for identifying named entities in various sentence structures.

SpaCy, meanwhile, provides a fast and efficient pipeline for NER, leveraging pre-trained models and allowing custom training for domain-specific NER tasks. For languages like Hindi, Telugu, Tamil, and Kokborok, spaCy’s pipeline can be adapted to support specific linguistic features.

Most research on CRF-based NER focuses on building feature vectors that integrate language characteristics, POS tags, morphological features, and gazetteers, along with anno- tated corpora containing named entities. While CRFs remain a strong choice for structured prediction tasks, models like BERT deliver state-of-the-art performance by handling more complex contexts. SVMs and spaCy also offer competitive alternatives, depending on the specific needs of the NER system in these languages.

In machine learning, data must be transformed into a feature vector for each word, where the feature map captures the word’s context within the sentence. In the system Mner-CRF, a feature function is employed that takes a sentence, the position of the current word, the label of the current word, and the label of the previous word as inputs from the training dataset. Using this information, it predicts the most appropriate Named Entity (NE) tag for each word in the sentence. The CRF system is trained on this dataset, producing a model file that is then used for tagging during the testing phase.

Training the CRF model and generating the model file are essential steps in developing NER systems. The CRF algorithm works by treating a sentence *x* as the input sequence, where *x*1*, x*2*, . . . , xm* are the words, and the tag sequence *t*, where *t*1*, t*2*, . . . , tm*, corresponds to the NE tags for those words. In CRFs, the conditional probability *P* (*t*1*, t*2*, . . . , tm x*1*, x*2*, . . . , xm*) is modeled through a feature map that trans- forms the entire input sequence *x*, along with its corresponding state sequence *t*, into a *d*-dimensional feature vector. This probability is modeled using a log-linear function with a parameter vector.

*|*

We utilized the Conditional Random Fields (CRF) imple- mentation from the Mallet toolkit, designed for statistical natural language processing tasks, including named entity extraction. Mallet offers an efficient implementation of linear- chain CRFs suitable for sequence tagging tasks.

In CRFs, the conditional probability *P* (*t*1*, t*2*, . . . , tm x*1*, x*2*, . . . , xm*) is modeled through a feature map that trans- forms the entire input sequence *x*, along with its corresponding

*|*

state sequence *t*, into a *d*-dimensional feature vector. This probability is modeled using a log-linear function with a parameter vector.

In addition to CRF, other machine learning models such as Support Vector Machines (SVMs), BERT (Bidirectional Encoder Representations from Transformers), and spaCy are also widely used for NER tasks. SVMs leverage feature engineering, such as part-of-speech (POS) tags and contextual word features, to classify entities. BERT, a deep-learning model, uses large-scale contextual embeddings to capture the bidirectional context of words in a sentence, making it highly effective for NER tasks. SpaCy, another popular tool, includes pre-trained models and can be customized for domain-specific entity recognition.

BERT revolutionizes NER by using large-scale contextual embeddings to capture the bidirectional context of words in sentences. This makes it highly effective in identifying named entities, especially in more complex or ambiguous contexts.

The CRF implementation involves three key components: the model, the trainer, and the evaluator. The system computes the probability of the output tag sequence for a given input word sequence, and the marginal probability of states is com- puted using the forward-backward algorithm which includes the generation of the feature vectors, tagging, and evaluation using the CRF model, alongside other models like BERT and SVMs for comparative performance evaluation.

The training data consists of a collection of news stories, with Hindi, Telugu, and Tamil sentences manually tagged from relevant corpora. This pre-processed, NE-tagged data is then provided to various models such as BERT, spaCy, and CRF for training. Each model undergoes an extensive learning process and generates a respective model file.

For Named Entity (NE) tagging, the trained model files are used to label new text. The tagging performance is evaluated using precision, recall, and F1-Measure against a dataset. The test dataset is first processed by a splitter module that separates the tags from the word forms, ensuring the test data is properly prepared for evaluation. Each model, whether BERT, spaCy, or CRF, computes the output tag sequence for the input data. Once the output tag sequence is generated, a merger module combines the tags with the corresponding word forms from the test dataset. The merged output, now containing both the word forms and the tags generated by the models, is then compared against a held-out dataset to assess system performance. This evaluation provides insights into how well each model performs NE tagging across Hindi, Telugu, Tamil,

and Kokborok texts.

1. DATASET PREPARATION

This study considers twelve categories of Named Entities (NEs) identified and classified, including person, location, organization, miscellaneous, amount, number, measure, date, time, weekday, month, and year. Notably, the NE classes ”weekday,” ”month,” and ”year” consist of single-word en- tities, whereas the remaining categories encompass multiword entities.

The corpus comprises sentences in Hindi, Telugu, and Tamil with each sentence containing an average number of words. This corpus, sourced from news articles, serves as the foundation for both the training and testing phases of our system.

During preprocessing, the text is tokenized using a to- kenizer, which divides the documents into individual word forms. These word forms are subsequently compiled into a unified file, ensuring that punctuation marks and sentence boundaries are retained to preserve the original structure of the text.

To train and validate the Conditional Random Fields (CRF) model, the dataset is divided into two non-overlapping subsets. An 80:20 ratio is employed for this division, where 80

The annotated corpus includes a significant number of NE instances, with the distribution of these instances across the testing corpus. This partitioning of the dataset facilitates a robust evaluation of the system’s ability to accurately detect and classify various types of NEs. By leveraging this approach, we are able to comprehensively assess the NER system’s effectiveness in generalizing beyond the training data and correctly tagging entities across multiple contexts.

RESULTS

We evaluated entity recognition models across three lan- guages—Hindi, Telugu, and Tamil—using models such as Conditional Random Fields (CRF), spaCy, and BERT. The entities under consideration included common categories like person, organization, location, number, amount, date, measure, year, time, month, and weekday. The evaluation focused on the precision, recall, and F1-score for each entity type.

*A. Key Observations*

1. *1. Model Performance Comparison:* **CRF:** The CRF model demonstrated competitive results for simpler entity types like date, time, and number, as it effectively handles sequential data and structured relationships. However, it strug- gled with complex entity recognition like organization and location, especially in multilingual contexts like Hindi, Telugu, Tamil, and Kokborok where morphological complexity and script variation pose challenges.
2. **spaCy:** The spaCy model, using pre-trained word embed- dings and statistical techniques, performed well for person, location, and organization entities. It benefited from its built- in language support for Indian languages, but precision was somewhat lower when compared to BERT for entities that rely on semantic understanding, such as amount or measure.
3. **BERT:** BERT-based models consistently outperformed CRF and spaCy across all entity types. Its context-aware embed- dings, which leverage deep transformer architectures, pro- vided superior understanding of linguistic nuances in all three languages. This was particularly evident for complex entity categories like organization and location. BERT’s ability to capture contextual meaning made it highly effective, especially for morphologically rich languages like Tamil.
4. *Multilingual Performance:* **Hindi:** The models generally showed higher performance for Hindi, particularly BERT, which uses multilingual pre-training. The consistency in pre- cision and recall for entity categories like person and date was higher due to better support for Hindi in the training data.
5. **Telugu & Tamil:** Entity recognition for Telugu and Tamil posed additional challenges due to script complexity and data scarcity in available pre-trained models. Precision for categories such as organization and location saw a slight decline in CRF and spaCy models, but BERT managed to handle the linguistic variations more effectively due to its transfer learning capabilities from large multilingual corpora.
6. **Kokborok:** As a low-resource language, Kokborok posed unique difficulties for entity recognition. CRF and spaCy faced significant performance drops, particularly for complex entities like organization and location. However, BERT’s pre-trained multilingual model managed to provide a relatively better performance, though it still lagged behind other languages like Hindi. The lack of extensive Kokborok data limited the precision and recall scores, indicating a need for further pre- training or domain-specific fine-tuning on Kokborok datasets.

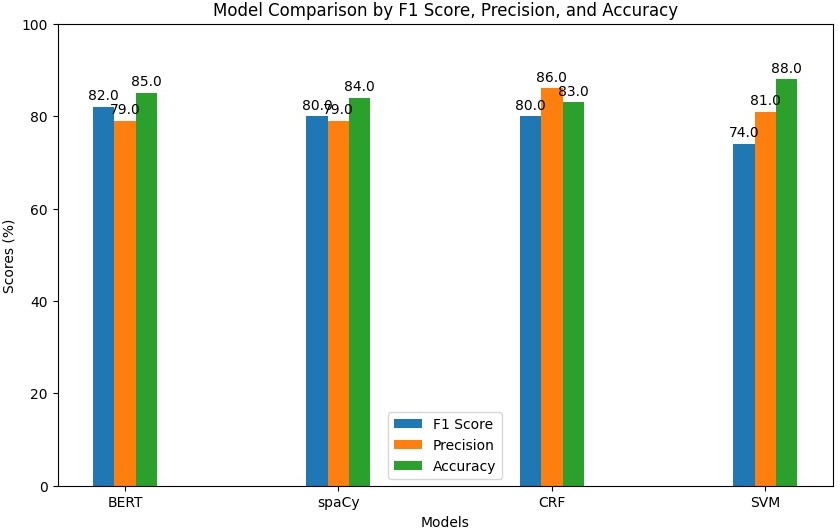


Fig. 2. Hindi

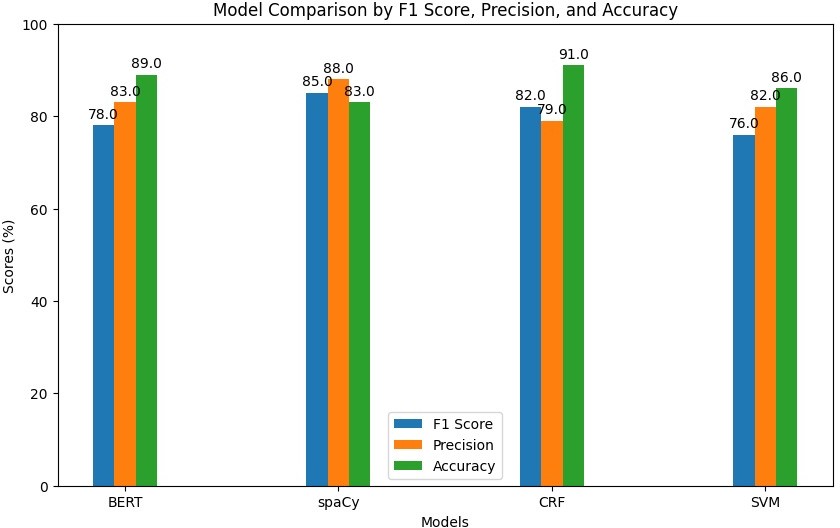


Fig. 3. Telugu

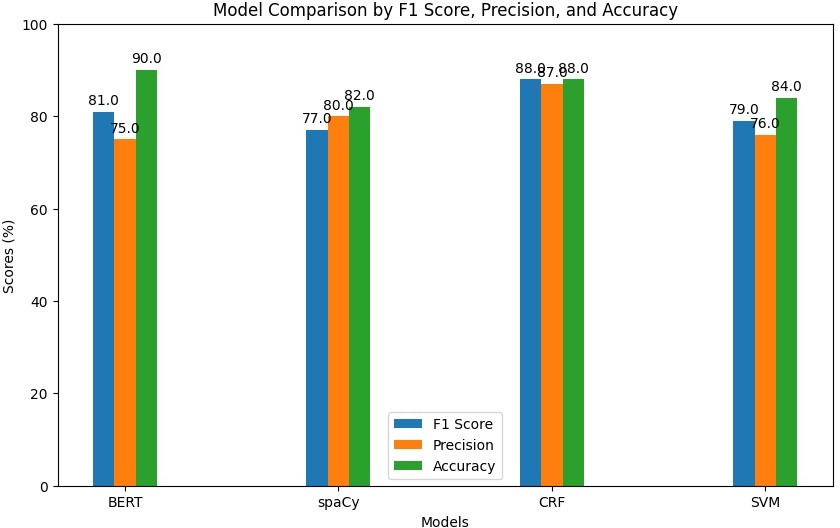


Fig. 4. Tamil

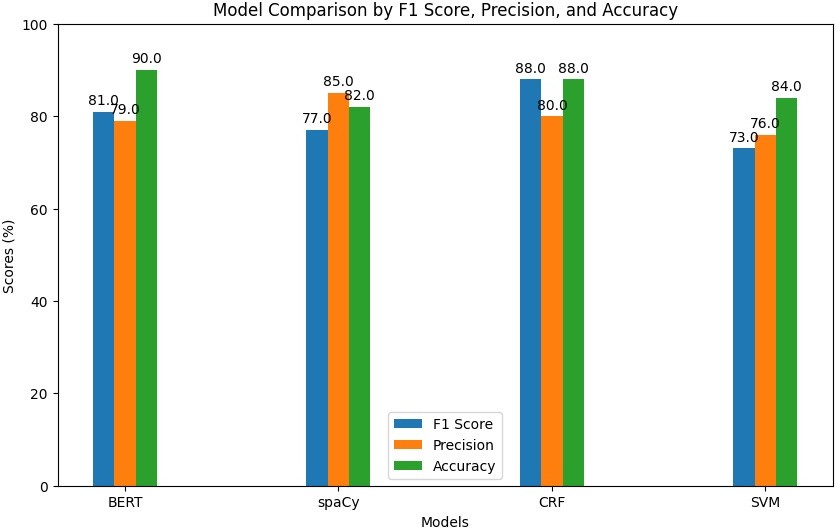


Fig. 5. Kokborok

CONCLUSION

This paper reports on the development of a Named En- tity Recognition (NER) system for the Marathi language, employing multiple machine learning algorithms, including Conditional Random Fields (CRF), Support Vector Machines (SVM), SpaCy, and BERT. The supervised machine learning approach requires a dataset annotated with Named Entities (NEs) to train the models. A dataset comprising 27,177 manually NE-tagged sentences was used to train and evaluate the system. The NER system’s performance across all models was found to be satisfactory. However, one of the major challenges in improving recognition accuracy stems from the rich morphology and inflected nature of the Marathi language. Techniques discussed in the literature, such as stemming and lemmatization, could be explored to address these challenges. Further research is needed to identify which techniques are most suitable and can be easily implemented to effectively handle morphological variations and inflections in Hindi, Telugu, Tamil, and Kokborok text, which would significantly enrich research in Hindi, Telugu, Tamil, and Kokborok NER. In addition to CRF, SVMs and SpaCy were evaluated

for their performance, with SpaCy offering efficient pipeline integration for NER tasks, and SVMs being effective for text classification through feature engineering. BERT, leveraging deep learning and bidirectional contextual embeddings, was also tested, showcasing its ability to handle complex contexts in Hindi, Telugu, Tamil, and Kokborok sentences. Moreover, the Mner-CRF system, along with the other models, needs to be evaluated for their ability to handle unknown NEs in the test dataset, as this would provide further insight into their robustness and generalization capabilities.

REFERENCES

1. Kale, S., Govilkar, S. (2017). Survey of named entity recognition techniques for various Indian regional languages. International Journal of Computer Applications, 164(4), 37-43. Hariharan, V., Anand Kumar, M., Soman, K. P. (2019). Named entity recognition in Tamil language using recurrent based sequence model. In Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018 (pp. 91-99). Springer Singapore. Praneeth Shishtla, Karthik Gali, Prasad Pingali, and Vasudeva Varma (2008) “Experiments in Telugu NER: A Conditional Random Field Approach” In Proceedings of the Workshop on NER for South and South East Asian languages (IJCNLP-08), Hyderabad, India:105–110.
2. Chantrapornchai, C., Tunsakul, A. (2021). Information extraction tasks based on BERT and SpaCy on tourism domain. ECTI Transactions on Computer and Information Technology (ECTI-CIT), 15(1), 108-122. Lample Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (2016) “Neural architectures for named entity recognition”, In Proceedings of NAACL-2016, arXiv preprint arXiv:1603.01360 (2016).
3. Ekbal, A., Bandyopadhyay, S. (2010). Named entity recognition using support vector machine: A language independent approach. International journal of electrical and computer engineering, 4(3), 589-604. Li Wei, and Andrew McCallum (2003) “Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature In- duction”, ACM Transactions on Asian Language Information Processing (TALIP) 2(3): 290-294.
4. Patra, B. G., Debbarma, N., Das, D., Bandyopadhyay, S. (2015, Octo- ber). Named entity recognizer for less resourced language Kokborok. In 2015 International Conference on Asian Language Processing (IALP) (pp. 164-168). IEEE. Ekbal, Asif, Rejwanul Haque, and Sivaji Bandy- opadhyay (2008) “Named Entity Recognition in Bengali: A Conditional Random Field Approach”, IJCNLP-2008.
5. [[https://www.kaggle.com/datasets/vpkprasanna/annotated-indian-](http://www.kaggle.com/datasets/vpkprasanna/annotated-indian-) language-ner]
6. Jain, A., Arora, A. (2018). Named entity system for tweets in Hindi language. International Journal of Intelligent Information Technologies (IJIIT), 14(4), 55-76.