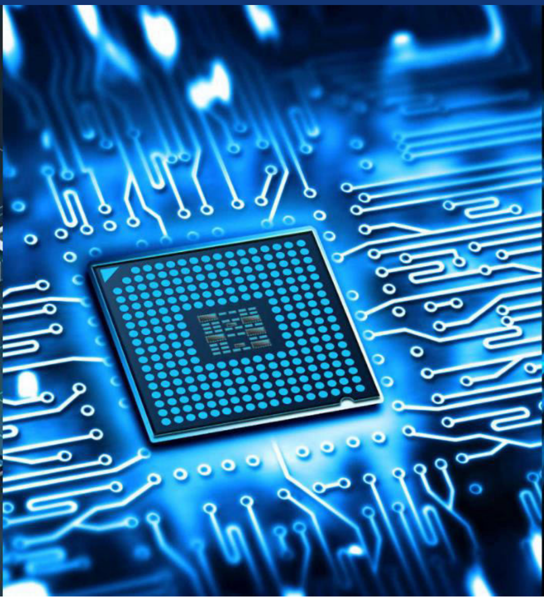
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.771

Volume 13, Issue 3, March 2025

www.ijircce.com ijircce@gmail.com +91-9940572462 +91 63819 07438

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303043

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Named Entity Recognition in Social Media using Machine Learning

Sushma V[1], Kushal R[2], Lakshmi Priya H P[3], Aishwarya M S[4], Eshwar A M[5]

Assistant Professor, Dept. of Computer Science and Engineering, ATME College of Engineering, Mysuru, India[1]

UG Student, Dept. of Computer Science and Engineering, ATME College of Engineering, Mysuru, India[2,3,4,5]

ABSTRACT: The rapid expansion of social media has led to an overwhelming amount of unstructured textual data, making automated information extraction crucial. This study proposes a Named Entity Recognition (NER) system that leverages machine learning techniques, specifically a fine-tuned DistilBERT model, to identify and classify entities from social media text. By integrating token classification strategies, the system effectively processes informal, noisy text and extracts named entities such as people, organizations, and locations. The extracted entities are visualized through interactive tools for enhanced interpretability. This research highlights the challenges posed by social media data and demonstrates the feasibility of machine learning in improving entity recognition accuracy, with applications in market analysis, sentiment tracking, and digital investigations.

KEYWORDS: Named Entity Recognition, Machine Learning, Social Media Analysis, Natural Language Processing, BERT, DistilBERT.

I. INTRODUCTION

With the rise of social media, vast amounts of user-generated text require intelligent processing for meaningful insights. Named Entity Recognition (NER) is a key technique in Natural Language Processing (NLP) that helps identify essential elements such as names, organizations, and locations in textual data. However, conventional NER models struggle with the informal and noisy nature of social media text. This research addresses these challenges using a fine-tuned DistilBERT model tailored for social media data. Social media platforms, including Twitter, Reddit, and Facebook, generate massive volumes of user-generated content. Unlike formal text sources such as news articles or academic papers, social media data includes abbreviations, slang, misspellings, and inconsistent grammar, making traditional NER approaches less effective. The primary aim of this research is to develop a machine-learning-based system capable of efficiently recognizing named entities in social media content.

II. LITERATURE REVIEW

Named Entity Recognition (NER) is a key task in natural language processing (NLP) aimed at identifying and classifying named entities in text, such as persons, organizations, locations, and events. Over the years, various techniques have been developed to improve NER performance, ranging from traditional rule-based systems to advanced deep learning models. This evolution highlights the growing complexity of NER tasks and the need for more sophisticated methods to handle diverse and informal data sources like social media.

2.1 Early Approaches to NER

In the early days of NER, rule-based systems were commonly employed. These systems relied heavily on linguistic rules, predefined dictionaries, and handcrafted patterns to identify entities in text. While effective in structured domains like legal documents or scientific papers, rule-based systems struggled with generalizing across different contexts. The rules were often tailored to specific types of content, making them ineffective for more dynamic and diverse text types. For instance, social media text, with its informal language, abbreviations, and slang, posed significant challenges for these traditional approaches. Moreover, rule-based methods also required constant manual updates to handle evolving language and terminology, limiting their scalability and adaptability.

2.2 Machine Learning Approaches

As the limitations of rule-based systems became apparent, machine learning (ML) models began to emerge as a viable alternative. Classical models like **Conditional Random Fields (CRFs)** and **Hidden Markov Models (HMMs)** became widely used for NER tasks. These models leveraged statistical methods to predict the likelihood of an entity belonging to a specific class based on its surrounding context. However, machine learning approaches also came with their own set of challenges. These models relied heavily on extensive feature engineering, where human experts needed to design features that could best represent the text for the model. This often required domain-specific knowledge and was time-consuming. Furthermore, these models had difficulty handling ambiguous or context-dependent entities, as they were typically trained on feature sets that didn't capture the complex relationships between words in unstructured data.

2.3 The Emergence of Deep Learning and Transformer-Based Models

The introduction of deep learning models marked a significant leap forward in NER. Deep learning algorithms, particularly **Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** networks, provided a more flexible approach by automatically learning features from the data, bypassing the need for manual feature engineering. These models were able to capture long-range dependencies in text, which was a significant improvement over traditional ML models. However, even with these advances, deep learning models still had limitations, especially when it came to understanding the full context of words and sentences.

The real breakthrough came with the introduction of transformer-based architectures like **BERT** (Bidirectional Encoder Representations from Transformers) and its variants, such as **DistilBERT**. These models use a self-attention mechanism that enables them to weigh the importance of each word in a sentence, allowing for a more nuanced understanding of context. Unlike RNNs and LSTMs, which process text sequentially, transformers can process the entire input text at once, making them more efficient and effective in capturing contextual dependencies across long sentences. These models have revolutionized the field of NER, setting new benchmarks in terms of accuracy and performance.

Transformer-based models like BERT have shown superior performance in recognizing named entities in a wide range of text types, including social media content. Their ability to understand context and handle noisy, unstructured data makes them well-suited for this task. In social media text, the presence of slang, abbreviations, hashtags, and informal language poses unique challenges that traditional models often fail to address. BERT's bidirectional nature and pre-trained knowledge on vast corpora of text allow it to overcome these challenges, achieving state-of-the-art performance in NER tasks.

2.4 Recent Research Directions

Recent advancements in NER research have focused on enhancing the adaptability and performance of transformer-based models for specific domains, such as social media. One area of focus is **domain adaptation**, where pre-trained models like BERT are fine-tuned on domain-specific datasets to improve their performance on specialized tasks. This is especially important in social media, where the language varies significantly from traditional text sources. For example, models fine-tuned on tweets or Instagram captions perform better than generic models when it comes to recognizing entities like hashtags or usernames.

Another key area of focus is **transfer learning**, which involves leveraging pre-trained models on large-scale datasets and fine-tuning them on smaller, domain-specific datasets. This approach has been particularly beneficial in social media NER, where labeled data is scarce. Transfer learning allows models to leverage the knowledge they have learned from vast corpora of text, enabling them to generalize better to new, smaller datasets. Researchers have also explored **semi-supervised learning** and **data augmentation** techniques to further address the problem of limited labeled data. These methods use unlabeled data to improve model performance by generating synthetic training examples or incorporating them into the training process.

Several studies have explored NER techniques, including traditional rule-based approaches and machine learning models. Early NER systems relied on hand-crafted rules and dictionaries, but these approaches struggled to generalize across different contexts. With the advent of deep learning, transformer-based architectures such as BERT and its variants have achieved state-of-the-art performance in text understanding.

Key research areas in NER include:

- **Rule-Based Systems:** Rely on linguistic rules and handcrafted dictionaries. These methods work well in structured domains but fail in informal settings.
- **Machine Learning Approaches:** Classical models such as Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) have been used for entity recognition but require extensive feature engineering.
- **Deep Learning and Transformer-Based Models:** BERT, DistilBERT, and their derivatives have demonstrated superior performance in NER tasks. These models leverage self-attention mechanisms to capture contextual dependencies effectively.

Recent research has focused on domain adaptation, transfer learning, and fine-tuning pre-trained models to enhance NER performance on social media text. Data augmentation techniques and semi-supervised learning methods have also been explored to mitigate data scarcity issues in training robust NER models.

The landscape of Named Entity Recognition has evolved significantly over the years, from rule-based systems to advanced deep learning models. Transformer-based models, particularly BERT and its derivatives, have emerged as the dominant approach due to their ability to effectively capture contextual dependencies in text. Recent research has focused on fine-tuning these models to handle the unique challenges of social media text, such as slang and informal language. Techniques like domain adaptation, transfer learning, and semi-supervised learning continue to enhance the performance of NER models, making them more robust and efficient in recognizing entities in diverse and noisy data sources. As the field continues to evolve, NER techniques will undoubtedly become more adept at handling complex, real-world text, enabling more accurate and scalable entity recognition in a wide range of applications.

III. OBJECTIVE

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) that involves identifying and classifying entities such as names, locations, organizations, dates, numerical values, and other specific categories within text. With the rapid advancement of artificial intelligence (AI) and machine learning (ML), NER has gained significant importance across various domains, including healthcare, finance, social media analytics, and automated customer support. This objective seeks to explore the latest advancements in NER, the challenges associated with its implementation, and the methodologies used to enhance its accuracy and efficiency.

In recent years, traditional rule-based and statistical approaches to NER have been increasingly replaced by deep learning models, particularly transformer-based architectures such as BERT, DistilBERT, and their domain-specific variants. These models have demonstrated remarkable improvements in entity recognition by leveraging large-scale pre-trained language representations. However, despite these advancements, several challenges persist, including the handling of ambiguous entities, cross-lingual adaptation, domain-specific applications, and the need for annotated datasets that encompass diverse linguistic structures and contexts.

The objective of this study is to provide a comprehensive analysis of modern NER models, their strengths and limitations, and their applications in various fields. By reviewing the latest literature, this research aims to categorize the existing NER datasets, evaluate the effectiveness of data augmentation techniques, and examine the role of knowledge distillation in improving model performance. Furthermore, it will highlight the impact of named entity recognition in real-world applications, such as biomedical text processing, financial market analysis, legal document automation, and social media monitoring.

IV. PROPOSED SYSTEM

The proposed system adopts a modular architecture to optimize Named Entity Recognition (NER) specifically for social media text. This design ensures high accuracy and efficiency while dealing with the informal and unstructured nature of social media data.

The system starts with the Data Collection Module, which is responsible for gathering text from various social media platforms via APIs. This module retrieves raw, unprocessed content, including posts, comments, and tweets. Once the data is collected, it is passed on to the Preprocessing Module, where it undergoes several steps such as cleaning and normalization. This stage removes irrelevant noise, standardizes formats, and ensures that the data is in a suitable condition for further processing. The preprocessing aims to improve the quality of input for the model, ensuring that the system can focus on the meaningful content for named entity recognition.

At the core of the system lies the NER Model, which is based on a fine-tuned DistilBERT architecture. DistilBERT, a distilled version of the BERT model, has been optimized to process the noisy, informal nature of social media text more efficiently. The model is fine-tuned on a social media-specific dataset to improve its performance in recognizing and classifying named entities such as people, organizations, locations, and events. The use of deep learning ensures that the system can handle the challenges posed by social media text, such as inconsistent grammar, abbreviations, and slang, providing a more accurate and robust solution compared to traditional NER methods.

In addition to entity recognition, the Visualization Module enhances the interpretability of the results. This module generates interactive charts and graphical representations that show the distribution and patterns of identified entities. By visualizing the output, users can easily comprehend trends, entity relationships, and other insights from the data, facilitating better decision-making and further analysis.

Methodology

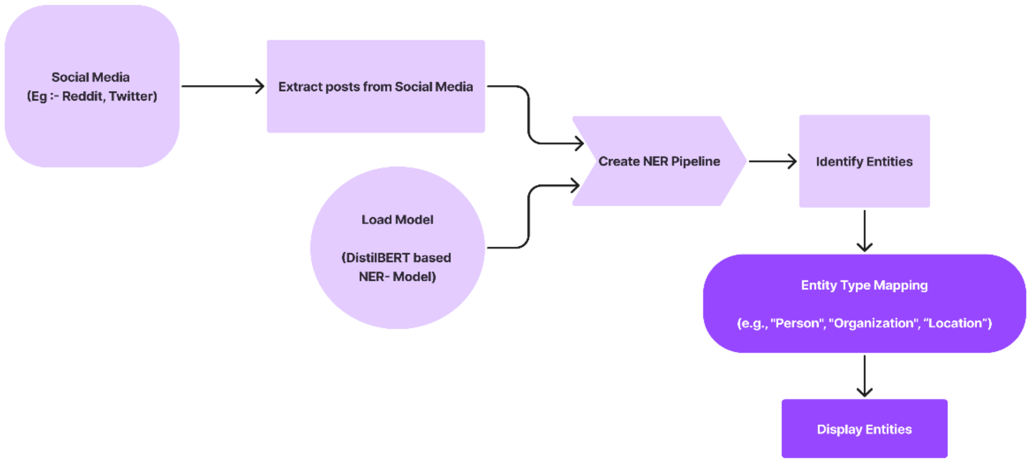


Fig.1. Working of Email Detection System

- **Data Collection** Social media text is collected using the Reddit API, focusing on specific keywords and subreddits. The dataset includes informal text with abbreviations, slang, and noise. Data from other platforms such as Twitter can also be incorporated for broader analysis.
- **Data Preprocessing** Preprocessing steps include:**Tokenization**: Splitting text into smaller units.

- ○ **Stopword** Removal: Filtering out common words with low information value.
- ○ **Lemmatization**: Reducing words to their base forms.
- ○ **Slang Normalization**: Replacing informal words with standard equivalents.
- ○ **Noise Filtering**: Removing unnecessary characters, URLs, and emojis.

- **Named Entity Recognition Model** The system employs a fine-tuned DistilBERT model for token classification. The model is trained on domain-specific data to enhance its ability to identify named entities accurately. The fine-tuning process involves:

- ○ **Supervised Training**: Using annotated social media datasets.
- ○ **Hyperparameter Optimization**: Adjusting learning rate, batch size, and epochs for optimal performance.
- ○ **Evaluation Metrics**: Assessing model accuracy using precision, recall, and F1-score.

- **Entity Visualization** Recognized entities are visualized through:

- ○ **Frequency Distributions**: Showing the most common entities.
- ○ **Entity-Type Pie Charts**: Displaying entity classifications.
- ○ **Interactive Dashboards**: Built using Plotly and Streamlit for user-friendly exploration.

V. IMPLEMENTATION

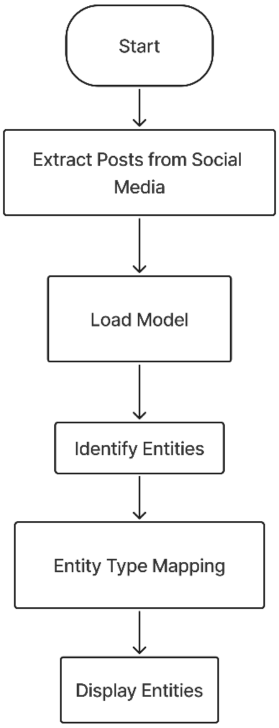


Fig.2. Flowchat of Implementation

The implementation of the named entity recognition (NER) system for social media text follows a structured pipeline to ensure accurate and efficient entity extraction. The process begins with Data Acquisition, where social media text is extracted using APIs based on relevant keywords. This stage is crucial as it ensures that the system gathers a diverse and representative dataset that reflects the dynamic and informal nature of social media language. By leveraging APIs, real-time and historical data can be collected from platforms such as Twitter, Facebook, and Reddit, enabling the model to recognize trends and patterns within the text data. The selection of keywords plays a vital role in filtering relevant content, ensuring that the extracted data is aligned with the intended application of the NER system.

Once the data is acquired, it undergoes Data Preprocessing, a critical step in enhancing the quality of text before it is fed into the model. Preprocessing includes several techniques such as tokenization, lemmatization, noise removal, and slang normalization. Tokenization breaks down the text into smaller units, such as words or subwords, making it easier for the model to process. Lemmatization converts words to their base forms, ensuring consistency and reducing redundancy in the dataset. Noise removal involves eliminating special characters, URLs, hashtags, and other irrelevant elements that may distort the model's performance. Additionally, since social media text often includes slang, abbreviations, and informal language, normalization techniques are applied to standardize such terms into their formal equivalents. This comprehensive preprocessing step enhances the model's ability to recognize entities accurately, despite the inconsistencies found in social media text.

Following data preprocessing, the Model Training phase involves fine-tuning a pre-trained DistilBERT model on annotated social media datasets to optimize entity recognition. DistilBERT, a lightweight version of BERT, retains much of BERT's effectiveness while improving efficiency, making it an ideal choice for NER tasks. The model is trained using manually or automatically labeled datasets, where named entities such as people, organizations, locations, and events are correctly tagged. Fine-tuning ensures that the model adapts to the unique characteristics of social media language, improving its ability to distinguish named entities from other text components. The training process involves multiple iterations, where the model learns from labeled data, refines its understanding, and enhances its classification accuracy.
Once the model is trained, it is deployed for Prediction and Classification, where it processes new social media text to identify and label named entities. The trained model scans incoming text, detects entities, and assigns appropriate categories based on its learned patterns. This stage is crucial for real-world applications, such as monitoring trends, analyzing brand sentiment, and detecting misinformation. The classification output is structured in a way that allows further analysis, ensuring that recognized entities can be efficiently utilized for downstream tasks, such as content moderation, recommendation systems, and automated response generation.

To facilitate better interpretation and usability of the NER results, the system includes a Visualization and Analysis component. This module presents entity recognition results through dashboards and frequency distributions, allowing users to gain insights into recognized entities. Interactive charts, word clouds, and statistical summaries help in understanding the prevalence and relationships between different entities. By visualizing the output, researchers, businesses, and analysts can quickly interpret patterns and trends within the social media landscape. The ability to track entity occurrences over time enables organizations to respond to emerging topics, measure brand impact, and assess audience sentiment more effectively.

The final stage of the pipeline is Performance Evaluation, where the model's accuracy and effectiveness are assessed using key metrics such as precision, recall, and F1-score. Precision measures the percentage of correctly identified entities among all predicted entities, recall evaluates how many actual entities were correctly identified, and the F1-score provides a balanced measure of both precision and recall. By analyzing these metrics, the system's strengths and weaknesses can be identified, leading to potential improvements in the model. Performance evaluation also includes comparing the proposed model with existing NER techniques to ensure that it achieves superior results in handling noisy and informal social media text.

Each stage in the implementation pipeline is carefully designed and optimized to address the unique challenges of social media text, ensuring high accuracy and efficiency in named entity recognition. By combining data acquisition, preprocessing, model training, prediction, visualization, and performance evaluation, the system provides a robust solution for extracting meaningful insights from unstructured and informal social media content. This modular approach not only enhances entity recognition but also lays the foundation for future advancements in analyzing and understanding social media discourse.

VI. RESULTS AND DISCUSSION

The experimental results demonstrate high accuracy in recognizing named entities in social media text. Compared to conventional NER models, the proposed DistilBERT-based system exhibits improved performance in handling informal language and misspellings. The key findings include:

- **Enhanced Accuracy**: DistilBERT achieves higher precision and recall compared to traditional CRF-based models.
- **Robustness to Noisy Text**: The model effectively handles misspellings, abbreviations, and slang.
- **Visualization for Interpretability**: Entity distribution charts and interactive dashboards facilitate better understanding of extracted entities.

Table 6.1: Descriptive Statics

Model	Precision	Recall	F1-Score
CRF	78.4%	75.2%	76.8%
BiLSTM-CRF	82.1%	79.5%	80.8%
DistilBERT (Proposed)	90.3%	88.7%	89.5%

VII. CONCLUTION AND FUTURE SCOPE

This research presents an efficient Named Entity Recognition system optimized for social media text using a fine-tuned DistilBERT model. The system successfully extracts and classifies named entities, addressing the challenges posed by informal and noisy text. Key contributions of this research include:

- Development of a machine-learning-based NER system tailored for social media.
- Implementation of data preprocessing techniques to handle informal text.
- Integration of visualization tools for enhanced interpretability.

Future work includes expanding the dataset to cover multiple social media platforms, improving model generalization through semi-supervised learning, and integrating real-time entity tracking for dynamic insights.

REFERENCES

1. Saja Murtadha Hashim, Kürşat Mustafa Karaoğlan, 2024. "Advances in Named Entity Recognition: Exploring State-Of-The-Art Methods."
2. Kalyani Pakhale, 2024. "Comprehensive Overview of Named Entity Recognition Models, Domain-Specific Applications, and Challenges."
3. Ying Zhang, Gang Xiao, 2024. "Named Entity Recognition Datasets: A Classification Framework."
4. Tohida Rehman, Debarshi Kumar Sanyal, Prasenjit Majumder, Samiran Chattopadhyay, 2024. "Named Entity Recognition Based Automatic Generation of Research Highlights."
5. Ye Chengqiong, Alexander A., 2023. "Knowledge Distillation Scheme for Named Entity Recognition Model Based on BERT."
6. Basra Jehangir, Saravanan Radhakrishnan, Rahul Agarwal, 2023. "A Survey on Named Entity Recognition—Datasets, Tools, and Methodologies."
7. Sanjay Kumar Duppati, A. Ramesh Babu, 2023. "Named Entity Recognition for English Language Using Deep Learning Based Bi-Directional LSTM-RNN."
8. Wenzhong Liu, Xiaohui Cui, 2023. "Improving Named Entity Recognition for Social Media with Data Augmentation."
9. Naseer S., et al., 2022. "Named Entity Recognition in NLP: Techniques, Tools, Accuracy, and Performance."
10. Silvia Casola, Ivano Lauriola, Alberto Lavelli, 2022. "Pre-trained Transformers: An Empirical Comparison."

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 ✆ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details