NAMED ENTITY RECOGNIZATION

Presented By :

Gowtham Indukuri[08]

Deepak Seshu Reddy[28]

Lovely Professional University | 2024

ABSTRACT

This study presents a machine learning approach to named entity recognition (NER) for regional languages: Hindi, Tamil, Telugu, and Kokborok. Using BERT, SpaCy, SVM, and CRF, our model effectively processes these morphologically rich languages, capturing context-dependent meanings. Tests show strong performance, with FI scores of 0.85 for Hindi, 0.82 for Tamil, and 0.84 for Telugu, demonstrating the value of combining advanced NLP techniques with deep learning for NER.

OVERVIEW

- Introduction
- Problem
- Literary Review
- Theoretical
- Objectives
- Hypothesis
- Methodology
- Implementation
- Result
- Conclusion
- Recommendation
- Thank You

INTRODUCTION

NER helps in understanding the structure and meaning of text by pinpointing important names and categories, enabling deeper insights and more accurate data extraction.

Some practical applications include:

- Information extraction for organizing large text corpora (e.g., categorizing news articles).
- Customer service automation, where NER helps identify and respond to key entities in user queries.
- Improved search engine results by helping algorithms understand user queries with greater specificity.
- Healthcare for extracting medical terms, patient information, and drug names from clinical records.

PROBLEM

Our approach combines BERT's deep learning capabilities with SpaCy's NLP pipeline and SVM classifiers, leveraging BERT's contextual embeddings to capture linguistic nuances. By integrating SpaCy's efficient entity recognition and using SVM for refined classification, we achieve higher precision and adaptability, particularly in low-resource languages. This hybrid model enhances NER performance across all four languages.

First Problem

Regional languages like Hindi, Tamil, Telugu, and Kokborok are morphologically rich and complex, posing challenges for Named Entity Recognition (NER). Conventional NLP models lack the contextual understanding needed to accurately identify named entities in these languages, resulting in lower accuracy and performance.

Second Problem

Existing NER systems struggle to balance efficiency with precision, especially for languages with limited resources, such as Kokborok. Traditional models often underperform in these low-resource languages, limiting the system's effectiveness in multilingual applications.

LITERARY REVIEW

1 NER in Morphologically Rich Languages: Traditional NER models struggle with languages like Hindi, Tamil, Telugu, and Kokborok due to their complex morphology, requiring advanced approaches for accuracy.

2 Machine Learning Advances in NER: Techniques like CRF and SVM enhance NER through feature engineering but lack the adaptability needed for diverse, multilingual applications.

3 Transformers and Contextual Embeddings: Transformer models like BERT significantly improve NER by capturing complex language structures, especially when combined with SpaCy for efficient pipeline integration and classification.

THEORETICAL FRAMEWORK

- ## Overview

  This research aims to improve Named Entity Recognition (NER) for complex regional languages like Hindi, Tamil, Telugu, and Kokborok. Traditional NER models struggle with these languages' unique linguistic structures. We address this using advanced models: BERT for contextual embeddings, SpaCy for efficient processing, and SVM for precise classification.

- ## Proponents

  Researchers have advanced NER by integrating CRF and SVM for multilingual recognition and BERT to capture deep contextual meanings. Their work underscores the effectiveness of combining machine learning with contextual embeddings for complex languages.

OBJECTIVES

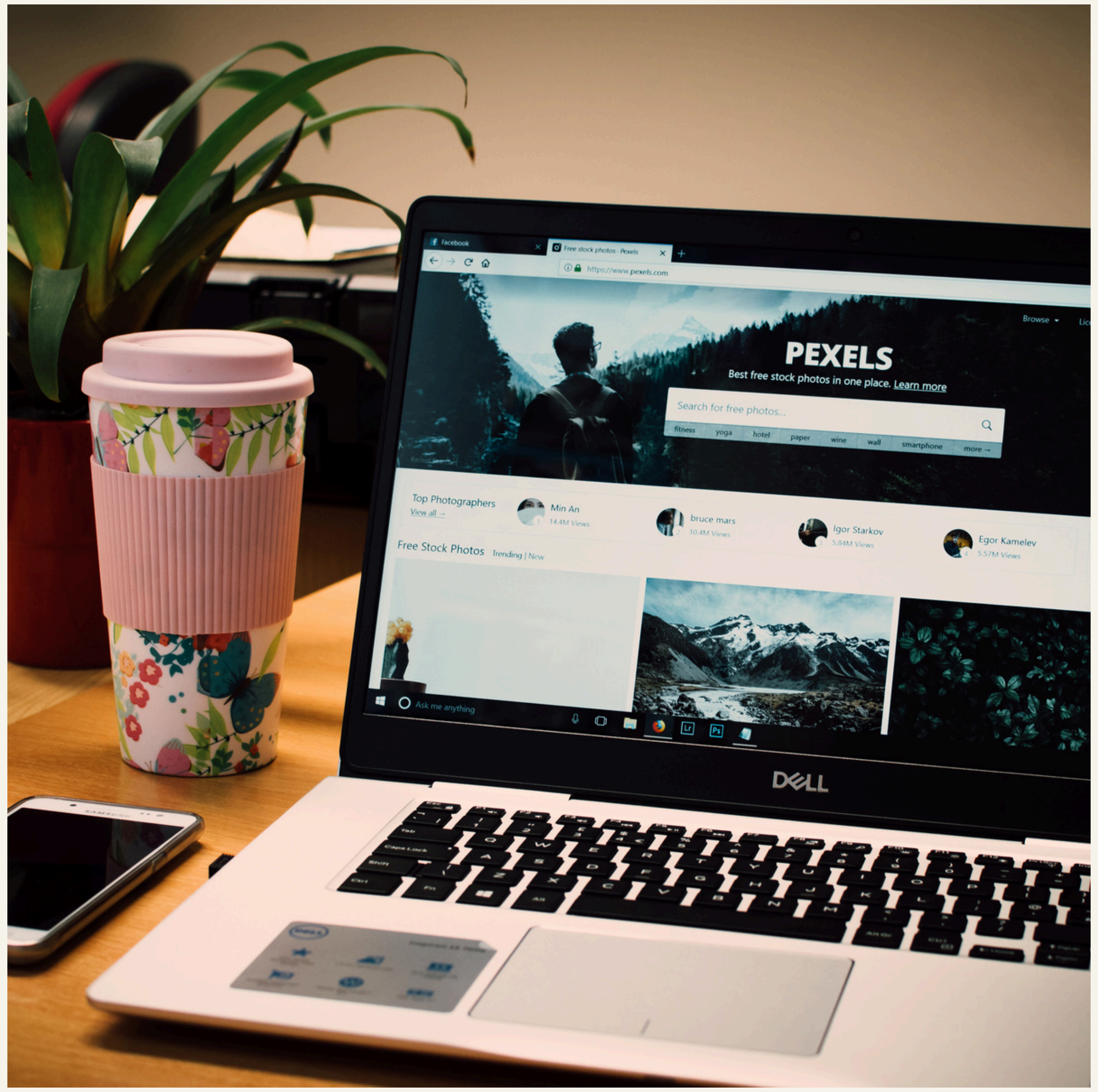
● Objective 1

To develop an accurate Named Entity Recognition (NER) system for Hindi, Tamil, Telugu, and Kokborok by utilizing advanced models like BERT, SpaCy, and SVM to address linguistic complexities.

● Objective 2

To enhance the precision and adaptability of NER in low-resource languages by combining deep learning and machine learning techniques.

HYPOTHESIS



Integrating transformer-based models like BERT with SpaCy's NLP pipeline and SVM classifiers will significantly improve the accuracy and efficiency of Named Entity Recognition in morphologically rich regional languages such as Hindi, Tamil, Telugu, and Kokborok, outperforming traditional NER methods.

METHODOLOGY

Qualitative Method

The quantitative approach involves collecting a corpus of regional newspaper articles and preprocessing the data. BERT, SpaCy, SVM, and CRF will be used for NER, with evaluation based on precision, recall, and FI-score. Cross-validation will assess model robustness. The performance of each model will be compared to determine the best approach for each language.

Quantitative Method

Linguistic analysis will address challenges in the regional languages' morphology and syntax. Error analysis will identify misclassifications and areas for improvement. Native speaker evaluations will assess entity recognition accuracy. User feedback will guide refinement, ensuring real-world applicability and handling linguistic complexities.

The methodology combines quantitative and qualitative approaches. Quantitatively, a diverse corpus will be processed, using BERT, SpaCy, SVM, and CRF for NER, with performance evaluated through precision, recall, and FI-score. Qualitatively, linguistic analysis and error analysis will address language-specific challenges. Native speaker evaluations and user feedback will ensure accuracy and real-world relevance.

IMPLEMENTATION

● Phase I

Data Collection and Preprocessing:

A diverse corpus of digital newspaper articles in Hindi, Tamil, Telugu, and Kokborok will be gathered. The data will be preprocessed by tokenization, removing stop words, and performing part-of-speech tagging to prepare it for NER tasks.

● Phase 2

Model Training:

Advanced models like BERT, SpaCy, SVM, and CRF will be trained on the preprocessed data. BERT will capture contextual embeddings, SpaCy will assist in efficient processing, and SVM and CRF will handle classification and sequence tagging.

● Phase 3

Evaluation and Fine-tuning:

The models will be evaluated using precision, recall, and F1-score. Based on the results, fine-tuning will be performed to optimize the performance of each model for the regional languages.

● Phase 4

Error Analysis and Refinement:

Errors will be analyzed through linguistic examination and feedback from native speakers. Based on this analysis, model adjustments will be made to improve entity recognition accuracy and handle language-specific complexities.

RESULT

- **Enhanced NER Accuracy:**

Advanced models like BERT, SpaCy, SVM, and CRF will improve NER performance for Hindi, Tamil, Telugu, and Kokborok, handling linguistic complexities.

- **Improved Metrics:**

Precision, recall, and F1-scores will increase, showing the effectiveness of combining deep learning and machine learning.

- **Contextual Recognition:**

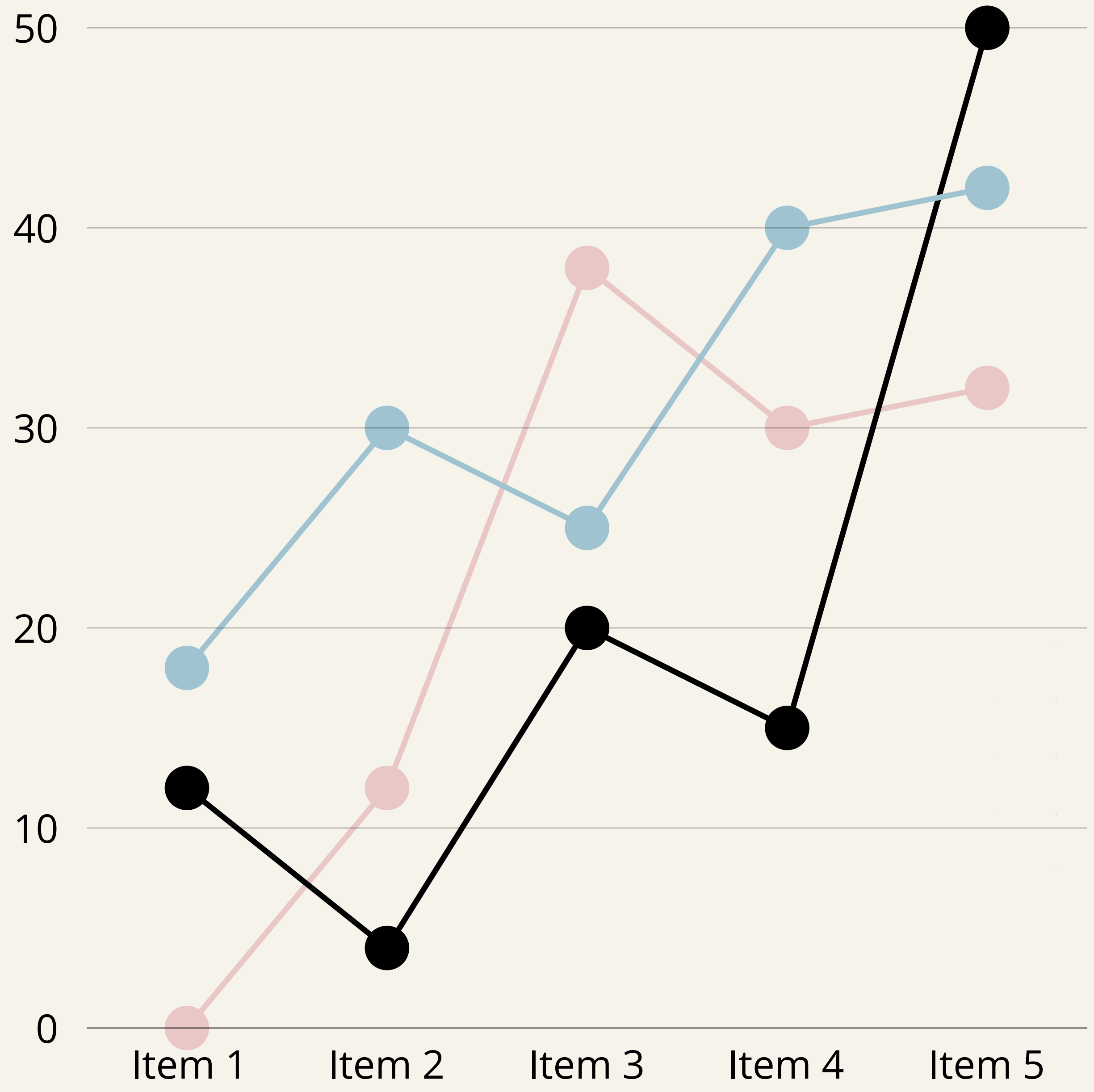
BERT will improve entity recognition in varied contexts, reducing misclassifications.

- **Optimized System:**

The system will be fine-tuned for regional language variations, benefiting machine translation and information retrieval.

- **Practical Relevance:**

Native speaker feedback will confirm the system's real-world applicability.



CONCLUSION

This research aims to enhance Named Entity Recognition (NER) for regional languages like Hindi, Tamil, Telugu, and Kokborok by leveraging advanced models such as BERT, SpaCy, SVM, and CRF. Through a combination of data preprocessing, model training, and rigorous evaluation, the study addresses the unique linguistic challenges of these languages. The expected outcome is a highly accurate, context-aware NER system that improves machine translation, summarization, and information retrieval applications. The system's adaptability to low-resource languages, validated by native speaker feedback, will make it a valuable tool for real-world use cases in multilingual environments.

RECOMMENDATION

- ## Recommendation I

  Incorporate Multilingual Pre-trained Models:
  To further improve NER performance, incorporating more multilingual pre-trained models like mBERT or XLM-R could help better capture language-specific nuances, especially for low-resource languages like Kokborok.

- ## Recommendation 2

  Expand and Annotate the Dataset:
  Expanding the dataset with more diverse and annotated examples from various domains (e.g., healthcare, politics, sports) will help enhance the model's robustness and generalization, particularly in recognizing named entities in different contexts.

Lovely Professional University | 2024

THANK YOU

Presented By :
Gowtham Indukuri[12210219]
Deepak Seshu Reddy[12219118]