WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

NER ANNOTATOR TOOL

Akshata Jadhav[1], Birajdar Dhaneshwari[2], Khairnar Saloni[3], Arati.S.Patil[4]

[1] Department of Computer Technology Sou. Venutai Chavan Polytechnic, Pune Email: akshatajadhav2209@gmail.com
[2] Department of Computer Technology Sou. Venutai Chavan Polytechnic, Pune Email:drbirajdar2006@gmail.com
[3] Department of Computer Technology Sou. Venutai Chavan Polytechnic, Pune Email:salonikhairnar2006@gmail.com
[4] Lecturer and project guide Department of Computer Technology Sou. Venutai Chavan Polytechnic, Pune Email: aratipatil708@gmail.com

ABSTRACT :

Named Entity Recognition (NER) is a key technology in natural language processing (NLP) that identifies important elements in text, like names, organizations, locations, and dates. This project focuses on improving NER by creating a simple and effective system for annotating text.

We start by looking at existing NER methods, such as Conditional Random Fields (CRFs) and newer models like BERT. This helps us understand what works well and where there are challenges, especially for less common languages and specific fields. Based on this research, we develop clear guidelines for annotators to help them accurately label different types of entities.

To make the annotation process easier, we create a semi-automated tool that uses machine learning techniques to assist human annotators. This combination improves efficiency and keeps accuracy high. A diverse group of annotators from various backgrounds ensures our system is culturally relevant. Overall, this project not only enhances NER capabilities but also offers a flexible framework that can be adapted for different languages and contexts. It contributes to advancements in automated text analysis, making it easier to extract valuable information from large amounts of text.

Keywords: Named Entity Recognition (NER) Natural Language Processing, Machine Learning , Text Analysis

1.INTRODUCTION :

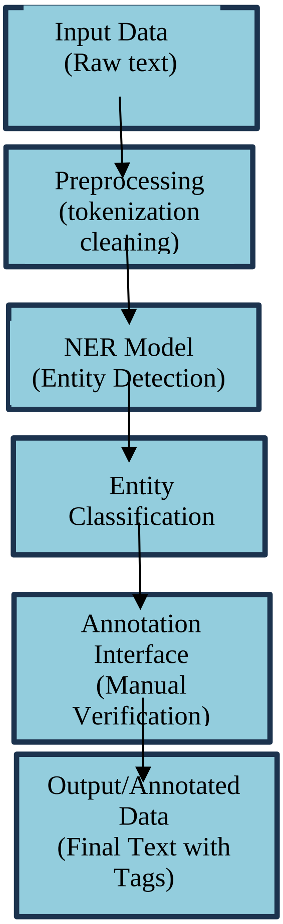
In the modern digital landscape, the explosion of unstructured text data has become a defining characteristic of the information age. From social media platforms and online news outlets to scientific journals and corporate documents, the volume of textual information generated daily is staggering. This deluge of data presents both immense opportunities and significant challenges for organizations and researchers. On one hand, it offers a wealth of insights that can drive innovation, inform decision-making, and enhance understanding across various domains. On the other hand, the sheer scale and complexity of this data make it increasingly difficult to process, analyze, and extract meaningful information manually.

Named Entity Recognition (NER), a critical component of natural language processing (NLP), has emerged as a powerful tool to address these challenges. NER focuses on identifying and classifying specific entities within text, such as names of people, organizations, locations, dates, monetary values, and more. By automating the extraction of these entities, NER enables systems to understand and structure unstructured text, paving the way for advanced applications like information retrieval, sentiment analysis, machine translation, automated summarization, and knowledge graph construction. Despite its transformative potential, the effectiveness of NER systems heavily relies on high-quality annotated datasets, which are often scarce, expensive, and time-consuming to produce.

The process of annotating text data for NER is a labor- intensive task that requires domain expertise, consistency, and significant effort. Manual annotation is prone to human error, inconsistencies, and scalability issues, especially when dealing with large datasets. Moreover, the lack of user-friendly tools for annotation further exacerbates the problem, making it difficult for researchers, data scientists, and even non-technical users to create high-quality labeled datasets. This bottleneck hinders the development and deployment of robust NER models, limiting their potential to unlock insights from unstructured text.

To address these challenges, this project focuses on the development of an advanced NER annotator tool designed to simplify and streamline the text annotation process. Our goal is to create an intuitive, user-friendly platform that empowers individuals—ranging from researchers and data scientists to domain experts and students—to annotate text data efficiently and accurately. By leveraging state-of-the-art NLP techniques and incorporating features such as pre-annotation suggestions, collaborative annotation, and customizable entity labels, our tool aims to reduce the time and effort required for manual annotation while improving the quality and consistency of labeled datasets.The proposed NER annotator is not just a tool for labeling text; it is a comprehensive solution designed to enhance the entire workflow of NER-based projects. It supports multiple file formats, integrates with existing NLP pipelines, and provides export options for seamless compatibility with machine learning frameworks. Additionally, the tool is designed to be adaptable to various domains, from healthcare and finance to social sciences and beyond, ensuring its relevance across diverse use cases.By democratizing access to high-quality text annotation, this project seeks to accelerate the development of NER models and their applications. Whether it’s extracting key information from legal documents, identifying disease-related terms in medical literature, or analyzing geopolitical events from news articles, our NER annotator aims to empower users to unlock the full potential of unstructured text data. Ultimately, this project aspires to contribute to the broader field of NLP by addressing a critical bottleneck in data preparation, enabling researchers and organizations to harness the power of NER for innovation and discovery.

2. BLOCK DIAGRAM :



3.PROPOSED METHODOLOGY :

The NER Annotator project follows a structured methodology to ensure efficiency and usability. It begins with Requirement Analysis, where user needs are identified through surveys, and existing annotation tools are studied to define essential features like pre-annotation suggestions, customizable labels, and collaboration support. In the Design & Planning phase, the system architecture is developed, UI wireframes are created, and the technology stack (Python for backend, React for frontend, and SpaCy/BERT for NER) is chosen. The Data Collection & Preprocessing stage involves gathering, cleaning, and tokenizing text data from various sources to prepare it for annotation and model training.

The Annotation Interface Development focuses on building an interactive UI using React/Angular, featuring entity labeling, real-time collaboration, and pre-annotation suggestions. In the NER Model Integration phase, a pre-trained model like SpaCy or BERT is implemented to assist with entity recognition and allow fine-tuning with user-annotated data. To support teamwork, the Collaboration & User Management module enables multi-user roles, real-time updates, and role-based access control. Data is securely managed in the Storage & Database Setup phase using PostgreSQL or MongoDB, with version control and encryption to ensure security.

For usability, the Output & Export functionality allows users to download annotated data in Excel format, making it easily accessible for further processing. Additionally, reports and analytics are generated to track annotation progress. The Testing & Validation phase ensures reliability through unit and integration testing, followed by user validation and feedback implementation. Finally, in the Deployment & User Training stage, the tool is deployed on a cloud platform or as a standalone application, with comprehensive documentation provided to assist users.

4. IMPLEMENTATION :

4.1 Flask Backend:

- Create API routes to handle form submissions (/submit-form), process JSON data (/process_data), and serve the UI (index.html). Use Pandas to structure annotations.

4.2 Frontend UI:

- Build an interactive annotation tool using HTML, CSS, and JavaScript. Allow users to select and label entities.

4.3 Data Handling:

- -Process JSON input, extract entities using pd.json_normalize(), and save annotations as Excel (.xlsx) files.

4.4 Annotation Features:

- Enable highlighting, editing, and pre-annotation with ML models.

5. APPLICATION :

A Named Entity Recognition (NER) annotator has several key applications. First, it can be used in data extraction, where it automates the process of identifying important details from large texts. Second, it aids in sentiment analysis, helping to assess public opinion about brands, products, or individuals. Third, it can enhance content recommendation systems by identifying key topics and entities in texts. Fourth, NER is useful in the healthcare industry, where it helps identify medical terms and conditions from clinical data. Fifth, it supports social media monitoring, tracking mentions of specific brands or events to inform marketing strategies. Finally, NER plays a crucial role in training machine learning models by providing labeled data that improves model accuracy.

6. SCOPE FOR FUTURE WORK :

The NER annotator tool can be further enhanced by incorporating multilingual support, improving accuracy through advanced machine learning models like BERT, and developing a user-friendly interface for non-technical users. Future work could also focus on enabling real-time annotation, allowing custom entity definitions, and integrating cloud-based collaboration features. Expanding file format compatibility (e.g., PDF, Word, Excel), creating a mobile app, and adding error detection mechanisms would increase its versatility. Additionally, automating dataset generation, optimizing performance for large datasets, and adapting the tool for educational purposes could broaden its applicability. Incorporating a feedback system and robust security features like encryption and user authentication would further refine the tool, making it more reliable and accessible for diverse use cases. These improvements aim to make the NER annotator tool more efficient, scalable, and user-centric.

7. CONCLUSION :

The NER Annotator Tool project demonstrates the practical application of Named Entity Recognition (NER) in extracting and categorizing key information from text. Developed as part of the final diploma project in computer technology, this tool highlights the importance of NER in areas like data analysis, language processing, and automation. Through this project, students gained hands-on experience in programming, machine learning, and user interface design. While the current version achieves basic functionality, there is significant scope for future improvements, such as multilingual support, better accuracy, and real-time annotation. This project not only enhances technical skills but also provides a foundation for building more advanced tools in the future. Overall, it serves as a valuable learning experience and a stepping stone for further innovation in the field of natural language processing.

8. ACKNOWLEDGMENT

We would like to express our sincere gratitude to all those who supported us in completing the NER Annotator Tool project. First and foremost, we thank our project guide and faculty members for their valuable guidance, encouragement, and feedback throughout the project. We are also thankful to our institution for providing the necessary resources and infrastructure to work on this project. Special thanks to our classmates and friends for their suggestions and moral support during challenging times. Lastly, we extend our gratitude to our families for their constant motivation and belief in us. This project has been a great learning experience, and we are thankful to everyone who contributed to its successful completion**.**

9. REFERENCES :

1. D. Nadeau, S. Sekine,” Named Entity Recognition: A Survey”, 2007.
2. Explosion AI (Matthew Honnibal, Ines Montani), “Prodigy: A New Annotation Tool for Efficient Machine Learning”, 2017.
3. Explosion AI (Matthew Honnibal, Ines Montani), “A Survey of Named Entity Recognition and Classification”, 2003.
4. Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, Jiawei Han, “Deep Learning for Named Entity Recognition: A Survey”, 2019.
5. SpaCy: Industrial-Strength Natural Language Processing in Python, Matthew Honnibal, Ines Montani, 2017.
6. Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, Jiawei Han “A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models”, 2020.