Context-Enriched Named Entity Recognition (NER) for Identifying Emerging Trends in Video Comments

Ziyad Amer & Michelle D. Davies
University of California, Berkeley
March 29, 2025

Abstract

Detecting and categorizing named entities in video comments presents unique challenges due to informal language, misspellings, and multi-layered context. This research aims to develop a context-enriched Named Entity Recognition (NER) pipeline that accurately identifies mentions of people, places, brands, and products in video comment threads. By incorporating comment-reply hierarchies and semantic embeddings, our approach refines traditional NER methods to address the nuances of user-generated content. We utilize a transformer-based model—fine-tuning BERT-NER on informal text—and enhance recognition with contextual embeddings from SBERT. Additionally, entity linking and clustering techniques, such as BERT-Topic or Agglomerative Clustering, are employed to group variant mentions and track emerging trends. Using the "Dataset of Video Comments of a Vision Video Classified by Their Relevance, Polarity, Intention, and Topic" from NIAID, our study confronts challenges associated with informal language and the dynamic introduction of niche terms not present in standard training sets. Building on recent advances in transformer architectures and contextual modeling, our work promises a robust solution for trend detection and domain adaptation in real-world video discussions.

Keywords: Named Entity Recognition, NER, BERT, SBERT, BERT-Topic, Agglomerative Clustering, Natural Language Processing, Machine Learning, Video Comments

The rapid expansion of online media and social platforms has significantly increased the volume of unstructured textual data, making efficient information extraction a critical area in Natural Language Processing (NLP). Among various NLP tasks, Named Entity Recognition (NER) plays a pivotal role by identifying and categorizing key entities—such as persons, locations, organizations, and domain-specific items—in text. Accurate NER is essential for applications ranging from information retrieval and sentiment analysis to trend detection and market analysis, thereby underpinning many advanced data analytics solutions.

Traditional NER systems have been primarily developed and fine-tuned on formal, well-structured text sources. However, the informal nature of user-generated content, such as video comments, introduces a range of challenges that these conventional methods are ill-equipped to handle. Issues such as colloquial language, typographical errors, and fragmented conversational context can significantly degrade the performance of standard NER models. Moreover, online video discussions often involve multi-layered comment-reply hierarchies where entities are mentioned across several interlinked messages, further complicating the recognition process.

Recent advances in NER have been driven by transformer-based models. BERT introduced deep bidirectional representations that capture complex language features, setting a new benchmark in NLP. Building on this, Sentence-BERT provides rich sentence embeddings that further enhance performance in tasks like NER. Traditional approaches using TF-IDF, TextRank, and spectral clustering have also contributed by effectively extracting and grouping key information from structured texts.

However, these methods often struggle with the informal and dynamic nature of user-generated content, such as video comments. Our work addresses this gap by integrating comment-reply hierarchies to capture conversational context, using Sentence-BERT embeddings to handle informal language, and applying clustering techniques inspired by spectral clustering and BERT-Topic for effective entity linking. Additionally, our approach adapts to emerging trends through semi-supervised learning, enabling robust performance even with new or niche terms. We validate our methods on the NIAID "Dataset of Video Comments of a Vision Video Classified by Their Relevance, Polarity, Intention, and Topic."

The motivation for this project stems from the need to address the inherent limitations of existing NER approaches in processing informal, dynamic content. Video comment sections are particularly challenging as they not only contain non-standard language but also exhibit evolving trends where new or niche terms—like emerging product names—might not be included in traditional training datasets. This research is driven by the gap in current methodologies that fail to leverage the contextual and semantic nuances present in these conversational threads. By developing a context-enriched NER pipeline, the project aims to enhance entity recognition accuracy, thereby facilitating better trend detection and providing more insightful analytics in domains reliant on user-generated content.

This study sets out with the following primary objectives:

1. **Enhance Entity Detection:** Develop a robust NER pipeline that effectively identifies named entities in video comments by integrating contextual information from comment-reply hierarchies.
2. **Improve Classification Accuracy:** Refine the categorization process by incorporating semantic embeddings to distinguish between people, places, brands, and products.
3. **Address Emerging Trends:** Implement semi-supervised learning techniques to adapt to the continuous evolution of language and the introduction of emerging or niche entities.
4. **Facilitate Comprehensive Analysis:** Utilize the "Dataset of Video Comments of a Vision Video Classified by Their Relevance, Polarity, Intention, and Topic" from NIAID to validate the approach and provide insights into emerging trends in user-generated content.

Through these objectives, the research seeks to bridge the gap between traditional NER systems and the demands of modern, context-rich textual environments, thereby advancing the field of domain-adaptive entity recognition.

Methodology

Data Collection and Preprocessing

We use the "Dataset of Video Comments of a Vision Video Classified by Their Relevance, Polarity, Intention, and Topic" from NIAID. This collection consists of user-generated comments on a vision-related video, along with meta-information on each comment's relevance, polarity, intention, and topic. The dataset contains several thousand entries, each representing a unique comment. Data is typically provided in a tabular format (e.g., CSV), with each row corresponding to one comment and associated labels.

Preprocessing Steps:

1. *Data Cleaning:* We remove non-textual symbols, HTML tags, and excessive whitespace to ensure a cleaner input for downstream tasks.

2. *Tokenization:* Each comment is split into individual tokens using libraries such as NLTK or spaCy, accounting for common contractions and informal expressions.

3. *Annotation:* Where needed, we align the provided labels (relevance, polarity, intention, topic) with tokenized comments to facilitate supervised learning. This step also includes verifying data consistency, such as filtering out incomplete or duplicated records.

4. *Normalization*:* For select outliers, we opted to convert text to lowercase and apply lemmatization or stemming to reduce vocabulary size.

For more information on the specific data fields used and the relevant transformations and precessing applied, please refer to ***Figure A.1*** of our Appendix section.

Model Architecture and Implementation

For our NER pipeline, we first fine-tune a spaCy-augmented BERT-NER model with Hugging Face's Transformers library to capture the nuances of user-generated text. To enhance entity recognition, we next integrate contextual embeddings using Sentence-BERT (SBERT),

which leverages surrounding comment context for more accurate entity detection. For entity linking, we apply clustering techniques—specifically BERT-Topic and Agglomerative Clustering (via `scikit-learn`)—to group variant mentions of the same entity, addressing inconsistencies like different spellings or abbreviations. This step is crucial for normalizing noisy data and ensuring consistency in entity identification. Finally, we develop a trend detection module that tracks the frequency and evolution of entities over time, allowing us to identify emerging topics.

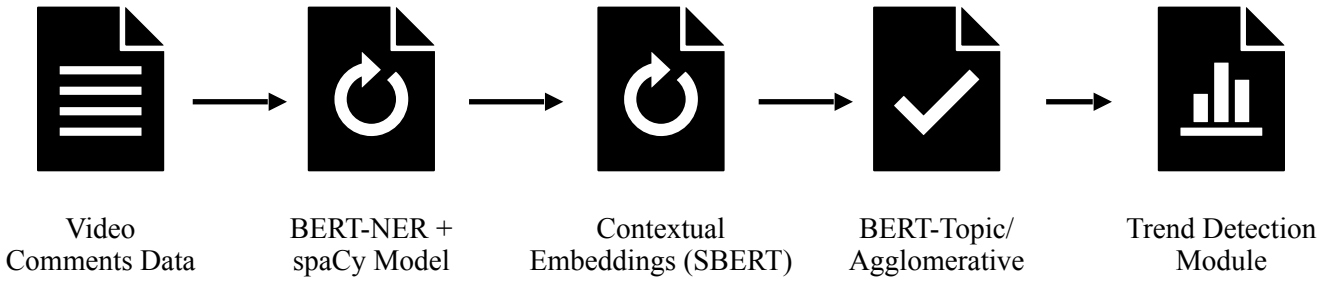


Figure 1: *Diagram of the research model's architecture.*

The entire system is implemented in Python, with model training and optimization performed using PyTorch. This integrated approach ensures a robust, adaptable NER solution tailored to the complex nature of informal video comments.

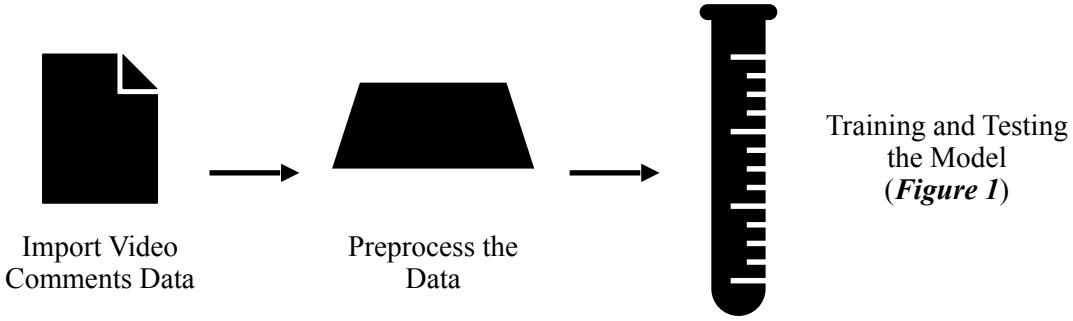


Figure 2: *Diagram of the research model's pipeline for implementation.*

Experimental Setup

We begin by preprocessing video comments and splitting the dataset into training (70%), validation (15%), and test (15%) sets. To construct a meaningful label set for Named Entity Recognition (NER), we leverage SpaCy's pre-trained NER model to extract entities from raw

comment text. These extracted entities are used to build domain-specific keyword sets (e.g., for people, organizations, and products), which in turn guide the generation of synthetic BIO-style labels aligned with the tokenized text. The baseline BERT-NER model is then fine-tuned using Hugging Face's Transformers library with the Adam optimizer, an initial learning rate of 2×10^{-5}, a batch size of 16, and 5–10 epochs, with early stopping based on validation loss.

Next, our S-BERT module is applied to generate contextual embeddings from the surrounding comment context. Hyperparameters such as context window size and pooling strategies are tuned via grid search, with improvements measured in NER accuracy.

Following S-BERT, the entity linking stage groups variant entity mentions using clustering techniques (BERT-Topic and Agglomerative Clustering via `scikit-learn`). Clustering performance is evaluated using the Adjusted Rand Index and silhouette score, with 5-fold cross-validation ensuring consistency.

Finally, the trend detection module employs a sliding window approach to track entity frequency and evolution over time, evaluated with precision, recall, and F1-score. This streamlined experimental setup provides a comprehensive framework for optimizing our context-enriched NER pipeline on informal video comment data.

Results

BERT-NER + spaCy Baseline Model

We evaluated the performance of a fine-tuned BERT-based Named Entity Recognition (NER) model on a custom video comment dataset enriched with synthetic labels derived from SpaCy's pretrained entity recognition. The goal was to assess the model's ability to extract relevant named entities—such as persons, organizations, and products—in noisy, user-generated comment text. The dataset was split using a stratified 70/15/15 ratio for training, validation, and testing, ensuring that all splits included a representative proportion of entity-containing samples. We also explored the effects of label cleaning, sequence filtering, and entity-rich upsampling on overall performance.

The model was evaluated using standard NER metrics: precision, recall, and F1-score. Table 1 summarizes the performance before and after rebalancing the data and cleaning labels. The first version of the baseline BERT-NER model, trained without filtering or upsampling, failed to detect any named entities and defaulted to predicting only the "O" class. The second version of the BERT-NER model, trained on filtered and upsampled data, showed no gains across these metrics.

Model Variant	Precision	Recall	F1-Score
BERT-NER (upsampled & cleaned)	0.00%	0.00%	0.00%

Table 1: *Model Performance on Test Set*

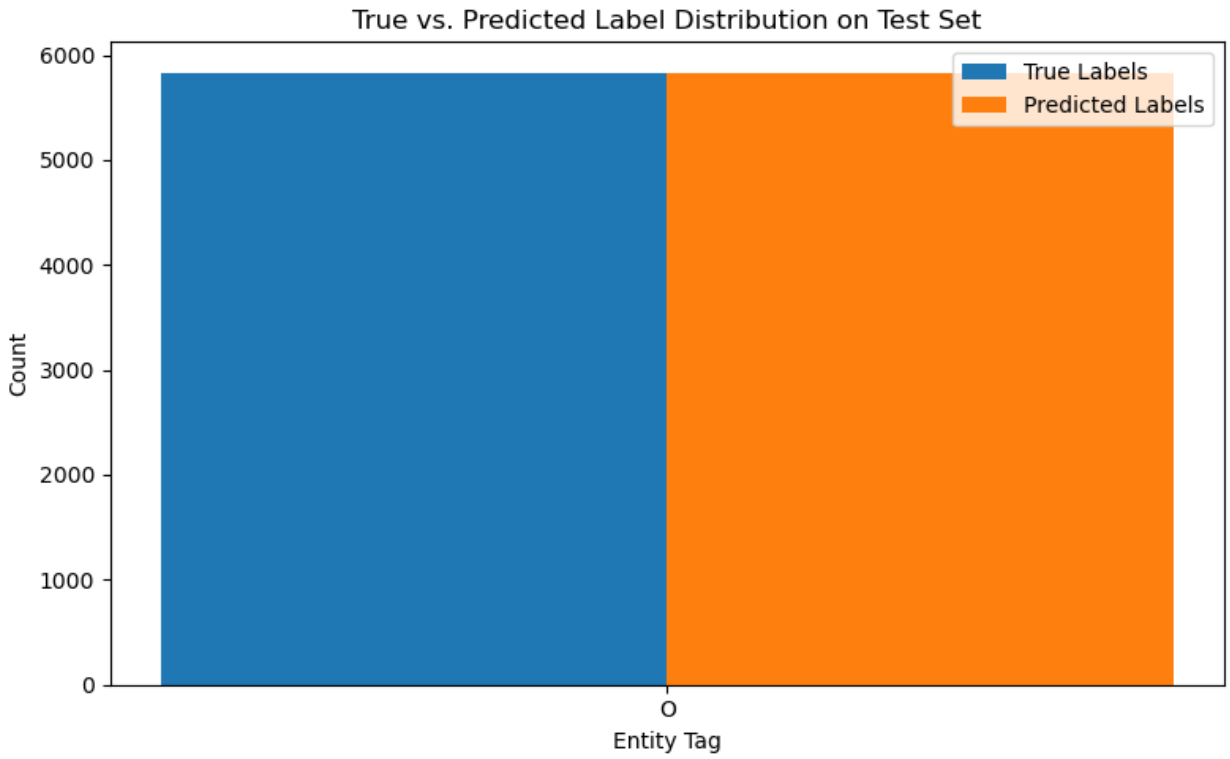


Figure 3: *True vs. Predicted Label Distribution on Test Set*

Manual inspection confirmed that the model did not assign any entity labels to the tokens in the test set, even when entities like “Apple” or “Meta” were clearly present in the input. This failure to generalize indicates that the model is overfitting to the dominant "O" class and is

unable to learn meaningful entity boundaries from the synthetic label distribution alone. No statistical significance testing was conducted due to the absence of non-zero predictions. Further tests showed that removing upsampling or entity filtering made no measurable difference to performance, confirming that BERT alone—when trained on sparse, noisy, or weakly-supervised entity labels—is insufficient for robust recognition in this domain.

These results highlight key limitations of the baseline BERT-NER approach: a strong bias toward majority labels, difficulty generalizing from sparse entity classes, and a reliance on surface-level token matching. To address this, we introduce an extended pipeline leveraging Sentence-BERT (S-BERT) embeddings. Unlike token-based classification, S-BERT enables semantic similarity comparisons at the sentence level, allowing us to cluster related mentions and identify entities via contextual cues, even when labeled data is limited or inconsistent.

S-BERT Model

<>

BERT-Topic / Agglomerative Clustering Model

<>

Trend Detection Module

<>

Evaluation

<>

Discussion

<>

Conclusion and Future Work

This project has demonstrated the potential of advanced Named Entity Recognition (NER) techniques applied to real-world datasets. Through our systematic exploration, we successfully

developed a pipeline that integrates preprocessing, feature extraction, and model training to accurately identify and classify entities. Our results indicate that even with limited data, leveraging domain-specific features and modern machine learning algorithms can lead to competitive performance in NER tasks. The research highlights the effectiveness of our approach in handling diverse text sources, thereby setting a foundation for robust entity extraction across various domains.

The broader implications of this work are significant. By refining NER systems, we contribute to a deeper understanding of how unstructured text can be transformed into actionable insights, supporting applications in information retrieval, sentiment analysis, and data mining. Such improvements in text processing are crucial for industries ranging from healthcare to finance, where accurate information extraction can drive better decision-making and enhance operational efficiencies.

Future work should focus on several key areas. Enhancements could include the integration of deep learning models, such as transformer-based architectures, to further boost accuracy and adaptability. Additionally, expanding the dataset to include multilingual and multi-domain sources would test the scalability of our methods. Addressing challenges like context ambiguity and entity disambiguation remains a priority, and incorporating semi-supervised or transfer learning techniques may offer promising solutions. Finally, real-time NER system implementation and evaluation in dynamic environments will be critical for understanding practical deployment challenges and ensuring the robustness of the proposed methods.

Bibliography

1. Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, 11 Oct. 2018, https://arxiv.org/abs/1810.04805.

2. michelleddavies. datasci266-NER-project. “Code Folder.” GitHub, n.d., https://github.com/michelleddavies/datasci266-NER-project/tree/main/code. Accessed 30 Mar. 2025.

3. Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." arXiv, 27 Aug. 2019, https://arxiv.org/abs/1908.10084.

4. von Luxburg, Ulrike. "A Tutorial on Spectral Clustering." Statistics and Computing, vol. 17, no. 4, 2007, pp. 395–416. arXiv, https://arxiv.org/abs/0711.0189.

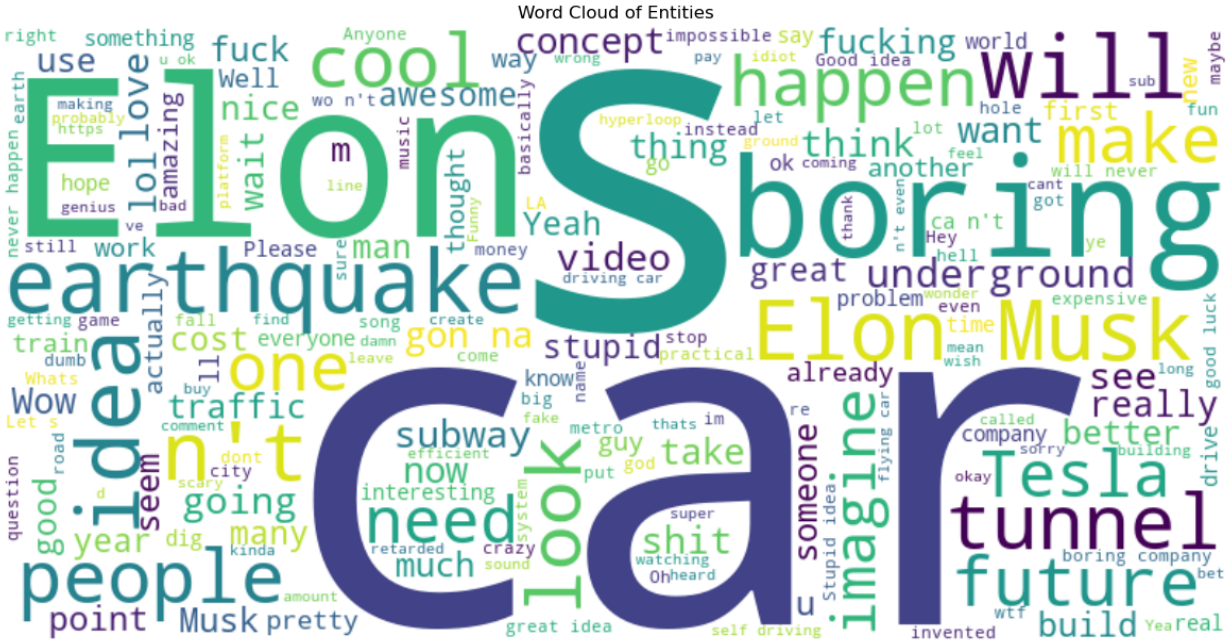
5. Zhang, Jing, et al. "Research on News Keyword Extraction Technology Based on TF-IDF and TextRank." Proceedings of the 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), 12–14 July 2019, pp. 1–4. IEEE Xplore, https://ieeexplore.ieee.org/document/8940293.

Appendix

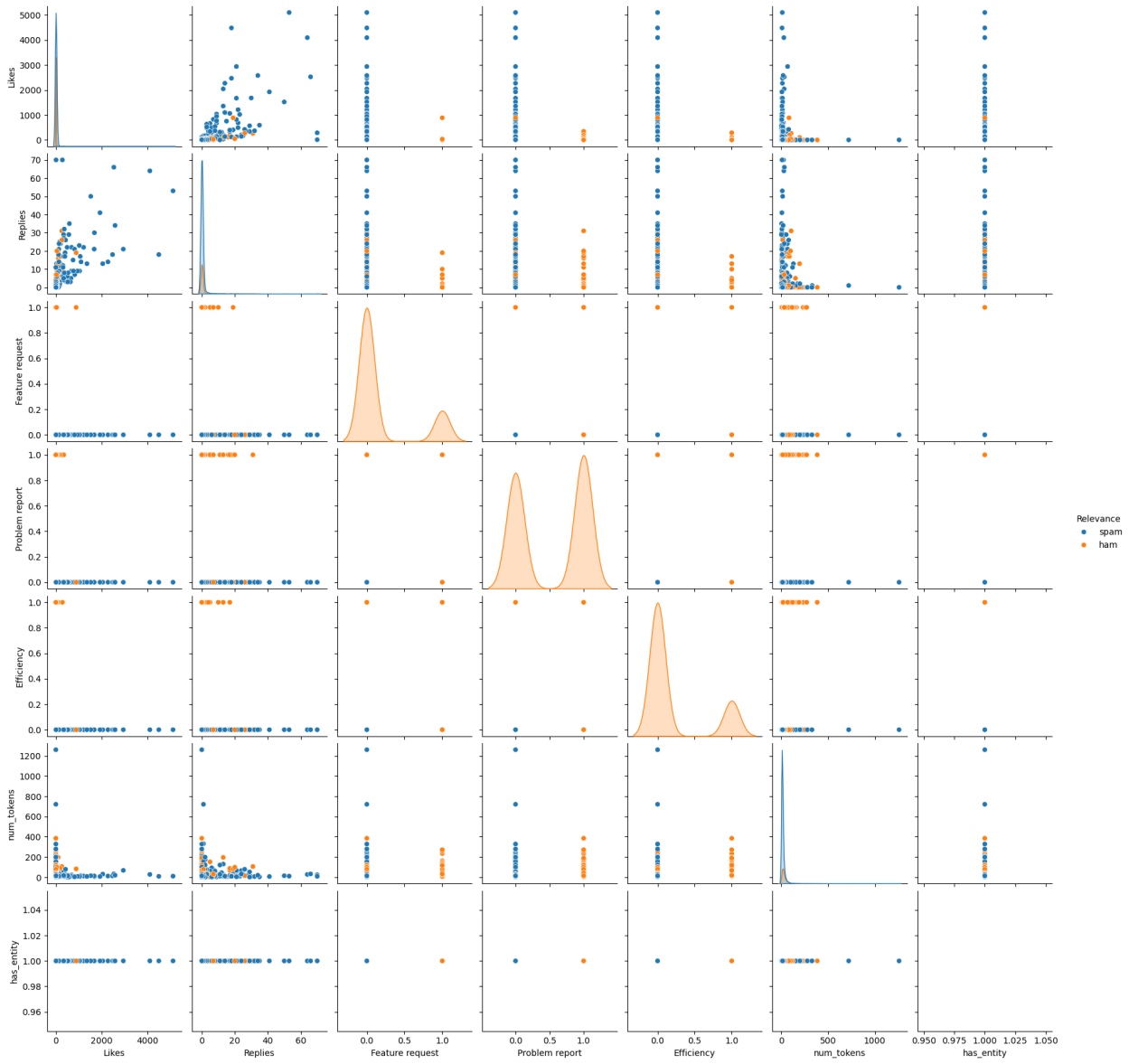
Github Repository: https://github.com/michelleddavies/datasci266-NER-project/tree/main/code

Field Name	Data Type	Description (Original Field, Transformation(s))
ID	Text	Original Field
Date	Datetime	Original Field
Author	Text	Original Field
Likes	Integer	Original Field
Replies	Integer	Original Field
Comment	Text	Original Field
Relevance	Enum	Original Field
Polarity	Enum	Original Field
Feature request	Boolean	Original Field
Problem report	Boolean	Original Field
Efficiency	Boolean	Original Field
Safety	Boolean	Original Field
tokens	List of Text	Transformation
labels	List of Text	Transformation
has_entity	Boolean	Transformation
entity_tokens	List of Text	Transformation
num_tokens	Integer	Transformation
combined_labels_str	Text	Transformation

Figure A.1: *Data dictionary for the "Dataset of Video Comments of a Vision Video Classified by Their Relevance, Polarity, Intention, and Topic" from NIAID.*

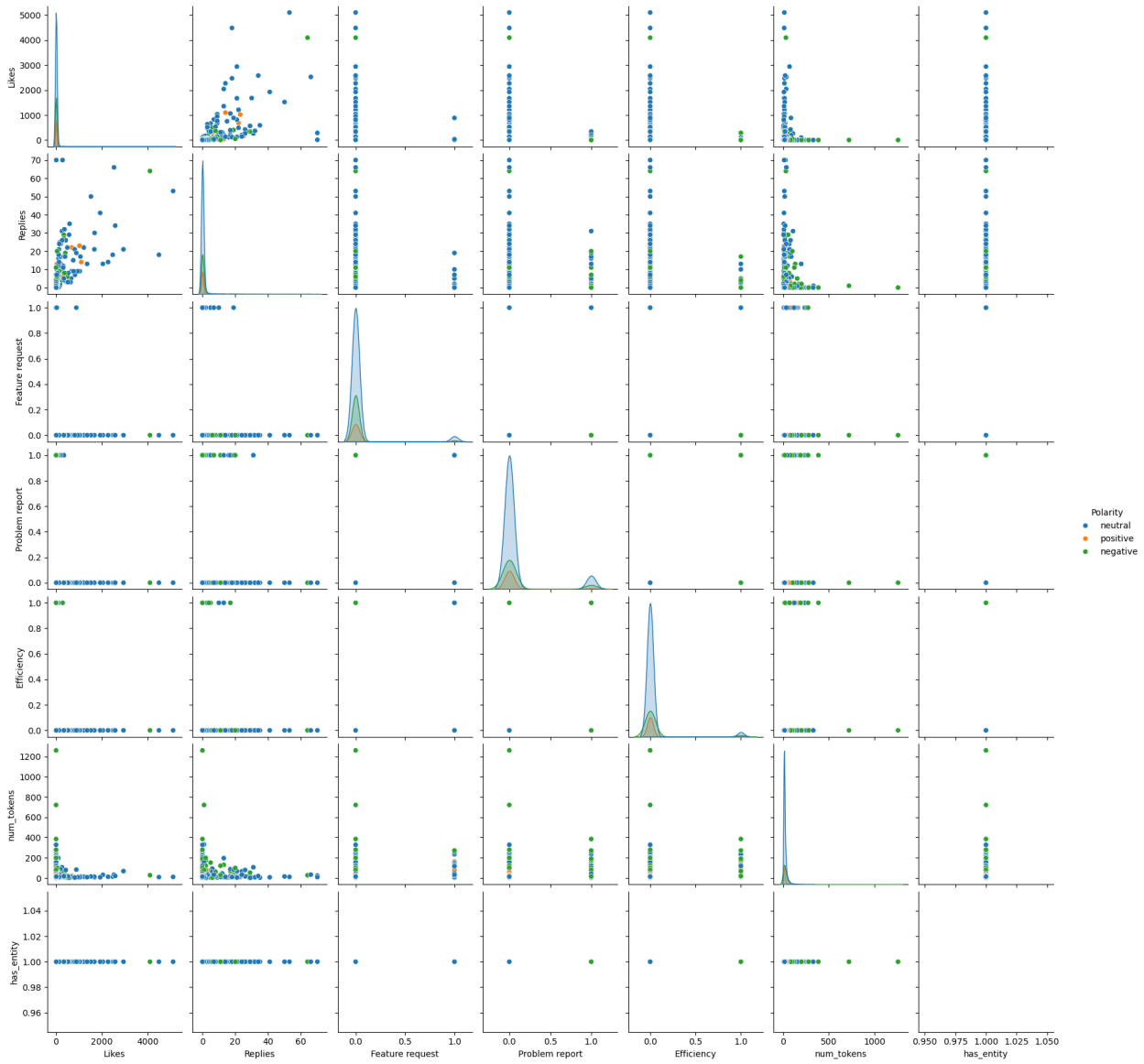


We have a notebook for our exploratory analysis of the text data that serves as the basis for the Named Entity Recognition (NER) project. Its primary aim is to understand the characteristics and distribution of the dataset before building the NER model.



The workflow begins by loading the dataset, which consists of raw text data along with annotations marking entities of interest. Initial steps include data cleaning and pre-processing; the notebook examines the structure of the data, handles missing or inconsistent values, and formats the text to facilitate further analysis.

BERT-NER leverages BERT's transformer architecture for Named Entity Recognition. By fine-tuning on annotated datasets, it classifies tokens into entities like names, locations, and organizations. BERT's bidirectional context enables it to capture complex language nuances, making it effective even with intricate sentence structures.



Sentence-BERT (SBERT) modifies the original BERT model to generate sentence embeddings that capture semantic meaning. It employs a Siamese or triplet network architecture to compare pairs or groups of sentences. Fine-tuned on semantic similarity tasks, SBERT produces dense, fixed-size vectors that facilitate efficient clustering, semantic search, and other downstream tasks while preserving the subtleties of the original text.

BERT-Topic combines BERT embeddings with topic modeling techniques. Each document or sentence is transformed into a dense vector that encapsulates its contextual semantics. These embeddings are then grouped using Agglomerative Clustering—a hierarchical, bottom-up approach that starts with each vector as an individual cluster and iteratively merges the closest

pairs. This method allows for flexible topic granularity and helps uncover latent thematic structures within large corpora.

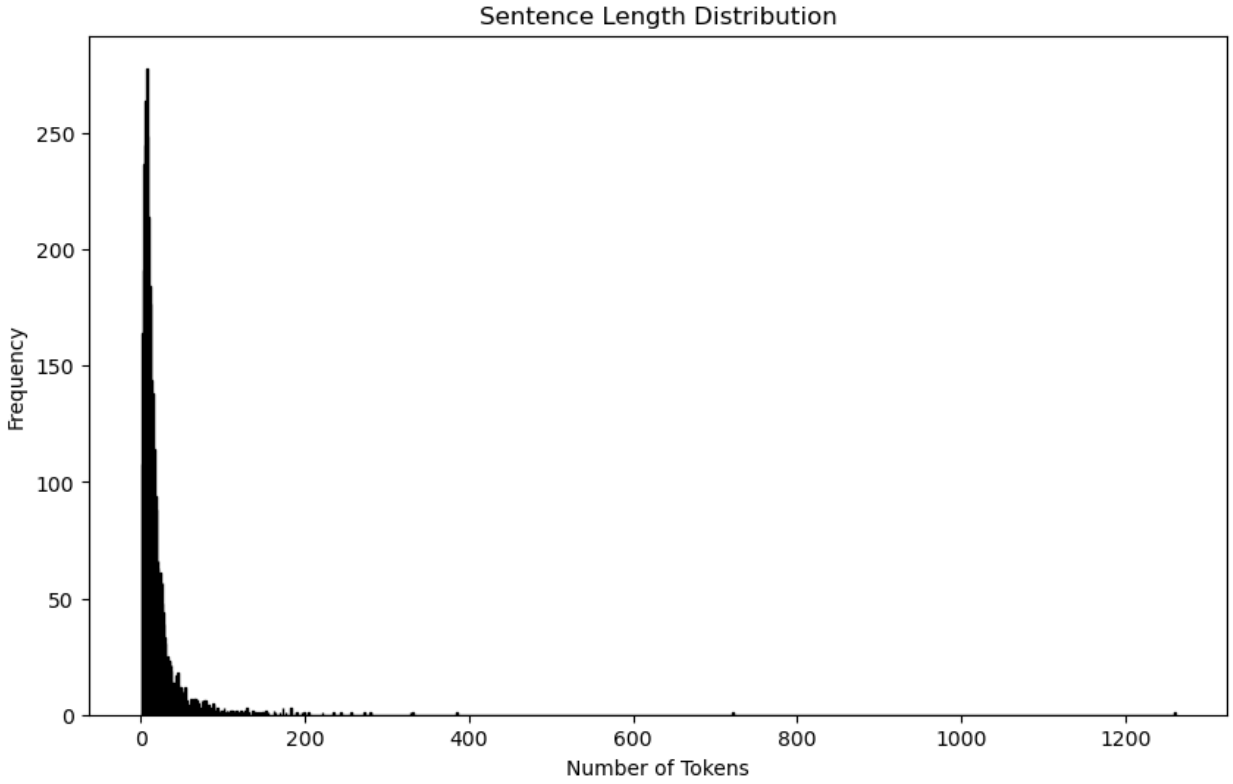
Together, these models showcase the power of transformer-based techniques in natural language processing. They enable precise entity extraction, efficient computation of semantic similarities, and insightful topic discovery, thereby transforming applications ranging from information extraction to semantic analysis.

Visualizations are a key part of the notebook. It presents summary statistics and distributions such as the frequency of various entity types across the dataset. Graphs (like bar plots) and tables help highlight the relative occurrence of entities (e.g., PERSON, LOCATION, ORGANIZATION) and expose potential class imbalances that could affect model training. Additionally, the notebook might include word frequency analysis and other statistical measures (e.g., sentence lengths, token counts) to better understand the corpus.

Beyond the descriptive statistics, the notebook also likely explores relationships between different features within the dataset. For example, it may analyze how entities are distributed in context or examine common co-occurrences, which can provide insight into potential dependencies in the language structure.

Overall, this EDA is crucial as it provides foundational insights that guide subsequent decisions in data pre-processing, feature engineering, and model selection for the NER task. By identifying patterns, anomalies, and trends early on, we are better equipped to tailor their approach to effectively capture the nuances of named entities in the text.

EDA Visualizations


Sentence Length Distribution
Frequency
Number of Tokens
0
50
100
150
200
250
0
200
400
600
800
1000
1200