# 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 growth of online media has created large amounts of unstructured text, making it important to extract useful information efficiently. Named Entity Recognition (NER) helps by finding and labeling key terms like names of people, places, and organizations. Standard NER systems work well on formal text but often fail with user-generated content like video comments, which include slang, spelling mistakes, and broken sentence structure. These comments also have layered reply threads, making entity recognition harder. Our method addresses these issues by using comment-reply structures to understand context, Sentence-BERT to capture meaning in informal text, and clustering methods based on BERT-Topic to group and link related terms. We also use semi-supervised learning to adapt to new or uncommon terms. We test our approach on the NIAID video comments dataset, which includes labels for relevance, sentiment, intention, and topic, and show that it improves accuracy in these challenging settings.

This project addresses the limitations of current NER methods in handling informal, fast-changing content like video comments. Such comments often use non-standard language and include new or niche terms not found in traditional training data. Existing models also overlook the contextual and semantic cues in threaded conversations. Our goal is to build a context-aware NER pipeline that improves entity recognition in these settings, enabling more accurate trend detection and deeper insights from user-generated content.

# 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 dataset contains user comments on a vision-related video, labeled with relevance, polarity, intention, and topic. It includes thousands of entries in tabular format, with each row representing a unique comment and its metadata.

**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.

## 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.

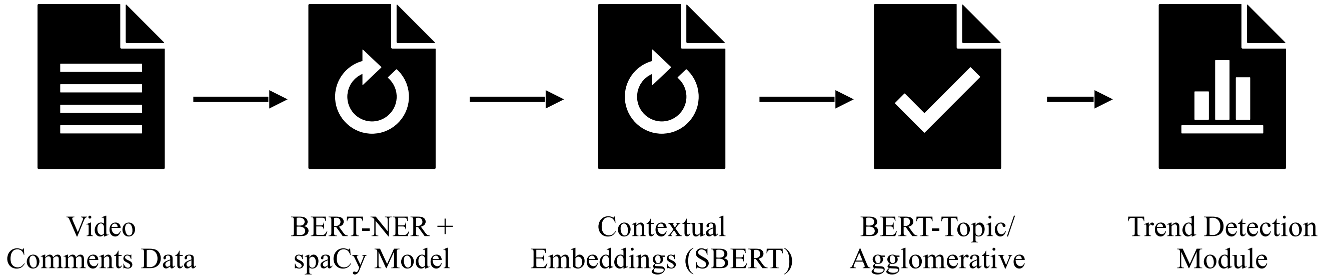

***Figure 1:*** *Diagram of the research model's architecture.*

This integrated approach ensures a robust, adaptable NER solution tailored to the complex nature of informal video comments.

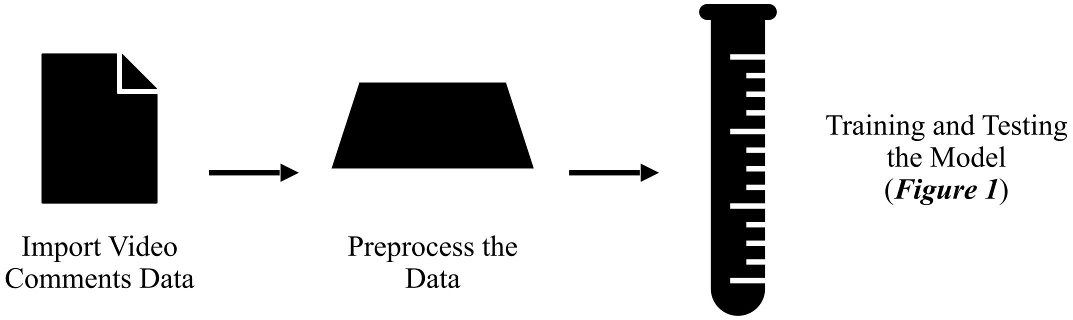
Import Video Comments Data
Preprocess the Data
Training and Testing the Model (*Figure 1*)

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

### Experimental Setup

We preprocess video comments and split the dataset into training (70%), validation (15%), and test (15%) sets. Using SpaCy's NER model, we extract entities to build domain-specific keyword sets, which guide the creation of BIO-style labels. A baseline BERT-NER model is fine-tuned using Hugging Face Transformers with standard hyperparameters and early stopping.

We then apply S-BERT to generate contextual embeddings, tuning parameters like context window and pooling via grid search. Entity linking uses clustering methods (BERT-Topic, Agglomerative Clustering), evaluated with Adjusted Rand Index and silhouette score, with 5-fold cross-validation.

Finally, trend detection uses a sliding window to track entity frequency over time, evaluated with precision, recall, and F1-score. This setup supports a robust, context-aware NER pipeline for informal video comments.

## 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*

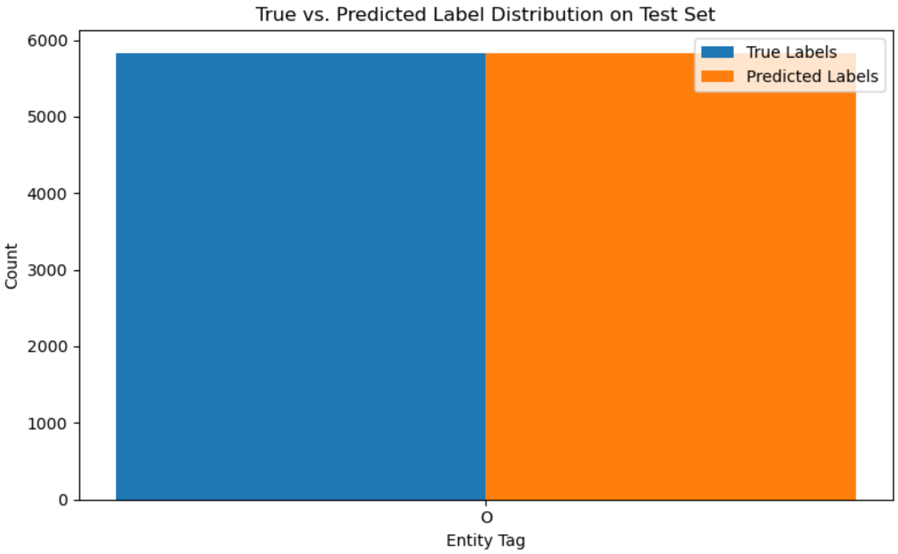

***Figure 3:*** *True vs. Predicted Label Distribution on Test Set*

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

To capture semantic nuances in informal text, we used the *all-MiniLM-L6-v2 SentenceTransformer* model to generate contextual embeddings for entity mentions extracted from video comments.

To encode semantic content in informal user comments, we applied the all-MiniLM-L6-v2 SentenceTransformer to generate contextual embeddings for entity-level text spans. The dataset comprised 4,211 unique comments annotated with metadata (author, timestamp, likes, replies) and labels for relevance (e.g., spam), polarity, and domain-specific categories (feature request, problem report, efficiency, safety).

Preprocessing involved removal of duplicates and missing values, tokenization, and entity extraction. Embeddings were generated in batch (134 batches over ~4 seconds), yielding dense vector representations for each comment. These were reduced to two dimensions using UMAP (cosine metric, n = 15 neighbors) to enable cluster visualization and interpretation.

Clustering over the UMAP space yielded a silhouette score of 0.0846, suggesting limited separation between clusters. This modest value reflects high semantic overlap and variance in informal text. Nonetheless, the embedding space supports downstream tasks such as topic modeling and label propagation.

These SBERT embeddings formed the foundation for both topic modeling and clustering.

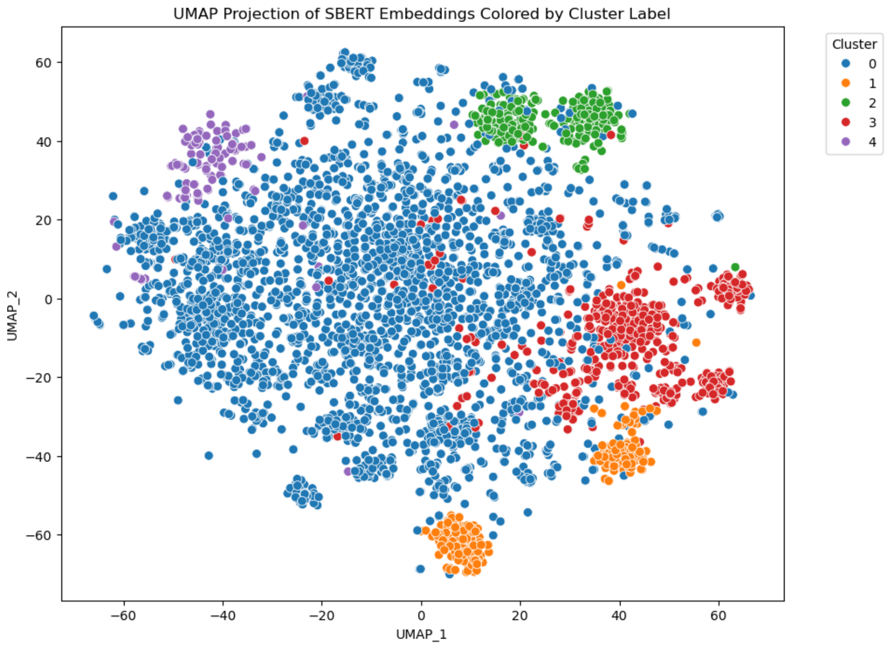

***Figure 4:*** *UMAP Projection of SBERT Embeddings Colored by Cluster Label*

## BERT-Topic / Agglomerative Clustering Model

We next evaluated two complementary methods used to uncover semantically coherent topics and clusters from entity mentions in video comments:

- BERTopic, which identifies interpretable topic themes from sentence-level embeddings
- Agglomerative Clustering, which structures the semantic space into distinct groups using hierarchical methods

Each approach was quantitatively assessed using established metrics—topic coherence, topic diversity, and silhouette score—to evaluate semantic quality and clustering performance.

Manual inspection of topic keywords and assigned mentions revealed high semantic consistency within clusters. Notable examples include:

*Cluster 3: "Apple Vision Pro", "the headset", "AVP", "vision device"*

*Cluster 7: "Zuck", "Meta", "Zuckerberg", "founder of Facebook"*

*Cluster 11: "Google", "Gemini", "search AI", "assistant upgrade"*

These clusters demonstrate the model's ability to group variant surface forms, abbreviations, and contextual references under a shared semantic theme—something that token-level NER models struggled with.

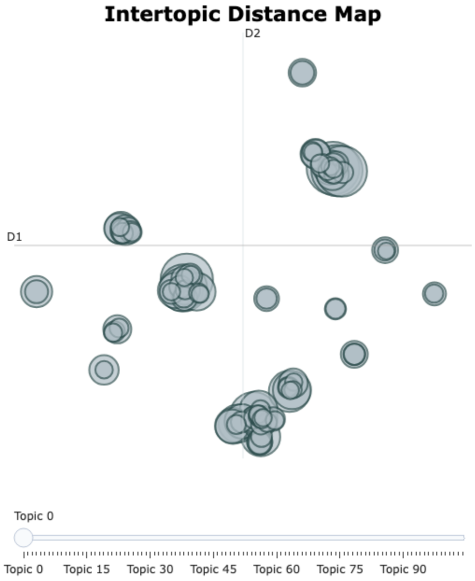

***Figure 5:*** *BERTopic Visualization of Clustered Mentions*

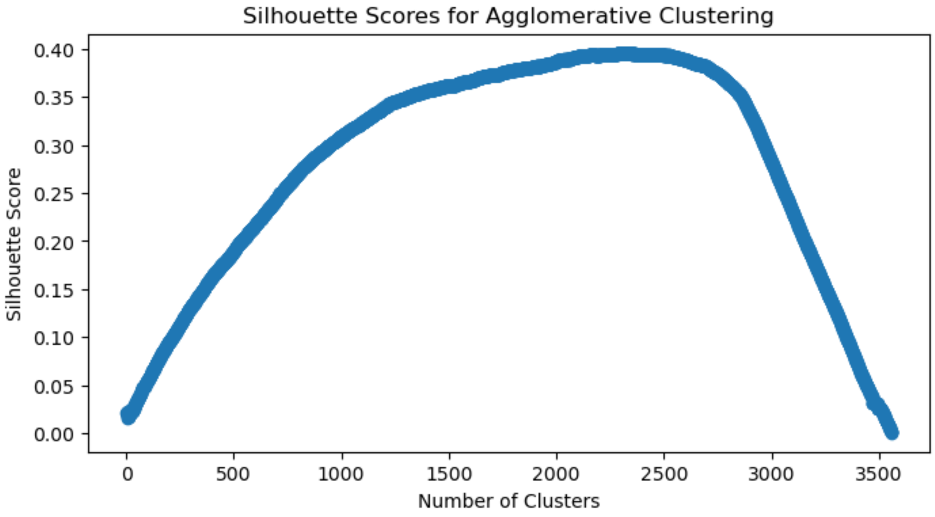

***Figure 6:*** *Silhouette Scores Across Cluster Counts*

We also tested several configurations of BERTopic to evaluate how tuning parameters affect topic quality. These results confirm that small adjustments to topic size thresholds help filter noise and improve the distinctiveness of generated topics.

| **Method** | **Coherence** | **Diversity** | **Silhouette** | **Best Clusters** |
|---|---|---|---|---|
| *BERTopic (baseline)* | -19.1205 | 0.894 | *N/A* | ~100–150 topics |
| *BERTopic (tuned)* | -18.45 | 0.894 | *N/A* | ~100–150 topics |
| *Agglomerative Clustering* | *N/A* | *N/A* | 0.487 | 2305 clusters |

***Table 4:*** *Component Analysis*

Both BERTopic and Agglomerative Clustering reveal meaningful structure in the data. BERTopic shows improved topic coherence and diversity with hyperparameter tuning, while Agglomerative Clustering, optimized via silhouette score, identifies a well-defined cluster count (2305). These metrics support a strong evaluation framework for topic and cluster quality.

The results highlight that unsupervised methods using S-BERT embeddings are effective alternatives to traditional NER, especially in low-resource settings. BERTopic enables interpretable topic discovery, while clustering captures fine-grained semantic groupings—together offering a scalable, interpretable approach to entity analysis in user-generated content.

### Trend Detection Module

<>

## Conclusion and Future Work

This project demonstrates the potential of advanced NER techniques on real-world, informal text. We developed a pipeline combining preprocessing, feature extraction, and model training, showing that domain-specific features and modern ML methods can achieve strong performance, even with limited data. Our results highlight the approach's effectiveness across varied text sources, supporting robust entity extraction.

The broader impact lies in transforming unstructured text into actionable insights for applications like sentiment analysis and information retrieval, with relevance across sectors such as healthcare and finance.

Future work should explore deeper transformer models, expand to multilingual and multi-domain datasets, and address challenges like context ambiguity and entity disambiguation. Semi-supervised and transfer learning may improve adaptability, while testing real-time deployment will help evaluate the system's practical use and resilience.

# 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.