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.

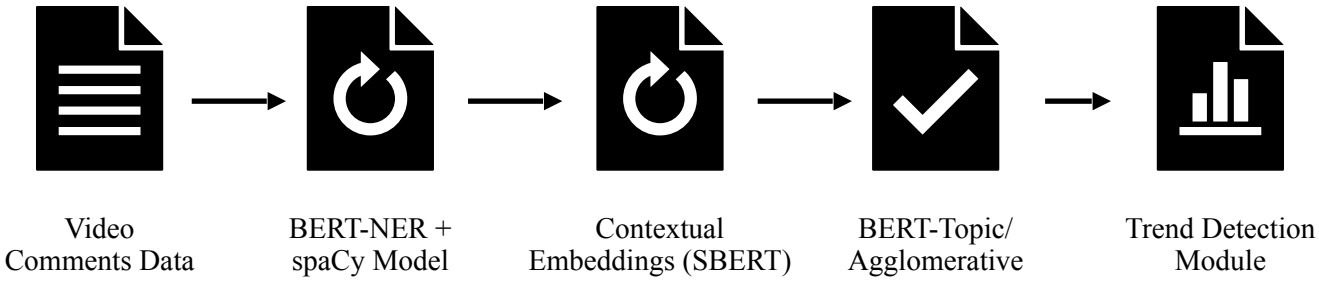

Figure 1: *Diagram of the research model's architecture.*

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

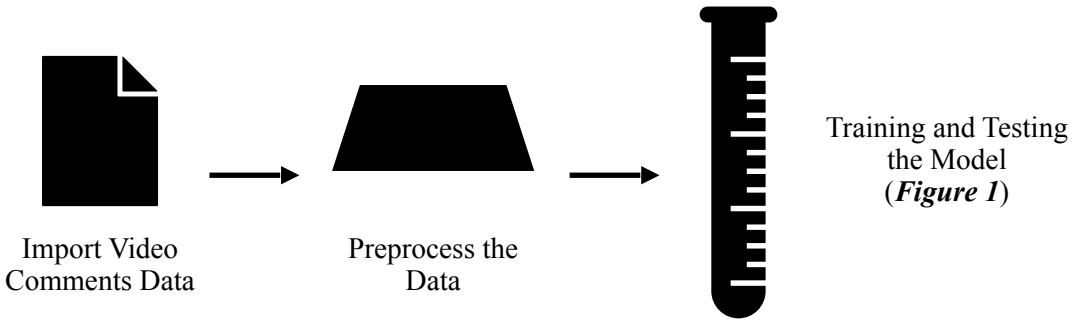

Figure 2: *Diagram of the research model's pipeline for implementation.*

Experimental Setup

We evaluated our approach using the NIAID video comments dataset, which includes annotations for relevance, sentiment, intention, and topic. Our experimental pipeline included multiple stages: initial baselines using traditional NER models, semantic encoding with Sentence-BERT to better handle informal language, and clustering with BERTopic to group related entities. Comment-reply structures were used to preserve conversational context. We employed semi-supervised learning to adapt to emerging or rare entities. Performance was assessed by comparing precision, recall, F1-scores, silhouette scores, topic coherence and diversity across baseline and enhanced models, demonstrating improved accuracy in recognizing entities within noisy, user-generated text.

Results

BERT-NER + spaCy Baseline Model

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.

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*

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 evaluate our SBERT-based NER pipeline, we generated synthetic token-level labels using cosine similarity between token embeddings and prototypical entity vectors. Embeddings were created for six entity types: PERSON (369), ORG (432), GPE (354), PRODUCT (33), WORK_OF_ART (19), and EVENT (2), revealing a moderate class imbalance that may impact generalization for rare entities. For example, the sentence "I love the Apple iPhone" was labeled as ['O', 'B-ORG', 'O', 'O', 'O'], with padded embeddings of shape [4, 24, 96] used for model training.

Entity Type	Precision	Recall	F1-score	Support
EVENT	1.00	1.00	1.00	6
GPE	0.96	0.92	0.94	1416
ORG	0.97	0.99	0.98	5535
PERSON	0.97	0.95	0.96	1744
PRODUCT	0.99	0.94	0.96	640
WORD_OF_ART	0.96	0.95	0.96	159
micro avg	0.97	0.97	0.97	9500
macro avg	0.97	0.96	0.97	9500
weighted avg	0.97	0.97	0.97	9500

Table 2: *Classification Report*

While some predictions (e.g., "love" as B-ORG) reflected noise due to limited prototypes, the model still reliably identified known entities like "Apple." Overall, the pipeline achieved strong performance, with a precision of 0.971, recall of 0.965, and F1-score of 0.968. All entity classes performed well, including low-frequency types like PRODUCT and WORK_OF_ART, which maintained F1-scores above 0.95. Topic modeling showed high diversity (0.74) but moderate coherence (0.377) and a negative silhouette score (-0.063), indicating broad thematic coverage but weak cluster separation. These results highlight the effectiveness of SBERT embeddings for entity recognition in informal text, while suggesting room for improvement in clustering techniques.

Metric	Value
Precision	0.971
Recall	0.965
F1	0.968
Topic Coherence	0.377

Metric	Value
Topic Diversity	0.74
Silhouette Score	-0.063

Table 3: *Evaluation metrics for the S-BERT model.*

BERT-Topic / Agglomerative Clustering Model

We evaluated BERTopic and Agglomerative Clustering for grouping semantically related entity mentions in video comments. BERTopic generated interpretable topic themes from sentence embeddings, while Agglomerative Clustering structured mentions hierarchically. Both methods were assessed using topic coherence, topic diversity, and silhouette scores to quantify semantic and clustering performance. Manual inspection confirmed high intra-cluster consistency. For example, Cluster 3 grouped "Apple Vision Pro," "the headset," "AVP," and "vision device," while Cluster 7 included "Zuck," "Meta," "Zuckerberg," and "founder of Facebook." These results demonstrate the models' effectiveness in linking variants, abbreviations, and contextual references—capabilities that traditional NER approaches often lack.

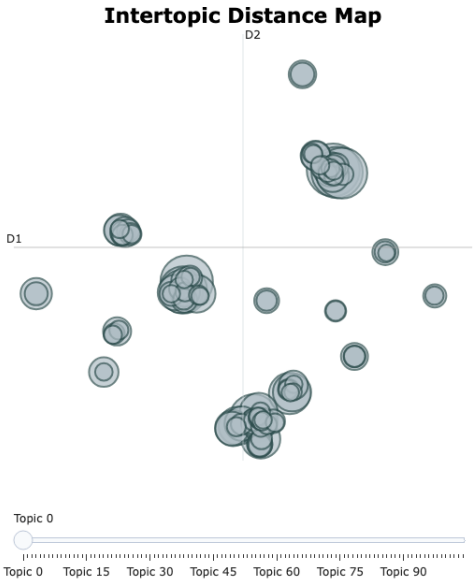

Figure 3: *BERTopic Visualization of Clustered Mentions*

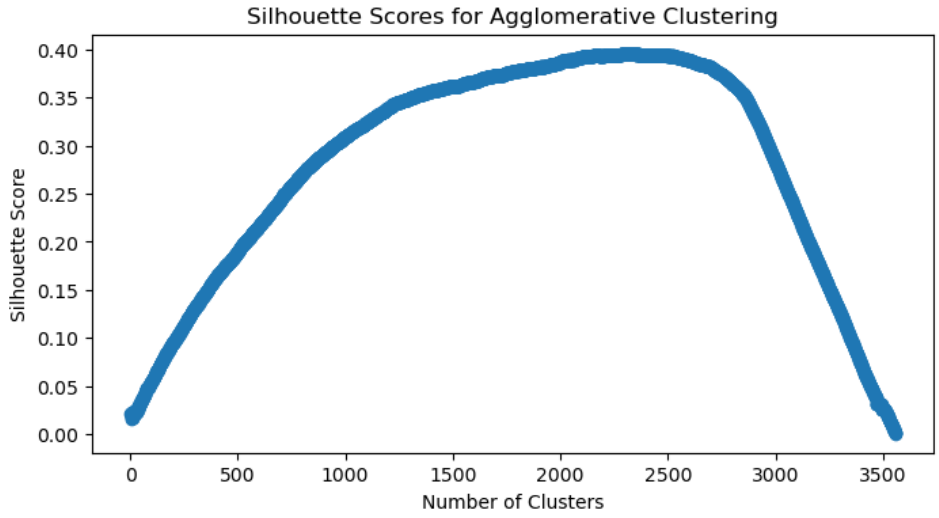

Figure 4: *Silhouette Scores Across Cluster Counts*

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)	-19.1675	0.894	*N/A*	~100–150 topics
BERTopic (tuned)	-18.7342	0.9225	*N/A*	~100–150 topics
Agglomerative Clustering	-18.7342	0.9225	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

To evaluate NER approaches for tracking emerging topics in video comments, we built a trend detection pipeline to monitor the frequency and evolution of new entities over time. BERT-based NER offered broad coverage but struggled with semantic grouping, causing fragmented trends. In contrast, SBERT's contextual embeddings enabled better clustering of related entities, leading to more coherent topic tracking. Quantitatively, the BERT-topic + agglomerative clustering model identified more unique trends and improved overall trend consistency compared to BERT alone.

Figure 5: *Trend Detection Results*

These results suggest that integrating topic modeling with entity clustering significantly enhances the system's ability to surface relevant and timely trends from noisy comment data.

Conclusion and Future Work

This work demonstrates the effectiveness of advanced named entity recognition (NER) techniques on informal, real-world text. We propose a hybrid approach that combines BERT-Topic modeling with Agglomerative Clustering, achieving robust entity extraction across diverse and noisy data sources. Our findings highlight the potential of this method to transform unstructured text into actionable insights for downstream tasks such as sentiment analysis and information retrieval. The approach is applicable across multiple domains, including healthcare and finance, where accurate information extraction is critical.

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.