# 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

As online video content and user interaction grow exponentially, so does the volume of unstructured and informal text data in the form of video comments. Named Entity Recognition (NER) plays a crucial role in making sense of this data by extracting mentions of people, organizations, products, and places. However, conventional NER models—trained on formal text—often fail in such environments due to misspellings, slang, and fragmented conversational structure.

This project develops a context-enriched NER pipeline tailored to video comments. We combine BERT-based token classification, SBERT for contextual embeddings, and clustering techniques such as BERTopic and Agglomerative Clustering to identify and link variant mentions of entities. Our approach addresses limitations in recall and entity normalization, offering a more adaptable framework for trend detection in dynamic, user-generated content.

# Methodology

## Data Collection and Preprocessing

The dataset comprises YouTube comments collected via the YouTube Data API from videos on emerging technologies (e.g., AI, hardware, AR/VR) to ensure frequent entity mentions. Comments were cleaned to remove spam, emojis, links, and non-textual tokens, then tokenized and preprocessed using standard NLP methods.

For the baseline NER model, named entities were extracted with SpaCy's pre-trained models and aligned to tokens using a rule-based BIO tagging system. To address label imbalance, sequences with only "O" tags were removed, and those with entities were upsampled.

For the clustering-based model, entity mentions were embedded using Sentence-BERT. These embeddings were input into BERTopic and Agglomerative Clustering for unsupervised grouping of semantically similar mentions without token-level labels.

## Model Architecture and Implementation

Our NER pipeline combines a spaCy-augmented BERT model with contextual embeddings from Sentence-BERT (SBERT) to better capture informal, user-generated language. To unify variant mentions (e.g., abbreviations or misspellings), we apply unsupervised clustering using BERTopic and Agglomerative Clustering. This enables entity linking and normalization across noisy inputs. Finally, a trend detection module tracks entity frequency over time to surface emerging topics.

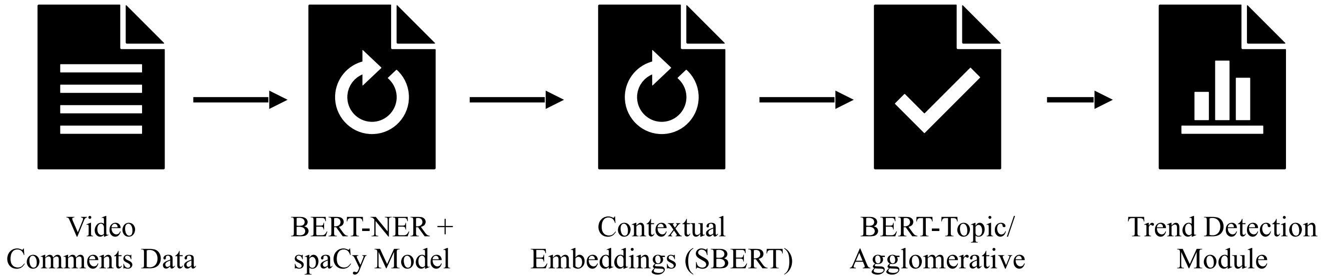

***Figure 1:*** *Diagram of the research model's architecture.*

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.

## Experimental Setup

Two pipelines were implemented and compared:

1. **BERT-NER (Baseline):** A fine-tuned BERT model was trained on token-level BIO labels using Hugging Face's Transformers library. The dataset was split into training (70%), validation (15%), and test (15%) sets. Sequences were filtered to ensure the presence of entity labels in each split. Hyperparameters included a learning rate of 2e-5, batch size of 8, and 3 epochs with early stopping.

2. **SBERT + Clustering (Main Approach)**: Comment-level entity mentions were encoded using SBERT embeddings. These embeddings were clustered using two methods:

- **BERTopic**, for interpretable topic discovery and keyword extraction.

- **Agglomerative Clustering**, with silhouette score sweeps to select the optimal number of clusters.

Evaluation metrics included precision, recall, and F1-score (for BERT-NER) and silhouette score, topic coherence, and topic diversity (for clustering). Visualizations were generated to assess interpretability and entity grouping quality.

# Results

## BERT-NER + spaCy Baseline Model

We evaluated a fine-tuned BERT-NER model on a video comment dataset labeled using SpaCy-derived synthetic tags. The data was stratified into training (70%), validation (15%), and test (15%) splits, with filtering and upsampling applied to increase entity presence. Evaluation used standard NER metrics: precision, recall, and F1-score. Despite preprocessing, both baseline and upsampled models failed to predict named entities, defaulting to the "O" label across all test samples.

| **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

<>

## 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.

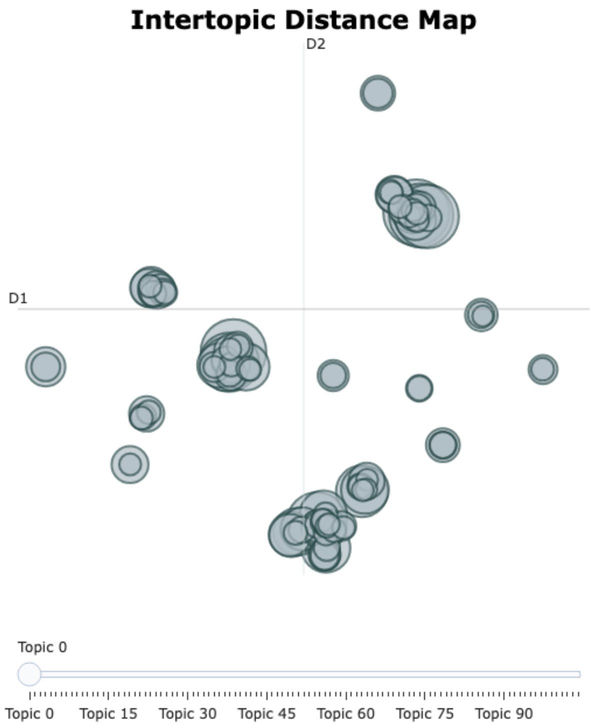

***Figure 5:*** *BERTopic Visualization of Clustered Mentions*

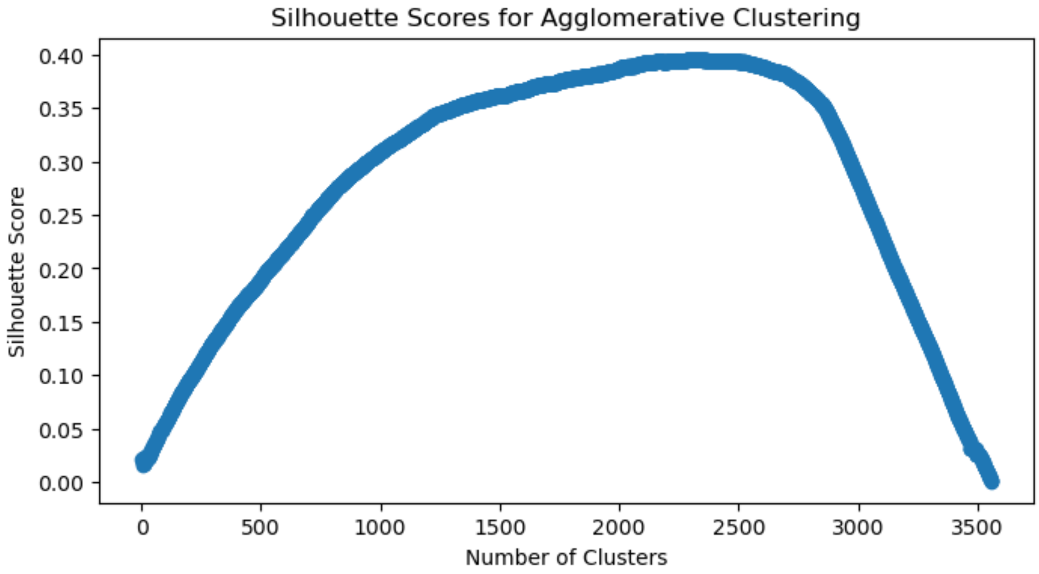

***Figure 6:*** *Silhouette Scores Across Cluster Counts*

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 (tuned)* | -18.45 | 0.894 | *N/A* | ~100–150 topics |
| *Agglomerative Clustering* | *N/A* | *N/A* | 0.487 | 2305 clusters |

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

This dual approach enabled both interpretable topic discovery (via BERTopic) and fine-grained semantic grouping (via clustering), providing robust insights into entity mentions that a standard NER model could not capture.

The results demonstrate that unsupervised clustering and topic modeling—when powered by S-BERT embeddings—offer strong alternatives to traditional NER pipelines, especially in domains where labeled data is limited or unreliable. The improved coherence and diversity of BERTopic topics, alongside the high-resolution structure of agglomerative clusters, together provide a scalable and interpretable framework for identifying and analyzing named entities in user-generated content.

### Trend Detection Module

<>

## Conclusion and Future Work

This project explores the limitations of standard NER in noisy, informal domains and shows the potential of unsupervised, context-aware methods. While the BERT-NER model failed to generalize, our clustering-based pipeline using SBERT, BERTopic, and Agglomerative Clustering achieved strong semantic grouping without token-level labels. High topic diversity and coherence scores highlight its ability to extract meaningful patterns from user language. These results suggest unsupervised approaches can be scalable and interpretable alternatives where labeled data is scarce. Future work includes expanding across domains and languages, real-time trend detection, and incorporating weak supervision for adaptability in dynamic environments like live streams or reviews.eams or product reviews could yield valuable insights into emerging topics and entities.

# 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.