DATASCI 266 Project Proposal - Section 2
Chelle Davies
Ziyad Amer

Option 1: Context-Enriched Named Entity Recognition (NER) for Identifying Emerging Trends in Video Comments

Objective

This research aims to develop a context-enriched NER pipeline to detect and categorize mentions of people, places, brands, and products in video comments. Unlike conventional NER, our approach will incorporate comment-reply hierarchies and semantic embeddings to refine entity recognition and identify emerging trends.

Significance & Challenges

- Traditional NER struggles with informal language and misspellings in user-generated comments.
- Named entities in video discussions often appear across multiple comments and replies, requiring a context-aware approach.
- Emerging or niche terms (e.g., newly introduced product names) may not be present in existing NER training datasets, necessitating semi-supervised learning for domain adaptation.

Dataset

We will use the "Dataset of Video Comments of a Vision Video Classified by Their Relevance, Polarity, Intention, and Topic" from NIAID.

Algorithms & Implementation

- Transformer-based NER models: Fine-tune BERT-NER or spaCy's Transformer NER for informal comment text.
- Contextual embeddings: Use SBERT (Sentence-BERT) to improve entity recognition by incorporating surrounding comment context.
- Entity linking and clustering: Use BERT-Topic or Agglomerative Clustering to group similar entity mentions (e.g., different spellings of the same product).
- Trend detection: Track the frequency and evolution of newly identified entities to uncover trending topics.

Related Work

1. Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, 11 Oct. 2018, arxiv.org/abs/1810.04805.
2. Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." *arXiv*, 27 Aug. 2019, https://arxiv.org/abs/1908.10084.
3. von Luxburg, Ulrike. "A Tutorial on Spectral Clustering." Statistics and Computing, vol. 17, no. 4, 2007, pp. 395–416. arXiv, arxiv.org/abs/0711.0189.

4. 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, ieeexplore.ieee.org/document/8940293.