Comparative Analysis of Gun Violence Coverage Across News Outlets

Overview

As an application of NLP, in this assignment, you will analyze how different news outlets portray victims and shooters in gun violence incidents. Using 100 news articles (25 each from CNN, Fox News, The New York Times, and The Wall Street Journal), you will extract, cluster, and statistically analyze the descriptive language used to characterize these entities. This assignment combines NLP techniques including coreference resolution, phrase extraction, embedding-based clustering, and statistical hypothesis testing.

Dataset: Google Drive Link

Task 1: Context Extraction with Coreference Resolution (20 points)

Objective: Identify all text segments that reference the victim(s) and shooter(s) in each article, resolving coreferences to create coherent contexts.

Requirements:

- Use a coreference resolution tool (e.g., spaCy's neuralcoref, AllenNLP's coref, or Hugging Face's coref models) to identify all mentions of victims and shooters
- Replace pronouns and other referring expressions with the actual entity names (e.g., "He fired multiple shots" → "The shooter fired multiple shots")
- Extract complete sentences that contain these resolved references
- Store the extracted contexts separately for victims and shooters for each article
- Brief documentation (max 300 words) explaining your approach and any challenges encountered

---

Task 2: Description Extraction (20 points)

Objective: From the contexts extracted in Task 1, identify specific phrases, words, or expressions used to describe the entities or their actions.

Requirements:

- Develop and implement a method to extract descriptive elements. You may use ONE or COMBINE multiple approaches:
  - **POS-based extraction**: Extract adjectives, adverbs, and descriptive noun phrases that **modify the entities (victims or shooters).**
  - **LLM-based extraction**: Use a language model (GPT, or open-source LLMs) to identify descriptive phrases
  - **Sentiment-based extraction**: Apply sentiment classifiers or emotion detection models to identify emotionally-charged or evaluative language
  - **Dependency parsing**: Use syntactic dependencies to find modifiers and predicates related to entities
- Justify (expected strengths and limitations) your chosen approach(es) with a clear rationale (max 300 words)
- Ensure extracted descriptions maintain enough context to be interpretable.

Task 3: Description Clustering (20 points)

Objective: Group semantically similar descriptions into clusters to identify patterns in how entities are portrayed.

Requirements:

- Choose an embedding approach:
  - **Option A**: Use Word2Vec, GloVe, or FastText embeddings for individual descriptions
  - **Option B**: Use BERT/RoBERTa token embeddings for descriptions
  - **Option C**: Use SBERT sentence embeddings on the full extracted context, then extract contextual embeddings for your descriptions
- Apply DBSCAN (or justify an alternative clustering algorithm like K-Means, Hierarchical, or HDBSCAN)
- Tune hyperparameters (epsilon, min_samples for DBSCAN) and document your process
- Create meaningful cluster labels based on manual inspection
- Visualize the clusters (t-SNE or UMAP projection recommended)
- Document (max 300 words) explaining your embedding choice and hyperparameter tuning process.

Example clusters:

- Age references: "18-year-old", "teenager", "minor", "juvenile"
- Violence descriptors: "opened fire", "shot multiple times", "killed"
- Victimhood framing: "innocent victim", "caught in crossfire", "bystander"

Task 4: Manual Cluster Evaluation (15 points)

Objective: Assess the quality of your clusters both lexically (surface form) and semantically (meaning).

Requirements:

- Manually review each cluster and evaluate:
  - **Lexical coherence**: Do items share similar words or morphological patterns?
  - **Semantic coherence**: Do items convey similar meanings or framings?
  - **Cluster purity**: Are there obvious misclassifications?
- Document problematic clusters and explain why they occurred
- Refine clusters if necessary (merge, split, or reassign) and justify your decisions

Task 5: Cross-Outlet Frequency Analysis (10 points)

Objective: Create comprehensive tables showing how frequently each cluster appears across news outlets and entity types.

Requirements:

- Create tables showing:
  - **Proportion table**: For each outlet and entity type, what percentage of descriptions fall into each cluster?
  - **Frequency table**: Raw counts of cluster occurrences by outlet and entity type
- Calculate both absolute frequencies and normalized proportions
- Create visualizations (heatmaps, grouped bar charts) to illustrate patterns

Task 6: Statistical Hypothesis Testing (15 points)

Objective: Determine if the observed differences in framing are statistically meaningful.

1. **Hypothesis Formulation:** For the top 3 most frequent clusters for ***each*** **entity (shooter vs victim)**, formulate a null hypothesis stating that there is no difference in the proportion of that frame used across the four news outlets.
2. **Statistical Test:** Conduct a **Chi-Squared Test of Homogeneity** for the distribution of these framing clusters across the four news outlets.

3. **Conclusion:** Based on the p-value, state whether you reject or fail to reject the null hypothesis, concluding whether a specific outlet exhibits a statistically significant **overuse** or **underuse** of a particular frame cluster when compared to the expected distribution.

Submission Requirements

Code Submission:

- Well-commented Jupyter notebook(s) or Python scripts
- Requirements.txt or environment.yml file
- README with setup and execution instructions

Dataset Submission:

- Final processed datasets (CSV/JSON format)
- Cluster assignments and labels