Milestone 2: Topic discovery for news articles, Winter 2025

Jakub Bazyluk
Warsaw University of Technology
jakub.bazyluk.stud@pw.edu.pl

Daniel Tytkowski
Warsaw University of Technology
daniel.tytkowski.stud@pw.edu.pl

Mikołaj Kita
Warsaw University of Technology
mikolaj.kita.stud@pw.edu.pl

Anna Wróblewska (Supervisor)
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This project studies automatic topic discovery for news articles and evaluates topic-level summaries along three axes: sentiment, factual content, and framing/bias. We design experiments to verify the reliability of automatic methods, propose guided methodologies to separate sentiment, factual information, and framing, and provide an open, reproducible pipeline. Our approach integrates embedding-based topic modeling, claim detection, and evidence-based verification, leveraging recent advances in multilingual sentiment analysis and seed-guided topic discovery.

1 Introduction

In today's world, news articles are published at a rapid pace across multiple sources, platforms, and countries. While this offers broad access to information, it also introduces challenges for consumers: distinguishing between objective reporting, subjective interpretation, and biased narratives has become increasingly difficult.

Topic modeling remains a fundamental task in Natural Language Processing. Historically, probabilistic algorithms such as Latent Dirichlet Allocation, Latent Semantic Indexing, and Nonnegative Matrix Factorization served as the standard. However, recent advancements have shifted toward embedding-based methods like BERTopic and Top2Vec, which leverage transformer architectures to capture contextual nuances that bag-of-words models miss.

The objective of this project is to design and evaluate an automated system that first extracts coherent news topics, then explicitly separates them into **Factual Claims** and **Narrative Frames**. We move beyond simple clustering by employing guided methods that force models to respect linguistic boundaries between fact and opinion.

2 Project goals

The primary goal of this project is to design a robust pipeline for analyzing news articles that combines topic identification, sentiment analysis, bias detection, subjectivity assessment, and framing recognition. Specifically, the project aims to:

1. **Automatically classify articles by sentiment, bias, and subjectivity**, providing a quantitative summary of content characteristics.
2. **Identify latent news topics** using embeddings and clustering, enabling thematic exploration across multiple sources.
3. **Assess framing and narrative patterns** in news articles through zero-shot or claim-extraction based classification methods.
4. **Design experiments to validate the reliability and accuracy of automatic methods,** including comparison to human labeled datasets, but mainly accessing the differences in clustering using different techniques.

3 Hypothesis

Several hypotheses can be formulated for this study:

1. **Framing influences topic differentiation** Articles covering similar topics might be framed differently on the source, so the event may be viewed from different perspective.
2. **Bias affects perceived topic boundaries** Highly biased articles may emphasize certain parts of a story while understating others, effectively influencing the thematic representation.

3. **Interaction of framing and bias amplifies differentiation** Combining framing probabilities and bias levels with embeddings will likely provide a richer, multidimensional view of the articles.

4 Data

4.1 BBC dataset

We use a custom dataset created from scratch by us. Using publicly available articles, we download them and store them in a dictionary.

- url: url to the article
- title: title of the article
- pubdate: publication date of the article
- topic: topic based on BBC segmentation
- text: article itself

All articles come from renowned source - BBC. It's possible that in the future we will enhance our dataset with new sources.

4.2 AG News Subset

The AG News dataset (3) is a benchmark for text classification. It contains news articles categorized into four classes: *World*, *Sports*, *Business*, and *Science/Technology*. Each example consists of a **title**, a **description**, and a corresponding **label**. The dataset is publicly available via the Hugging Face Datasets library (3). The labels in this dataset are treated as ground truths for later work.

5 Approach & research methodology

5.1 Claim extraction

We propose a method of summarizing articles in a coherent form of tabular data. Each news in our BBC dataset is split into sentences and each sentence is processed to get a dictionary of key-value pairs for each action happening in this sentence with given keys:

- actor: a person that performs an action
- action: self-explanatory
- object: receipient of an action - for example: I meet you. In this case 'you' is object
- location: location of an action based on NER
- time: time of an action based on NE

We use spacy transformers model **en_core_web_trf** and careful algorithm to extract data. It's based on POS and NER. As an example of possibilities of spacy model we show a tree of annotated sentence: 'US President Donald Trump's overseas envoy will travel to Germany this weekend to meet Ukrainian President Volodymyr Zelensky and European leaders for more talks on ending the war.'. We provide example in Figure 1

Claim extraction will potentially be used in future parts of our work to enhance results and further exploit framing techniques.

5.2 Embeddings pipeline

In this project, we propose a multi-stage pipeline for automatic analysis of news articles, aimed at identifying topics and summarizing them with respect to sentiment, bias, subjectivity, and framing.

5.2.1 Dataset Collection

We begin by collecting a large-scale dataset of news articles from multiple sources. Using data from different publishers ensures diversity in writing style, political orientation, and editorial perspective, which is essential for analyzing bias and framing influence.

5.2.2 Text Embedding

Each article is represented using dense vector embeddings computed from the article content. These embeddings capture the semantic structure of the text and are later used for topic identification, clustering, and similarity analysis. The model used for embeddings is

5.2.3 Sentiment Analysis

Sentiment is measured using the `twitter-roberta-base-sentiment` model. Each article is assigned one of three sentiment labels:

- Negative (0)
- Neutral (1)
- Positive (2)

This step provides a high-level emotional characterization of the news content.

5.2.4 Bias Detection

To assess media bias, we apply the `newsmediabias/UnBIAS-classifier`. The model categorizes articles into three classes:

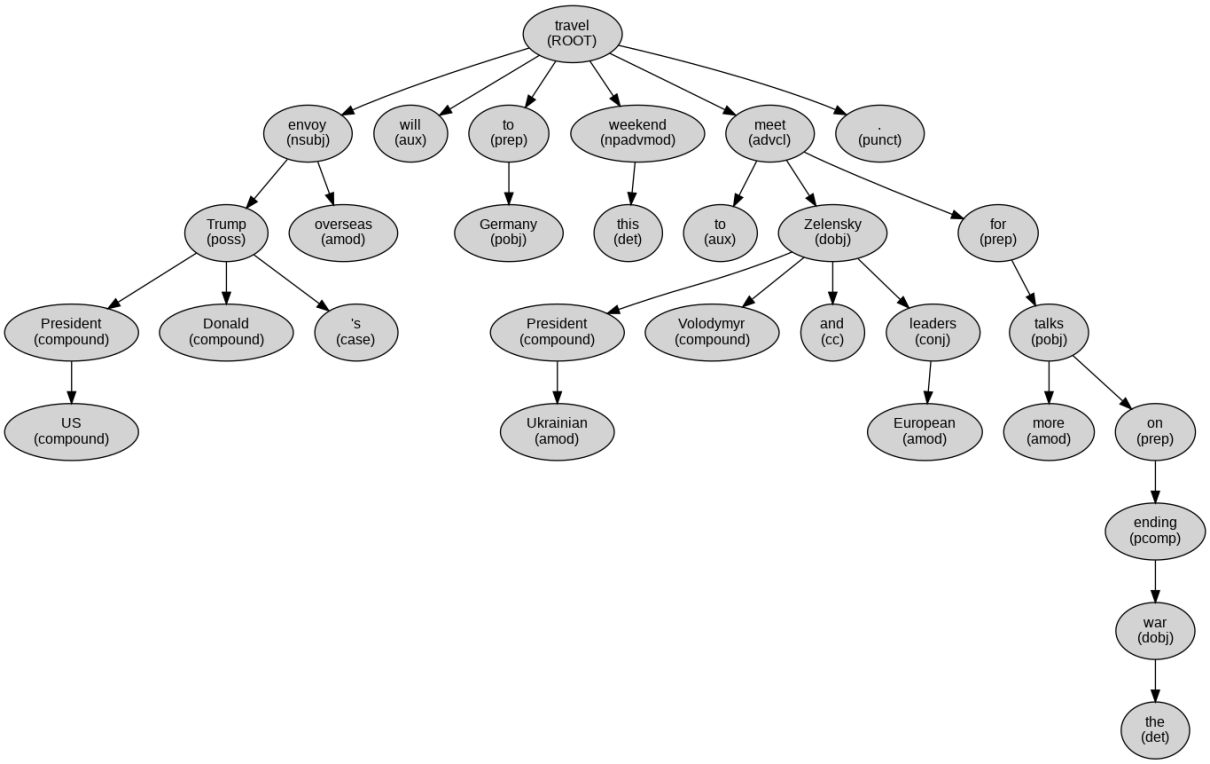

Figure 1: Claim extraction graph

- Neutral
- Slightly Biased
- Highly Biased

The predicted categories are converted into numerical labels and used as indicators of editorial bias.

5.2.5 Subjectivity Classification

Subjectivity is measured using the `GroNLP/mdebertav3-subjectivity-multilingual` model. Articles (or their constituent sentences) are classified as either:

- Objective (OBJ)
- Subjective (SUBJ)

This step helps distinguish factual reporting from opinionated or interpretative content.

5.2.6 Framing Analysis

Framing analysis is the most challenging component of the pipeline. Since framing is difficult to define with a fixed label set, we adopt a guided classification approach using `cross-encoder/nli-deberta-v3-small`. A predefined set of hypothetical framing labels (Corporate/Markets, Social Impact/Labor, Neutral/Reporting and Non-Economic) is provided to the model, which assigns a single discrete label.

5.2.7 Output Representation

For each article, the pipeline returns:

- A semantic embedding vector
- Sentiment label
- Bias label
- Subjectivity label
- Framing probability vector

This structured representation enables downstream analyses such as topic clustering, cross-source comparison, and evaluation of how sentiment, bias, and framing influence topic formation.

5.3 Topic Discovery Methodology

5.3.1 Feature Construction

For each news article, multiple embedding representations are constructed in order to capture different nature of the content:

- $emb_{content}$ – semantic embedding of the article text

- $emb_{sentiment}$ – embedding or numerical representation of sentiment
- emb_{bias} – embedding or numerical representation of media bias
- $emb_{subjectivity}$ – embedding or numerical representation of subjectivity
- $emb_{framing}$ – vector of framing probabilities

Each embedding encodes complementary information about the article beyond its lexical semantics.

5.3.2 Baseline Topic Clustering

1. Baseline clustering is performed using $emb_{content}$ only.
2. This configuration serves as the baseline and reflects topic structure derived purely from semantic similarity.
3. Cluster coherence and interpretability are evaluated.

5.3.3 Topic Clustering with Sentiment

1. A joint representation is created by concatenating:

$$emb_{content} \oplus emb_{sentiment}$$

2. Topic clustering is performed on the augmented embedding.
3. The resulting clusters are compared with the baseline to identify:
   - Changes in cluster assignments
   - Separation of emotionally distinct narratives
   - Variations in cluster coherence

5.3.4 Topic Clustering with Sentiment and Bias

1. The embedding space is further extended as:

$$emb_{content} \oplus emb_{sentiment} \oplus emb_{bias}$$

2. Topic clustering is performed on the extended representation.
3. Differences relative to previous configurations are analyzed, focusing on:
   - Bias-driven subtopic formation
   - Cluster splitting or merging
   - Source-dependent narrative separation

5.3.5 Topic Clustering with Full Context

1. A final embedding is constructed by concatenating all available representations:

$$emb_{final} = emb_{content} \oplus emb_{sentiment} \oplus$$

$$\oplus emb_{bias} \oplus emb_{subjectivity} \oplus emb_{framing}$$

2. Topic clustering is performed on the full gathered information.
3. This setting aims to capture both topical similarity and narrative perspective.

5.3.6 Comparative Evaluation

For each clustering configuration, the following analyses are conducted:

- Quantitative comparison of clusterings using stability and similarity metrics (e.g., Adjusted Rand Index, Normalized Mutual Information).
- Qualitative inspection of representative articles from each cluster.
- Assessment of how sentiment, bias, and framing alter topic boundaries.

5.3.7 Expected Outcome

We hypothesize that augmenting semantic embeddings with sentiment and bias information will significantly influence the topic clustering structure. Each step in topic discovery will refer to our hypothesis. In particular:

- Sentiment will separate emotionally divergent fragments within the same topic.
- Bias will differentiate editorial perspectives.
- Full-context embeddings will reveal topic structures that are not observable using content alone.

Pipeline like that should demonstrate that topic discovery in news articles is influenced not only by semantic content but also by sentiment, bias, and framing.

6 Results

6.1 Preprocessed dataset

After whole process of preparing dataset according to **embedding pipeline** we are left with 12000 embedded articles.

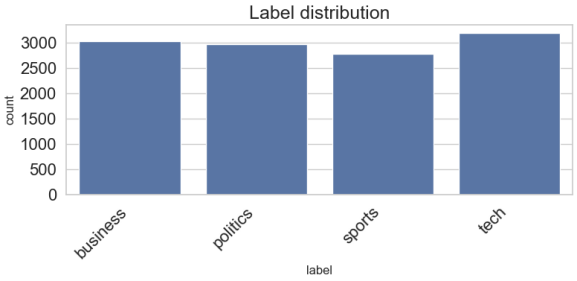

Figure 2: Default label distribution

Each article has default generic category label retrieved directly from data sources. The distribution is more or less even.

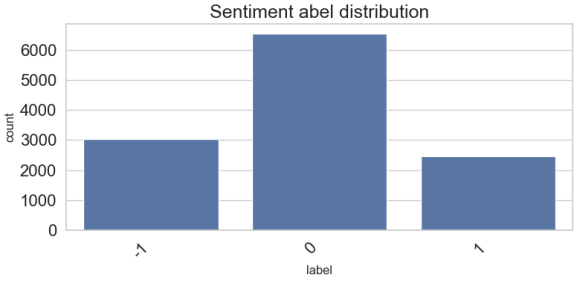

Figure 3: Sentiment labels distribution

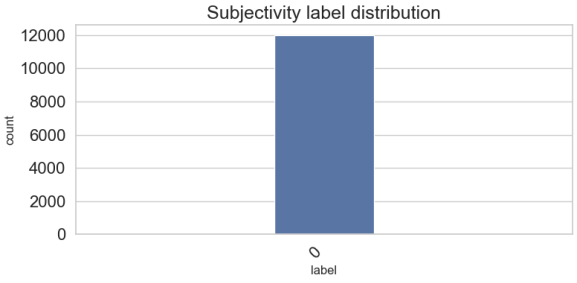

Figure 4: Subjectivity labels distribution

We are working mainly on neutral data according to the classification done by `GroNLP/mdebertav3-subjectivity-multilingual`.

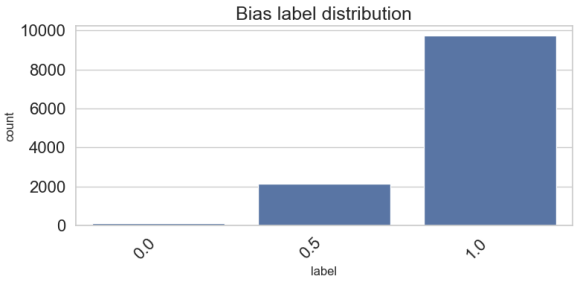

Figure 5: Bias labels distribution

We divided bias into 3 categories:

- Neutral (0)
- Slightly Biased (0.5)
- Highly Biased (1)

There is a contradiction with subjectivity and bias, which are presented by two different models. While `GroNLP/mdebertav3-subjectivity-multilingual` claims that data is objective, the model responsible for bias `newsmediabias/UnBIAS-classifier` classified the majority of data as highly biased.

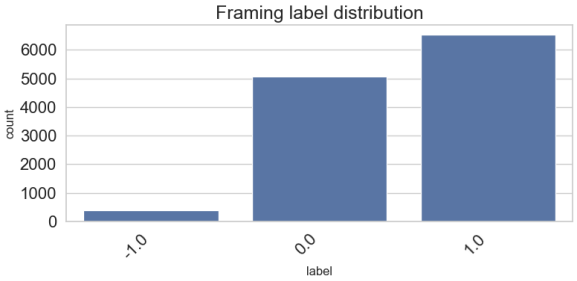

Figure 6: Framing labels distribution

Framing is divided into 3 distinct values for now:

- Corporate & Markets (1)
- Social Impact & Labor (-1)
- Neutral/Reporting (0)
- Non-Economic (0)

Semantic Article Representation Each article is represented as a dense vector of 384 numerical features containing its semantic meaning. The embeddings are generated using a pretrained sentence-level transformer model. In this

space, semantically similar articles are positioned closer together, while conceptually distinct articles should be placed farther apart. This representation enables effective comparison of articles based on meaning rather than just lexical similarity and serves as the primary input for downstream tasks such as topic clustering and similarity analysis.

Derived Narrative Features On top of the semantic representation, we incorporate a set of derived features that capture additional narrative characteristics of each article. These features include sentiment, subjectivity, media bias, and framing, which are extracted using pretrained transformer-based models. While semantic embeddings encode what the article is about, the derived features describe how the information is presented. Together, they provide a richer, multidimensional representation that enables analysis of both topical similarity and narrative perspective, and allow us to study how sentiment, bias, and framing influence topic discovery.

6.2 Clustering

t-SNE Visualization of Article Clusters To explore the structure of the article embeddings, we applied t-distributed Stochastic Neighbor Embedding to project high-dimensional representations into two dimensions. Figure 10 shows the resulting visualization, where each point represents an article and colors correspond to the clusters obtained through our topic discovery pipeline. The t-SNE plots reveal that semantic embeddings alone produce coherent clusters reflecting topical similarity. When derived features such as sentiment, subjectivity, bias, and framing are incorporated, the cluster structure becomes more nuanced, highlighting distinctions driven by narrative perspective as well as content. This visualization confirms that augmenting semantic embeddings with additional features can reveal latent patterns that are not apparent from text semantics alone.

7 Summary

We developed a new dataset that integrates topic discovery with bias, sentiment, framing, and subjectivity analysis. While human annotation is the gold standard, we utilized various NLP models to generate these metrics, as manual labeling would be impossible for us. The resulting dataset is available at `https://huggingface.co/datasets/mkita/topic-discovery-for-news-articles`. Our results shows that even a relatively simple approach of incorporating derived features can lead to significant differences in the discovered topics. By augmenting semantic embeddings with sentiment, subjectivity, bias, and framing information, we observed changes in cluster composition and separation. Articles that were previously grouped together based solely on semantic similarity were split into distinct clusters when narrative characteristics were considered, revealing patterns of opinion, emotional tone, and editorial perspective.

8 Next steps

Future work will focus on developing a classification model trained on this augmented dataset. By utilizing a subset of human-annotated labels as ground truth, we can evaluate whether the inclusion of bias, sentiment, and framing features enhances the model's predictive accuracy. This will allow us to quantify the extent to which narrative dimensions improve automated topic labeling compared to standard text-based approaches.

References

Zhang, Xiang, et al. "Character-level convolutional networks for text classification." *NIPS* 2015.

Hugging Face Datasets: `https://huggingface.co/datasets/ag_news`

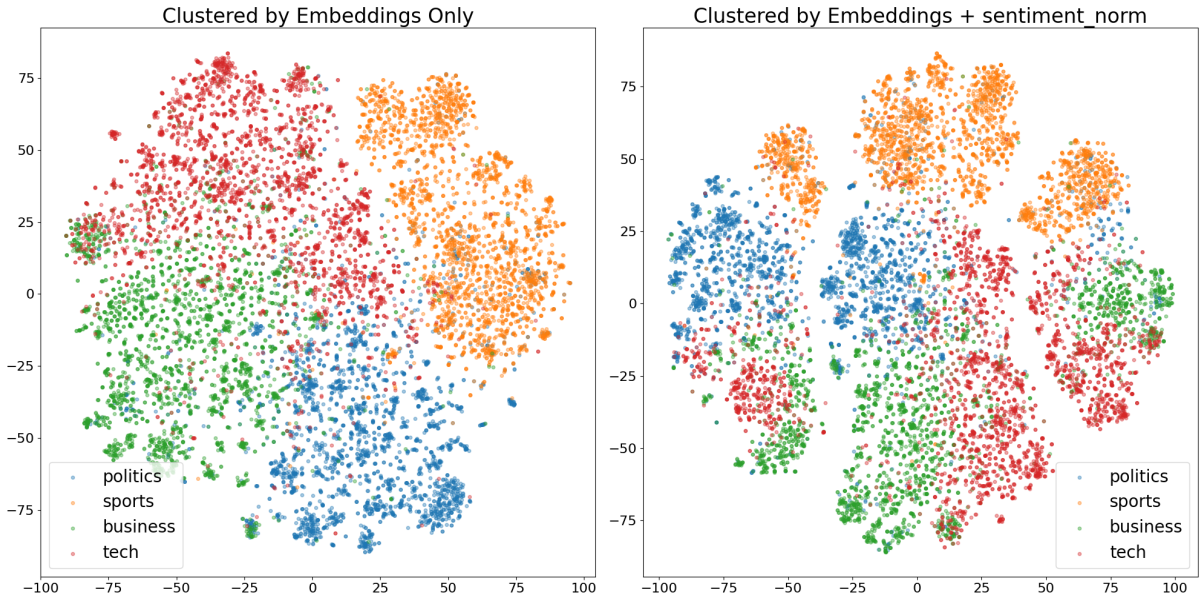

Figure 7: Sentiment influence on clusters

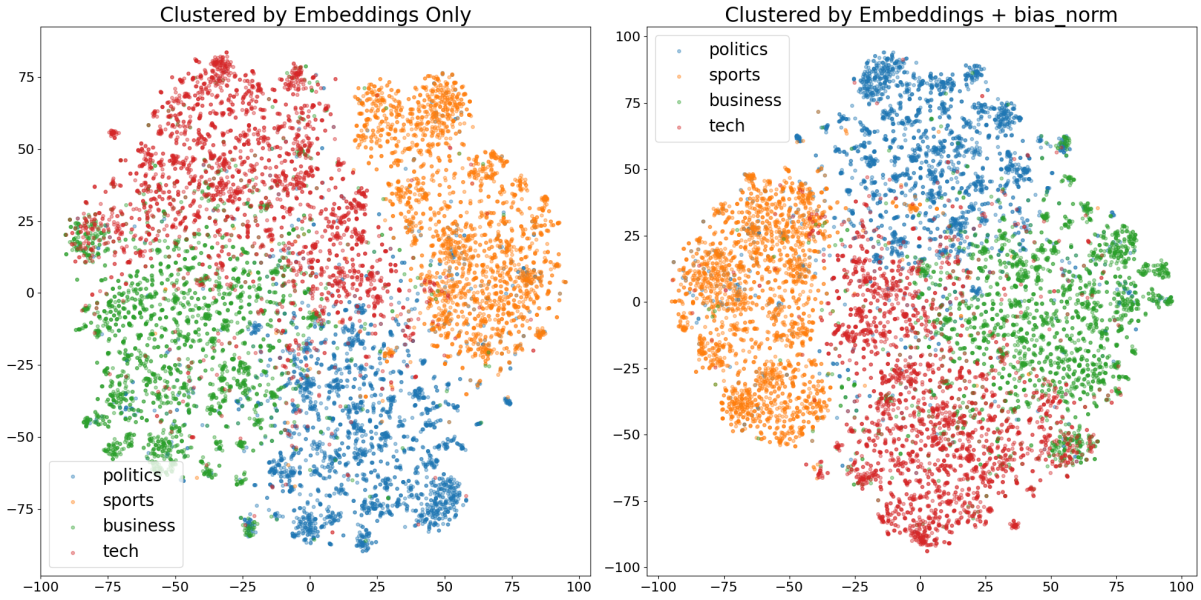

Figure 8: Bias influence on clusters

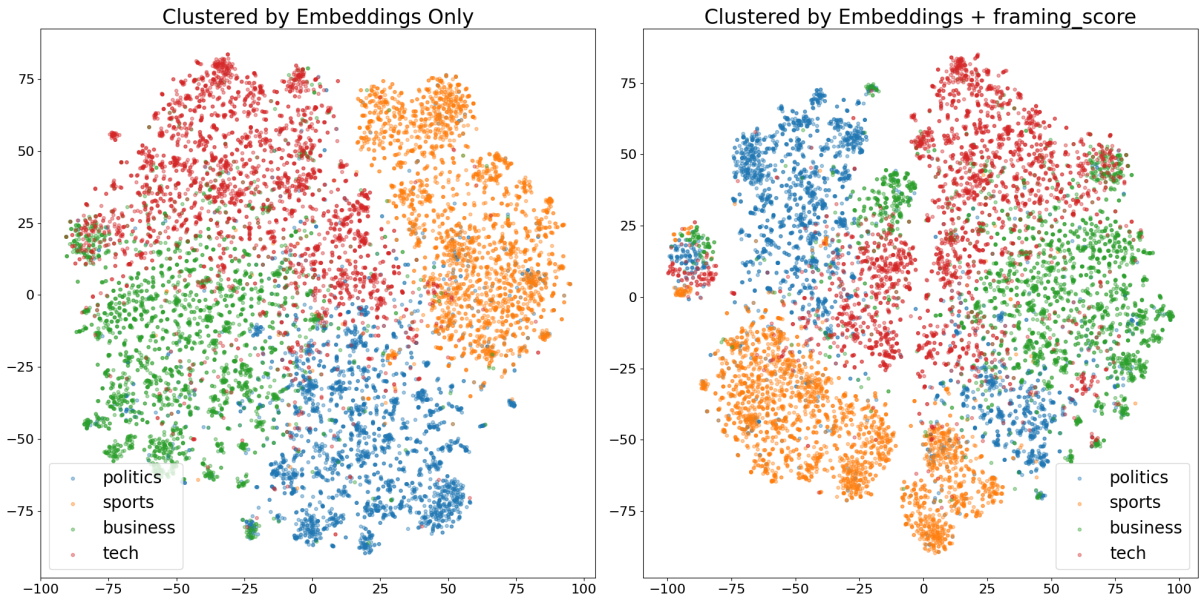

Figure 9: Framing influence on clusters

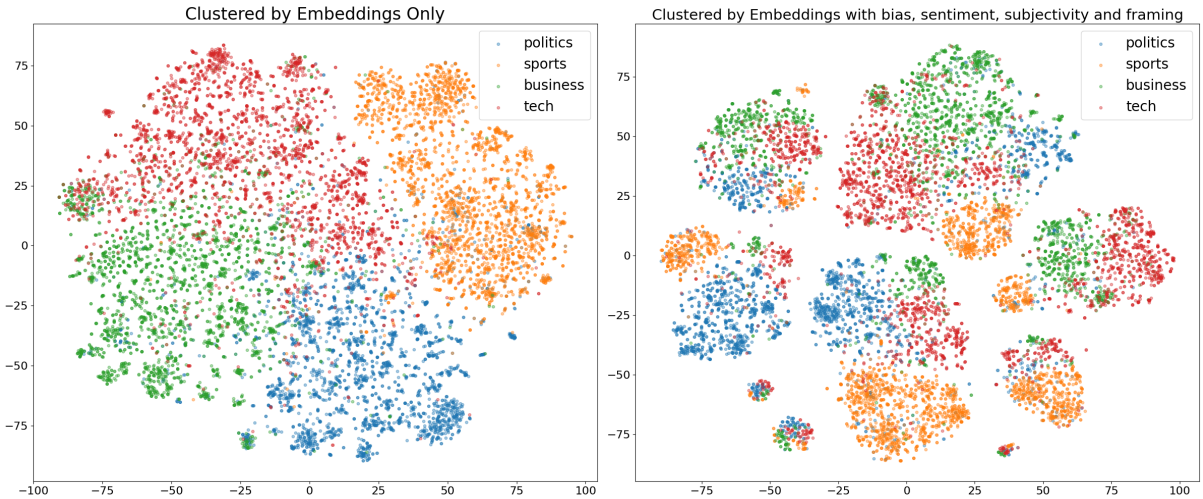

Figure 10: Combined influence on clusters