Topic discovery for news articles

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

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

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

Abstract

This project investigates scalable methods for structured information extraction and nuance-aware analysis of contemporary news text. We introduce a curated dataset of BBC news articles collected from official RSS feeds and propose a transparent, dependency-based pipeline for extracting structured claims as actor–action–object tuples. Unlike supervised approaches, the proposed method operates without labeled training data and prioritizes interpretability, faithfulness to the source text, and robustness across domains. To address the lack of gold-standard annotations, we design a fully automatic evaluation framework that assesses structural validity, faithfulness, coverage, and redundancy at scale.

In addition, we extend the analysis beyond structural claims by constructing an enriched news dataset augmented with sentiment, bias, subjectivity, and framing signals derived from pretrained language models. Using this dataset, we explore topic discovery and representation learning through visualization and ablation studies. Results show that incorporating nuance dimensions yields consistent performance gains in downstream classification tasks, highlighting their potential value for more complex news analysis scenarios. Overall, this work demonstrates how interpretable extraction methods and automated nuance modeling can jointly support scalable, reproducible analysis of real-world news corpora.

1 Introduction

The news media play a central role in shaping public understanding of events, policies, and social dynamics. As the volume of online journalism continues to grow, there is increasing demand for automated methods to extract, structure, and analyze information from news text to support downstream tasks such as fact-checking, verification, bias analysis, and knowledge graph construction. Meeting this demand requires approaches that scale to large corpora while remaining faithful to the linguistic and semantic complexity of real-world reporting.

A core challenge in news analysis lies in representing journalistic content at an appropriate level of abstraction. While document-level labels and summaries are useful for high-level categorization, many applications require finer-grained representations that capture who did what to whom, under which conditions. Structured claims offer a natural intermediate representation, enabling news text to be decomposed into explicit, machine-readable assertions. However, existing claim and argument extraction methods often rely on supervised learning and narrowly annotated datasets, limiting their portability, transparency, and reproducibility across domains.

At the same time, understanding news content extends beyond factual structure alone. Sentiment, bias, subjectivity, and framing play a critical role in how information is presented and interpreted. Recent advances in pretrained language models have enabled automatic inference of such nuance dimensions at scale, yet their integration with structural representations of news content remains underexplored. Bridging this gap offers an opportunity to combine explicit claim-level structure with higher-level narrative and editorial signals.

In this work, we address these challenges

through a two-part study. First, we propose a dependency-based claim extraction pipeline applied to a curated dataset of contemporary BBC news articles. The pipeline is deterministic, interpretable, and does not require labeled training data. To evaluate its outputs at scale, we introduce an automatic evaluation framework based on structural and lexical proxy metrics, enabling reproducible comparison without human annotation. Second, we construct an enriched news dataset augmented with automatically derived sentiment, bias, subjectivity, and framing features, and investigate their utility for topic discovery and classification through visualization and ablation studies.

Together, these contributions demonstrate a practical and extensible approach to large-scale news analysis that balances structural rigor, interpretability, and semantic nuance. By combining dependency-based claim extraction with automated perspective modeling, this project highlights how complementary NLP techniques can support deeper, more reliable analysis of real-world news text.

2 Literature Overview

This work lies at the intersection of news text analysis, information extraction, and structured claim representation. Prior research in these areas has explored a range of datasets and extraction methodologies, from supervised argument mining to rule-based and neural information extraction systems. However, existing approaches exhibit limitations in terms of dataset accessibility, domain specificity, and evaluation scalability.

Several publicly available datasets have been proposed for studying claims, arguments, and factual statements in text. In the context of news media, datasets such as Reuters-21578 (1) have been widely used for topic classification and document clustering. However, these datasets primarily provide document-level labels and do not support fine-grained claim or event extraction. More recent resources for fact-checking and claim verification, such as FEVER (2) and related benchmarks, focus on sentence-level claims paired with evidence, but rely on manually curated or simplified claims rather than naturally occurring journalistic text.

As a result, there remains a gap between richly annotated but narrow datasets and large-scale news corpora that preserve the complexity of real-world reporting while remaining suitable for structured information extraction.

Existing claim extraction approaches can be broadly categorized into supervised neural models and rule-based or hybrid systems. Supervised methods often frame claim or argument extraction as a sequence-labeling or span-prediction task using transformer-based architectures. While these models achieve strong performance when trained on high-quality annotations, their effectiveness depends heavily on domain-specific labeled data. It is difficult to evaluate or reproduce in low-resource settings.

Rule-based and syntactic approaches, by contrast, leverage dependency parsing and linguistic heuristics to identify predicate-argument structures (3). Such methods are more transparent and interpretable, and they can be applied to new domains without retraining. Prior work has shown that dependency-based extraction is effective for event extraction and relational tuple construction, particularly when full semantic correctness is less critical than structural consistency and faithfulness to the source text.

Evaluation remains a major challenge in claim extraction. Many studies rely on human annotation or task-specific metrics, which limit scalability and make iterative system development costly. Proxy metrics and automatic structural checks have therefore been proposed as practical alternatives for large-scale experimentation, particularly when claims are intended as intermediate representations rather than final user-facing outputs.

2.1 Contributions of This Work

This project builds on prior work in syntactic information extraction while addressing several of the limitations identified above. The main contributions are as follows:

- We introduce a curated, continuously extensible dataset of contemporary BBC news articles.
- We propose a transparent, dependency-based claim extraction pipeline that represents claims as structured actor-action-object tuples. The approach prioritizes interpretability and faithfulness to the source text, making it suitable for downstream tasks such as fact-checking, verification, and knowledge graph

construction.

- We design a fully automatic evaluation framework that measures structural quality, faithfulness, coverage, and redundancy without requiring gold-standard annotations. This enables scalable experimentation and reproducible comparison across system variants.

3 BBC Dataset

For this study, we constructed a curated dataset of BBC News Articles derived from official BBC News RSS feeds. The dataset comprises several hundred contemporary news articles collected across multiple topical categories, including World, UK, Business, Politics, Health, Science and Environment, Technology, and General News. Articles were sourced exclusively from verified BBC RSS endpoints to ensure editorial reliability and content authenticity.

Each dataset entry includes the article URL, headline, publication timestamp, topical label, and the fully extracted article text. Publication dates were standardized into Unix timestamps to enable consistent analysis. To ensure textual quality, article content was extracted from the original HTML pages using the Trafilatura library, which removes boilerplate elements such as navigation menus, advertisements, and non-editorial content. Only items identified as genuine news articles—excluding promotional material, applications, podcasts, and duplicate links—were retained.

To prevent redundancy, articles were deduplicated based on their canonical URLs, and the dataset was incrementally updated by appending only newly discovered articles. The final dataset was serialized in JSON format, facilitating reproducibility, extensibility, and straightforward integration with downstream machine learning and natural language processing pipelines.

3.1 EDA

We performed a detailed linguistic analysis using the spaCy NLP library (with the $en\_core\_web\_trf$ transformer model) to extract syntactic, semantic, and named entity information. The analysis included:

- Token-Level Analysis: Each token was examined for its lemma, part-of-speech (POS) tag, morphological features, syntactic dependency, and head word. This provided insight into the grammatical structure and functional role of each word.

- Dependency Parsing: Dependency relations were visualized both in textual form and graphically using Graphviz and displaCy.

- Named Entity Recognition (NER): Entities such as persons, geopolitical locations, and organizations were identified and labeled according to standard ontology (e.g., PERSON, GPE, ORG).

Overall, this EDA provides a comprehensive foundation for understanding the text's grammatical structure, semantic content, and entity relationships, supporting further NLP tasks such as information extraction, knowledge graph construction, and automated summarization.

4 Experimental Procedure, Results, and Analysis

This section describes the experimental setup used to evaluate the proposed structured claim extraction pipeline.

4.1 Solution Plan and Experimental Setting

The overall goal of the experiments is to evaluate the effectiveness and robustness of a dependency-based structured claim extraction pipeline applied to real-world news text. We first detail the extraction algorithm, which is based on syntactic dependency parsing, and then introduce the fully automatic evaluation metrics used to assess extraction quality in the absence of human annotations.

All experiments were conducted on the curated BBC News Articles Dataset described in Section 3. Articles were segmented into sentences using spaCy's sentence boundary detection. Each sentence was independently processed to avoid cross-sentence leakage and to ensure reproducibility.

Linguistic preprocessing was performed using the `en_core_web_trf` spaCy model, which provides transformer-based tokenization, part-of-speech tagging, dependency parsing, and named entity recognition. Default model parameters were used throughout to ensure comparability and to avoid overfitting the pipeline to a specific dataset.

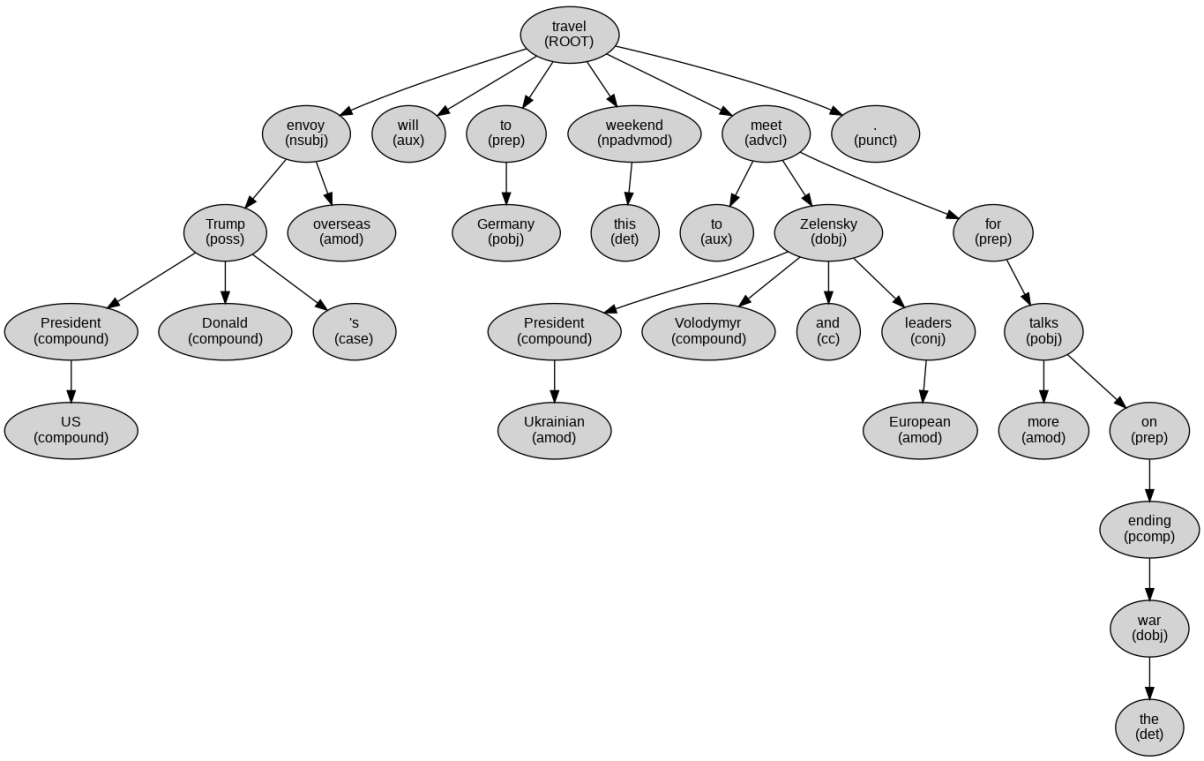

Figure 1: Example of linguistic EDA

The claim extraction algorithm was applied deterministically to each parsed sentence. No supervised learning, fine-tuning, or external knowledge bases were used. This design choice allows the evaluation to focus on the structural capabilities of syntactic extraction rather than learned semantic generalization.

4.2 Claim Extraction Algorithm

Our claim extraction algorithm operates on dependency-parsed text using the spaCy natural language processing library. Given a document, we segment it into sentences and process each sentence independently. For each sentence, we extract zero or more structured claims, where each claim is represented as a tuple:

$$c = \langle a, v, o \rangle$$

where a denotes the *actor*, v the *action*, and o the *object*.

4.2.1 Actor Identification

Actors are identified as the entities responsible for the described event or state. For each sentence, the algorithm first locates the main grammatical subject and then expands it to form a complete actor mention.

To obtain a coherent actor span, the system includes descriptive information attached to the subject, such as modifiers, numerical expressions, and appositive or descriptive phrases. This expansion allows the actor to capture full noun phrases rather than isolated tokens, preserving multi-word entities and descriptive references.

The final actor representation is constructed as an ordered sequence of collected tokens, yielding a natural-language phrase corresponding to the entity or entities participating in the claim.

4.2.2 Action Extraction

Actions represent the central event or relation expressed in the sentence.

To produce an informative action phrase, the algorithm augments the core predicate with auxiliary elements that express tense, modality, or aspect, as well as closely attached particles and prepositional components. This results in an action representation that reflects how the event is presented in the source sentence rather than a simplified verb lemma.

4.2.3 Object Construction

Objects correspond to the target of the extracted action. The algorithm identifies elements that

function as complements of the action and expands them to include their associated descriptive material.

The resulting object representation thus captures both simple and complex outcomes of the action, allowing claims to express events, relations, or embedded propositions in a unified format.

Each extracted claim is finally represented as a structured tuple consisting of an actor, an action, and an object.

4.2.4 Automatic Evaluation Metrics

Because no gold-standard annotations are available, we adopt a fully automatic evaluation framework that relies on proxy metrics to assess extraction quality. This choice is motivated by prior work showing that large-scale manual annotation for claim and argument extraction is expensive, domain-specific, and difficult to maintain across evolving systems (4).

Our evaluation framework is designed to capture multiple complementary aspects of claim extraction quality. Structural metrics assess whether extracted claims conform to a predefined schema, ensuring that outputs are well-formed and linguistically plausible.

All metrics are computed automatically at the sentence and corpus levels and are used primarily for comparative analysis across system versions rather than as absolute measures of semantic correctness. This evaluation setup enables scalable experimentation and reproducible system comparison.

4.2.5 Well-Formedness

Well-formedness measures whether extracted claims contain a non-empty actor and action. We do not include the object because it is naturally missing in many valid sentences. Formally, for a set of extracted claims C, the well-formedness rate is defined as:

$$\text{WF} = \frac{1}{|C|} \sum_{c \in C} \mathbb{I}[a_c \neq \emptyset \wedge v_c \neq \emptyset]$$

where $\mathbb{I}$ is the indicator function.

4.2.6 Schema Validity

Schema validity verifies that each claim adheres to basic linguistic constraints:

- The actor contains at least one noun or pronoun.
- The action contains at least one verb.
- The object exceeds a minimum token length.

The schema validity score is computed analogously to the well-formedness score as the proportion of claims that satisfy all constraints.

4.2.7 Faithfulness to Source

Faithfulness measures how strongly each claim is grounded in its source sentence. For each claim slot $s \in \{a, v, o\}$, we compute token overlap with the sentence:

$$F_s = \frac{|T(s) \cap T(S)|}{|T(s)|}$$

where $T(\cdot)$ denotes the set of lowercase alphanumeric tokens and S is the source sentence. The overall faithfulness of a claim is the mean of the slot-level scores:

$$F(c) = \frac{1}{3} \sum_{s} F_s$$

4.2.8 Coverage

Coverage quantifies how much of the sentence content is captured by the extracted claims:

$$\text{Coverage} = \frac{|T(C) \cap T(S)|}{|T(S)|}$$

where $T(C)$ is the union of tokens across all claim slots in the sentence. This metric penalizes both under-extraction and trivial copying.

4.2.9 Redundancy

Redundancy measures semantic overlap among claims extracted from the same sentence. Each claim is converted into a textual representation and embedded using a sentence transformer. Pairwise cosine similarities are computed, and pairs exceeding a similarity threshold (0.9) are counted as redundant:

$$\text{Redundancy} = \frac{\#\text{redundant claim pairs}}{\#\text{total claim pairs}}$$

4.3 Results

Table 1 reports the results of the proposed evaluation metrics.

The results indicate that the claim extraction pipeline is structurally stable, with a well-formedness rate of 1.00, meaning that all extracted claims contain non-empty actor and action fields.

Metric	Score
Well-Formedness	1.00
Schema Validity	0.45
Average Faithfulness	0.89
Coverage	0.86
Redundancy	0.05
Claims per Sentence	2.04

Table 1: Claim extraction evaluation

This suggests that the extraction process reliably produces complete claim representations and does not fail catastrophically at the schema level.

Faithfulness to the source text is high, with an average score of 0.89, indicating that the extracted claims are strongly grounded in their originating sentences. This suggests that the system rarely introduces hallucinated content and generally preserves lexical alignment with the source text, a desirable property for downstream tasks such as fact-checking and verification.

The coverage score of 0.86 shows that the extracted claims capture a substantial portion of sentence content. While high coverage reflects effective extraction of salient information, it may also indicate a tendency toward verbose claims that closely mirror the original sentence structure. This trade-off is acceptable in the context of claim extraction, where preserving contextual completeness is often preferable to aggressive abstraction.

The schema validity rate of 0.45 reveals a notable limitation of the current approach. Although claims are well-formed, fewer than half fully satisfy the linguistic constraints required for a clean actor–action–object structure. Manual inspection suggests that this is primarily due to underspecified or weak action phrases. This finding highlights action extraction as the main bottleneck of the pipeline.

Redundancy remains low at 0.05, indicating that the system rarely produces semantically duplicate claims from the same sentence. This suggests that the extraction algorithm strikes a reasonable balance between recall and fragmentation, avoiding excessive over-generation.

Finally, the system extracts approximately 2 claims per sentence. This aligns with expectations for news text, where sentences frequently encode multiple events or relations, and suggests that the pipeline can decompose complex sentences into multiple structured assertions.

Compared to supervised claim extraction approaches reported in prior work, the proposed method trades semantic precision for transparency and scalability. While supervised models may achieve higher task-specific accuracy, they require extensive labeled data and are difficult to adapt across domains. In contrast, the proposed pipeline operates without training data, offers interpretable outputs, and enables rapid iteration and error analysis.

From a computational perspective, the dominant cost arises from transformer-based parsing rather than claim extraction itself. Average processing time per article remains within practical limits for offline analysis, and memory consumption scales linearly with document length. This suggests that the approach is suitable for medium- to large-scale news corpora, particularly in research settings where interpretability is prioritized.

Overall, the results demonstrate that the proposed system produces faithful, low-redundancy claims with strong structural consistency. At the same time, the analysis clearly identifies action extraction as the main area for future improvement, motivating extensions such as richer predicate modeling or hybrid syntactic–semantic approaches.s

5 Automatic sentiment, bias, and framing recognition for news datasets

As an additional part of our project, we designed an automated pipeline to identify news topics and assess them across sentiment, biases, factuality, and framing. The resulting dataset is hosted on HuggingFace at mkita/topic-discovery-for-news-articles-test.

5.1 Base dataset

We base our dataset on the AG News dataset, a benchmark for text classification. The dataset comprises 120,000 training and 7,600 testing samples categorized into four distinct 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 (5). The labels in this dataset are treated as ground truths for later work.

5.2 Feature Augmentation

We utilized specialized NLP models to enrich the dataset with four nuance dimensions. The features were extracted as follows:

5.2.1 Text Embedding

Each article is represented using dense vector embeddings to capture its semantic structure. We utilized the all-MiniLM-L6-v2 model provided by the *Sentence-Transformers* framework (6). This model is a distilled version of the MiniLM architecture (7), specifically fine-tuned on a sentence-pair dataset to optimize for semantic-similarity tasks. It maps each article to a 384-dimensional dense vector space.

5.2.2 Sentiment

Sentiment is measured using the twitter-roberta-base-sentiment model, based on the RoBERTa model architecture (8). Each article is assigned one of three sentiment labels:

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

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

5.2.3 Bias Detection

To quantify editorial leanings, we employed the UnBIAS-classifier, which is based on the BERT model architecture (9). This model evaluates the linguistic markers of prejudice within the text, categorizing articles into three classes: *Neutral*, *Slightly Biased*, and *Highly Biased*. These categorical outputs were mapped to numerical values $(0, 0.5, 1)$.

5.2.4 Subjectivity Classification

Distinguishing between objective reporting and interpretative journalism is critical for nuance analysis. We utilized the mdebertav3-subjectivity-multilingual model, which is based on a DeBERTa-based classifier (10), fine-tuned to detect the presence of editorial voice. Articles are classified as either *Objective* or *Subjective*. This binary feature serves as a proxy for identifying opinion-heavy content versus purely factual reporting.

5.2.5 Framing Analysis

Media framing—the process by which certain aspects of a story are emphasized to promote a particular interpretation—is a high-level semantic task. Given the lack of a universal label set for news framing, we adopted a **Zero-Shot NLI** approach using `nli-deberta-v3-small`.

The model treats the article text as a premise and evaluates the entailment probability against a set of frames:

- Corporate & Markets
- Social Impact and Labor
- Neutral/Reporting
- Non-Economic

This method allows for flexible categorization without the need for a specifically labeled framing dataset.

5.3 Exploratory Data Analysis of New Dataset

The dataset has equal proportions in terms of labels:

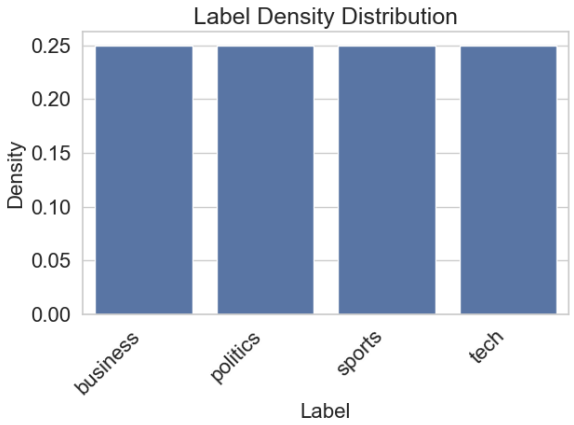

Figure 2: Label distribution

Initial analysis indicates that approximately 55% of the articles exhibit a neutral sentiment. The remaining portion is split between negative and positive sentiments, with negative sentiment having a slightly higher share, suggesting a slight skew toward critical reporting within the dataset.

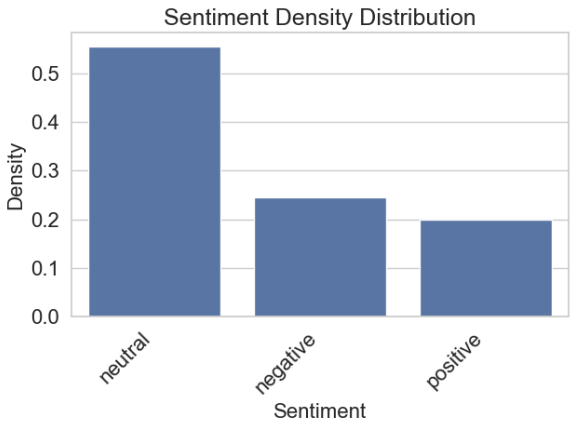

Figure 3: Sentiment labels distribution

Notably, the automated pipeline classified the majority of the articles as biased, as shown in Figure 4. This result highlights a significant challenge in automated nuance detection: the potential for models to be oversensitive to specific linguistic markers. Nevertheless, the labels were maintained without manual intervention to preserve the integrity of the model's output for further evaluation.

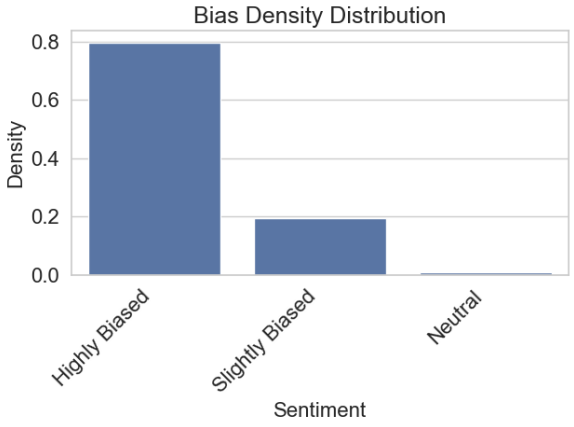

Figure 4: Bias labels distribution

The distribution of frames aligns with the composition of the AG News dataset. As illustrated in Figure 5, a plurality of articles are framed within the context of Corporate & Markets, whereas a significantly smaller portion focuses on Labor and Social Impact. This trend reflects the business and technology focus prevalent in the underlying source.

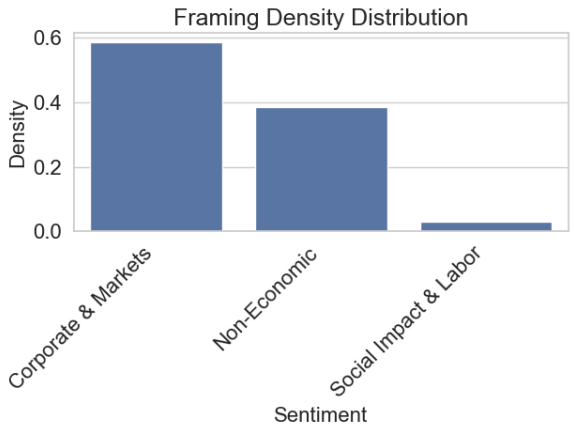

Figure 5: Framing labels distribution

5.4 Topic Discovery Methodology

5.4.1 Feature Construction

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

- $emb_{content}$ – semantic embedding of the article text
- $emb_{sentiment}$ – numerical representation of sentiment
- emb_{bias} – numerical representation of bias
- $emb_{subjectivity}$ – numerical representation of subjectivity
- $emb_{framing}$ – vector of framing probabilities

5.4.2 t-SNE Visualization of Article Clusters

To explore the structure of article embeddings, we applied t-distributed Stochastic Neighbor Embedding to project high-dimensional representations into 2D. Figure 6 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 yield coherent clusters that reflect topical similarity.

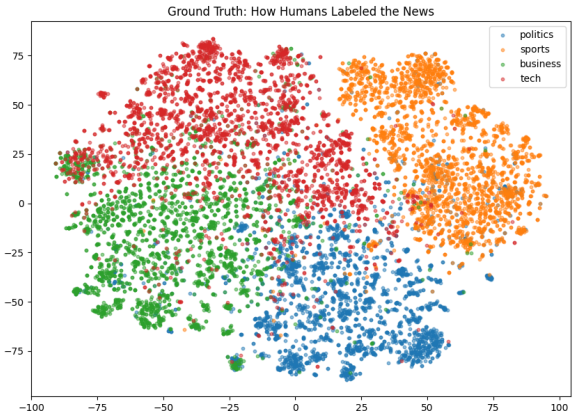

Figure 6: Topic structure derived from semantic similarity

When derived features such as sentiment, subjectivity, bias, and framing are incorporated, the cluster structure becomes more nuanced, highlighting distinctions driven by both narrative perspective and content.

Figure 7: Topic structure derived from nuance dimensions and semantic similarity

Visualization in Figure 7 confirms that augmenting semantic embeddings with additional features can reveal latent patterns that are not apparent from text semantics alone.

5.5 Ablations Studies

To evaluate the empirical utility of the machine-learning generated nuance dimensions, we conducted an ablation study using a Multi-Layer Perceptron classifier. The objective was to determine the marginal contribution of each feature set toward the downstream task of news classification.

5.5.1 Experiment Setup

The MLP was trained using the 384-dimensional dense embeddings as the baseline feature set. We then iteratively introduced normalized feature vectors for each nuance dimension to observe their impact on test accuracy. All models were evaluated on the held-out AG News test set to ensure generalizability. The model was trained on a single Nvidia RTX 3090 GPU.

Table 2: Ablation Study: Impact of Feature Augmentation on Classification Accuracy

Feature Configuration	**Test Accuracy**
emb + Sentiment	**0.913**
emb + Bias	0.898
emb + Subjectivity	0.898
emb + Bias + Subjectivity + Framing	0.898
emb + Subjectivity + Framing	0.896
emb + Bias + Framing	0.895
emb + Bias + Subjectivity	0.894
emb + Sentiment + Framing	0.893
emb + Framing	0.891
Full Feature Vector	0.890
Baseline (emb)	**0.888**

The inclusion of sentiment yielded the most significant performance gain, increasing the baseline accuracy from 0.888 to 0.913. This suggests that the emotional content of a news article correlates with its thematic category. Crucially, integrating these automated nuance dimensions did not degrade model performance relative to the baseline in any configuration.

6 Workload distribution

Table 3: Team Effort Summary

Topic	**Person**	**Time Spent**
Claim Extraction	J. Bazyluk	10h
Literature Analysis	J. Bazyluk	4h
Literature Analysis	M. Kita	3h
Claim Extraction Idea	J. Bazyluk	5h
Original Dataset Creation	M. Kita	8h
BBC Dataset Preparation	J. Bazyluk	4h
Ablation Studies	M. Kita	5h
EDA	J. Bazyluk	1h
Report	J. Bazyluk	10h
Report	M. Kita	7h

7 Rebuttal

- Overall literature review is sparse and lacks more detailed presentation.
  The literature review has been expanded to provide a more detailed discussion of relevant datasets, extraction methodologies, and evaluation approaches, and to position the contributions of this work within existing research clearly.

- You identified that the main weakness of your extraction system is the lack of action informativeness. Are you considering integrating generative approaches to summarize or rephrase predicates to make the action tuple more descriptive?
  This is an interesting and promising direction; within the scope of the current work, we attempted to mitigate action informativeness through rule-based refinement, while integrating generative predicate summarization is left as future work.

- The evaluation metrics should be explained (formula, word explanation, etc.), e.g., ”topic coherence”, ”factuality”, ”topic stability”. The claim ”recent advancements have shifted toward embedding-based methods like BERTopic and Top2Vec” is relatively outdated given the advances in LLMs.
  The evaluation metrics will be clarified by providing formal definitions and explanations. Additionally, the statement regarding recent advances toward embedding-based methods (e.g., BERTopic, Top2Vec) has been removed, as it is no longer representative of the current state of the field given recent LLM-based developments. This direction was ultimately not pursued and has therefore been deleted from the project.

- The dataset is not clearly defined. FEVER is mentioned, but it is unclear whether it will be used. It is stated only in the evaluation that it will be used.
  We provide a clear and comprehensive description of datasets.

8 Conclusions

This work presented a scalable and interpretable framework for automated news analysis that combines structured claim extraction with nuance-aware content modeling. We introduced a curated dataset of contemporary BBC news articles and proposed a deterministic, dependency-based pipeline for extracting actor–action–object claims from real-world journalistic text. The pipeline operates without labeled data and emphasizes transparency and faithfulness to the source, making it suitable for reproducible large-scale analysis.

To evaluate extraction quality in the absence of gold-standard annotations, we designed a fully automatic evaluation framework capturing structural validity, faithfulness, coverage, and redundancy. Results show that the system reliably produces well-formed, highly faithful, and low-redundancy claims and can decompose complex sentences into multiple assertions. However, the analysis also identifies a key limitation: the extracted action components are often underspecified, resulting in reduced schema validity despite otherwise strong structural performance.

In addition, we constructed an enriched news dataset augmented with sentiment, bias, subjectivity, and framing signals derived from pretrained models. While performance gains on a standard topic classification benchmark such as AG News are modest, the consistency of these features across ablation settings is notable. This stability suggests that such nuance dimensions capture complementary information that may provide greater value in more complex downstream tasks, particularly where thematic content alone is insufficient and expert annotation is costly.

9 Future Work

The primary direction for future work is improving the informativeness of the extracted action component. In the current system, action phrases are refined using rule-based syntactic heuristics, which preserve faithfulness but often lack semantic richness. While limited rule-based refinements were explored within the scope of this project, more expressive action representations remain an open challenge.

A promising extension is the integration of generative approaches to summarize or rephrase predicates into more descriptive action phrases, conditioned on the source sentence and extracted arguments. A hybrid pipeline combining deterministic actor and object extraction with generative predicate modeling could improve semantic clarity while maintaining interpretability. Integrating

such generative predicate summarization is therefore left as future work.

Beyond claim extraction, future research could further explore the role of sentiment, bias, subjectivity, and framing features in downstream tasks such as claim verification, stance detection, or narrative analysis, where stable nuance dimensions are likely to offer higher marginal utility. Additional directions include extending the pipeline to support cross-sentence relations, incorporating multilingual data, and conducting limited human-in-the-loop evaluation to calibrate automatic metrics better.

Overall, this work lays the groundwork for interpretable, scalable news analysis while highlighting clear opportunities for enhancing semantic expressiveness through hybrid syntactic–generative approaches.

References

[1] D. Lewis, "Reuters-21578 Text Categorization Collection," UCI Machine Learning Repository, 1987. [Online]. Available: https://doi.org/10.24432/C52G6M

[2] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," in *NAACL-HLT*, 2018.

[3] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Commun. ACM*, vol. 51, no. 12, p. 68–74, Dec. 2008. [Online]. Available: https://doi.org/10.1145/1409360.1409378

[4] M. Lippi and P. Torroni, "Argumentation mining: State of the art and emerging trends," *ACM Trans. Internet Technol.*, vol. 16, no. 2, Mar. 2016. [Online]. Available: https://doi.org/10.1145/2850417

[5] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush, and T. Wolf, "Datasets: A community library for natural language processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 175–184. [Online]. Available: https://aclanthology.org/2021.emnlp-demo.21

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[7] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: deep self-attention

distillation for task-agnostic compression of pre-trained transformers," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198953378

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[10] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=m_filter59Zy