Narrative Structure Extraction in Disinformation and Trustworthy News: A Comparison of LLM, KG, and KG-Augmented Pipelines

Justina Mandravickaitė
Vytautas Magnus University, Lithuania
justina.mandravickaite@vdu.lt

Abstract

Identifying narrative structures such as characters, events, causes, and frames in political news is essential to detecting bias and understanding political dynamics, among other areas. Large language models (LLMs), while performing well on a variety of natural language processing (NLP) tasks, may hallucinate, whereas pure knowledge graph (KG) methods, while excelling at text structuring and information extraction, suffer from sparsity. In this study, three pipelines for extracting narrative structures from disinformation and trustworthy news are evaluated: (1) LLM-only prompt-based extraction, (2) KG-only graph-based queries and (3) an augmented approach combining LLM prompts augmented with appropriate KG data. The results were evaluated intrinsically as well as extrinsically. For intrinsic evaluation, element coverage, fuzzy overlap, coherence, coverage gain and error reduction were measured, while extrinsic evaluation used matching with Wikidata and downstream classification. The augmented pipeline improved the coverage and coherence of narrative elements, but also boosted the classification of disinformation, as it outperformed both LLM-only and KG-only pipelines.

1 Introduction

Narrative extraction from political news is important for revealing how disinformation manipulates the stories by, e.g., assigning blame (Rauchfleisch and Jungherr, 2024), inflating or downplaying events (Keith Norambuena et al., 2023) or constructing false causal chains (Lei and Huang, 2023) in order to shape public perception. Identification of narrative structures aids in revealing rhetorical strategies that fact-checking alone may miss (Byram, 2022).

Large language models (LLMs), while perform well in diverse NLP tasks, including information extraction, may hallucinate relations or entities (Mckenna et al., 2023; Li et al., 2024a). Meanwhile, knowledge graphs (KGs) can suffer from incomplete coverage (Wang et al., 2022) or outdated ontologies (Zhou et al., 2024; Hegde et al., 2025), among other issues, which may result in, e.g., missing events (Zhuang et al., 2023) or misclassified roles (Lu et al., 2024). While many pipelines are assessed using intrinsic metrics (e.g., coherence (German et al., 2025) or reconstruction accuracy (Keith Norambuena et al., 2023)) or individual extrinsic tasks (e.g., downstream classification (Das et al., 2024)), there is an opportunity to develop more comprehensive validation protocols that combine both approaches to provide a more comprehensive evaluation.

To address these issues, in this study, three pipelines for extracting narrative structures from disinformation and trustworthy news are evaluated: (1) LLM-only prompt-based extraction, (2) KG-only graph-based queries and (3) an augmented approach combining LLM prompts augmented with appropriate KG data. The proposed evaluation framework combines intrinsic and extrinsic evaluation. It integrates measures of narrative coverage, fuzzy overlap, coherence, coverage gain and error reduction (intrinsic), alongside downstream tasks such as matching with Wikidata and disinformation classification (extrinsic).

2 Related Work

Advances in narrative extraction for political and social science have relied on structured representations such as event schemas to analyze social processes, events and framing (Zhang et al., 2019; Halterman, 2020; Li et al., 2024b). Transformer-based multilingual models have achieved high accuracy in crisis event detection across languages, illustrating usefulness of schema-guided extraction (Hettiarachchi et al., 2021).

Schema refinement has emerged for extracting

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 86–103, 2025

events without predefined templates, such as Liberal Event Extraction (LEE) which jointly extracts events and induces schemas (Li and Geng, 2024). Integrating schemas with cultural norms and commonsense knowledge was used to support framing analysis (Li et al., 2024b). Visualization tools have been used to reveal how discrete events affect outcomes over time (Shen et al., 2024).

Also, graph-based methods have been proposed, such as using AMR (abstract meaning representation) to extract actors, events, and perspectives from digital media(Pournaki and Willaert, 2024). Knowledge Graph–based approaches enhance semantic precision and reveal causal relationships in narrative analysis (de Kok et al., 2024; Romanou et al., 2023). In bias detection, integrating frame-based knowledge with text models improved bias and stance detection (Li, 2021). Also, domain-specific KGs for news facilitate advanced bias detection and narrative synthesis (Yang et al., 2024).

Event-centric narrative extraction has been a trend for a while (Voskarides et al., 2021; Keith Norambuena et al., 2023). Advances and challenges in multimodal event extraction has been introduced as well (Hürriyetoğlu et al., 2024).

Recent research revealed that hybrid methods mitigate weaknesses of purely symbolic (such as KG-based) or neural approaches (such as LLM-based) (Panchendrarajan and Zubiaga, 2024; Zhu, 2024). For example, multi-agent approaches has been recommended for KG construction and reasoning, (Zhu et al., 2024). Also, at least several LLM–KG integration frameworks have been outlined, such as SymAgent, a neural-symbolic agent for multi-step reasoning and automatic KG updates (Liu et al., 2025)or MindMap which merges KGs and LLMs for improved inference transparency (Wen et al., 2024).

Different proposed directions also involve merging LLMs with relation extraction to build semantically rich KGs (Santini, 2024) and optimize LLM reasoning over KGs via selective triple selection (Wang, 2025). Also, recent developments in processing long narratives address past limitations by integrating dynamic KGs with LLMs to improve story comprehension (Andrus et al., 2022).

Narrative extraction evaluation includes variety of metrics and approaches,such as intrinsic metrics for event and attribution extraction, stressing reproducible practices (Zhang et al., 2019). Also, the complexity of assessing narrative coherence has been highlighted (Santana et al., 2023). Furthermore, studies have evaluated narrative elements via Multiple Choice Narrative Cloze (Hatzel and Biemann, 2023), accuracy metrics and downstream performance (Tang et al., 2021), structural coherence checks and user studies (Norambuena and Mitra, 2021).

This paper reports comparative evaluation of LLM-only, KG-only and Augmented pipelines for extracting narrative structures in political news, assessed with both intrinsic metrics and extrinsic grounding. Unlike typical KG–LLM work, in this paper narrative extraction gains were directly linked to improved disinformation classification while maintaining interpretable representations.

3 Methods

3.1 Data

For this study a part of the dataset for multilingual detection of pro-Kremlin disinformation in news articles (Leite et al., 2024), containing data on disinformation and trustworthy news articles, was used. This dataset consists of 18 249 articles in 42 languages published January 2015-July 2023. As the full text of the news articles was not publicly available, to reconstruct the dataset URL links of the articles were used with DiffBot API[1] (free for academic purposes) to acquire them. Only English-language articles were selected, i.e. a total of 6,546 articles (425 – disinformation and 6121 – trustworthy). Some articles were no longer available or have been modified. So, after filtering and cleaning the final dataset used in this study was made of 308 disinformation news articles and 302 – trustworthy news articles. Labeling news articles as 'disinformation' and 'trustworthy' is based on the original dataset, i.e., article labels were reused as they were assigned in the original dataset (Leite et al., 2024). More details of the final dataset are presented in Appendix B.

While the data is of modest size, the experimentation targets pipeline comparison under controlled conditions, not training a large downstream model. Although the final dataset is $\sim$ 300 articles per class, intrinsic evaluations operate per article and summarize many extracted elements. Also, extrinsic evaluation analyses use simple, regularized classifier (logistic-regression) with grouped, repeated cross-validation to reduce variance.

[1]Accessible at https://www.diffbot.com/

3.2 Pipelines

Based on existing narratology and framing research, the following narratives structures were extracted and compared in three pipelines (LLM-only, KG-only and Augmented): characters (entities – persons and organizations), events (predicates indicating what happens to/among characters over time), causal links (predicates that express cause/effect), framing (predicates that indicate attribution/association) (Hellman, 2024; Heddaya et al., 2024; Wang et al., 2025).

For this study, the first step, before applying any of the pipelines, was text summarization. This step was applied to reduce noise and to use computational resources more efficiently in the later stages. Extractive summarization was used for this task as it extracts the most important sentences from the given text, preserving factuality and original wording (Hofmann-Coyle et al., 2022). Article summaries were used in all three pipelines that were comprised of different steps.

Summarization. The cleaned news articles were summarized using two extractive summarization approaches – LexRank Summarizer (Erkan and Radev, 2004) and BERT Extractive Summarizer (Miller, 2019). To choose from two summary-candidates, mean ROUGE-1/2/L F1 scores (Lin, 2004) were computed against the original article text. ROUGE-1/2/L F1 score favors coverage of source content and discourages off-source material and was used as a heuristic to choose between two summary-candidates. Therefore, summaries generated by the approach that had higher ROUGE values, showing higher source-overlap, were used for the experiments.

The length of the extractive summary was tied to text length: texts under 300 words were summarized in two sentences; texts of 300–799 words long – in four sentences and texts of 800 words or longer – in six sentences. To ensure quality, summaries were also manually inspected, correcting such issues as occasional incomplete final sentences. These summaries were used with all three narrative extractive pipelines. The examples of the summaries are presented in Appendix A.

LLM-only pipeline. Characters, events, causal links and framing were extracted via prompting. For this task, Mistral Small 3.1 24b[2] was used. The decision was made to use a smaller model, not the largest and most popular one like OpenAI's ChatGPT to manage computational and financial costs while maintaining competitive performance. The prompt used for extraction is presented in Appendix C.

KG-only pipeline. This pipeline consisted of several components, introduced below.

1. *Relational triple extraction.* The pre-trained REBEL (Relation Extraction By End-to-end Language generation) (Cabot and Navigli, 2021) model was applied for this task, extracting entity-relation-entity triples from summaries of news articles. These triples were used for graph construction. REBEL combines Named Entity Recognition (NER) and Relation Classification (RC) into a single task and covers >200 relationship types. Extracted triples were aggregated per article and manually validated.

2. *Mapping and inferring relations.* Relations of REBEL schema that imply framing, events and causality were mapped and inferred manually (for details, see Appendix D). This enrichment was necessary to make outputs of all there pipelines comparable.

3. *Building RDF graphs.* In this step, one rdflib graph per article (i.e. from only that article's triples) was built. In this graph structure entities (nodes) of the triples were linked by their relations (edges). Python library rdflib[3] was used for this task.

4. *Querying narrative elements.* REBEL triples (entity-relation-entity) were grouped into four relation families that were used to compute structure-level features: PERSON_RELS (relations that assert roles, affiliations or actor–actor ties), EVENT_RELS (relations that indicate actions/occurrences or changes; also used to collect event nodes (instance_of = Event)), CAUSAL_RELS (explicit cause/effect predicates and a small set of implied causal links as a result from mapping and inferring relations in previous step), FRAMING_RELS) (attribution/stance and portrayal; REBEL labels were mapped to short paraphrases (e.g., described_as → "is portrayed as"), see Appendix D). These narra-

[2]Accessible at https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503

[3]Accessible at https://github.com/RDFLib/rdflib

tive element types were extracted via querying over RDF graphs.

Augmented pipeline. For extraction of narrative elements (characters, events, causal links and framing) from article summaries, prompts were augmented with relevant data from article-level KGs to provide context. The prompt used for extraction of narrative elements with Augmented pipeline is presented in Appendix C.

3.3 Evaluation

The evaluation framework of this study combines intrinsic as well as extrinsic evaluation to get a more comprehensive assessment of extraction of narrative structures. The special focus in this evaluation is paid to exploring the use of an LLM and KG in combination. Also, formal paired significance tests were applied only for metrics that are directly comparable across pipelines across pipelines and quantify extraction quality rather than volume, i.e., overlap (fuzzy Jaccard), coverage gain and error reduction, and extrinsic F1 in the classification. For other quantities, descriptive statistics are reported, avoiding inferential claims.

3.3.1 Intrinsic Evaluation

Intrinsic evaluation of outputs of all three pipelines consisted of the following components:

Coverage: raw counts as well as average counts per element were calculated. This measures recall potential by element type and is simple, interpretable, surfaces systematic under-extraction.

Fuzzy-Jaccard overlap: paraphrase-aware intersection / union between pipelines in terms of extracted narrative elements was calculated (Cross et al., 2020). For this task, *multi-qa-MiniLM-L6-cos-v1*[4] model and cosine similarity (Gunawan et al., 2018) with threshold 0.7 were used. Fuzzy Jaccard overlap shows cross-pipeline agreement on which elements were extracted, while tolerating paraphrases. This measure is scale-free, directly comparable and allows for semantic matching that avoids penalizing lexical variation.

Mean coherence: embedding-based sentence similarity was calculated for LLM-only and Augmented pipelines. Items of the lists of extracted narrative elements were treated as standalone "sentences." For this task, *all-MiniLM-L6-v2*[5] model was used and average adjacent-sentence cosine similarity was computed between items in the lists of narrative elements extracted from the same summaries of the news articles. For KG-only pipeline, graph-based (structural) coherence metrics (density (Shang and Huang, 2024), avg. path length and largest component ratio) were calculated for the graphs, created from relational triples, belonging to each article in the dataset. Semantic (embedding-based) coherence may show internal thematic consistency of extracted elements. Meanwhile, structural (graph-based) coherence may reveal narrative connectedness or fragmentation in the triple graph.

Orthogonal metrics (coverage gain & error reduction): these metrics were used for LLM-only and Augmented pipelines to compare and evaluate coverage of narrative elements and find unmatched elements, which may indicate model hallucination. These metrics were calculated using the following formulas:

Coverage gain for an element type = | Augmented pipeline | – | LLM-only pipeline |

Error reduction = | LLM-only pipeline unmatched elements | – | Augmented pipeline unmatched elements

Orthogonal metrics were calculated per element type. Coverage gain is the corpus-level difference in total extracted items. Unmatched elements are items produced by a pipeline that have no fuzzy match in the other pipeline. Error reduction refers to the decrease in unmatched elements when switching from LLM-only to Augmented pipeline. A ratio > 1 of Coverage gain and Error reduction means that Augmented pipeline adds more new correct elements than LLM-only pipeline removes. Meanwhile, a ratio < 1 means Augmented pipeline introduces more "errors" (unmatched elements) than additional coverage, indicating added noise. Orthogonal metrics capture the augmentation trade-off between newly matched and unmatched elements. In other words, these metrics directly quantify, per element type, whether augmentation was justified.

3.3.2 Extrinsic Evaluation

Loose matching with Wikidata: For a grounding / checking consistency, narrative elements, extracted with all three pipelines, were matched with loosely Wikidata. Linking extracted elements known knowledge base tests whether outputs correspond to real entities, reducing hallucinations and improving interpretability and Wikidata is widely

[4]Accessible at https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1

[5]Accessible at https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

used for such tasks (Möller et al., 2022). A stratified sample of 300 elements from each pipeline was taken. The unique labels were used to query against Wikidata[6] to determine if they matched real-world entities. For this task, Wikidata Search API (based on the MediaWiki API)[7] was used to perform a relevance-based search (or loose matching), i.e., matches were based on labels, aliases, redirects, and included some tolerance for minor typo. The results were ranked by relevance.

Downstream classification: Logistic-regression classifier was built for this task (see Appendix E for details). Counts of narrative elements and coherence (for LLM-only and Augmented pipelines – embedding-based metric and for KG-only pipeline – graph-based metrics) were used as features. For evaluation, 5-fold and 10-fold cross-validation(Wong and Yeh, 2019) was performed and F1-score (Naidu et al., 2023) was used to assess the results. Classification was used as an extrinsic evaluation to assess which narrative extraction pipeline more effectively captures the signal needed to distinguish disinformation from trustworthy news.

Paired tests (paired t-test (Rainio et al., 2024) and Wilcoxon signed-rank test (Rey and Neuhäuser, 2011)) were used to evaluate the statistical significance of the results by performing pairwise comparisons on the pipelines.

Cohen's d (Goulet-Pelletier and Cousineau, 2018) was calculated to measure the effect size, i.e. how big is the difference in classification performance by LLM-only, KG-only and Augmented pipelines when distinguishing disinformation from trustworthy news.

4 Results

4.1 Intrinsic Quality

4.1.1 Raw Counts and Mean Coverage of Narrative Elements

Table 1 shows descriptive counts of extracted narrative elements per pipeline. These raw counts summarize extraction tendencies and are not subjected to formal significance testing. As Table 1 shows, KG-only pipeline lags behind on every element, reflecting that raw triple extraction alone captures only a small part of narrative content.

Meanwhile, Augmented pipeline slightly outperformed LLM-only pipeline on characters and events for both classes. As for causal links, slightly higher counts resulted from Augmented pipeline for disinformation. However, for trustworthy news LLM-only pipeline provided higher counts, though the difference, in comparison to Augmented pipeline, is very small. Meanwhile, for framing more elements were extracted with LLM-only pipeline for both disinformation and trustworthy news.

Trustworthy articles generally provided higher counts and means than disinformation across all pipelines. This could be likely because longer, more structured reporting offers richer narratives.

The augmented pipeline mostly maintained LLM-only breadth of coverage while adding a modest boost, such as on actors and events, over LLM-only. KG alone was too sparse for standalone narrative extraction.

4.1.2 Overlap of Narrative Elements by Pipeline

To assess overlap of extracted narrative elements between pipeline pairs on the same article and element type per class, fuzzy Jaccard was used. As presented in Table 2, there almost no overlap for **llm∩kg**, meaning each alone extracts mostly disjoint sets of elements. Meanwhile, **llm∩augmented** overlap more substantial (0.46–0.47 for characters, 0.13–0.18 for events, up to 0.15 for framing). This shows that Augmented pipeline not only retained most LLM-only elements but also enriched them with additional ones from KG-only.

However, **kg∩augmented** overlap was low, though not so low as for **llm∩kg**, which means that Augmented pipeline's additional KG-based data was a small part of its output.

4.1.3 Coherence Assessment

To assess semantic continuity (for LLM-only and Augmented pipelines) and structural cohesion (for KG-only pipeline) within output for the same article, mean coherence measures were calculated. Embedding-based (semantic) coherence was slightly higher in Augmented pipeline than LLM-only one for disinformation (0.411 vs. 0.400) and there was no difference in coherence in terms of trustworthy news (Table 3).

Graph-based (structural) coherence in the KG-only pipeline showed that, in disinformation arti-

[6]Accessible at https://www.wikidata.org/wiki/Wikidata:Main_Page

[7]Accessible at https://www.wikidata.org/w/api.php?action=help

Element type	Pipeline	Disinformation	Trustworthy
Characters / actors	LLM-only	2226 (7.25)	2453 (8.10)
	KG-only	1147 (3.74)	1228 (4.05)
	Augmented	**2377 (7.74**	**2541 (8.39)**
Events	LLM-only	1826 (5.95)	2043 (6.74)
	KG-only	244 (0.79)	291 (0.96)
	Augmented	**1849 (6.02)**	**2118 (6.99)**
Causal links	LLM-only	1287 (4.19)	**1362 (4.50)**
	KG-only	71 (0.23)	77 (0.25)
	Augmented	**1298 (4.23)**	1355 (4.47)
Framing (actors+events)	LLM-only	**2401 (7.82)**	**2462 (8.13)**
	KG-only	106 (0.35)	158 (0.52)
	Augmented	2227 (7.25)	2274 (7.50)

Table 1: Raw counts and means (in brackets)

Element	llm∩kg (D/T)	llm∩augmented (D/T)	kg∩augmented (D/T)
Characters/actors	0.14 / 0.13	**0.47 / 0.46**	0.19 / 0.19
Events	0.008 / 0.006	**0.18 / 0.13**	0.013 / 0.014
Causal links	0 / 0	**0.076 / 0.058**	0.002 / 0.002
Framing	0.002 / 0.001	**0.15 / 0.12**	0.001 / 0.003

Table 2: Mean Fuzzy Jaccard (D – disinformation, T – trustworthy news)

cles, nearly 43 % of nodes belonged to the largest connected component, compared to 38 % in trustworthy articles. The average shortest path length was approximately 1.2 in both cases, and graph density was higher for disinformation (0.23 vs. 0.20). These results suggested that narrative elements extracted from disinformation articles were more interconnected. The similar average path lengths indicate that, despite differences in overall connectedness, the "distance between ideas" was similar across both categories.

The augmented pipeline's modestly higher coherence suggested that incorporating KG-derived facts helped to make the narrative elements semantically tighter, even if the gain was small. In contrast, KG-only graphs presented measurable structural coherence, though their metrics were not directly comparable to embedding-based measures.

4.1.4 Coverage Gain and Error Reduction

Orthogonal metrics (coverage gain and error reduction) were computed only for the LLM-only and Augmented pipelines. The KG-only pipeline was excluded because its extraction process was based on querying structured relational graphs which is very different from the prompt-based extraction in the LLM-only and Augmented pipelines.

In terms for disinformation, for **actors/characters** and **events**, the Augmented pipeline gives a modest but consistent net boost in coverage (2–7 %) with minimal noise (Table 4).

For **causal links**, the Augmented pipeline captured just as many or very slightly more matched causal links as it introduces unmatched ones.

For **framing edges**, the KG-augmented prompts added quantity but at the cost of introducing more unmatched framing elements, suggesting the prompting strategy for framing relations may need further refinement (e.g. more precise examples or stricter filtering).

Regarding trustworthy news, Augmented pipeline also provided slightly more new **actors/characters** and **events** than mismatches removed (Table 5).

For causal links, augmentation slightly favored mismatch reduction over added coverage. Finally, for framing edges, augmentation primarily improved precision in reducing mismatches than in-

Pipeline	Metric	Disinformation	Trustworthy
LLM-only	Embedding-based coherence	0.400	0.396
Augmented	Embedding-based coherence	**0.411**	**0.396**
KG-only	Largest component ratio	0.427	0.382
	Avg. path length	1.222	1.244
	Density	0.228	0.195

Table 3: Mean coherence by pipeline

Element	Coverage Gain	Error Reduction	Gain / Reduction
Characters/actors	1151	1079	1.07
Events	1736	1703	1.02
Causal links	1233	1226	1.01
Framing edges	2291	2460	0.93

Table 4: LLM-only vs. Augmented pipeline coverage gain & error reduction: disinformation news

Element	Coverage Gain	Error Reduction	Gain / Reduction
Characters/actors	1289	1252	1.03
Events	1996	1933	1.03
Causal links	1277	1297	0.98
Framing edges	2306	2567	0.90

Table 5: LLM-only vs. Augmented pipeline coverage gain & error reduction: trustworthy news

creasing in new elements. This, again, suggests that the prompting strategy for framing relations may need further tuning.

Overall, these intrinsic evaluations supported the assumption that Augmented pipeline outperforms pure LLM-based extraction on core narrative elements, especially actors and events. Meanwhile, framing may need a more targeted prompt design or post-extraction filtering.

4.2 Extrinsic Quality

4.2.1 Matching with Wikidata

KG-only extractions of narrative elements aligned with Wikidata at a very high rate ($\sim$ 70 %), reflecting that REBEL triples largely take out canonical entities and relations that already exist in Wikidata (Table 6). Meanwhile, LLM-only was far less grounded ($\sim$ 15–20 %), since free-form prompts often generate paraphrases, alternate names, or relations that are not directly findable in Wikidata. Finally, Augmented pipeline stayed in between the aforementioned pipelines, reflecting LLM's grounding by including the KG triples in the output.

As for matching per element type, Augmented pipeline outperformed LLM-only in terms of characters/actors but still trailed behind KG-only as $\sim 60\%$ of sampled characters/actors extracted by this pipeline matched Wikidata entities (Table 7).

However, all pipelines struggled in grounding events in Wikidata, though again Augmented pipeline had a higher match rate than LLM-only and KG-only had more matches than both other pipelines.

Finally, causal links and framing had very low Wikidata coverage overall ($\leq 3\%$), reflecting that these relations are rarely modeled as explicit triples in Wikidata. KG-only did not cover these elements at all (Table 7).

4.2.2 Downstream Classification

Document classification was used as an extrinsic evaluation to examine whether pipeline-derived narrative features, rather than surface text, may help distinguish between disinformation and trustworthy news.

Augmented pipeline achieved the highest mean F1-score in both 5- and 10-fold cross-validation, 0.666 and 0.657, accordingly (Table 8). The improvement is moderate in effect size (d $\sim$ 0.6–0.8)

Pipeline	Trustworthy	Disinformation
LLM-only	44 / 300 (14.7 %)	61 / 300 (20.3 %)
Augmented	71 / 300 (23.7 %)	77 / 300 (25.7 %)
KG-only	211 / 300 (70.3 %)	210 / 300 (70.0 %)

Table 6: Overall Wikidata matches

Element type	Pipeline	Trustworthy (out of sample)	Disinfo (out of sample)
Characters/actors	LLM-only	31 / 300 (10.3 %)	47 / 300 (15.7 %)
	Augmented	60 / 300 (20.0 %)	51 / 300 (17.0 %)
	KG-only	178 / 300 (59.3 %)	176 / 300 (58.7 %)
Events	LLM-only	4 / 300 (1.3 %)	6 / 300 (2.0 %)
	Augmented	6 / 300 (2.0 %)	14 / 300 (4.7 %)
	KG-only	33 / 300 (11.0 %)	34 / 300 (11.3 %)
Causal links	LLM-only	7 / 300 (2.3 %)	4 / 300 (1.3 %)
	Augmented	5 / 300 (1.7 %)	9 / 300 (3.0 %)
	KG-only	—	—
Framing	LLM-only	—	4 / 300 (1.3 %)
	Augmented	—	3 / 300 (1.0 %)
	KG-only	—	—

Table 7: Wikidata maches by element type

(Table 11 and shows a clear trend toward statistical significance in the more stable 10-fold cross-validation ($p \approx 0.08$) as showed in Table 10.

Furthermore, Augmented pipeline significantly outperformed KG-only in both cross-validations (t-test $p < 0.05$ for both splits; Wilcoxon significant in 10-fold cross-validation), as presented in Tables 9 and 10, and large effect sizes ($d \sim 0.9$–1.9) as showed in Table 11.

Finally, LLM-only vs. KG-only did not show significant statistical difference in either 5- or 10-fold cross-validation ($p > 0.05$ for all tests) and the effect size was small.

5 Conclusions

The Augmented pipeline showed intrinsic gains in narrative-element coverage and coherence, which also were translated to meaningful downstream improvements in disinformation classification as Augmented pipeline outperformed LLM-only pipeline by a moderate margin and KG-only pipeline – by a large margin, measured in effect size.

Augmented pipeline matched or slightly exceeded LLM-only on characters, events, and causal links while KG-only lagged behind. Also, Augmented pipeline retained most all LLM-only extractions for characters/actors and added KG-informed ones, whereas LLM-only and KG-only barely overlapped.

Moreover, Augmented pipeline's embedding-based coherence was marginally higher than LLM-only, suggesting KG-based augmentation tightened the narrative. Furthermore, Augmented pipeline gained a boost in valid characters and events beyond LLM-only, with minimal noise. Causal links, on the other hand, were similar in quantity, while framing was in need of refinement.

In addition, Augmented pipeline had higher match rate to Wikidata than LLM-only, showing a balance between creative extraction and factual grounding, though lagging behind KG-only due to its schema closely matching Wikidata.

Finally, Augmented pipeline outperformed the other two pipelines in downstream classification – achieved 0.666 ± 0.022 (5-fold cross-validation) and 0.657 ± 0.032 (10-fold cross-validation), versus 0.619/0.607 for LLM-only and 0.591/0.595 for KG-only. In statistical tests Augmented pipeline significantly outperformed KG-only pipeline ($p < 0.05$) and showed a moderate effect size over LLM-only pipeline (Cohen's $d \sim 0.6$–0.8), with a trend toward significance in 10-fold cross-validation ($p \approx 0.08$).

Future work includes experimentation with both smaller- and larger-scale open-source and propri-

Validation	LLM-only	KG-only	Augmented
5-fold	0.619 ± 0.060	0.591 ± 0.025	**0.666 ± 0.02**2
10-fold	0.607 ± 0.082	0.595 ± 0.061	**0.657 ± 0.032**

Table 8: Mean F1-scores)

Pipeline pairs	**Paired t-test**	**Wilcoxon signed-rank**
LLM-only vs KG-only	t = 1.358, p = 0.246	W = 2.000, p = 0.188
LLM-only vs Augmented	t = -1.708, p = 0.163	W = 3.000, p = 0.312
KG-only vs Augmented	t = -4.259, p = 0.013	W = 0.000, p = 0.062

Table 9: Significance testing (5-fold validation)

Pipeline pairs	**Paired t-test**	**Wilcoxon signed-rank**
LLM-only vs KG-only	t = 0.330, p = 0.749	W = 25.000, p = 0.846
LLM-only vs Augmented	t = -1.986, p = 0.078	W = 10.0, p = 0.084
KG-only vs Augmented	t = -2.799, p = 0.021	W = 3.0, p = 0.020

Table 10: Significance testing (10-fold validation)

Comparison	**5-fold cross-validation**	**10-fold cross-validation**
LLM-only vs. KG-only	+0.61 (moderate)	+0.10 (small)
LLM-only vs. Augmented	–0.76 (moderate)	–0.63 (moderate)
KG-only vs. Augmented	–1.90 (**large**)	–0.89 (**large**)

Table 11: Effect Sizes (Cohen's d)

etary models, compare multiple KG-building methods (e.g. pipeline vs. joint extraction) and prompt-engineering strategies as well. The extension to non-English sources by leveraging cross-lingual embeddings and integrating language-specific KGs is also planned. To better capture narrative nuances, human-in-the-loop coherence judgments, causal-structure metrics, fact-checking efficiency, etc. are among future plans as well.

Limitations

Despite encouraging results, this study has several limitations. Narrative quality was evaluated through proxy metrics (coverage, overlap, coherence), but lacked manually annotated ground truth, preventing a direct assessment of factual correctness, hallucination, or omission.

Also, all news articles were in English and focused on disinformation and trustworthy reporting, therefore, results may not generalize to other political genres or languages with different discourse structures.

Furthermore, single LLM (*Mistral Small 3.1 24b*) was used for narrative extraction. Additional experiments are needed to explore whether the observed benefits of Augmented pipeline persist across other LLMs.

Moreover, the downstream task used logistic regression with hand-crafted narrative features. While this helped with interpretability, more complex classifiers might capture additional signals.

In addition, fuzzy Jaccard overlap metric relies on sentence embeddings and thresholds. While it captures surface variation, it may under- or overestimate semantic similarity, especially in terms of figurative or culturally specific language.

Finally, The Augmented pipeline relied on KG-derived context included into the prompt. However, different prompt writing techniques and order of elements making this pipeline (e.g., using LLM outputs to filter KG content) were not tested and the results may be different.

Acknowledgments

This research was funded by the Research Council of Lithuania (LMTLT), grant agreement No. S-PD-24-88.

References

Berkeley R Andrus, Yeganeh Nasiri, Shilong Cui, Benjamin Cullen, and Nancy Fulda. 2022. Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10436–10444.

Katra Byram. 2022. Narrative as social action: making rhetorical narrative theory accountable to context. *Poetics Today*, 43:455–478.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Valerie Cross, Valeria Mokrenko, Keeley Crockett, and Naeemeh Adel. 2020. Using fuzzy set similarity in sentence similarity measures. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE.

Rohan Das, Aditya Chandra, I-Ta Lee, and María Leonor Pacheco. 2024. Media framing through the lens of event-centric narratives. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 85–98.

Mike de Kok, Youssra Rebboud, Pasquale Lisena, Raphael Troncy, and Ilaria Tiddi. 2024. From nodes to narratives: A knowledge graph-based storytelling approach. In *TEXT2STORY 2024, 7th International Workshop on Narrative Extraction from Texts (Text2Story), colocated with ECIR 2024*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Fausto German, Brian Keith, and Chris North. 2025. Narrative trails: A method for coherent storyline extraction via maximum capacity path optimization. *arXiv preprint arXiv:2503.15681*.

Jean-Christophe Goulet-Pelletier and Denis Cousineau. 2018. A review of effect sizes and their confidence intervals, part i: The Cohen's d family. *The Quantitative Methods for Psychology*, 14(4):242–265.

Dani Gunawan, CA Sembiring, and Mohammad Andri Budiman. 2018. The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of physics: conference series*, volume 978, page 012120. IOP Publishing.

Andy Halterman. 2020. Extracting political events from text using syntax and semantics. *Technical report MIT*.

Hans Ole Hatzel and Chris Biemann. 2023. Narrative cloze as a training objective: Towards modeling stories using narrative chain embeddings. *Proceedings of the The 5th Workshop on Narrative Understanding*.

Mourad Heddaya, Qingcheng Zeng, Alexander Zentefis, Rob Voigt, and Chenhao Tan. 2024. Causal micro-narratives. In *Proceedings of the 6th Workshop on Narrative Understanding*, pages 67–84.

Harshad Hegde, Jennifer Vendetti, Damien Goutte-Gattat, J Harry Caufield, John B Graybeal, Nomi L Harris, Naouel Karam, Christian Kindermann, Nicolas Matentzoglu, James A Overton, Mark A Musen, and Christopher J Mungall. 2025. A change language for ontologies and knowledge graphs. *Database: The Journal of Biological Databases and Curation*, 2025:baae133.

Maria Hellman. 2024. Narrative analysis and framing analysis of disinformation. In *Security, Disinformation and Harmful Narratives: RT and Sputnik News Coverage about Sweden*, pages 101–121. Springer.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2021. Daai at case 2021 task 1: Transformer-based multilingual socio-political and crisis event detection. *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*.

Ella Hofmann-Coyle, Mayank Kulkarni, Lingjue Xie, Mounica Maddela, and Daniel Preoţiuc-Pietro. 2022. Extractive entity-centric summarization as sentence selection using bi-encoders. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 326–333.

Ali Hürriyetoğlu, Surendrabikram Thapa, Gökçe Uludoğan, Somaiyeh Dehghan, and Hristo Tanev. 2024. A concise report of the 7th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 248–255.

Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. 2023. A survey on event-based news narrative extraction. *ACM Computing Surveys*, 55(14s):1–39.

Yuanyuan Lei and Ruihong Huang. 2023. Identifying conspiracy theories news based on event relation graph. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9811–9822.

João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. Euvsdisinfo: A dataset for multilingual detection of pro-Kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5380–5384.

Chang Li. 2021. *Improving Stance and Bias Detection in Text by Modeling Social Context*. Ph.D. thesis, Purdue University.

Haochen Li and Di Geng. 2024. Prompt-based graph model for joint liberal event extraction and event schema induction. *arXiv preprint arXiv:2403.12526*.

Jin Li, Ruifan Yuan, Yu Tian, and Jingsong Li. 2024a. Towards instruction-tuned verification for improving biomedical information extraction with large language models. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 6685–6692.

Sha Li, Revanth Gangi Reddy, Khanh Nguyen, Qingyun Wang, Yi Fung, Chi Han, Jiawei Han, Kartik Natarajan, Clare Voss, and Heng Ji. 2024b. Schema-guided culture-aware complex event simulation with multi-agent role-play. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 372–381.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ben Liu, Jihai Zhang, Fangquan Lin, Cheng Yang, Min Peng, and Wotao Yin. 2025. Symagent: A neural-symbolic self-learning agent framework for complex reasoning over knowledge graphs. In *Proceedings of the ACM on Web Conference 2025*, pages 98–108.

Zhengdong Lu, Ziqian Zeng, Jianwei Wang, Hanlin Wang, Weikai Lu, and Huiping Zhuang. 2024. Zero-shot event argument extraction by disentangling trigger from argument and role. *International Journal of Machine Learning and Cybernetics*, pages 1–19.

Nick Mckenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774.

Derek Miller. 2019. Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Cedric Möller, Jens Lehmann, and Ricardo Usbeck. 2022. Survey on english entity linking on wikidata: Datasets and approaches. *Semantic Web*, 13(6):925–966.

Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. 2023. A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference*, pages 15–25. Springer.

Brian Felipe Keith Norambuena and Tanushree Mitra. 2021. Narrative maps. *Proceedings of the ACM on Human-Computer Interaction*, 4:1 – 33.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Synergizing machine learning & symbolic methods: A survey on hybrid approaches to natural language processing. *Expert Systems with Applications*, 251:124097.

Armin Pournaki and Tom Willaert. 2024. A graph-based approach to extracting narrative signals from public discourse. *arXiv preprint arXiv:2411.00702*.

Oona Rainio, Jarmo Teuho, and Riku Klén. 2024. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

Adrian Rauchfleisch and Andreas Jungherr. 2024. Blame and obligation: The importance of libertarianism and political orientation in the public assessment of disinformation in the United States. *Policy & Internet*, 16(4):801–817.

Denise Rey and Markus Neuhäuser. 2011. Wilcoxon-signed-rank test. In *International encyclopedia of statistical science*, pages 1658–1659. Springer.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Léo Laugier, Karl Aberer, and Antoine Bosselut. 2023. Crab: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(8):8393–8435.

Cristian Santini. 2024. Combining language models for knowledge extraction from Italian TEI editions. *Frontiers in Computer Science*, 6:1472512.

Wenbo Shang and Xin Huang. 2024. A survey of large language models on generative graph analytics: Query, learning, and applications. *arXiv preprint arXiv:2404.14809*.

Zhu Shen, Ambarish Chattopadhyay, Yuzhou Lin, and Jose R Zubizarreta. 2024. An anatomy of event studies: hypothetical experiments, exact decomposition, and weighting diagnostics. *arXiv preprint arXiv:2410.17399*.

Jialong Tang, Hongyu Lin, M. Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. *ArXiv*, abs/2106.08629.

Nikos Voskarides, Edgar Meij, Sabrina Sauer, and Maarten de Rijke. 2021. News article retrieval in context for event-centric narrative creation. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 103–112.

Guan Wang, Rebecca Frederick, Jinglong Duan, William BL Wong, Verica Rupar, Weihua Li, and Quan Bai. 2025. Detecting misinformation through framing theory: the frame element-based model. *Journal of Computational Social Science*, 8(3):72.

Shaofei Wang. 2025. Enhancing in-context learning of large language models for knowledge graph reasoning via rule-and-reinforce selected triples. *Applied Sciences*, 15(3):1088.

Zihao Wang, Hang Yin, and Yangqiu Song. 2022. Logical queries on knowledge graphs: emerging interface of incomplete relational data. *IEEE Data Eng. Bull.*, 45(4):3–18.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388.

Tzu-Tsung Wong and Po-Yang Yeh. 2019. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1586–1594.

Peiyi Yang, Bin Song, and Zhiyong Zhang. 2024. Research on knowledge graph construction methods for news domain. *Academic Journal of Science and Technology*, 11(1):58–64.

Hao Zhang, Frank Boons, and Riza Batista-Navarro. 2019. Whose story is it anyway? Automatic extraction of accounts from news articles. *Information processing & management*, 56(5):1837–1848.

Tong Zhou, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Cogmg: Collaborative augmentation between large language model and knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 365–373.

Shenzhe Zhu. 2024. Exploring knowledge graph-based neural-symbolic system from application perspective. *arXiv e-prints*, pages arXiv–2405.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

Ling Zhuang, Hao Fei, and Po Hu. 2023. Knowledge-enhanced event relation extraction via event ontology prompt. *Information Fusion*, 100:101919.

A Examples of News Article Summaries

In this section a few examples of news article summaries are presented together with their ROUGE scores. The summary extracted with the method that achieved higher ROUGE score, which reflects higher overlap with original text, was used in the experiments (Table 12).

Title	LexRank summary	BERT extractive summary	ROUGE (LexRank)	ROUGE (BERT)
EU supports Zelen-sky's peace plan – Michel	Ukraine is The European Union and The European Union is Ukraine. In Africa, in Asia, China and in Latin America, in every meeting with foreign leaders, support for Ukraine is the first issue The European Union raise, European Council President Charles Michel said. The damages caused by Russia to Ukraine must be compensated and all those – without exception – responsible must be and will be held accountable, European Council President Charles Michel said. The European Union have frozen over EUR 300 billion of Russian assets. Since day one of Russia's attack, Europeans have understood that Russia's attack is more than an attack on Ukraine. The European Union is more united than ever.	The European Union supports the Peace Plan proposed by President of Ukraine Volodymyr Zelensky, European Council President Charles Michel has said. You this goal an end to this war of aggression and The European Union shares this goal with you. The damages caused by Russia to Ukraine must be compensated and all those – without exception – responsible must be and will be held accountable, European Council President Charles Michel said. European Council President Charles Michel am strongly convinced over EUR 300 billion of Russian assets should and will be used for Ukraine's reconstruction. Since day one of Russia's attack, Europeans have understood that Russia's attack is more than an attack on Ukraine.	0.291	**0.292**

Title	LexRank summary	BERT extractive summary	ROUGE (LexRank)	ROUGE (BERT)
International Space Station preparing for 'space taxis'	Two astronauts will prepare the International Space Station for two new docks for the spacecraft that will ferry astronauts and gear to and from the International Space Station. Two new docks for the spacecraft that will ferry astronauts and gear to and from the station will allow both Boeing CST's and SpaceX's Dragon spacecraft to dock on the American side of the ISS, even though the two companies' ships are designed differently. The astronauts will be setting up cables on the first two walks and antennae on the third walk scheduled for March 1. It will be the first time vessels have docked on the American side of the ISS since NASA ended the Space Shuttle program. NASA awarded Boeing a $4.2 billion contract in September to develop a transportation capable of carrying human passengers.	A spacewalk outside the ISS planned for Friday was postponed by a day, after "added analysis of spacesuits" the astronauts will wear, according to a NASA statement. NASA confirmed it needed more time to look at suits but did not give additional information. Boeing is building two new docks for the spacecraft that will ferry astronauts and gear to and from the station. SpaceX will carry two new docks to the ISS on cargo mission CRS-7. If all goes according to plan, it will be the first time vessels have docked on the American side of the ISS since NASA ended the Shuttle program. It will also allow NASA to increase crew size and scientific research.	**0.622**	0.59

Title	LexRank summary	BERT extractive summary	ROUGE (LexRank)	ROUGE (BERT)
The Maduro Diet: Food v. Freedom in Venezuela	Nicolas Maduro's regime has been unable to control inflation, which has disintegrated Venezuela's economy. As a result, many poor Venezuelans are now dependent on CLAP deliveries to put food on the table. A politically motivated food system called CLAP has become an essential part of nearly half the population's diet. Those overseeing the program earn a 200% profit per box. Though the regime aims to reduce dependence on imports, 90% of CLAP boxes come from Mexico. The international community should link external pressure with renewed political opposition to bring about a democratic transition.	As Venezuela's humanitarian crisis deepens, Maduro's regime has exploited it to tighten political control. Poverty has risen from 55% in 1998 to nearly 90% today. Though the regime claims to reduce food import dependence, 90% of CLAP boxes come from Mexico. The weaponization of the CLAP program is a sign of a desperate regime. The use of cryptocurrency-based aid models like EatBCH may help.	**0.294**	0.212

Table 12: Examples of news article summaries

B Statistics of the Final Dataset

This section presents a basic analysis of the final dataset used for experiments. Table 13 presents article distribution per year in the final dataset. For trustworthy articles, 2020 had the largest number of articles (109), while for disinformation the articles were distributed more equally, with 2019 and 2020 having the largest numbers (46 each).

Year	Disinformation	Trustworthy
2015	18	3
2016	26	9
2017	29	14
2018	43	24
2019	46	52
2020	46	109
2021	28	22
2022	43	52
2023	28	17

Table 13: Article distribution 2015-2023.

Table 14 lists publishers with the highest numbers of articles in the final dataset, where for disinformation the main publisher was RT (166 articles) and for trustworthy news the articles were more equally distributed with BBC having the largest number of articles (28).

Publishers	
Disinformation	**Trustworthy**
RT (166)	BBC (28)
TASS (51)	Polygraph.info (25)
Sputnik (15)	The Guardian (24)
Geopolitica.ru (12)	Radio Free Europe (11)
Global research (11)	DW.com (10)

Table 14: Publishers with the largest number of articles (number in brackets) in the final dataset.

Similarities and differences in topics across disinformation and trustworthy articles in the final dataset can be seen in word clouds (Fig. 1) representing the most frequent 25 words, excluding stopwords, of each part, with 'russian' being the most frequent word word in both groups of articles. Also, in both groups Russian-Ukrainian war related word forms are prominent, with 'said' being the most frequent word form. Furthermore, in trustworthy articles 'coronavirus' is among the most frequent word forms, while in disinformation articles this word form is not among most frequent ones, indicating difference in topic coverage.

C Prompts Used for LLM-only and Augmented Extractions of Narrative Elements

For LLM-only extraction of narrative elements, the following prompt was used:

```
prompt_text = (
"Extract the following narrative
elements from the news article:\n"
"- Characters (Who are the key actors?)\n"
"- Events (What happened?)\n"
"- Causal Links (Why did it happen?)\n"
"- Framing:\n"
"    - Actor Framing: How are key actors
(individuals, organizations) portrayed?
(e.g., hero, villain, expert, victim)\n"
"    - Event Framing: How is the event
presented? (e.g., crisis, scandal,
breakthrough, tragedy)\n\n"
f"News summary: {best_summary}")
```

For Augmented pipeline, the prompt below was used:

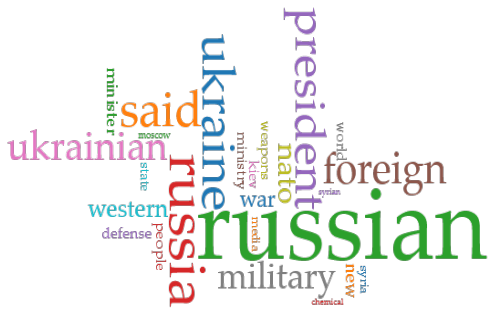
```
prompt_text = (
"Extract the following narrative elements
from the news article:\n"
"- Characters (Who are the key actors?)\n"
"- Events (What happened?)\n"
"- Causal Links (Why did it happen?)\n"
"- Framing:\n"
"    - Actor Framing: How are key actors
portrayed? (e.g., hero, villain)\n"
"    - Event Framing: How is the event
presented? (e.g., crisis, breakthrough)\n\n"
f"News summary:\n{best_summary}\n\n"
f"Knowledge Graph context:\n{kg_context}")
```

D REBEL Relation Sets/Mappings

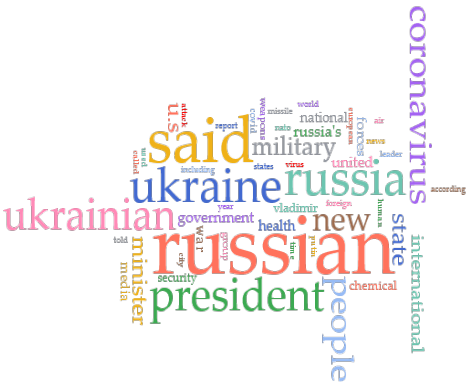
The relations of REBEL schema that imply framing, events and causality were mapped and inferred manually:

Relations suggesting events or time-anchored nodes:

- Participation-based: "participant_in", "participant", "winner", "candidate", "candidacy_in_election".
- Achievement or outcome: "award_received", "notable_work", "nominated_for".
- Conflict or disruption: "conflict", "significant_event", "victory", "replaced_by", "replaces".
- Time anchors: "inception", "start_time", "end_time", "point_in_time", "publication_date", "date_of_birth", "date_of_death", "work_period_(start)", "work_period_start".



(a) Most frequent word forms in disinformation articles



(b) Most frequent word forms in trustworthy articles

Figure 1: Most frequent word forms

- Organizational change: "founded_by", "dissolved_abolished_or_demolished_date", "location_of_formation", "established_date".
- Media & Creative: "original_broadcaster", "composer", "director", "producer", "production_company"

Relations suggesting causality:

- Explicit causality: "has_cause", "causes", "has_effect", "influenced_by", "leads_to", "responds_with".
- Implied causality: "participant_in", "participant", "winner", "founded_by", "replaced_by", "replaces", "award_received", "influenced_by", "significant_event", "conflict", "start_time", "end_time"

Relations suggesting framing:

- Explicit framing:
  - "described_as": "is portrayed as"
  - "depicts": "is represented as"
  - "notable_work": "is known for"
  - "award_received": "is recognized for"
  - "member_of_political_party": "is affiliated with"
  - "replaced_by": "is portrayed as outdated or replaced by"
  - "founded_by": "was founded by"
- Causal interpretation framing:
  - "has_cause": "is caused by"
  - "causes": "causes",
  - "has_effect": "results in"
  - "influenced_by": "is influenced by"
  - "leads_to": "leads to"
  - "responds_with": "is responded to with"
  - "participant_in": "participated in"
  - "participant": "participated in"
  - "winner": "wins due to"
  - "candidate": "candidacy in"
  - "significant_event": "is triggered by"
  - "conflict": "is involved in conflict"
  - "start_time": "starts at"
  - "end_time": "ends at"

E Downstream Classification Setup

The downstream classification setup for extrinsic evaluation consisted of the following:

- For each pipeline, a feature matrix was formed.
- Feature sets:
  - LLM-only: counts of narrative elements and values of embedding-based coherence measure produced by LLM-only pipeline from extractive summaries of news articles.
  - KG-only: counts of narrative elements and values of graph-based coherence measures (density, avg. path length, largest component ratio) produced by KG-only pipeline from extractive summaries of news articles.

  - Augmented: counts of narrative elements and values of embedding-based coherence measure produced by Augmented pipeline from extractive summaries of news articles.

- For classification, a scikit-learn pipeline was used:

  StandardScaler() → LogisticRegression(solver='liblinear', penalty='l2', C=1.0, max_iter=100) (defaults shown).

- Evaluation was performed with StratifiedKFold (5 and 10 splits, shuffle=True, random_state=42) using F1 as the primary metric via cross_val_score. Mean ± SD across folds are reported as well. Significance is assessed on fold-wise F1 with paired t and Wilcoxon signed-rank tests. Effect sizes were evaluated with Cohen's d.