Narrative Feature Extraction and Clustering from Humanitarian Journalism

CS333 Final Project, Wellesley College

Kexuan (Coco) Zhang
kz108@wllelesy.edu
Instructed by: Professor Carolyn Anderson

December 2025

1 Introduction

Research Question To what extent can humanitarian reporting from outlets (BBC News vs the New Humanitarian) be distinguished using narrative features alone, and which features most strongly drive separation (as evidence of focalisation differences)?

Modern humanitarianism institutionalised during the two World Wars. Through globalisation and decolonisation efforts, it creates emergency responses to underdeveloped regions while headquartering in the global north. This paper aims to decode the impact of this western-centric institutional set-up of humanitarianism through discourse analysis powered by natural language processing.

Language, the main driver of narratives, articulates a fundamental "semiotic dimension [...] of social practices", which empowers discourse analysis to capture the identity of social actors and their interactions.[1] Referring to narratologist Rimmon-Keenan's concept in structural narratology, focaliser, a "sustained inside view", this paper evaluates the hypothesis of there being a consistent focaliser behind humanitarian narratives: western institutions of power.[2] The power of discourses in changing our status quo motivates me to understand the impact of western-centric humanitarianism by auditing humanitarian discourses.

This phenomenon exists in various narrative forms including journalism, public speeches, and policy statements. For example, the USAID localisation agenda addresses "shifting power and resources toward local actors" with actions such as "empowering local champions", where words such as "local" and "empower" establish western institutions as an in-group with power hierarchy over local communities, the out-group.[6]

Computational methods are introduced to discourse analysis to sample statistically significant semantical or lexical patterns in the dataset. Computational linguistics has been critiqued for lacking capacities for contextual intepretation, which language models with larger attention windows are trying to take into account.

Discourse analysis conducted by Baker, et al. on portrayal of refugee and asylum seekers in the UK press outlines the methodological importance of identifying nomination, 'construction of in-groups and out-groups' and perspectivation 'positioning speakers' point of view' in a text.[4] In a case with a finite set of stakeholders, Orgad and Seu's studies on NGO–beneficiaries-audience relationship takes a discursive approach to building a structural representation between the three.[3]

By utilising Natural Language Processing tools, this paper aims to engineer and learn features representations of actors and their perspectives to understand the narrative framing by different journals on humanitarian events. It focuses on identifying narrative style differences between journalism outlets with different global positioning. The objective of this project is to compare global news from BBC News, a public broadcaster, and the New Humanitarian (TNH), an independent non-profit newsroom through tasks including text classification and semantic role labeling. On one hand, BBC News focuses on informing the mass public on highlighted global affairs. On the other hand, the New Humanitarian sets itself apart from public outlets with a commitment to reporting at the heart of humanitarian crises to address ethics and accountability. An example of comparing outlet styles is to train and perform classification on a labeled and mixed dataset combining both BBC and TNH articles to identify how lexically and semantically different the two outlets are.

2 Data and Methodology

2.1 Dataset

This project used two datasets, a 2004-2005 BBC News dataset created by D. Greene and P. Cunningham for Machine Learning research and a 2004-2005 the New Humanitarian dataset I manually collected. Both include only the titles and texts of published articles.

1. BBC News Dataset (BBC):

To incorporate the perspective of a major Western news outlet on global affairs, I sought a publicly available corpus that is both widely recognized and historically grounded. Several prominent datasets derived from sources such as The New York Times and CNN were considered; however, many are no longer accessible. In particular, the New York Times Annotated Corpus (1987–2007), which has been widely used in prior research, has been withdrawn from public distribution. As a result, the BBC News Dataset was selected as a practical alternative.[5] Although it spans only a two-year period, the dataset contains 2,225 documents, approximately 20 percent of which are labeled under the Politics category, making it a suitable resource for analyzing political and international reporting.

Noticing BBC News' inclination towards domestic reporting, I filtered out 180 articles that included mentions of foreign countries, international institutions, and key affair indicators such as 'conflict' and 'diplomacy'. The project mainly worked with this smaller filtered dataset with focus on international affairs.

2. The New Humanitarian dataset (TNH):

To complement mainstream Western news coverage, I would like to include a nontraditional outlet with an alternative narrative style. The New Humanitarian was selected due to its diverse authorship, which includes on-site journalists, stakeholder interviewees, and think tank specialists, offering perspectives distinct from those of major international media organizations. As the outlet does not provide a readily available public corpus, data collection relied on a combination of web scraping and manual curation. After automated scraping attempts using the BeautifulSoup library proved unsuccessful, the dataset used in this study was manually curated from The New Humanitarian website.

By extracting top frequent words addressing country names, institutions, and affairs from the BBC News dataset, I extracted five sets of keywords in the field of war, diplomacy, refugee, security, and international institution (see Appendix A). Combining these five sets of keywords with a time frame limit between 2004 and 2005, I created a collection of 180 articles with an even spread across the five fields.

2.2 Task Description

The project consists of two parts: the first is a classification task that trains models to classify BBC and TNH articles on humanitarian events, the second is a semantic role labeling task that identifies the narrative tone and actors within each sentence space, and summarises the trend over each journal.

Task A: Journal Classification

The two datasets are merged and preprocessed by annotating each article with its source and removing duplicates. The resulting corpus is partitioned into five equally sized folds with a randomized mix of BBC and TNH articles. Text classification models are evaluated using five-fold cross-validation, yielding performance estimates that are averaged across all data points.

Task B: Semantic Role Labeling

Semantic role labeling is performed on the BBC and TNH dataset separately. Due to a lack of semantic role annotation on these data, for each article in the dataset, we use spaCy for parsing and role labeling on a sentence-by-sentence basis. By aggregating per-article labeling results including actor count, actor type, and active/passive voice, we compare entities and tones used across the two datasets.

3 Model

3.1 Text Classification: TF-IDF

For the text classification task, the project used a TF-IDF regression model with a pipeline of TF-IDF vectorisation of the text and learned weights of word features through logistic regression. The model focuses on understanding how separable the BBC and TNH dataset are with their lexical choices.

The task ustilised a baseline TF-IDF model without any entity masking or topic masking, which assigned the most weight to features such as names (eg. "blair", "howard") in the BBC dataset and "IRIN", the former name of TNH. This indicates that the model might overperform by identifying name entities and topic-specific vocabulary as shortcuts. The project introduces control by training two additional models on texts with different levels of entity masking to investigate what narrative signals persist beyond entity mapping. This is similar to conducting ablation studies on high-performance features to diagnose the true indicators of model performance.

The two improved models include preprocessing the raw dataset using: 1) spaCy NER-based masking and 2) topic-based masking.

1. SpaCy Name Entity Recognition (NER) based masking

The spaCy EntityRecogniser processes unannotated text to categorise key information in the sentence. Among other state-of-the-art NER systems, I chose spaCy for its fast statistical approach and compatibility with the regular syntax in my datasets, which is compiled from professional journalism that follows standard English grammar. To mask out actors specific to each journal that can be give-away in categorisation, the project used NER to mask out any person, organisation, or geopolitical groups in the dataset.

2. Topic-based masking

The features learned from the NER-based masking still reveals a key difference in the top stakeholders mentioned in the two different journals. The BBC dataset addresses domestic governors including "prime minister" and "foreign secretary", while the TNH dataset addresses "aid worker" and "security council". Therefore, on top of the NER-based masking, we use the spaCy parser for English to mask out bigram and unigram level catch phrases that directly associate with the set of actors each journal addresses. Despite having the risk to overcorrect the text and eliminate useful categorisation markers, the project applies this improved model to test if lexical markers beyond topics.

Names and institutions dominate the baseline model but disappear after entity masking, while humanitarian terms ("refugee", "peace") persist across all conditions. This demonstrates that much of the baseline model's performance is driven by topical and institutional lexical cues, rather than narrative structure. The additional masking in improved models effectively eliminates obvious keywords as shown by the top features extracted from each TF-IDF model (graph visualised in Appendix B).

3.2 Semantic Role Labeling: SpaCy

Based on spaCy part-of-speech tagging, the project uses the trained pipeline to label semantic roles of tokens in each sentence in the dataset. By capturing the predicate (root verb) in each sentence space, the model parses tokens in the sentence space by actors into subject and object, while marking the active/passive voice.

The baseline model directly conducts semantic role labeling on each sentence and picks out the full noun phrase such as "one of UKIP 's two London Assembly members , Damien Hockney who is now Veritas ' deputy leader". The modifiers in the phrase make it difficult to compute counts of key actors. Therefore, the project works with an improved model that identifies the key word across noun phrases by stripping away propositions and adjectives within the clause. This model shows more precise role labeling. For example, from the previous noun phrase, "London Assembly members" will be labeled as the main actor in the sentence. This allows us to better identify what type of actors are being narrated as actors or recipients of actions and who the key players are.

The model uses a standard semantic role labeling framework to identify agents (ARG0) and patients "he will use the Upper House to advocate its reform", "use" is the root of the sentence and forms a predicate where "he" is the agent that uses, "the Upper House" is the patient that is being used. For a sentence in passive tone such as "Lord Kinnock was accompanied by Lords Leader Baroness Amos and Baroness Royall of Blaisdon", "Lord Kinnock" is the patient and "Lords Leader" is the actor in the predicate "ARG0 accompanies ARG1".[7]

4 Metric

The project investigates narrative features and their similarities and differences across the two corpora. Rather than optimizing against a single task-specific benchmark, model performance is evaluated qualitatively through the interpretability of learned representations and their ability to capture distinctive narrative characteristics. Accordingly, model features are iteratively refined, as described in the model section, to extract increasingly discriminative signals from the two datasets.

The study comprises two tasks with distinct evaluation paradigms.

1. F1 Score and Top Features in Text Classification. In the text classification task, a TF–IDF baseline is used to examine lexical differences between BBC and TNH reporting, treating each corpus as a bag of words. Classification performance, measured via the confusion matrix and F1 score, provides an estimate of how separable the two datasets are based on surface-level lexical cues. To further interpret model behavior, the highest-weighted features in the trained classifier are analyzed to identify the words most influential for distinguishing between sources.

2. Voice Rate and Actor Analysis via Semantic Role Labeling. In the semantic role labeling task, explicit role annotations are not available. Instead, spaCy is used to automatically label predicate–argument structures, enabling the extraction of features such as passive voice rate, passive constructions without explicit agents, and frequently mentioned actors across sentence roles. Aggregating these features at the article level allows for an analysis of how agency is distributed in each corpus, providing insights into whether key actors are framed as active or passive participants within the narrative.

5 Results

5.1 Task Classification Task

5.1.1 Model Performance Score

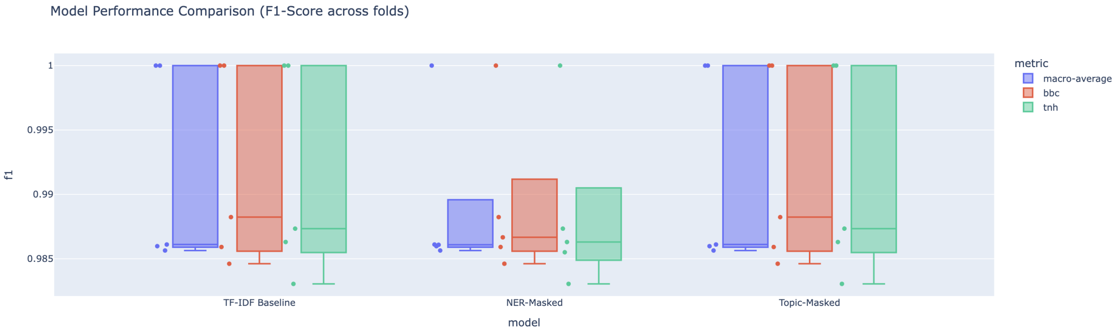

Figure 1: Model-Performance Comparison (F1)

Lexical TF-IDF model achieves near-ceiling performance on the dataset. TF-IDF Baseline and Topic-Masked models are both achieve near $F1 = 0.990.01$ across macroaverage, BBC, and TNH data. NER-Masked drops a bit but maintains a high performance with $F1_{max} = 0.99$. The drop under NER masking is evidence that named entities (people/orgs/geopolitical groups) are a big shortcut feature. This includes features such as “Blair”, “UNHCR”, “IRIN”. The performance scores show topical and genre artifacts as the biggest driver in classification. We need to understand the data on more contextually on a sentence structure level to extract features are informative even when topical cues are controlled.

The confusion matrix from the five training batches in the baseline TF-IDF matrix: [[37 1] [0 34]], [[31 2] [0 39]], [[39 0] [0 33]], [[37 1] [0 34]] [[32 0] [0 40]]. Further evidences the model’s well performance at the classification task with precision and recall all close to 100%. The next few graphs on top features extracted from each TF-IDF model will show the specific stylistic framing differences the models learned:

5.1.2 TF-IDF Feature Extraction

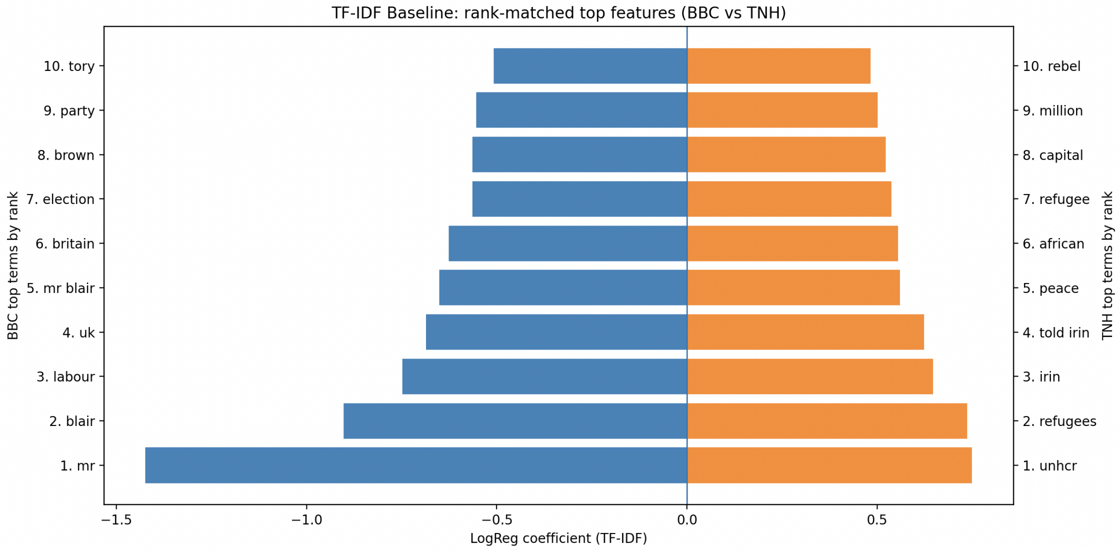

Figure 2: Top BBC and TNH Feature Extracted from Baseline Model

In the baseline TF-IDF model, the BBC side (left with negative coefficients) is dominated by honorific and UK institutional politics such as “mr”, “blair”, “labour”, “uk”, “election”, “tory”, “party”, which indicates the reporting centers around UK domestic impact and decision-making on international politics. On the other hand, TNH side (right with positive coefficients) is dominated by vocabulary around actors in humanitarian crisis including “unhcr”, “refugees”, “irin”, “african”, etc. This suggest that TNH takes a humanitarian intervention frame. Both observations align with the mission of the two publications. To observe deeper lexical cues, the NER-model masks the top entities and observes what remains.

The NER-based masking successfully removes specific names. BBC remains strongly signaled by institutional-role nouns such as “prime minister” and “chancellor” and quotes (“[PERSON] said”), while TNH remains signaled by conflict/humanitarian descriptors such as “refugees” and “peace”. This solidifies that BBC has a narrative frame that represents the voice of the institution / authority implied by titles and quoting while TNH foregrounds the affected population of humanitarian crises more often.

The topic-based filtering further indicates that BBC adopts a rhetorical style of institutional public discourse with formal reference conventions such as “Mr”. Despite masking out strong indicators such as “refugee”, TNH shows a narrative focus on humanitarian situation with actors such as “rebel”, numerics of impact, “million”, and geographical marker such as “north/south”. The features extracted from strong classification models show the areas of key differences, which includes voice, agency, attribution, and participant roles, which will be further investigated from a semantical perspective with the semantic role labeling task.

5.2 Semantic Role Labeling Task

5.2.1 Active/Passive Voice Detection

The project calculates both the rate of passive voice and the rate of passive voice without an agent in the sentence. Examples for the second case include “Refugees are displaced” and “Aid was delivered” where the actor is being omitted or blurred. Agentless passives can be used as a device to avoid direct attribution of responsibility.

With positive sentence framing being the norm in reporting, journalists tend to avoid using passives for clarity, especially constructing a passive scenario without explicitly mentioning the agent. The New Humanitarian shows a higher rate of using both passives and agentless passives in its reporting, which means that TNH frames scenarios where events hap-

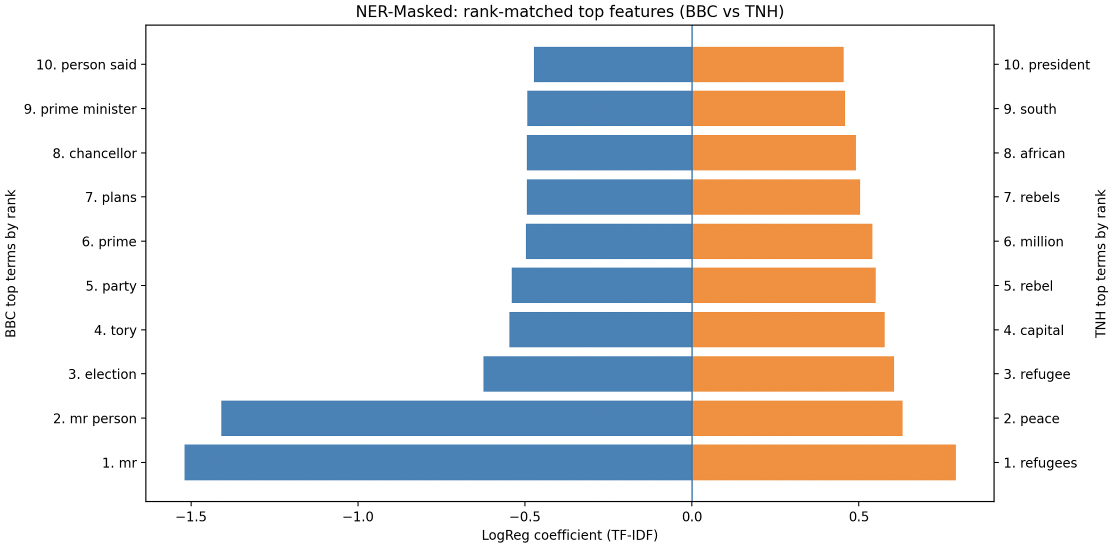

Figure 3: Top BBC and TNH Feature Extracted from NER-masked Model

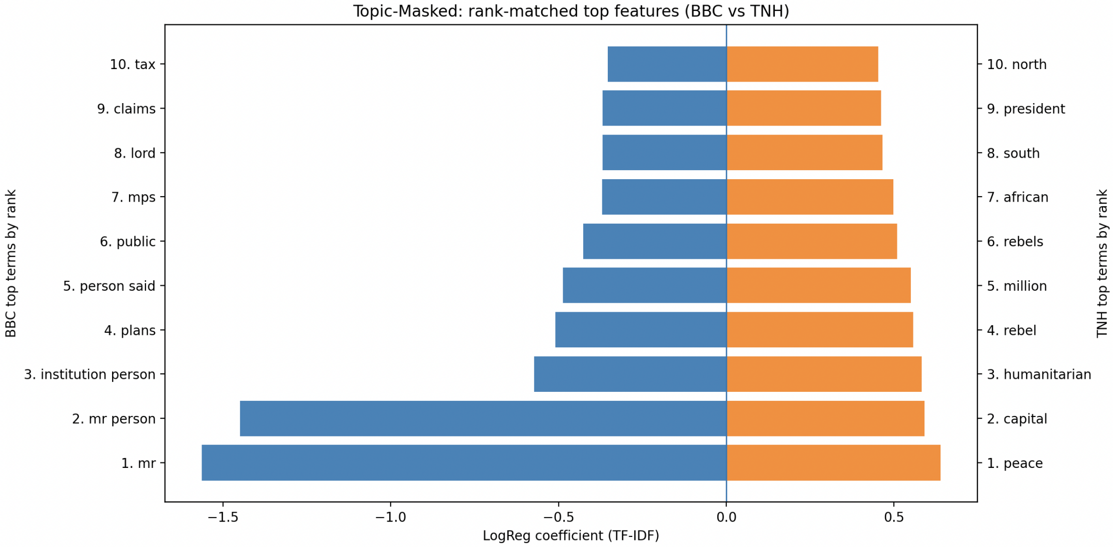

Figure 4: Top BBC and TNH Feature Extracted from Topic-masked Model

pen to people without clearly naming the actor. This is foregrounds conditions/outcomes and victimhood over oversimplified causes. We will further investigate the actors in both BBC and TNH datasets to understand their choice of voice framing.

5.2.2 Actor Type Distribution: Agent and Patient

Using the spaCy part-of-speech labeling, top labels for actors include the following:

- PERSON: People, eg. “Blair”
- ORG: Companies, agencies, institutions, etc, eg. “UNHCR”
- GPE: Countries, cities, states, eg. “Germany”

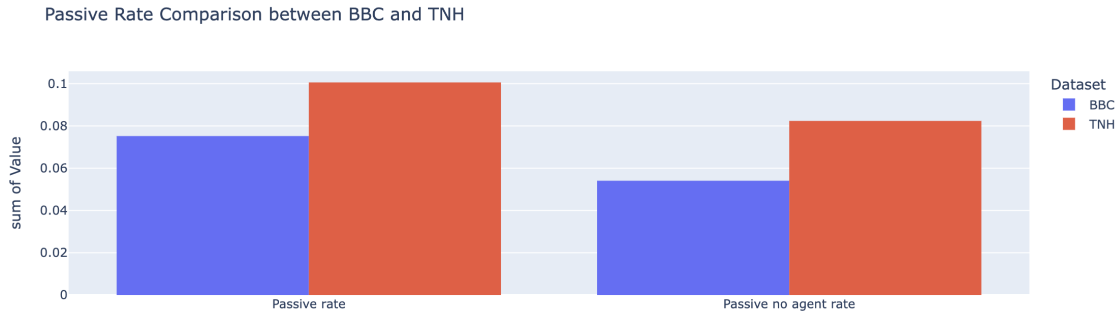

Figure 5: Passive and Passive-No-Agent Rate in BBC and TNH dataset

- NORP: Nationalities or religious or political groups, eg. “Afghans”
- PRON: Pronoun, eg. “he”
- NP: Noun Phrase, eg. “three major parties in the run”

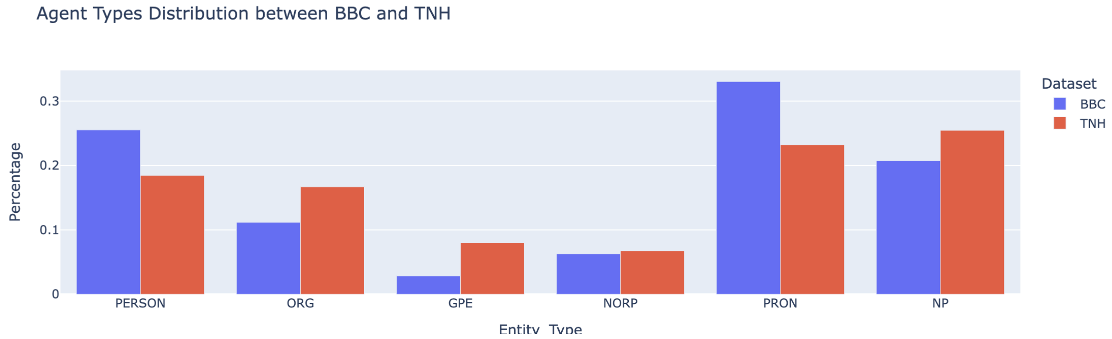

Figure 6: Agent (ARG0) Type Distribution in BBC and TNH dataset

In the BBC dataset, agents are oriented around figures and officials with a concentration in PERSON and PRON agents, which forms a reportorial narration about people’s actions. For the TNH dataset, the agents look more institutional or geopolitical with high ORG and GPE rate, which means that actions are attributed more to institutions and states. The agent distribution in TNH matches a humanitarian frame where institutions and regions are frequently described as causal forces.

NP (noun phrases) dominate for both BBC and TNH dataset as many patients are full noun phrases with modifiers. This graph still shows relative differences where BBC shows more PERSON and PRON patients and TNH shows more ORG/GPE/NORP patients. This suggests that TNH’s “affected entities” are more frequently groups, places, institutions, nationalities such as “Iraq”, “the UN”, “the government”, whereas BBC’s patients include more individual people/pronouns.

As the agent and patient type distributions capture who gets grammatical agency and who gets positioned as the affected party, the top actors further investigate the actors in these spaces and evidences our observations here.

5.2.3 Top Actors: Agent and Patient

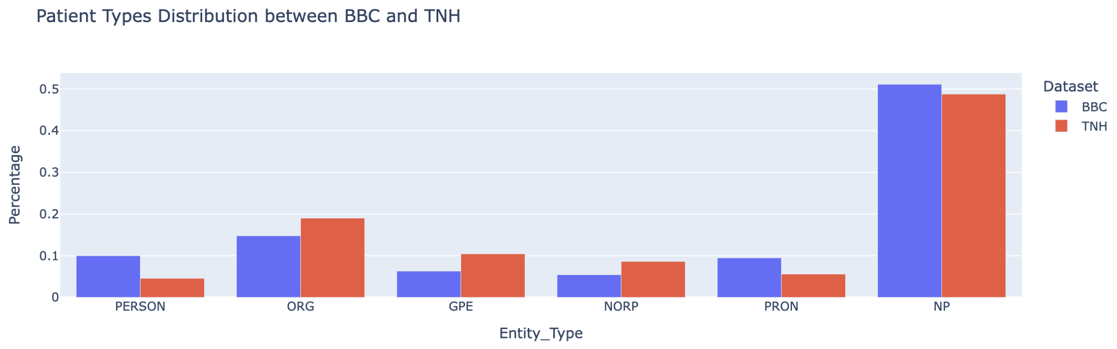

Figure 7: Patient (ARG1) Type Distribution in BBC and TNH dataset

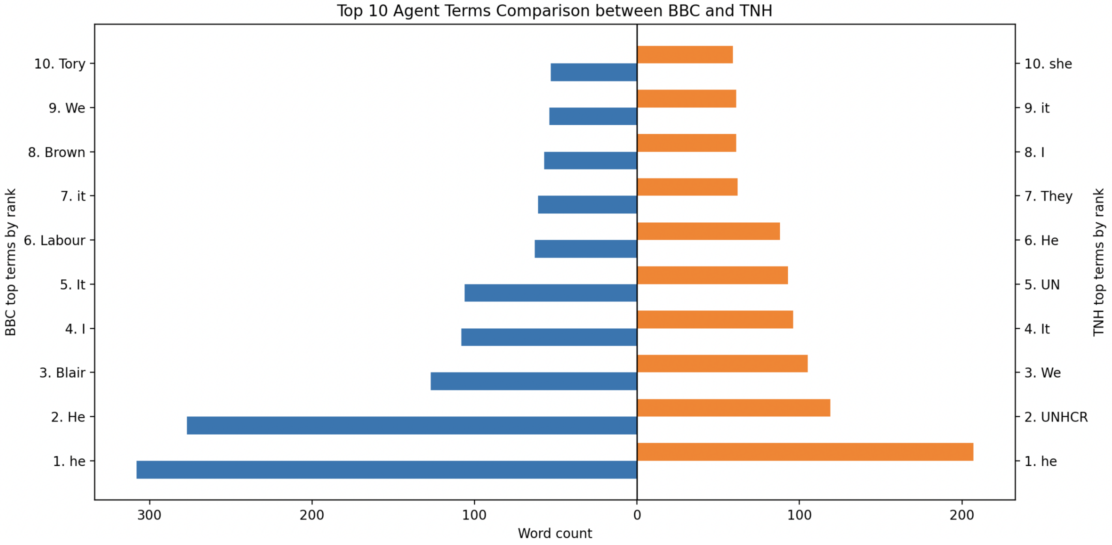

Figure 8: Top Agent Terms in BBC and TNH dataset

As the agent and patient type distributions capture who gets grammatical agency and who gets positioned as the affected party, the top actors further investigate the actors in these spaces and evidences our observations. The high ranking of pronouns in the TNH articles indicate that there's fewer actors in each article with more sentences addressing actors mentioned priorly in the article. While the notable agents in the TNH dataset are "UNHCR" and "UN", the BBC dataset maintains its focus on British domestic political actors such as "Tory" and "Labour", which aligns with our observations that BBC's global reporting springs from the domestic ecosystem.

The TNH patients strongly feature places and populations including "Afghani", "Iraqis", "Iraq","Cote d'Ivoire", and "Afghan". While, BBC patients contain a series of BBC-specific broadcast artifacts: "BBC Radio 4's", "BBC News", plus "UK", "Blair", "Brown", etc. Therefore, we find the TNH patients more focusing on stakeholders in the crisis world with the affected people and places. In contrast, BBC focuses more in the British media ecosystem and UK political scene. The broadcast provenance in the BBC dataset aligns with BBC's enterprise positioning as a broadcaster, which is a dataset style that doesn't relate to narrative focalisation.

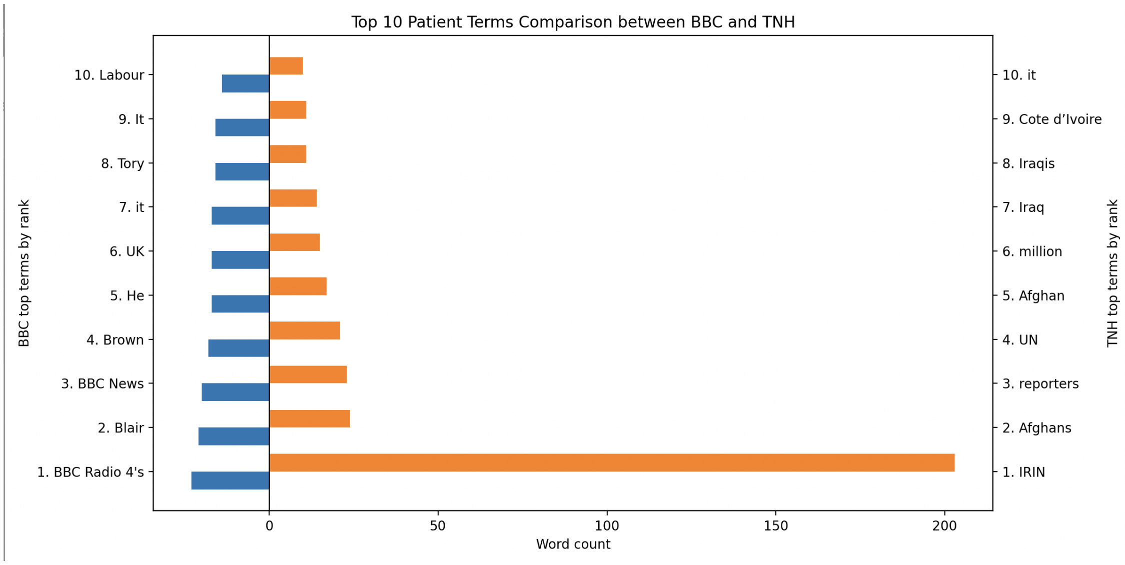

Figure 9: Top Patient Terms in BBC and TNH dataset

6 Conclusion

6.1 Findings

Through the project, we found that the BBC and TNH dataset massively differs from each other evidenced by the high-performance of classification models. And through our models, we found TF-IDF features, active/passive voices, and actor in sentence spaces as helpful features that drive separation between the two datasets, which marks the success in the text classification and semantic role labeling models as we derived explainable findings from them.

Through feature coefficients from the TF-IDF models and Semantic Role Labeling results, we found significant differences within the narrative structures. The stakeholder framing differs where TNH engages with affected population and portrays them as passive recipients of actions, especially the case where they don't directly attribute the action to explicit actors. Their institutional and geopolitical event framing aligns as a typical humanitarian journal while BBC's more framing focuses on elite British actors.

The differences in narrative structures spring from their different missions as journals where BBC News engage with international affairs with an anchor on actions of British political actors and its impact on the domestic space, while the New Humanitarian specialise in humanitarian reporting while engaging with affected populations. Their missions result in different narrative signatures in their reporting such as the significant usage of honorifics "mr" in BBC dataset and the high pronoun agent frequency in the TNH dataset that signals focused reporting and deep-dives on a few actors within one article.

6.2 Limitations and Future Work

Dataset: Despite temporal alignment (2004–2005), the BBC and TNH corpora differ substantially in journalistic mandate. This creates structural topical asymmetry, which is evidenced by the high-performance in TF-IDF models with near-ceiling performance even under masking. In the future, we can consider finding or collecting corpus that has event-based alignment to implement better control between the two datasets. Our dataset would also benefit from more careful cleaning of outlet metadata or transcription conventions to avoid classification models from recognising it as a shortcut.

Model Design The TF-IDF based text classification currently works under linear regression model, which works the best with additive and independent features. Since focalisation is often a combination of features such as passive voice + institutional agent omission, neural models that take explicit narrative features as inputs may show more intricate relations between features. My current SRL extraction uses spaCy dependency parsing and relies on rule-based passive detection and ARG0/ARG1 actor recovery, which can lead to noisy spans and inaccurate identification where noun phrases in modifiers can be counted as agents or patients, which can be improved in the future by exploring neural SRL with language models such as SpanBERT and Longformer.

Metric Design Aggregate metrics for semantic role labeling reduces noise but misses out on article-level differences in each dataset that can indicate different genre subtypes such as news vs analysis and crises severity. Some example metrics that future research can take into account include measuring the ratio of Western agents to local patients and the number of agentless passives per 1,000 words.

7 Appendix

7.1 Appendix A: Search Word Combination in the New Humanitarian Advanced Search

War & military affairs:("war" OR "military" OR "troops" OR "forces" OR "conflict") AND ("United States" OR "European Union" OR "Iraq" OR "China" OR "Pakistan" OR "Israel")

Diplomacy & negotiations: ("talks" OR "negotiations" OR "peace" OR "foreign policy") AND ("United States" OR "European Union" OR "China" OR "France" OR "Germany")

Refugees, asylum, humanitarian: ("asylum" OR "refugees" OR "refugee" OR "crisis") AND ("Africa" OR "Iraq" OR "Pakistan" OR "Turkey" OR "Israel")

Security issues: ("nuclear" OR "sanctions" OR "security council") AND ("United States" OR "China" OR "Iraq" OR "Pakistan")

International institutions: ("United Nations" OR "Security Council" OR "World Bank" OR "NATO") AND ("war" OR "peace" OR "conflict" OR "sanctions")

7.2 Appendix B: Top TF-IDF Features Before vs After Masking

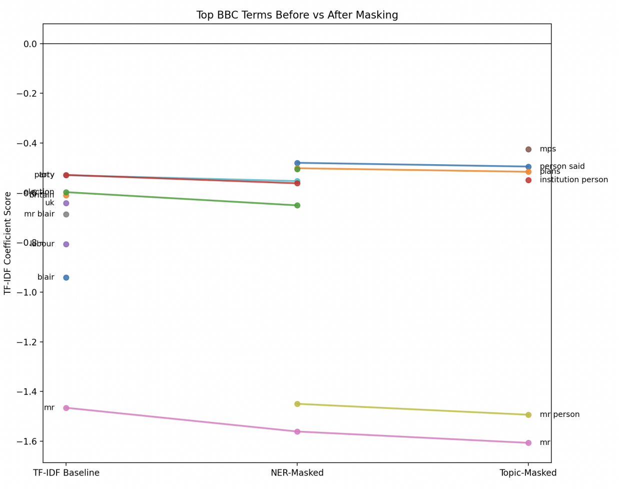

Figure 10: Top Words in BBC dataset Before vs. After Masking

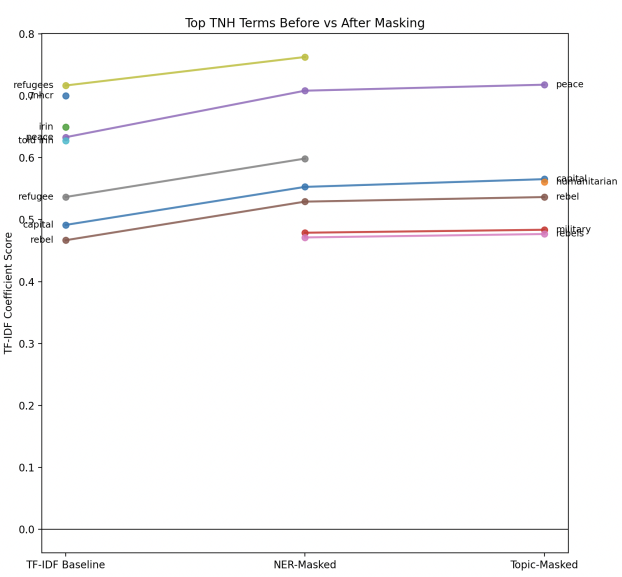

Figure 11: Top Words in TNH dataset Before vs. After Masking

References

[1] Norman Fairclough. (2012). Critical discourse analysis. The Routledge handbook of discourse analysis, 7: 11-22.

[2] Rimmon-Kenan, Shlomith. (2002). Narrative Fiction: Contemporary Poetics. Routledge.

[3] Orgad, S., & Seu, B. I. (2014). "Intimacy at a distance" in humanitarian communication. Media, Culture & Society, 36(7), 916-934.

[4] P. Baker, et al. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press, Discourse & Society, 19 (3): 273–306.

[5] D. Greene and P. Cunningham. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering, Proc. ICML 2006.

[6] Hans, Ananta. Localization, 28 Apr. 2025, D-Lab, Massachussetts Institute of Technology, Cambridge, MA.

[7] MIT Computer Science and Artificial Intelligence Laboratory. "Semantic Roles." MIT Workbench, 2020.