DS Project Report: Exploring Patterns of Sentiment and Content in News

Leevi Takala, Aleena Ahmad, Rennze Fabe

Motivation

News content and the way it is presented can impact public opinions. This is especially important when the news discusses unfamiliar or potentially vulnerable groups or communities.

This raises the question: **Are there any noticeable patterns in how news content is written when discussing different groups? Does the sentiment, language used, or content differ significantly across groups?**

If so, **can we detect and understand these patterns?** Can **we identify the groups that might be vulnerable to harmful patterns?** We wanted to see if, using word and sentiment analysis, we can find distinguishable patterns in news reporting for different groups.

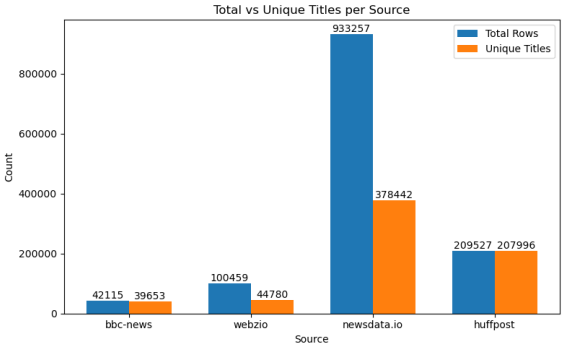
Data Collection

We found several datasets that contained news information: the Global News dataset, the HuffPost dataset , the Webhose dataset, and a BBC dataset. We chose to work with all of these to maximize the amount of data used. The number of records for each of the datasets is below:

Global News	HuffPost	Webhose	BBC
933,257	209,527	100,476	42,329

Exploratory Data Analysis

Since all the data was in CSV and JSON formats, we were able to look at the data in standard text and spreadsheet programs to quickly see what kind of data was included in the datasets and how it was formatted. We used Python and Pandas for initial overview. We first checked to see how many titles were unique.



We dropped columns not useful for our analysis, such as authors, image links, etc. However, many

datasets had ‘category’ columns. We kept mainly headlines, text, date, and categories. The categorizes of all datasets are visualized to determine predominant ones.

We also looked at how many values were missing for each column, which values appeared in fields like news source, category and author, and we looked at what the distributions of values were like for different fields.

Since for a lot of articles the text was missing, we limited our scope to working with just titles. The datasets had URLs that point to the original articles, but scraping the full articles was infeasible.

Preprocessing

We dropped articles that were missing the title and articles that were duplicates. Articles that were in languages other than English were also removed by checking if the titles contained foreign characters.

For the Webhose dataset, we converted all the data into CSV format as it originally consisted of thousands of JSON files, making it difficult to use. We also converted the titles into lowercase letters and removed punctuation.

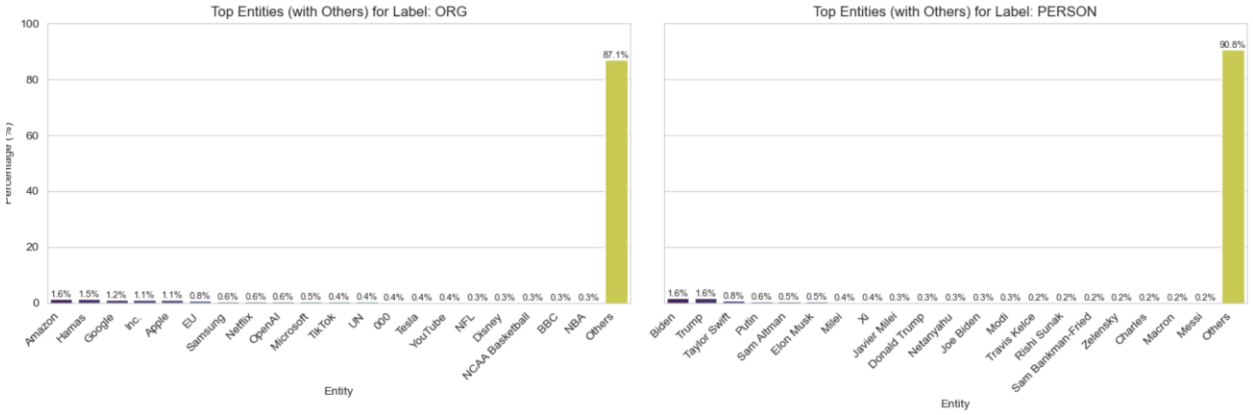
Learning Tasks

Named Entity Recognition

Our main goal was to detect and analyze any patterns of sentiment or content words that were biased towards certain groups. For this, the first step was to extract the groups and entities present in our news, against which we could detect patterns. To do so, we first tried a **regex-based** approach. We manually created a list of ‘vulnerable’ groups consisting of traditionally marginalized groups based on ethnicity, religion, sexuality, and physical ability.

However, this approach was impractical as it was difficult to write comprehensive patterns that could capture all instances of these groups. We then shifted to Named Entity Recognition. For this, we used Spacy’s en_core_web_trf model.

The NER model returned 18 unique labels (Date, Person, GPE, Cardinal, etc.) From these, we chose to work with articles with the labels of GPE, NORP, and PERSON as these related to people. To understand the most common entities within these labels, we visualized the top 10 entities for each label.



The top 10 entities constituted very small percentages of the total entities mentioned (as shown in distribution chart of ORG and PERSON label), implying that instead of a small group of entities being very frequently mentioned, there were many entities mentioned often in the articles. Therefore, instead of limiting ourselves to any top-k entities, we chose to keep all.

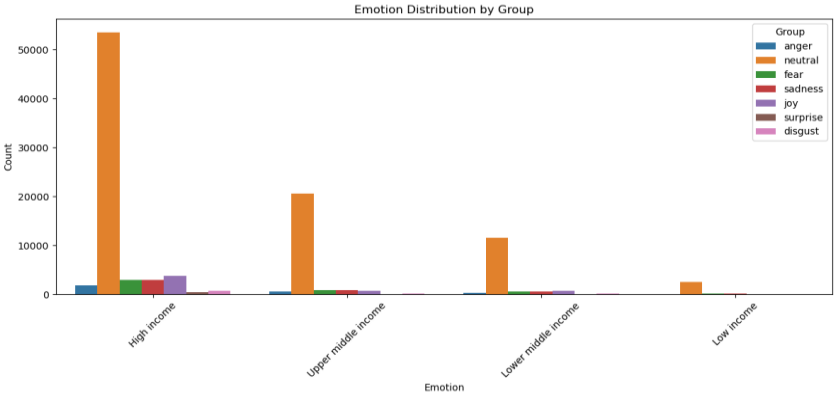
To aggregate these entities into groups that could correspond to 'vulnerable' and 'non-vulnerable', we incorporated data regarding countries' socioeconomic status. We used World Bank's 2024 classification of countries into income groups (low, lower middle, upper middle, and high income) and region groups (Europe, Middle East, etc.). These income and region groups were then used for our analysis.

Analysis & Results

The second step was extracting and generating features that could provide insight into bias or sensationalism. We could not train a model, as our data did not have labels of 'biased/unbiased' articles. Hence, we decided to extract linguistic elements for our analysis. We tried a number of methods, but not all were successful. These are summarized below:

Emotion Classification

We used an Emotion Text classifer found on HuggingFace for more fine-grained labels than simple sentiment analysis.

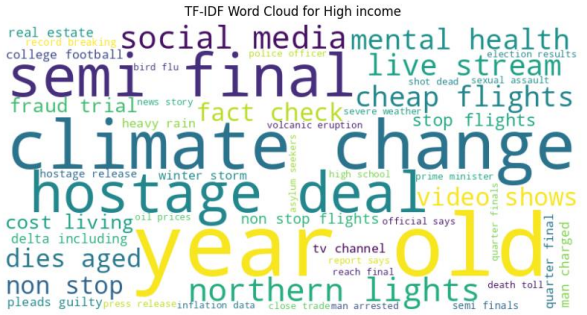


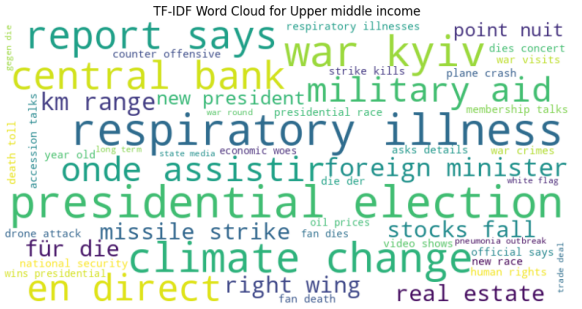
Once all articles were labelled, we visualized distribution of emotions for the income groups. Most articles for each group are neutral, and no group was overwhelmingly associated with any negative emotions. Hence, this implied that for our chosen groups, **the patterns of sentiment did not indicate disproportionate negativity towards vulnerable groups.**

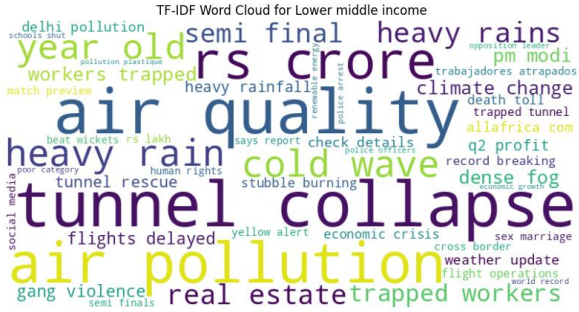
Most Common Words

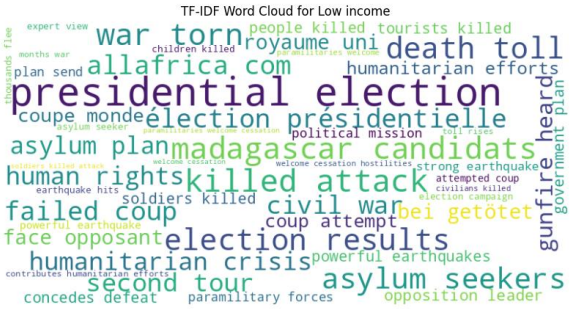
We looked at the words most associated with each group to see if any group was overwhelmingly associated with negative topics. We first extracted the useful content words from titles, removing stop words. Then we found the most common words for each entity. However, these were directly not useful since many consisted of the names of countries and places.

To solve this, we filtered the most common words to exclude place names and kept adjectives, verbs, and nouns. We first visualized unigrams, which did not show much difference in content between different groups. We then used bigrams and trigrams, which showed greater variations in topic, as words were contextualized. For these words, we computed TF-IDF scores across groups, so that our visualizations could use uniquely significant words for each group.









As we move from higher to lower income countries, there is a shift in the content being discussed.

- The phrases for higher income countries cover a wide range of topics, from climate change, mental health, sports, social media, to some social issues of illness and death.
- For lower income countries, the phrases overwhelmingly concern topics like war, crises, coups, violence, death, etc. In fact, it is difficult to find words that are neutral or positive.

While this difference in content words alone does not suggest a conscious effort by news media to negatively frame certain groups, it does show that negative topics are more commonly associated

with lower income countries. While it can be argued that this stems from these crises and violence being prevalent for lower income countries, it still creates two entirely different images in readers' minds regarding these groups.

Reflection on initial Mini Canvas

One of the greatest challenges during the project was choosing how to define our vulnerable groups as well as how to extract them from the data. We could have planned this better in our initial canvas. It also wasn't immediately clear what kind of modelling we would use for the data. The motivation behind our project was good. We believe we chose an important topic, and we already knew from our canvas which data sources to use. Data collection, preprocessing, and exploratory data analysis all went mostly as planned according to our canvas. We also knew initially that we would publish our results in the form of a blog. We believe this was the best way to communicate our results. However, we could not produce as many results as expected, and although we were able to find some patterns, namely a difference in word-use between income groups, the patterns were not as clear as we had hoped. We realized that detecting 'bias' or 'negative framing' was quite nuanced and could be hidden. A simple sentiment analysis model would not detect that, and perhaps the use of LLMs could aid in this task in the future. In the end, we deviated from the plan outlined in our canvas in several ways. Firstly, we used news headlines in our analysis, rather than full text articles. Secondly, we grouped entities based on income as opposed to the vulnerable groups we initially planned. In addition, we used an existing emotion classifier, rather than training our own. Moreover, the classifier was not a binary classifier as we initially planned to use. Lastly, we did not include any Finnish articles in the analysis. In the future, our project could be expanded by using full articles for analysis, as well as exploring alternative methods for modelling sentiment and vulnerable groups. These improvements could lead to more specific and interesting results.