SECU0057 Project: Exploring hate crime media coverage across major UK-based media outlets.

Jie Yi Yang Ye
UCL Department of Security and Crime Science

Abstract

This study investigates how UK media outlets frame hate crime through the analysis of 1581 articles across four major media organisations. Combining web scraping, text mining and machine learning techniques from SECU0057: Applied Data Science (2024/25). Reporting styles, sentiment, and readability are analysed to identify systematic differences in coverage. The results reveals that the outlets employ distinct editorial approaches when covering hate crime related incidents.

1 Introduction

The impact of media coverage on hate crime incidents can vary drastically depending on how a story is framed, the language used, and the depth of reporting. While responsible journalism can raise awareness and encourage constructive policy responses, sensational or biased coverage may reinforce harmful stereotypes or incite moral panic (Cohen, 2011).

Reis et al. (2015) conducted an experiment using different sentiment analysis methods on 69,907 headlines generated by The New York Times, BBC, Reuters, and Dailymail. They found that positive and negative headlines garnered greater interest than the headlines with more neutral tones. As most online news media outlets rely heavily on the revenues generated from the clicks made by their readers, hence news outlets are monetarily incentivised to make the tone of their headlines more polar, in other words, with stronger sentiments.

However, a neutral tone can create the illusion of merely reporting the truth, even while subtly creating a false narrative (Dijk, 1991). This effect can be achieved through the strategic use of numbers and figures despite the actual figures being vague or the scale of measurement left unclear. For instance, the Sun's editorial "1 in 5 Brit Muslims' sympathy for jihadis" was extremely misleading, as it only reveals in the last line of the column that only 1003 participants were interviewed which evidently is not representative of the 2.7 million Muslims in the UK at the time of publishing (Office for National Statistics, 2016). Sidoni (2018) highlights that this neutral tone of the article can be further reinforced by using quoted opinions from political figures, religious leaders or other entities that are deemed 'trustworthy' in the eyes of the public.

Using techniques acquired from SECU0057: Applied Data Science (2024/25), this project seeks to address the following research question: *How do major UK-based media outlets differ in their reporting of hate crime incidents, and what patterns emerge in terms of framing and sentiment?* Addressing this question is both important and relevant as media narratives play a powerful role in shaping public perception, and gaining insight into major news outlets' reporting patterns, editorial stances, and target audiences is essential in tackling hate crime.

2 Methodology

2.1 Data Collection

The news organizations selected for this project are The Independent, The Guardian, The Sun, and BBC (British Broadcasting Corporation). These outlets were chosen arbitrarily from a list of mainstream UK-based news sources. While this selection provides a representative snapshot, it is important to acknowledge that expanding the range of news outlets in future studies may yield more balanced insights, potentially reducing bias and enhancing the generalizability of the findings.

To locate articles relevant to the project, the search functionality on each news organization's official website was employed. The initial search query was designed to be comprehensive, incorporating a wide range of keywords associated with hate crimes:

“Hate crime” OR "Islamophobia" OR "anti-semitism" OR "homophobia" OR "transphobia" OR "racism" OR "sinophobia" OR "afrophobia" OR "xenophobia" OR "ableism" OR "misogyny" OR "anti-LGBTQ+" OR "anti-Asian racism" OR "anti-Indigenous" OR "anti-Roma" OR "anti-Catholic" OR "anti-Protestant" OR "anti-refugee" OR "anti-immigrant".

However, the search functionality on these sites were not equipped to process such queries. As a result, a simplified query, *“hate crime”*, was used to yield broader, more manageable results.

For each selected news outlet, RSelenium was used for entering the query into the site’s search bar and navigating through the resulting pages. Rvest was then used to scrape the URLs of the search results. Once the relevant HTML elements (such as the article title, publication date, and main body text) have been manually identified using the inspect option on the webpages, a loop was executed to navigate (RSelenium) to each of the URL’s webpages, extract the required content (Rvest) and save it to a dataframe.

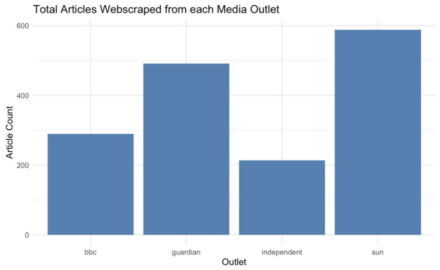


Figure 1: Bar plot for article count for each media outlet.

I managed to web scrape 289 articles from BBC, 491 articles from The Guardian, 213 articles from The Independent, and 588 articles from The Sun.

2.2 Data Mining

2.2.1 Word Frequency Analysis

To visualize which words dominate the coverage of each news outlet, word clouds are generated using the `textplot_wordcloud()` function from the quanteda.textplots package. The main body text of each article is pre-processed using the quanteda and dplyr packages: the text is tokenized, punctuation and stop words are removed, and the cleaned tokens are converted into a document-feature matrix (DFM). Word clouds are then created by applying `textplot_wordcloud()` to DFM subsets corresponding to each individual outlet.

2.2.2 Sentiment Analysis

The `sentiment_by()` function from the sentimentr package is used to calculate the overall sentiment scores of the articles. This package is preferred over syuzhet and textdata due to its ability to handle valence shifters such as negations and amplifiers. The function automatically splits the text into sentences, then into tokens, uses a built-in sentiment lexicon to assign sentiment scores to the tokens, and finally summarizes the sentence-level sentiment scores for each document.

To assess the emotional tone of news articles between outlets, the National Research Council Canada (NRC) Emotion Lexicon (Mohammad and Turney, 2013) is used to categorize words of each article into eight distinct emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

Firstly, the lexicon is converted into a dictionary format that associates each sentiment with the given list of words. The choice of the dictionary data structure makes the lookup of sentiments associated with the words significantly easier. The tokenized news corpus is then processed into a document-feature matrix (DFM) using the quanteda package. The NRC dictionary was then applied to this DFM, resulting in a new matrix where each document (article) is scored for the presence of each emotion based on the lexicon.

To account for variations in article length, the emotion scores were normalized by dividing the count of words associated with each emotion by the total word count for the respective article. This normalization was then scaled by a factor of 1000 to make the emotion scores more interpretable. This process ensures that longer articles do not dominate the emotion scores simply due to their length.

Finally, the normalized emotion scores for each article were aggregated by news outlets using the `group_by()` function. The average emotion scores for each outlet were calculated by taking the mean of the normalized scores across all articles from the same outlet.

2.2.3 Readability Complexity

To assess the complexity of language used by different news outlets, readability scores were calculated for all articles using the `textstat_readability()` function from the quanteda.textstats package. The resulting scores were grouped by news outlets to compute average readability levels, allowing for meaningful comparisons of writing complexity across sources. This provides insight into whether certain outlets tend to use more sophisticated or more accessible language in their reporting.

Three readability metrics were used to ensure a robust and comprehensive assessment: Flesch Reading Ease (based on syllable counts and sentence length), Flesch-Kincaid Grade Level (an education-level approximation), and Coleman-Liau Index (based on character density and sentence length).

2.3 Machine Learning

The k-means clustering algorithm was used to explore underlying patterns in how different news outlets frame hate crime incidents. The algorithm is chosen for its ability to handle large numerical datasets (Fahad et al., 2014) and is a good baseline model for unsupervised machine learning tasks.

Features used include readability scores, sentiment scores and word count. The goal is to identify natural groupings of articles that may reflect different styles and tones across media sources. To preprocess the data, readability scores, sentiment scores, and word count are normalised since k-means is sensitive to feature scale.

Finally, grid search, within the cluster sum of squares (WCSS) and elbow method (Zumel and Mount, n.d) is used to determine the optimal hyperparameters k for the algorithm.

3 Results

3.1 Word Frequency Analysis

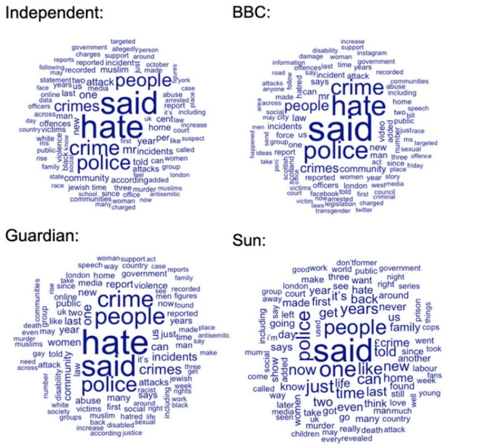
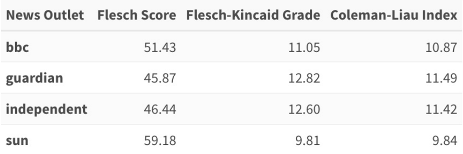


Figure 2: Word cloud for each media outlet

From the above word clouds, it is evident that all four news outlets, Independent, BBC, Guardian, and The Sun, share a core set of frequently used terms, including "people," "said," "police," and "crime." These commonalities suggest a consistent focus on individuals and law enforcement in discussions around hate crimes. However, there are some notable differences, such as the absence of religion-related terms such as "Muslim," "Jew," and "antisemitic" in The Sun's and BBC's word cloud, which are present in the rest. This may reflect differing editorial emphases or framing choices among the publications.

3.2 Readability Complexity



News Outlet	Flesch Score	Flesch-Kincaid Grade	Coleman-Liau Index
bbc	51.43	11.05	10.87
guardian	45.87	12.82	11.49
independent	46.44	12.60	11.42
sun	59.18	9.81	9.84

Figure 3: Average readability score for each outlet

The readability scores across the four news outlets reveal clear differences in text complexity (see figure 3). The Sun has the highest average Flesch Reading Ease score (59.18), suggesting its articles are the easiest to read. This is further supported by its significantly lower Flesch-Kincaid Grade Level (9.81) and Coleman-Liau Index (9.84), which indicate that The Sun's content requires a lower educational level to comprehend compared to the other outlets.

In contrast, The Guardian and The Independent exhibit similar readability profiles, with Flesch scores of 45.87 and 46.44, respectively, and higher Flesch-Kincaid (12.82 and 12.60) and Coleman-Liau scores (11.49 and 11.42). These results imply their articles are more complex and demand a higher reading proficiency. The BBC occupies a middle ground though still notably higher than The Sun.

3.3 Sentiment Analysis

Overall, the majority of articles exhibit negative sentiment scores, which aligns with the typically sombre nature of news reporting related to crime. However, the distribution and variability of sentiment differ notably among outlets.

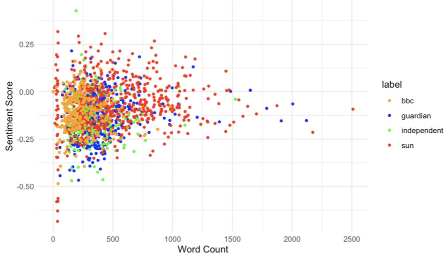


Figure 4: Scatter plot for document sentiment score against word count

The Sun and BBC display the greatest variation in sentiment scores, with The Sun spanning both the highest positive values (near 0.25) and strongly negative scores (below -0.25). This suggests a wider emotional range in their coverage, possibly reflecting a mix of sensationalist and neutral reporting. The BBC's scores, while also dispersed, cluster more moderately around neutral to slightly negative values. In contrast, The Guardian and The Independent show tighter clustering of sentiment scores, predominantly in the negative range (-0.25 to -0.50). This consistency may indicate a more uniform editorial tone, potentially tied to analytical or investigative content. Notably, no outlet demonstrates a clear relationship between article length (word count) and sentiment, as scores are distributed broadly across all word counts without a visible trend.

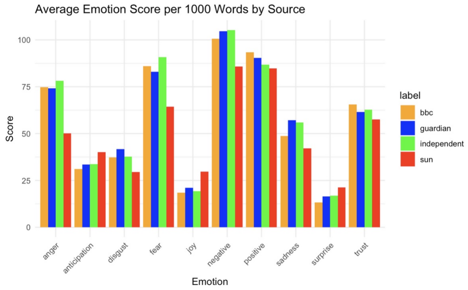
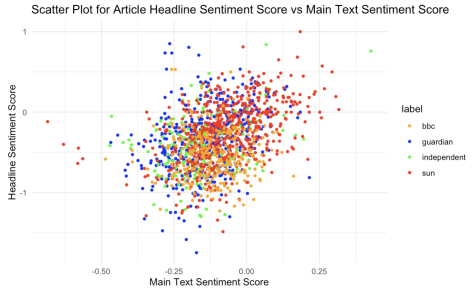


Figure 5: Bar plot for average emotion scores per 1,000 words in each media outlet.

Figure 5 presents the average emotion scores per 1,000 words for articles covering hate crime across four news outlets. All four outlets exhibit high scores for negative emotions (fear, anger, sadness), reflecting the distressing nature of hate crime reporting. Notably, positive emotions (joy, surprise) are consistently the least prominent across all outlets, underscoring the gravity of the subject matter. In addition, the emotion score for trust is also relatively high across all outlets, indicating that authoritative figures are frequently mentioned across all articles, this coincides with the observation from word cloud earlier.



x

Figure 6: Scatter plot for article headline sentiment score against content sentiment score

Figure 6 reveals a notable difference between headline and content sentiment, suggesting that a large majority of the article headlines from the dataset amplify the emotional tone of the main article content. This behaviour aligns with the findings of Reis et al. (2015), reflecting editorial strategies to attract readership, even when the article's actual content adopts a more neutral or nuanced stance.

3.4 K-means

The model is fit ten times for each k and the the mean WCSS is computed and plotted in figure 7.

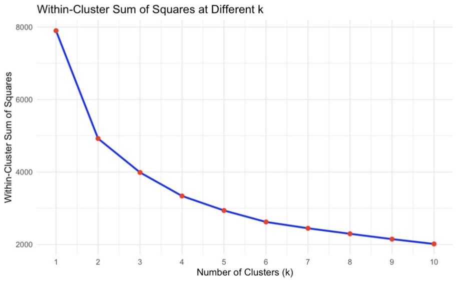


Figure 7: Line plot of WCSS against k.

Using the elbow method, k = 4 is determined to be the point of inflection for the k-means model fit using overall article sentiment score, readability score and article word count. Beyond k = 4, the marginal improvement in WCSS becomes minimal, indicating that additional clusters would likely lead to overfitting without providing substantial new insights into the underlying patterns in the data.

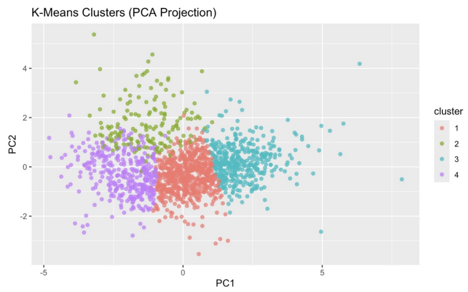


Figure 8: Principal Component projection of four-means result.

In figure 8, it can be observed that three clusters (1, 3, and 4) demonstrate clear separation along Principal Component 1 (PC1), suggesting these groups represent fundamentally different combinations of the features (sentiment, readability, and word count). Cluster 2 shows greater overlap with both clusters 1 and 4 along PC1, indicating more similarity in their underlying characteristics. This partial separation implies that while clusters 1, 3, and 4 represent distinct article types, cluster 2 may capture intermediate cases between the other clusters.

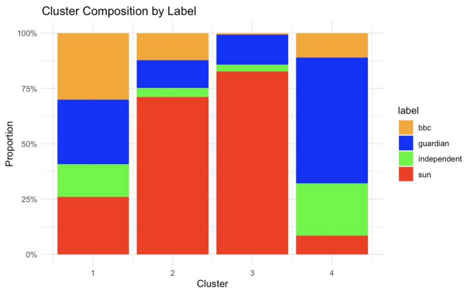
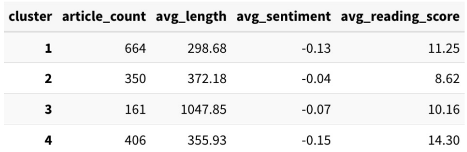


Figure 9: Bar plot showing cluster proportion with respect to media outlet.

Furthermore it can be observed in figure 9, that cluster 1 emerges as the most heterogeneous group, containing articles from all four outlets in relatively balanced proportions. This suggests these articles share common features that transcend outlet-specific styles, possibly representing standard news reporting that follows conventional journalistic practices across the media landscape. Notably, the BBC shows limited representation in Cluster 3. In addition, it can also be observed that clusters 2 and 3 are primarily composed of articles from The Sun. Cluster 4 consists mainly of articles from The Guardian and The Independent.



cluster	article_count	avg_length	avg_sentiment	avg_reading_score
1	664	298.68	-0.13	11.25
2	350	372.18	-0.04	8.62
3	161	1047.85	-0.07	10.16
4	406	355.93	-0.15	14.30

Figure 10: Summary table of cluster features.

The largest cluster, 1 (664 articles), represents a professional standard across all outlets, featuring moderately short articles (299 words average) with slightly negative sentiment and academic-level readability (Flesch-Kincaid 11.25). This suggests most hate crime related articles follows conventional journalistic norms. Cluster 2 (350 articles), dominated by The Sun, shows notably different characteristics, shorter articles with near-neutral average sentiment and significantly lower reading difficulty (8.62), reflecting the tabloid's emphasis on accessible content. The small but distinctive Cluster 3 (161 articles), also Sun-dominated, contains unusually long-form articles (1,048 words average) with intermediate readability, likely representing special reports or investigations. In contrast, Cluster 4 (406 articles) consists primarily of The Guardian and Independent featuring sophisticated language (14.30) and the strongest negative sentiment, indicating more analytical, emotionally charged coverage.

4 Discussion and Limitation

4.1 Methodological Constraint

While word clouds offer an intuitive visualisation of term frequency, they lack contextual nuance. Similarly, the sentimentr package, though efficient, may misclassify sentiment in cases of sarcasm or irony, a known challenge in lexicon-based sentiment analysis (Liu, 2012). This can be remedied by incorporating transformer-based models, like BERT (Devlin et al., 2019), in the sentiment analysis process. The readability metrics used in this project also present limitations, as these were designed for the context of educational text decades ago (Benoit et al., 2019) rather than journalistic or politically charged content. Modern alternatives, such as cohesion-based metrics (Graesser et al., 2011), might better capture the complexity of media language. The use of the elbow method for determining optimal clusters in k-means, while practical, is subjective. Hence, complementary metrics like the silhouette score could enhance robustness.

4.2 Dataset Constraints

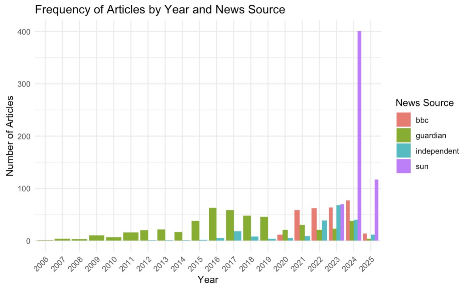


Figure 11: Bar plot for articles frequency by publishing year and news outlet.

Figure 11 shows that the dataset exhibits notable imbalances in article distribution across outlets and also in temporal coverage. Since some outlets are overrepresented compared to others, the analysis may inadvertently amplify the linguistic patterns of certain publications while under-representing others. Additionally, the lack of a uniform timeframe means that external events, such as reforms, crises, could disproportionately influence the sentiment or readability of certain outlets, confounding direct comparisons. Future work could mitigate this by ensuring a balanced sample collected within a more fixed timeframe.

Furthermore, the project focuses on a subset of outlets and topics, which may limit broader conclusions about media media reporting patterns. For instance, other regional publications might exhibit different patterns.

5 Conclusion

This project revealed key differences in how UK-based media outlets report on hate crimes, with The Sun favouring accessible, emotionally varied language, while The Guardian and The Independent employed more complex, consistently negative framing. Sentiment and readability analyses highlighted distinct editorial styles, and clustering identified outlet-specific reporting patterns. Despite methodological limitations, the findings underscores a significant difference in editorial approaches across media outlets when covering hate crime.

Future work could enhance robustness and generalisability of findings with transformer-based models, balanced datasets, and expand the scope with longitudinal analysis to further reporting trends.

References

Kenneth Benoit, Kevin Munger, and Arthur Spirling. 2019. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2):491–508.

British Broadcasting Corporation. Bbc news. Public service broadcaster.

Stanley Cohen. 2011. *Folk Devils and Moral Panics*, 1st edition. Routledge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Teun A. van Dijk. 1991. *Racism and the press*. Critical studies in racism and migration. Routledge, London ; New York.

Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras. 2014. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(03):267–279. Publisher: IEEE Computer Society.

Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. 2011. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, 40(5):223–234. Publisher: American Educational Research Association.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Office for National Statistics. 2016. 2011 UK censuses.

Julio Reis, Fabrício Benevenuto, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news.

MG Sidoni. 2018. Direct hate speech vs. indirect fear speech. A multimodal critical discourse analysis of the Sun's editorial "1 in 5 Brit Muslims' sympathy for jihadis". *Lingue Linguaggi 28 (2018), 267-292*.

The Guardian. The guardian. Online news publication.

The Independent. The independent. Online news publication.

The Sun. The sun. Online news publication.

Nina Zumel and John Mount. n.d. Practical Data Science with R.