Latent Semantic Analysis and Sentiment Analysis

of BBC News Corpus Data

Nino Miljkovic

June 22, 2025

ABSTRACT

This paper presents an analytical investigation into a large-scale news corpus using two complementary text mining methodologies: Latent Semantic Analysis (LSA) and Sentiment Analysis. Applied to a dataset of approximately 39,700 BBC news article descriptions sourced from Kaggle, the study employs Singular Value Decomposition (SVD) with Term Frequency-Inverse Document Frequency (TF-IDF) weighting to identify latent thematic structures within the corpus. Four dominant topic clusters are identified, namely politics, sports, Premier League football, and the conflict in Ukraine. Subsequent sentiment analysis reveals that the corpus exhibits a slight net positive skew, with sports coverage generating predominantly positive affect and conflict reporting generating predominantly negative affect. These findings demonstrate the complementary utility of LSA and sentiment analysis in characterizing both the topical and emotional dimensions of large unstructured text corpora.

1. INTRODUCTION

The exponential growth of digital news media has produced vast repositories of unstructured textual data, presenting both opportunities and methodological challenges for researchers interested in understanding patterns of media coverage, public discourse, and editorial tone. Automated text mining techniques offer scalable solutions to these challenges, enabling researchers to extract meaningful structure from corpora that would be impractical to analyze by hand.

This study applies two well-established natural language processing techniques to a corpus of BBC news article descriptions. The first technique, Latent Semantic Analysis (LSA), is an unsupervised dimensionality reduction approach that reveals the hidden thematic architecture of a text corpus by decomposing a high-dimensional term-document matrix into a lower-dimensional latent semantic space. The second technique, Sentiment Analysis, assigns quantitative affective scores

to documents on the basis of their constituent vocabulary, enabling a systematic characterization of the emotional valence associated with different news topics.

The primary objectives of this investigation are twofold: (1) to uncover the dominant latent themes present in BBC news coverage through LSA and SVD, and (2) to quantify and interpret the sentiment profiles associated with those themes. Together, these analyses provide a richer, multi-dimensional understanding of the corpus than either technique could yield in isolation.

2. DATASET

Source: The dataset was obtained from Kaggle, a publicly accessible platform for data science resources, and is available at https://www.kaggle.com/datasets/gpreda/bbc-news.

Dataset Name: bbc_news.csv

Description: The BBC News dataset was compiled by data scientist Gabriel Preda through automated collection of BBC news articles via their RSS feeds, employing the *requests_html* library and BeautifulSoup for web scraping. The dataset is regularly updated and represents a broad cross-section of BBC editorial output across multiple topic domains.

The dataset is structured as a CSV file containing approximately 39,700 entries, each corresponding to a distinct news article. Each entry comprises five fields: a headline title, a publication date, a globally unique identifier (GUID), a URL linking to the full article, and a brief textual description of the article's content. The description field, which constitutes the primary analytical unit of this study, provides rich, naturally occurring prose suitable for text mining applications including LSA and sentiment classification.

The copy of the dataset is provided as a supplementary file accompanying this report.

3. METHODOLOGY

3.1 Data Loading and Preprocessing

The CSV file was downloaded, inspected for structural integrity, and subsequently loaded into JMP Pro for analysis. The description column, containing the primary unstructured text, was imported into JMP Pro's Text Explorer module, with the GUID column assigned as the document identifier, as shown in Figure 1.

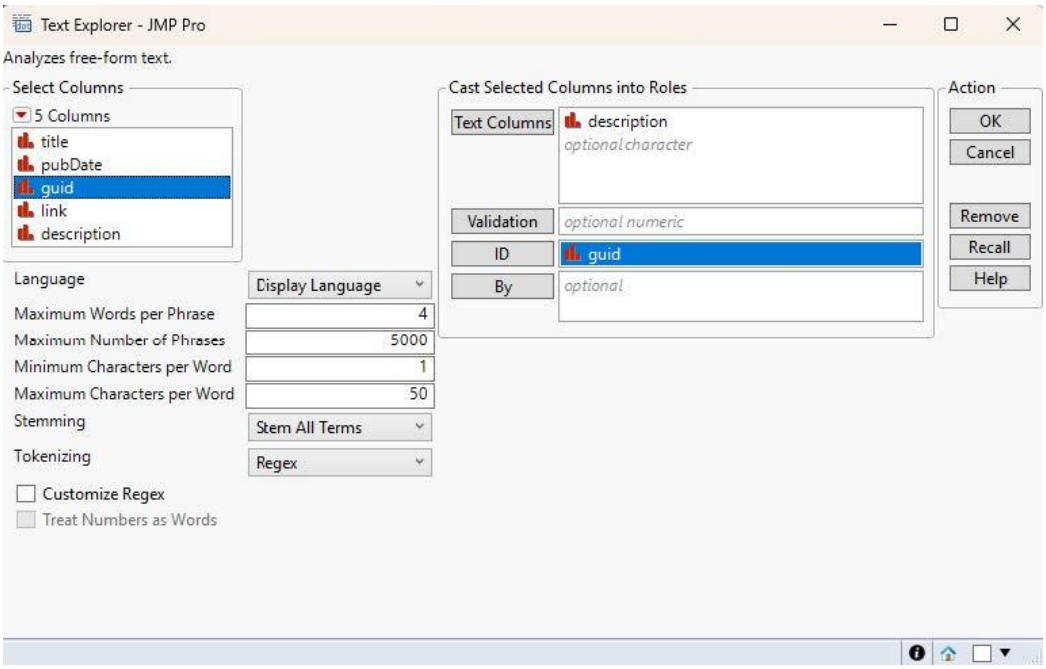

Figure 1: JMP Pro Text Explorer configuration: description selected as Text Column, guid as ID, Stem All Terms stemming applied

The following preprocessing pipeline was applied sequentially:

- **Tokenization:** Text was automatically tokenized into individual terms and multi-word phrases using JMP's built-in regular expression tokenizer. This process segmented raw prose into discrete analytical units suitable for statistical modeling.
- **Stop Word Removal:** Common English function words were filtered out using JMP's built-in stop word lexicon, enabled alongside the built-in phrases option, as shown in Figure 2.
- **Stemming:** The Stem All Terms algorithm was applied, reducing inflected word forms to their morphological root (e.g., "running" becomes "run"). Stemming consolidates semantically related term variants, improving the coherence of resulting term clusters and reducing effective vocabulary size.
- **Punctuation and Special Character Removal:** All punctuation marks and non-alphabetic characters were stripped during the tokenization phase, ensuring a clean token set for subsequent analysis.

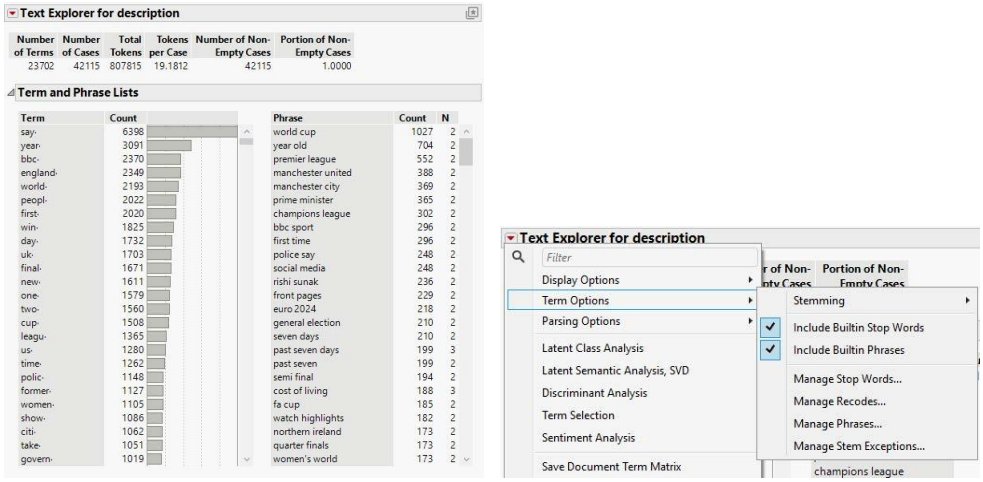

Figure 2: Text Explorer output showing term and phrase frequency lists (left) and Term Options configuration with built-in stop words and phrases enabled (right)

It should be noted that lemmatization was not applied in this study. While lemmatization is generally considered more linguistically precise than stemming, stemming was deemed sufficient for the dimensionality reduction objectives of this LSA task. Upon completion of preprocessing, the Text Explorer reported 23,702 unique terms across 42,115 cases, with an average of approximately 19.18 tokens per case. The prevalence of sports and political terminology in the highest-frequency vocabulary slots provided an early indication of the major thematic domains within the corpus.

3.2 Latent Semantic Analysis via Singular Value Decomposition

Following preprocessing, Latent Semantic Analysis was conducted by applying Singular Value Decomposition to the term-document matrix. SVD decomposes the TF-IDF weighted term-document matrix X into three matrices: X = UΣV^T, where U contains the left singular vectors (term vectors), V the right singular vectors (document vectors), and Σ is a diagonal matrix of singular values. JMP Pro was configured to extract 100 singular vectors with TF-IDF weighting, and the data were centered and scaled to highlight latent patterns, as shown in Figure 3.

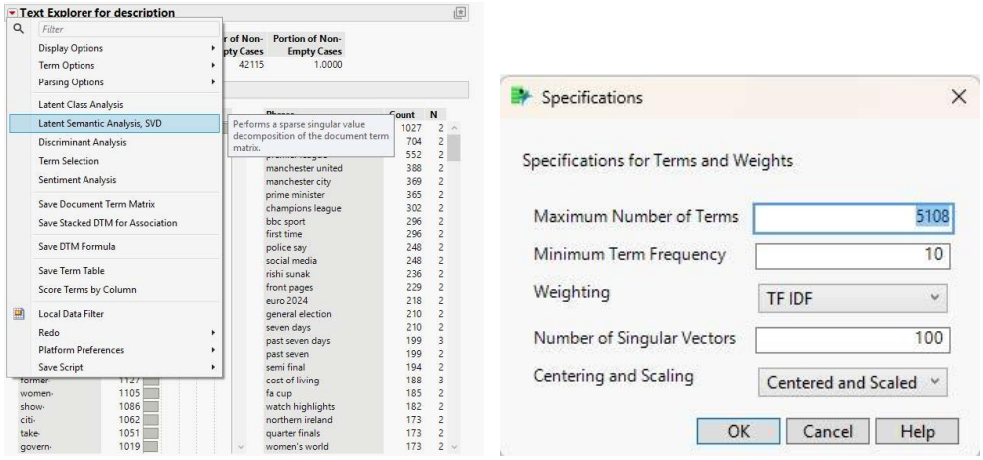

Figure 3: Latent Semantic Analysis SVD menu selection (left) and specifications dialog showing 100 singular vectors, TF-IDF weighting, and Centered and Scaled normalization (right)

The resulting SVD output was visualized as two scatter plots shown in Figure 4: a document plot (Doc Singular Vectors 1 and 2) and a term plot (Term Singular Vectors 1 and 2). The left plot depicts documents clustered around similar themes, while the right plot visualizes the distribution of terms along those same dimensions, where densely clustered terms indicate shared context.

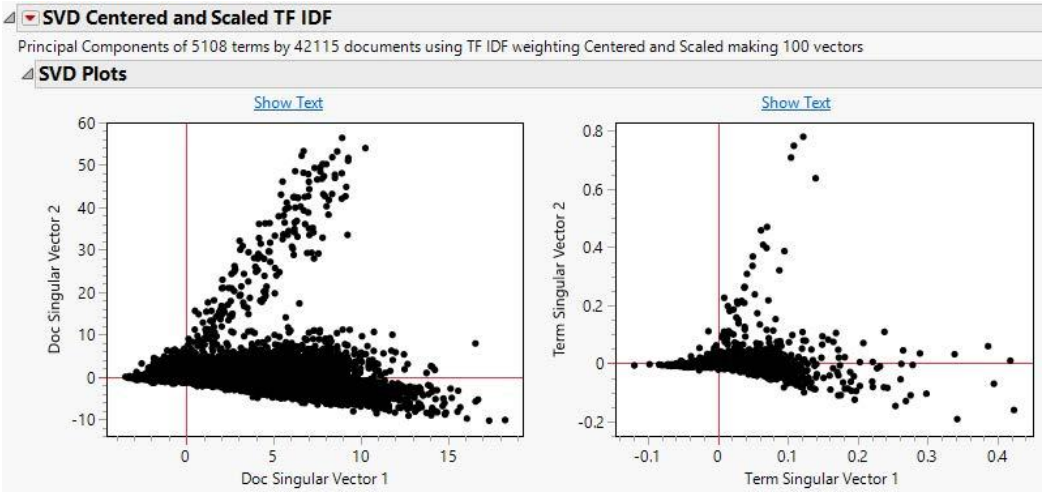

Figure 4: SVD document plot (left) and term plot (right): principal components of 5,108 terms by 42,115 documents using TF-IDF weighting, centered and scaled, with 100 vectors

4. RESULTS AND INTERPRETATION

4.1 Latent Semantic Analysis Results

By reducing dimensionality, the SVD plots highlight distinct topic groups, which can be revealed by highlighting quadrants and using the Show Text function to examine the associated textual data. Four dominant latent themes were identified through this process.

1. **Politics (Global and Local):** Articles clustered in the top-left quadrant of the document plot were found to predominantly concern political topics, including domestic UK governance, international diplomatic relations, and policy debates. As shown in Figure 5, representative article descriptions in this region referenced political figures such as Boris Johnson and policy themes such as refugee legislation, economic sanctions, and cost-of-living interventions.

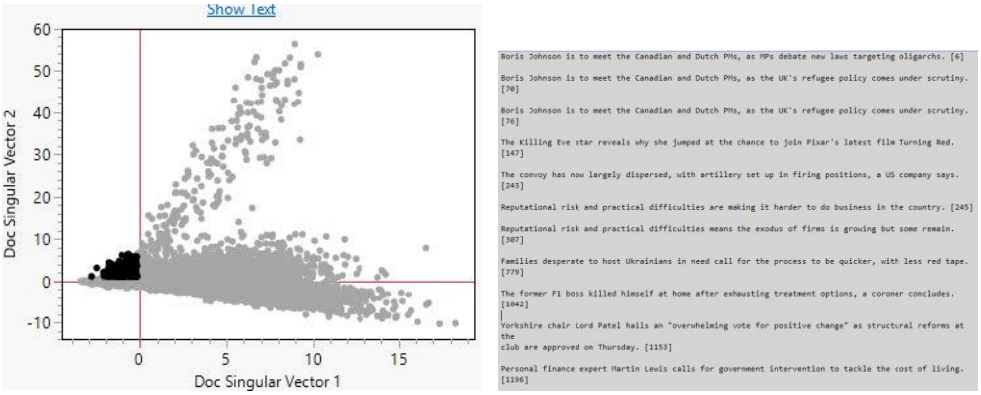

Figure 5: Top-left quadrant selection in the SVD document plot (left) and Show Text output revealing political article content (right)

2. **Sports (General and Formula One):** The top-right quadrant contained a heterogeneous collection of sports articles covering disciplines including tennis, rugby, Formula One, and the Paralympic Games (Figure 6). A distinct sub-cluster within this quadrant dealt exclusively with Formula One racing (Figure 7), with descriptions referencing drivers such as Max Verstappen and Charles Leclerc. The internal coherence of this sub-cluster illustrates SVD's capacity to recover hierarchically structured thematic organization.

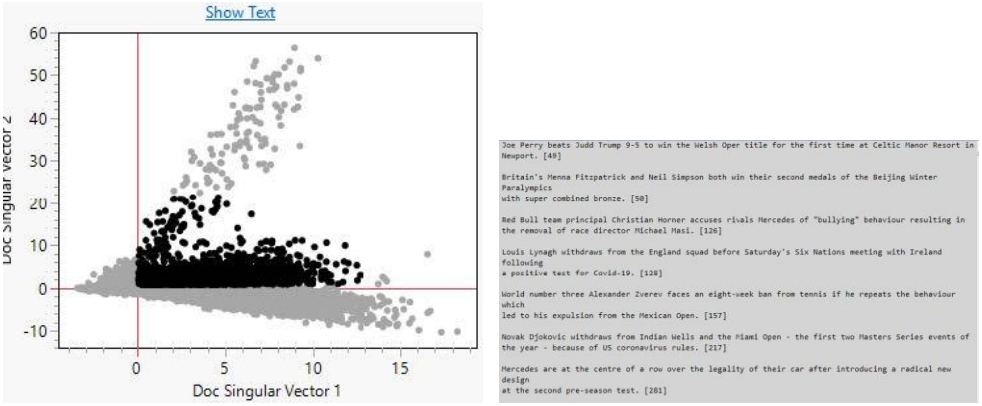

Figure 6: Top-right quadrant selection showing general sports articles (left) and Show Text output (right)

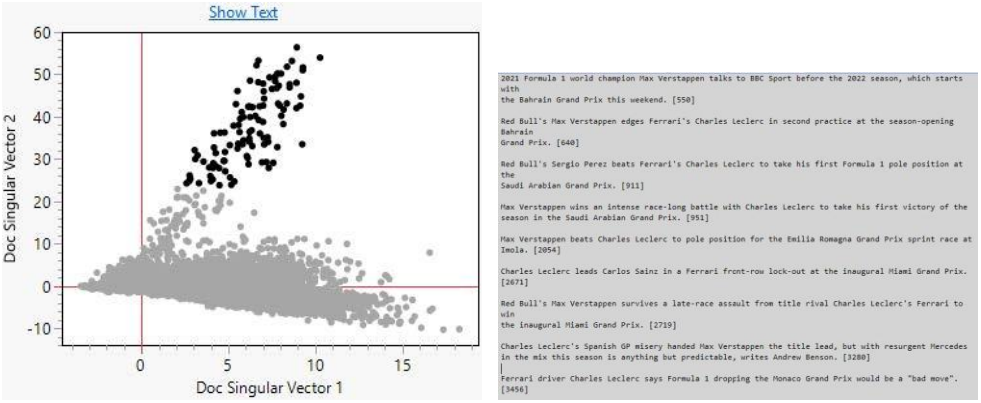

Figure 7: Formula One sub-cluster within the top-right quadrant (left) and Show Text output confirming exclusive Formula One coverage (right)

3. **Premier League Football:** Articles in the bottom-right quadrant shared a specific focus on association football, predominantly Premier League coverage (Figure 8). Descriptions referenced clubs including Manchester United, Manchester City, and Arsenal, as well as match analyses, managerial commentary, and transfer-related news.

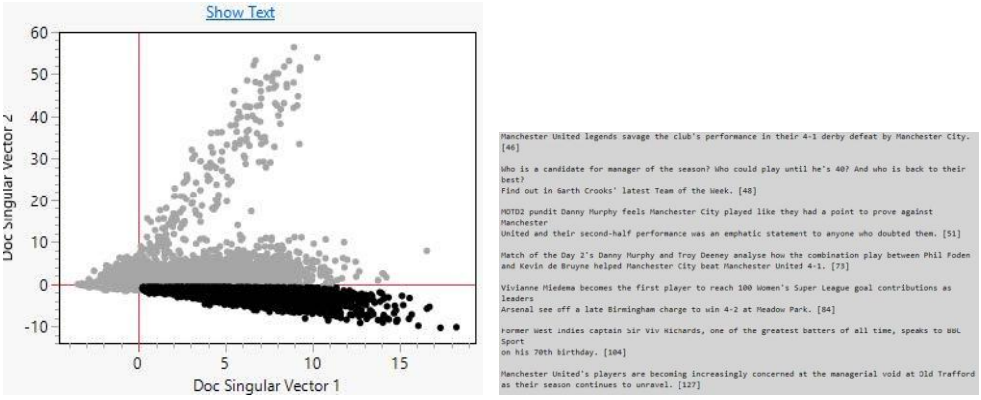

Figure 8: Bottom-right quadrant selection (left) and Show Text output revealing Premier League football content (right)

4. **Conflict in Ukraine:** The bottom-left quadrant contained articles unified by their coverage of the Russian invasion of Ukraine (Figure 9). Descriptions referenced civilian evacuations, military activity in cities including Kyiv and Kharkiv, and the humanitarian consequences of the conflict. The thematic coherence of this cluster was particularly pronounced, reflecting the intensity and specificity of the BBC's Ukraine-related reporting during the period covered by the dataset.

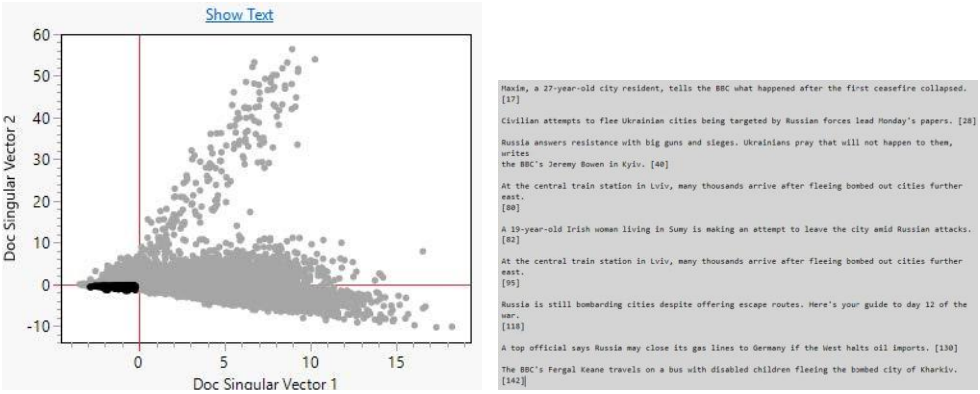

Figure 9: Bottom-left quadrant selection (left) and Show Text output confirming Ukraine conflict article content (right)

After examining all four quadrants, the four major themes in the BBC corpus can be established as: (1) Politics (global and local), (2) Sports including a Formula One sub-cluster, (3) Premier League and other football news, and (4) the Conflict in Ukraine. The SVD Term Plot quadrants further depict these four themes by having vocabulary related to each theme clustered in the corresponding quadrant. By clearly revealing these notable theme groups, Latent Semantic Analysis has helped reduce dimensionality and improve interpretability, thus setting the stage for Sentiment Analysis.

4.2 Sentiment Analysis Results

Sentiment Analysis was conducted using JMP Pro's built-in system that calculates net sentiment scores for documents. The full Sentiment Summary output is shown in Figure 10.

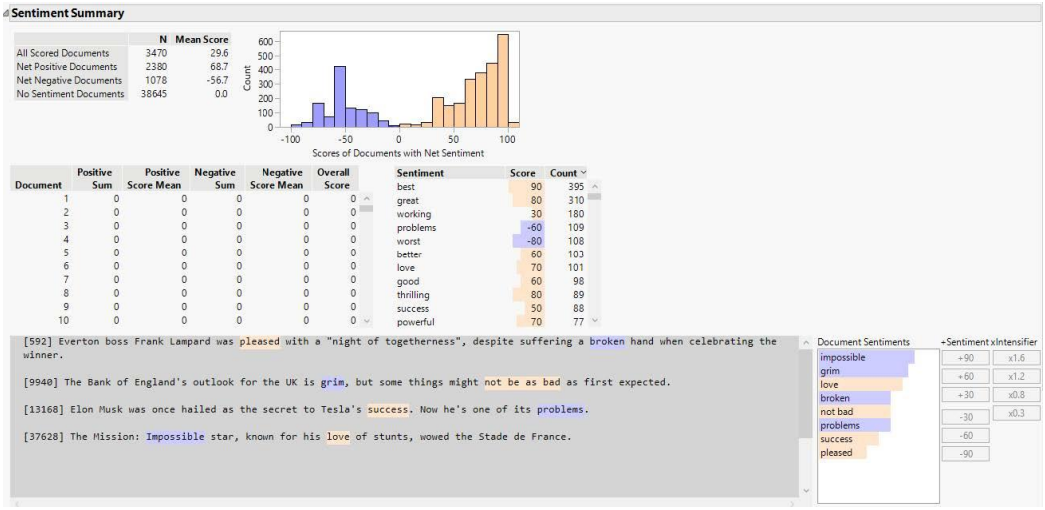

Figure 10: Sentiment Summary output: histogram of net sentiment scores, document-level scores table, and sample documents with highlighted sentiment terms

Of the 42,115 documents in the corpus, 3,470 received a non-zero sentiment score. Of these, 2,380 documents (68.6%) were assigned a net positive sentiment, while 1,078 (31.1%) received a net negative sentiment. The mean net positive score was 68.7 and the mean net negative score was -

56.7. The sentiment score distribution exhibited a slight rightward (positive) skew, with the most frequent positive scores in the +60 to +90 range and the negative peak at approximately -60.

The most frequently occurring sentiment-bearing terms are presented in Figure 11. Words like "best", "great", "better", and "love" dominate the positive category, while "problems" and "worst" are the most frequent negative terms.

Sentiment	Score	Count
best	90	395
great	80	310
working	30	180
problems	-60	109
worst	-80	108
better	60	103
love	70	101
good	60	98
thrilling	80	89
success	50	88
powerful	70	77

Figure 11: Top sentiment terms ranked by score and document count

By selecting bins of the histogram, different portions of the sentiment distribution can be investigated in greater detail. The net positive peak bins are examined first (Figure 12), followed by the net negative bins (Figure 13).

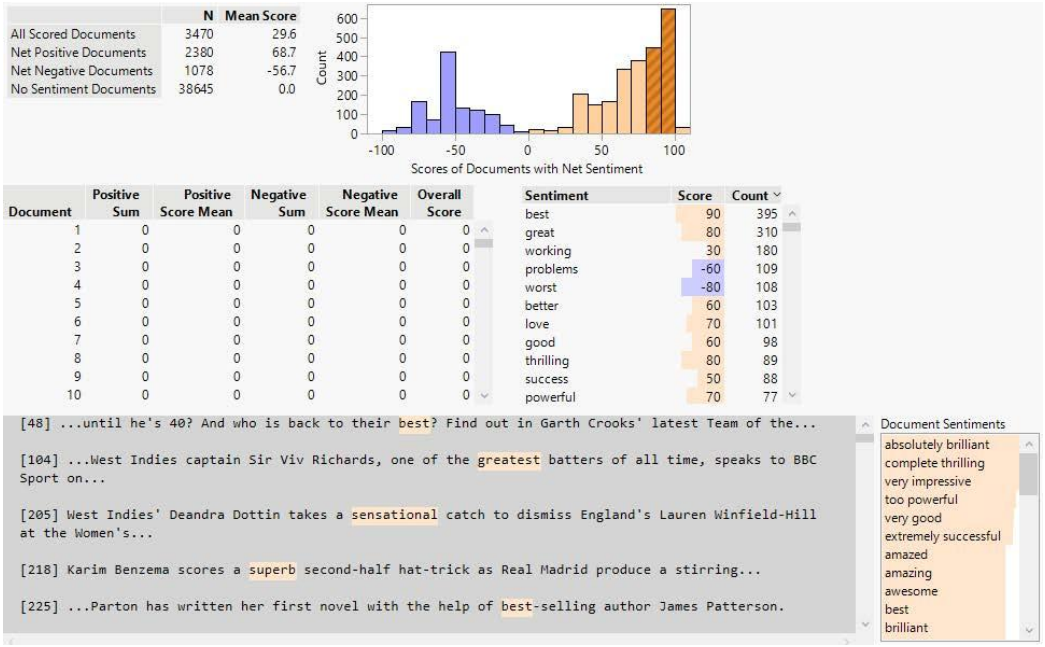

Figure 12: Net positive peak bin selection: highlighted histogram bins, sample documents with highlighted positive terms, and Document Sentiments panel showing intensified positive vocabulary

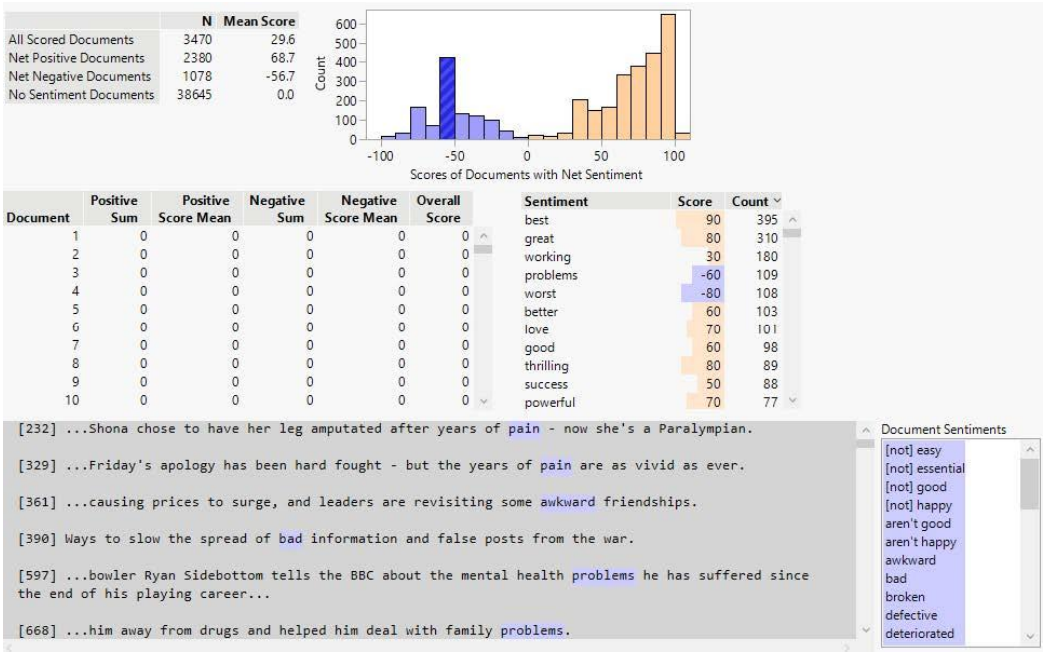

Figure 13: Net negative bin selection: highlighted histogram bins, sample documents with highlighted negative terms, and Document Sentiments panel showing negation constructions and negative vocabulary

4.3 Theme-Level Sentiment Profiles

A particularly illuminating aspect of the analysis involved cross-referencing the thematic clusters identified through LSA with their corresponding sentiment profiles. Two clusters were examined in detail.

Sports Cluster: Figure 14 shows the sports cluster selected in the SVD plot alongside its corresponding sentiment histogram. The distribution exhibits a clear rightward skew, confirming that sports reporting carries predominantly positive affective valence. The Document Sentiments panel (Figure 15) lists characteristic positive descriptors such as "so great", "absolutely brilliant", "excellent", and "sensational". Negative sentiment terms, while present, are substantially outnumbered by positive ones.

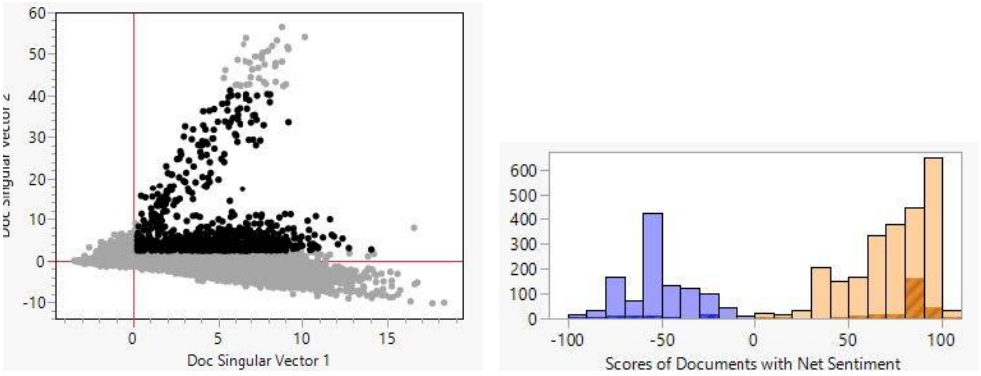

Figure 14: Sports cluster selected in the SVD document plot (left) and corresponding sentiment histogram showing positive skew (right)

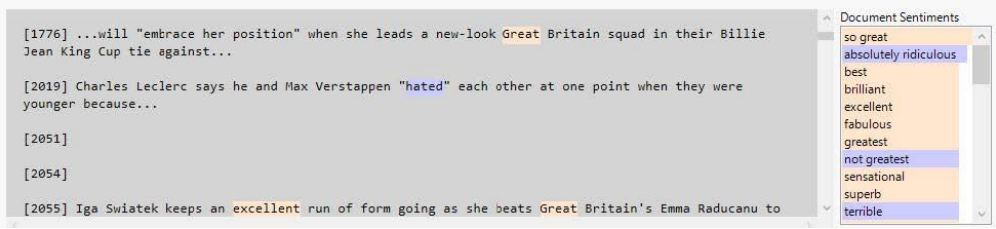

Figure 15: Document Sentiments panel for the Sports cluster, listing characteristic positive vocabulary including "so great", "absolutely brilliant", and "sensational"

Ukraine Conflict Cluster: In marked contrast, Figure 16 shows the Ukraine conflict cluster alongside its sentiment histogram, which is dominated by the negative range. This cluster accounted for a disproportionately large share of the corpus's total negative sentiment. Frequently flagged negative terms included "pain", "awkward", "bad", "broken", and negation constructions such as "not happy" and "not good". Cases of positive sentiment were present but comparatively rare, likely corresponding to articles reporting on diplomatic progress or individual acts of resilience within the broader conflict narrative.

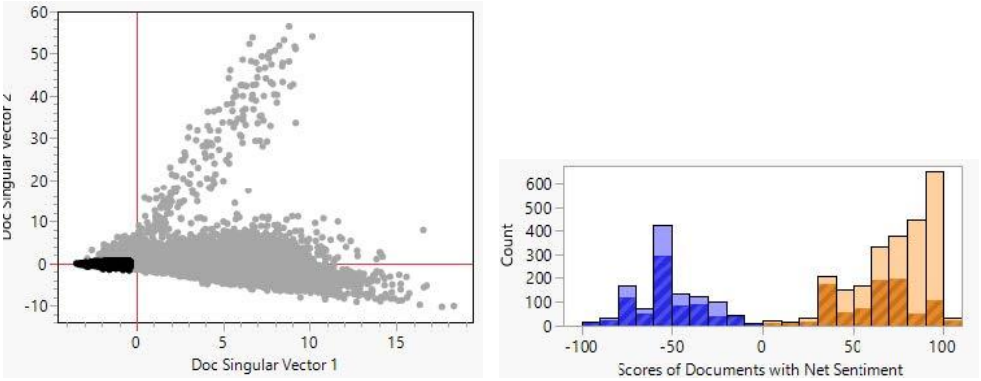

Figure 16: Ukraine conflict cluster selected in the SVD document plot (left) and corresponding sentiment histogram showing negative skew (right)

5. DISCUSSION

The results of this investigation illustrate the complementary strengths of LSA and sentiment analysis as components of a unified text mining pipeline. LSA, through its application of SVD to TF-IDF weighted term-document matrices, successfully uncovered four substantively meaningful and interpretable theme clusters within a corpus of nearly 40,000 news articles. The identification of these clusters without recourse to any labeled training data underscores the utility of unsupervised dimensionality reduction as an exploratory tool in large-scale media analysis.

The sentiment analysis component enriched this thematic structure with affective information, revealing that the emotional tone of BBC coverage is not uniform across topic domains. Sports journalism carries a strong positive affective signature, consistent with its audience-facing, narrative-driven conventions. Conflict journalism, by contrast, carries a predominantly negative

one, as would be expected from reporting on humanitarian crisis and armed conflict. These findings align with prior research on sentiment variation across news genres.

Several methodological considerations merit reflection. The reliance on JMP Pro's built-in lexicon for sentiment scoring introduces a degree of opacity into the classification process, as the specific sentiment lexicon and scoring algorithm are not fully documented. Lexicon-based sentiment analysis is also known to be sensitive to domain-specific vocabulary, irony, and negation structures, all of which are frequently present in news prose. Future work might profitably compare JMP's built-in scorer with established lexica such as VADER, LIWC, or AFINN to assess the robustness of the findings. Additionally, the choice to apply stemming rather than lemmatization, while pragmatically justified, may introduce minor inconsistencies in term conflation for morphologically complex words.

6. CONCLUSION

This study has demonstrated that the combined application of Latent Semantic Analysis and Sentiment Analysis constitutes a powerful and efficient framework for exploring large unstructured news corpora. Applied to a corpus of approximately 39,700 BBC news article descriptions, LSA identified four coherent latent themes corresponding to politics, general sports (with a Formula One sub-cluster), Premier League football, and the conflict in Ukraine. Sentiment analysis subsequently revealed that these themes carry distinct and interpretable affective profiles, with sports coverage trending positive and conflict coverage trending negative.

The two techniques proved mutually reinforcing: the thematic structure recovered by LSA provided a principled basis for interpreting sentiment variation, while the sentiment profiles added an additional interpretive dimension to the latent clusters. This analytical synergy suggests that combined LSA-sentiment pipelines represent a productive direction for future research in computational journalism, media studies, and information retrieval.

Future work could extend this analysis through the application of more granular topic models such as Latent Dirichlet Allocation, the use of transformer-based sentiment classifiers fine-tuned on news data, and a longitudinal analysis of how topic prevalence and sentiment have evolved over the timespan covered by the dataset.

REFERENCES

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188-230.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Preda, G. (2023). *BBC News dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/gpreda/bbc-news