HW6 – Text Mining

Fall 2025

MUSA 5000

Jun Luu

Alex Stauffer

Tessa Vu

Introduction

This report examines a corpus of news articles from the Reuters-21578 dataset, one of the most widely used benchmark collections for text classification and sentiment analysis research (Malvarez, 2015). It contains news stories from 1987 and covers a variety of topics within the business, economics, and trade domains. Understanding the overarching themes and language patterns with text and sentiment analysis provides valuable insight into historical and/or current financial events that are framed and communicated.

Beyond the academic context, news media shapes public perception of economic conditions, influences behaviors from investors and consumer spending, as well as the fact that it reflects broader social, political, and other concerns related to quality of life. Text mining techniques help uncover additional patterns to manual reading, such as preprocessing text data, visualizing frequent terms from word clouds, and sentiment analysis to understand the complex emotions related to news coverage during this specific period.

Methods

Of the many text mining techniques available, this report practices corpus creation, text preprocessing, stop words removal, stemming and lemmatization, and document term matrix (DTM), creating word clouds, conducting sentiment analysis, and k-means clustering.

The Reuters-21578 dataset was imported into a Quarto (.qmd) environment, and the analysis was restricted to the first 100 documents to satisfy and explore a little beyond assignment requirements while maintaining computational efficiency. These documents were aggregated into a corpus, defined as a structured collection of text documents suitable for text mining analysis. To verify corpus integrity and structure, text from a single document was inspected prior to analysis.

Text preprocessing was conducted to standardize and clean the corpus prior to analysis. All text was converted to lowercase to ensure consistency throughout the 100 documents. Four cleaning functions were utilized to clean the corpus: (1) transform characters such as "@" or "/" into spaces if conjoined with words, (2) the removal of apostrophe, (3) removal of numbers, and (4) removal of punctuation of the corpus. Next, stopwords are removed. Stopwords include common English articles (e.g., "the," "a"), conjunctions (e.g., "and"), and frequently occurring function words (e.g., "to," "of") that carry little semantic meaning. Here is an additional list of stop words that were removed that held little analytical value: "said", "says", "also", "reuter", "reuters", "will", "can", "may", "one", "two", "three", "four", "five", "six", "seven", "eight", "nine", "ten", "mln", "dlrs", "pct", "cts", "year", "years", "month", "months", "week", "weeks", "day", "days", "inc", "corp", "ltd", "company", "companies". Extra whitespace is removed from the documents.

Stemming was applied to reduce words to their common root forms, consolidating different grammatical tenses into a single term. For example, the words "running," "ran," and "runs" were all reduced to the root form "run." This process reduces vocabulary size and improves the interpretability of term frequency patterns by treating related word forms as equivalent. Additionally, stemming removed common word suffixes and endings like "-es", "-ed", and "-ing".

Following preprocessing, a document-term matrix (DTM) was constructed. A DTM is a representation of how frequently different terms appear in each of the documents. To examine dominant vocabulary across the corpus, term frequency analysis was conducted using the document–term matrix. Term frequencies were aggregated by summing the occurrence of each term across all documents, producing corpus-level frequency counts. These counts were ranked in descending order to identify the most frequently occurring terms, which informed subsequent visualizations and descriptive analysis. A histogram was generated to show the distribution of term frequency across all terms. A word cloud was generated to show the term that appears the most, with the higher frequency using a larger font.

Sentiment analysis was conducted using a lexicon-based approach implemented through the syuzhet R package. Sentiment lexicons consist of predefined word lists annotated with sentiment polarity or emotional categories. Four lexicons are available within the package: NRC, AFINN, Bing, and Syuzhet (Jockers). The NRC lexicon was selected as the primary sentiment framework due to its ability to capture multiple emotional dimensions beyond simple positive and negative polarity (anger, joy, fear, sadness, etc.). NRC sentiment scores were computed for all terms in the corpus and combined with term frequency data, classified as "negative", "neutral", or "positive", and then visualized. Sentiment scores were weighted by term frequency to estimate the overall emotional composition of the corpus. For comparison and validation purposes, sentiment classifications derived from the NRC lexicon were compared against results from the AFINN, Bing, and Syuzhet lexicons, with terms categorized as negative, neutral, or positive.

The final component of the analysis involved text clustering using k-means methodology. To improve clustering performance and reduce noise, sparsely occurring terms were removed from the DTM. Given the high sparsity of the matrix, only terms appearing in at least 5% of documents were retained. Columns with zero variance were also removed. The elbow method, based on within-cluster sum of squares (WCSS), was used to identify an appropriate number of clusters, with values of k e evaluated from 1 to 10. Higher WCSS values indicate that documents are far from their respective cluster centers, resulting in less coherent or "messier" clusters, while lower WCSS values indicate that documents are closer to their cluster centers, producing tighter and more internally consistent groupings. Based on the elbow plot, $k = 4$ was selected. K-means clustering was then performed on the reduced matrix. A data frame containing document identifiers and cluster assignments was created, and cluster centers were extracted and analyzed to identify the most representative terms associated with each cluster.

Results

3. In the Results, describe your results. For example, what are the findings of the data cleaning, word cloud, sentiment analysis?

Documents	100
Terms	2,027
Non- / Sparse Entries	3,924 / 198,776
Sparsity	98%
Maximal Term Length	17
Weighting	Term Frequency (TF)

Table 1: Document Term Matrix Summary

Table 1 summarizes the structure of the document–term matrix constructed from the cleaned Reuters corpus. The matrix contains 100 documents and 2,027 unique terms. Of the total possible document–term combinations, 3,924 entries are non-zero (meaning 3,924 document-term pairs where a word appears), while 198,776 entries are zero, resulting in an overall sparsity of approximately 98%. Only approximately 2% of all document-term combinations occur. This high level of sparsity is typical for text data, as most terms appear in only a small subset of documents. The longest term in the corpus contains 17 characters. All entries in the matrix are weighted using raw term frequency (TF).

Docs	bankamerica	billion	loss	march	net	new	profit	share	shr	stock
1	0	0	0	4	0	9	0	0	0	0
16	12	2	0	0	0	0	0	0	0	5
18	0	2	0	0	0	0	0	4	0	0
26	0	3	0	1	0	4	0	0	0	0
28	0	0	1	0	0	0	0	0	0	0
4	12	2	0	0	0	0	0	0	0	5
45	0	0	0	1	0	0	0	3	0	1
54	0	0	0	0	0	3	0	0	0	0
61	0	2	0	0	0	0	0	0	0	5
79	0	0	0	0	0	9	1	0	0	2

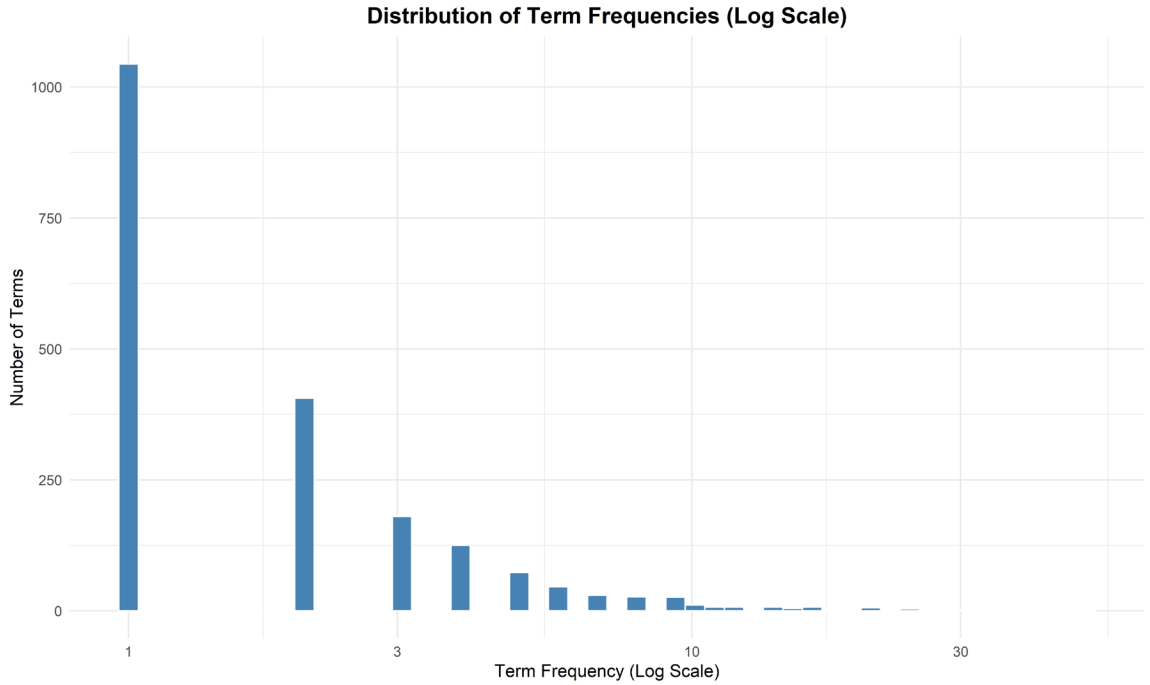
Table 2: Document Term Matrix Sample Terms

Table 2 presents a sample of terms from the document–term matrix to illustrate how term frequencies are distributed across individual documents. Each row represents a document, and each column represents a selected term, with cell values indicating the number of times a term appears within a document; zero values indicate the absence of the term. The prevalence of zero values highlights the sparsity of the matrix, while variation across documents reflects differences in article content. For example, in Document 1, the terms “march” and “new” appear 4 and 9 times, respectively. In Document 16, the terms “bankamerica,” “billion,” and “stock” appear 12, 2, and 5 times, respectively. Document 16 and Document 4 share similar contents. Similar patterns are observed across the remaining documents.

Term	Frequency
billion	50
new	45
share	32
shr	31
march	29
net	29
stock	29
profit	25
bankamerica	24
loss	24

Table 3: Top 20 Most Frequent Terms

Table 3 reports the most frequently occurring terms across the corpus, based on aggregated term frequencies from the document–term matrix. The most common terms include "billion," "new," "share," and "shr," (perhaps an abbreviation for 'share'), reflecting the financial and market-oriented focus of the Reuters articles. Frequencies represent total term counts across all documents.



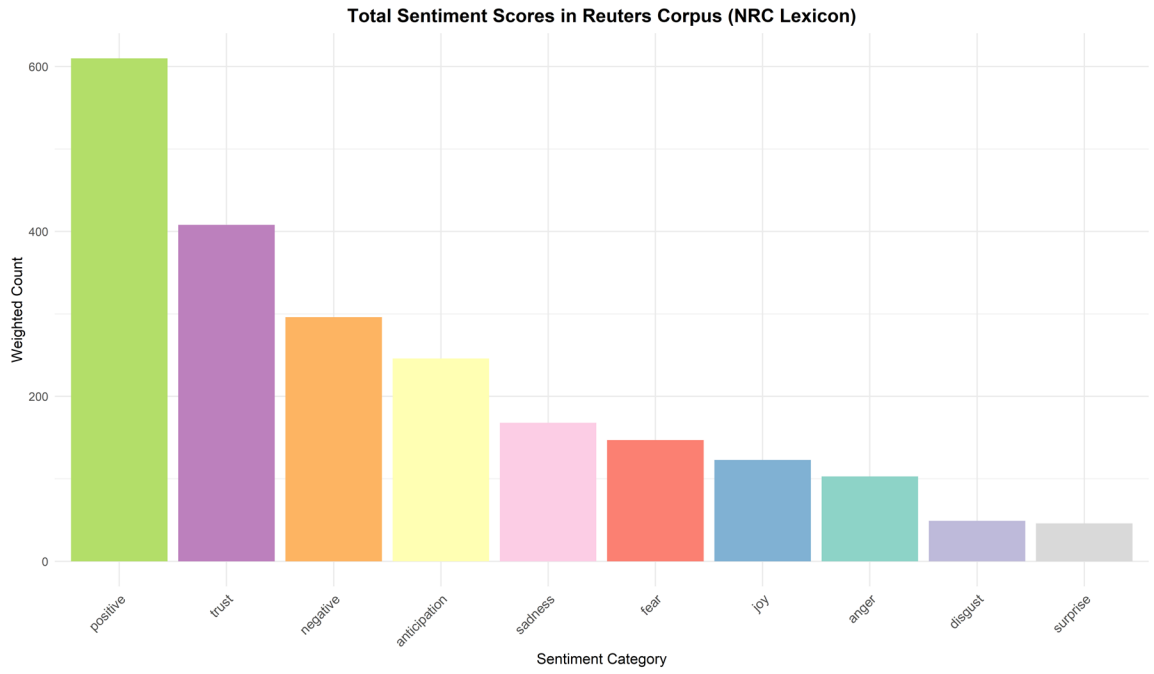
Plot 1: Histogram of Logged Term Frequencies

Plot 1 presents a histogram of aggregated term frequencies across the corpus on a logarithmic scale. The distribution is highly right-skewed, with most terms occurring at very low frequencies (one to two occurrences), and only a limited number appearing frequently, corresponding to the high-frequency terms reported in Table 3. This pattern is characteristic of text data and reflects

the highly sparse nature of the document–term matrix (98% sparsity), where a small set of common terms dominates overall word usage.

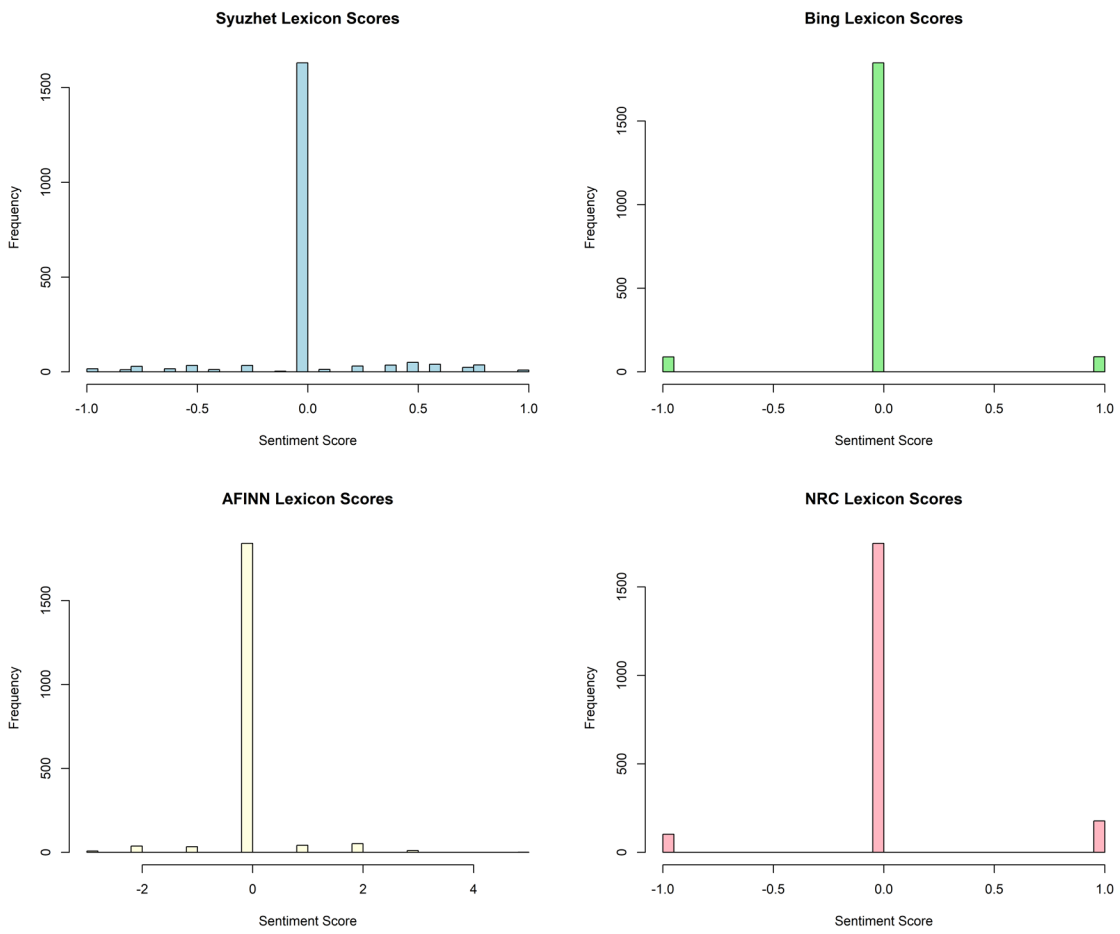
Plot 2: Word Cloud

Plot 2 presents a word cloud visualizing aggregated term frequencies across the corpus. Words are sized proportionally to their total frequency, with larger-sized terms indicating more frequent usage across documents. Consistent with Table 3 and Plot 1, high-frequency financial terms such as “billion,” “new,” “share,” and “march” are visually prominent, reinforcing the financial and market-oriented focus of the Reuters articles.



Plot 3: NRC Lexicon Sentiment Scores

Plot 3 displays total sentiment scores across the corpus using the NRC lexicon. Sentiment scores are weighted by term frequency, such that sentiment-associated words contribute proportionally to their overall frequency in the corpus. Positive and trust-related sentiments are the most prevalent categories, followed by negative and anticipation. Emotions such as sadness, fear, joy, and anger appear less frequently, while disgust and surprise are relatively rare. Overall, the distribution indicates that sentiment-bearing language in the Reuters corpus is dominated by neutral-to-positive categories.



Plot 4: Lexicon Histograms

Plot 4 compares the distribution of term-level sentiment scores across four lexicons. The majority of terms receive a sentiment score of zero, indicating that most words in the corpus are not associated with sentiment labels. Non-zero sentiment scores occur less frequently and differ in scale across lexicons, reflecting differences in how sentiment is encoded. The AFINN lexicon exhibits a wider range of sentiment intensity, while the Bing and NRC lexicons assign discrete sentiment categories. Overall, the distributions highlight that sentiment-bearing words constitute a relatively small subset of the corpus vocabulary.

Lexicon	Negative (%)	Neutral (%)	Positive (%)
Syuzhet	7.70	80.41	11.89
Bing	4.39	91.17	4.44
AFINN	3.95	90.82	5.23
NRC	5.08	86.14	8.78

Table 4: Sentiment Comparison

Table 4 compares the distribution of negative, neutral, and positive sentiment classifications across four sentiment lexicons. The Syuzhet lexicon assigns the highest proportion of positive sentiment (11.89%) and also the highest proportion of negative sentiment (7.70%) among the lexicons. The Bing lexicon classifies the corpus as predominantly neutral, with nearly equal proportions of negative (4.39%) and positive (4.44%) sentiment. The AFINN and NRC lexicons similarly indicate that most terms are neutral, with smaller shares classified as positive or negative. AFINN and NRC lexicons both lean slightly more positive albeit different distributions.

Term	Frequency	AFINN Score
profit	25	2
share	32	1
outstanding	4	5
shares	18	1
approval	9	2
growth	9	2
interest	15	1
rose	12	1
assets	6	2
ease	5	2

Table 5: Top 10 Most Positive Terms

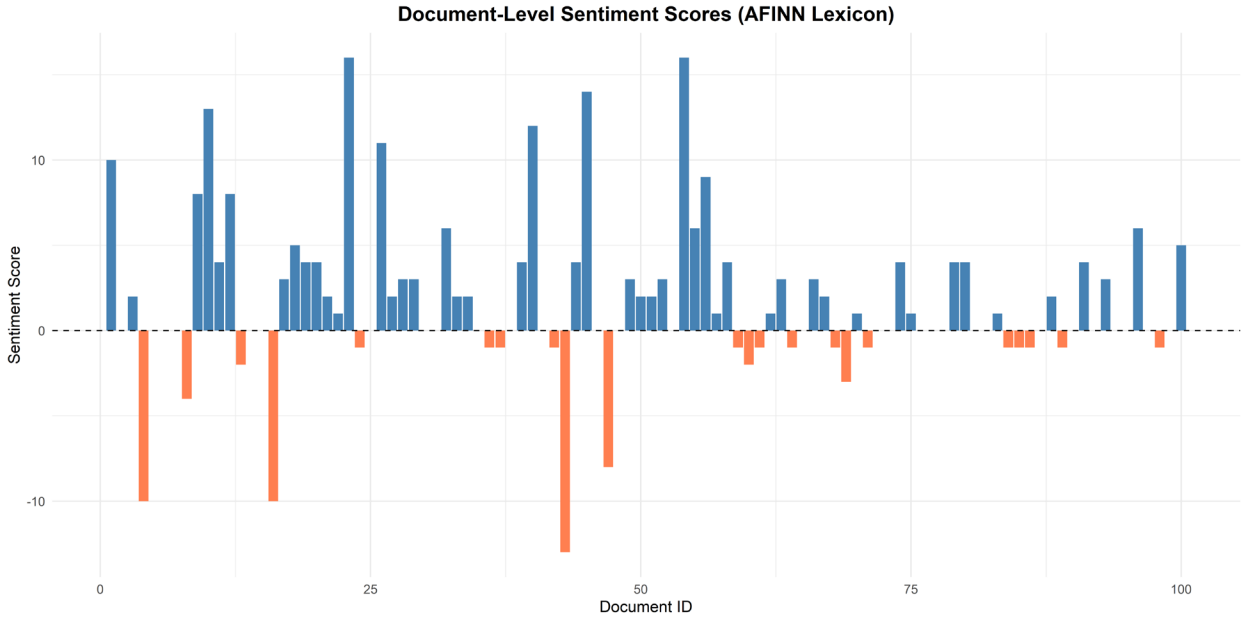
Table 5 reports the ten most positive terms in the corpus based on the AFINN sentiment lexicon. The AFINN lexicon assigns numeric sentiment scores ranging from –5 (most negative) to +5 (most positive), allowing individual terms to be ranked by sentiment intensity. The most positive term, “outstanding,” receives the highest possible score (+5), while other frequently occurring terms such as “profit,” “growth,” and “approval” receive moderate-to-low positive scores. Many of these terms also appear frequently in the corpus, indicating that positive sentiment in the Reuters articles is primarily associated with financial performance, market activity, and favorable economic outcomes.

Term	Frequency	AFINN Score
loss	24	-3
debt	21	-2
losses	4	-3
drop	9	-1
pay	7	-1
avoid	6	-1
falling	6	-1
waste	6	-1
criticism	3	-2

depressed	3	-2

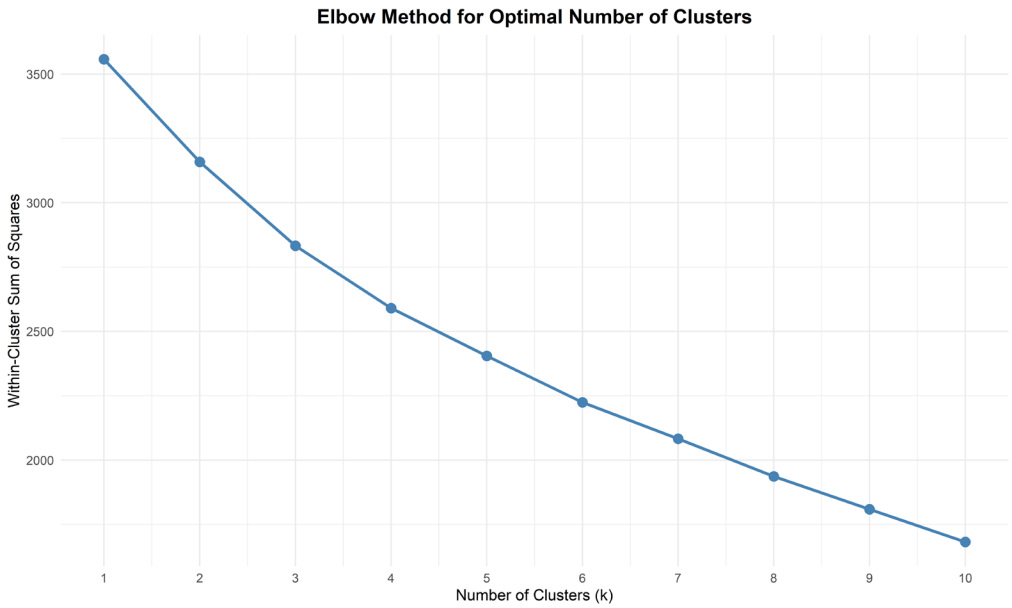
Table 6: Top 10 Most Negative Terms

Table 6 presents the ten most negative terms in the corpus based on the AFINN sentiment lexicon. To reiterate, the AFINN lexicon assigns numeric sentiment scores ranging from –5 (most negative) to +5 (most positive), enabling terms to be ranked by negative sentiment intensity. The most negative terms include "loss" and "losses," which receive the strongest negative scores (–3), followed by terms such as "debt," "drop," and "criticism." Many of these terms appear frequently across the corpus, indicating that negative sentiment in the Reuters articles is largely associated with financial decline, risk, and adverse economic conditions.



Plot 5: Document-Level Sentiment Scores (AFINN Lexicon)

Plot 5 displays document-level sentiment scores calculated using the AFINN Lexicon. Each bar represents the total sentiment score for an individual document, with positive values indicating net positive sentiment and negative values indicating net negative sentiment. Overall, sentiment scores are more frequently more positive than neutral or negative, suggesting that the corpus exhibits a mildly positive emotional tone. Several documents display negative sentiment, most notably among earlier documents and a pronounced negative spike near Document 50. Positive sentiment is most dominant across the corpus under the AFINN Lexicon.



Plot 6: Elbow Method for Optimal Number of Clusters

Plot 6 presents the elbow plot used to help evaluate the appropriate number of clusters for k-means analysis among corpus documents. The y-axis represents within-cluster sum of squares (WCSS) and the x-axis denotes the number of clusters k . As the number of clusters increases from $k = 1$ to $k = 4$, WCSS decreases sharply, indicating improvements in cluster cohesion as documents are distributed into more meaningful groups. Around $k = 5$ is where the decrease becomes more linear and gradual, suggesting diminishing returns if the number of k clusters continues to increase. Beyond this point, further increases in k yield only marginal reductions in WCSS and do not meaningfully improve how well the data are grouped.

Based on this pattern, $k = 4$ was selected as the optimal number of clusters. This choice captures most of the reduction in within-cluster dispersion from the cluster centroid, while avoiding over-segmentation that occurs after the elbow, where additional clusters provide limited analytical value.

Plot 7: Document Distribution Across K-Means Clusters

Plot 7 shows the distribution of corpus documents across four k-means clusters $k = 4$. Cluster 4 contains the majority of documents (n = 91), while Cluster 2 contains 5 documents, and Cluster 1 and Cluster 3 contain 2 documents each, summing to the full set of 100 documents analyzed. This highly imbalanced distribution reflects the strong similarity among most articles in the corpus following text preprocessing and sparsity reduction. By removing rare terms, documents share a more common vocabulary, leading k-means to assign the majority of observations to a single dominant cluster. The smaller clusters likely capture niche or atypical articles characterized by distinct term usage relative to the broader corpus. Given the Reuters-21578 source, the dominant cluster plausibly represents articles with shared financial terminology and standardized reporting language, while the smaller clusters reflect more specialized content.

Discussion

The text analysis of Reuters articles from 1987 reveals important patterns about how financial news was written. The most common words—"billion," "profit," and "stock"—show that coverage focused on corporate performance and market activity. However, sentiment analysis tells a different story. Although the NRC Lexicon found more positive language (600 instances) than negative language (150 instances), most articles (80–91%) were relatively neutral in tone. This means financial journalists used objective language rather than emotional language when reporting on business news (Mohammad & Turney, 2013). We see this balance when comparing positive terms like "profit," "growth," and "approval" with negative terms like "loss," "debt," and "drop." This pattern reflects journalistic standards: financial reporting prioritizes factual information over emotional framing. Understanding this matters because it shows how media shapes public perception of economics—not through emotional language, but through which topics receive coverage and how those topics are presented (Nielsen, 2011).

In the future, analyzing how sentiment changed throughout 1987 could reveal how news coverage reacted to specific financial events or market movements. Second, using clustering techniques could show whether articles about losses form separate groups from articles about growth. Third, summarizing article content would add context to the sentiment numbers, helping identify which economic stories dominated the news whether stories about managing crises, policy changes, or industry transformations. Together, these approaches would provide a fuller picture of how financial media conveys economic meaning.

References

Martinez, Miguel. 2015. “Classifying Reuters-21578 Collection With Python: Representing the Data.” The Practical Academic. March 20, 2015. https://miguelmalvarez.com/2015/03/20/classifying-reuters-21578-collection-with-python-representing-the-data/.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29(3), 436–465.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.