Loading and Cleaning the Datasets

We loaded four major sentiment lexicons: AFINN, Bing Liu's Opinion Lexicon, the NRC Emotion Lexicon, and the Loughran-McDonald Financial Sentiment Dictionary. Each of these datasets contained sentiment information for various English words. we merged the four lexicons into a single Data Frame. Some words appeared only in certain dictionaries. As a result, the unified Data Frame had multiple columns: word, value (AFINN score), sentiment (Loughran sentiment), sentiment_x (Bing sentiment), and sentiment_y (NRC emotional categories).

Filling Missing Data Using GloVe Embeddings

To handle all missing fields — **`value`**, **`sentiment_x`**, and **`sentiment_y`** — we adopted a unified strategy based on semantic similarity using pre-trained GloVe word embeddings. For missing **`value`** scores, we first loaded the GloVe 6B 200-dimensional embeddings, and for missing **`sentiment_x`** and **`sentiment_y`**, we used GloVe Twitter 27B embeddings, as these are more suited for emotion and opinion words found in informal text. For each missing entry, we first checked whether the word's GloVe embedding was available. If it was, we calculated its similarity to all words that had known values or sentiment labels, using **cosine similarity** — a metric that measures the cosine of the angle between two vectors and captures semantic closeness regardless of word frequency or word form. Specifically, the cosine similarity between two vectors A and B is given by:

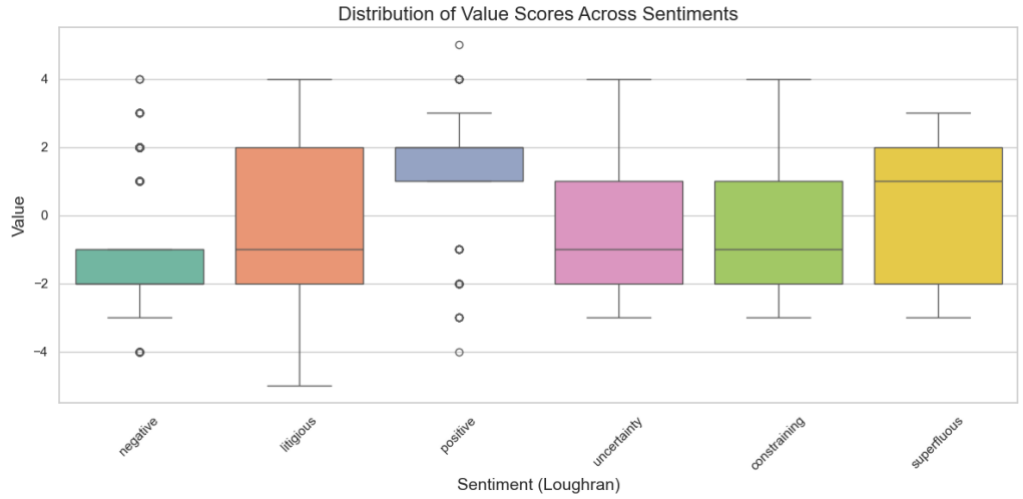
$$\textbf{cosine\_similarity (A,B)} = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{||\boldsymbol{A}|| \times ||\boldsymbol{B}||}$$

where A·B is the dot product, and ||A|| and ||B|| are their magnitudes. Using this method, for each missing **`value`**, we found the closest word and assigned its **`value`** score; for each missing **`sentiment_x`**, we assigned the most similar known sentiment label; and for each missing **`sentiment_y`**, which could have multiple emotions, we assigned the relevant list of emotion labels from the most semantically similar known word. After completing the imputations, we rounded the filled **`value`** scores to the nearest integer, clipped them within [-5, 5] for consistency, and handled the multi-label structure in **`sentiment_y`** carefully by later exploding it for analysis.

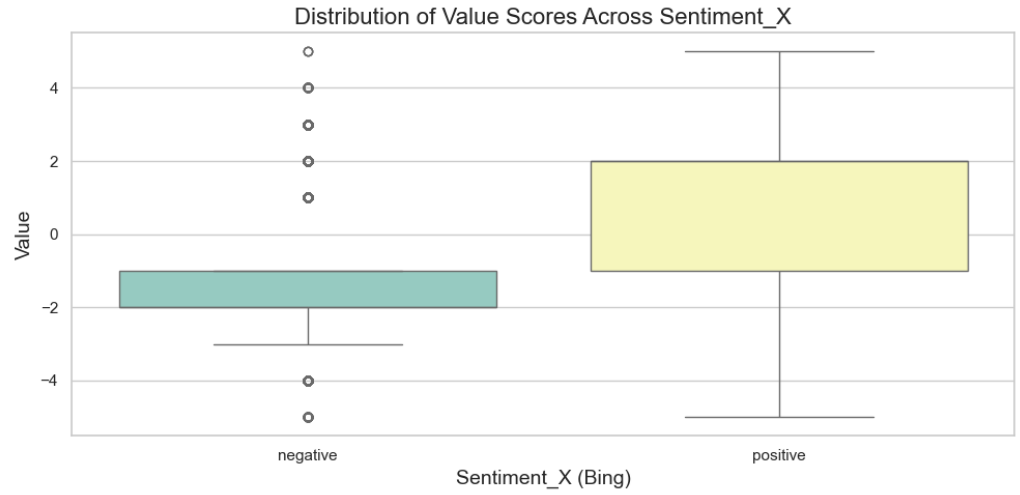
Deduplication and Final Cleaning

After imputing missing data, we cleaned the dataset further by removing duplicate words, ensuring that each word appeared only once with its enriched sentiment information. This made the dataset structured, easy to handle, and ready for visualization and further use.

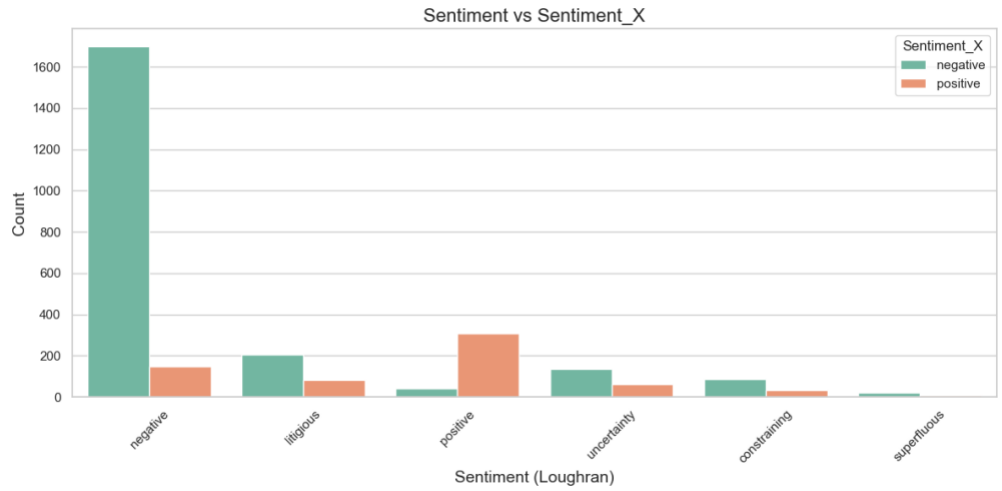
Exploratory Data Analysis and Visualization



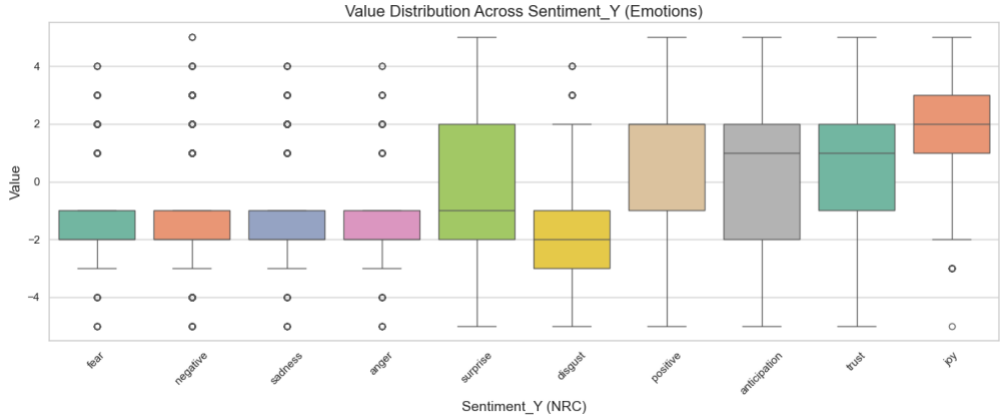
Value scores across Loughran categories aligned with financial sentiment expectations, with negative terms showing negative values and positive ones showing positive values. Overall, the visualizations confirm that the missing data imputation preserved natural, emotional, and financial sentiment structures accurately across all datasets.



The Value scores separated cleanly across Sentiment_X, with negative words having negative scores and positive words having positive scores, showing minimal overlap. This again confirms that the filled Value and Sentiment_X fields are internally consistent.



Similarly, analyzing Sentiment (Loughran financial categories) against Sentiment_X showed strong polarity consistency. Words marked negative or uncertain in the financial lexicon aligned with negative labels in Bing, while positive financial terms matched positive labels. Specialized terms like litigious and constraining were also more negative, reflecting their context-specific meanings.



The Value scores across Sentiment_Y categories further validated our approach. Words associated with negative emotions had lower Value medians, while positive emotions corresponded to higher medians. This clear separation between negative and positive emotional groups indicates that the imputed Value scores retained the correct sentiment direction.