Title:

Navigating Market Trends: Analyzing Sentiments in Finance Tweets

Abstract:

This study aimed to analyze public sentiment towards Fortune 500 companies using Twitter data from 2015 to 2020. We collected and preprocessed tweets, applying the Loughran-McDonald sentiment lexicon to classify sentiments. A logistic regression model with Lasso regularization was developed for sentiment prediction. Key insights were visualized through word clouds and Pareto charts, revealing significant terms and sentiment distributions. The model demonstrated robust performance with an accuracy of 82.7% and an ROC AUC score of 0.9134. Terms such as "growth" and "profit" were influential in swaying tweets positively, while words like "loss" and "decline" contributed to negative sentiment classifications. Results indicated that prominent companies like Alphabet Inc., General Motors, and Texas Instruments were frequently subjects of negative sentiment. The findings underscore the complexity of public sentiment and highlight the importance of nuanced analysis for financial market stakeholders. Limitations include potential biases in tweet samples and the need for more sophisticated models to capture sentiment nuances. Future research should focus on integrating advanced natural language processing techniques to enhance sentiment analysis accuracy.

Background and Significance:

Sentiment analysis has gained considerable attention in the field of finance and investment, primarily due to its potential to influence market movements. Social media platforms, especially Twitter, serve as a real-time barometer for public opinion, making them invaluable for understanding sentiment towards companies and their financial performance. This project focuses on sentiment analysis of tweets related to Fortune 500 companies, with the goal of uncovering insights that can inform investment decisions and corporate strategies.

The financial markets are significantly influenced by public sentiment, as reflected in social media discourse. Positive sentiment can boost stock prices, while negative sentiment can lead to declines. For investors and financial analysts, understanding these sentiment trends can provide a competitive edge. Companies can also benefit by monitoring public perception to manage their reputations and strategize accordingly.

Several studies have demonstrated the utility of sentiment analysis in financial contexts. Bollen et al. (2011) showed that public mood states, as measured by Twitter updates, are predictive of stock market movements. Similarly, Zhang et al. (2011) found that sentiment derived from microblogging sites like Twitter can be used to predict stock market returns. The Loughran-McDonald lexicon, which is tailored to financial texts, has been widely adopted for sentiment analysis in this domain (Loughran & McDonald, 2011). This lexicon categorizes words into sentiment-related categories such as positive, negative, uncertainty, and others, making it particularly useful for financial sentiment analysis.

Building on the foundation of existing research, this study hypothesizes that sentiment analysis of tweets can reveal significant patterns in public perception of Fortune 500 companies. These patterns, in turn, may correlate with market performance, providing valuable insights for stakeholders. The project aims to test this hypothesis by developing a robust sentiment analysis model and evaluating its effectiveness in predicting sentiment from twitter data.

The primary objectives of this study are:

1. To investigate the public sentiment towards Fortune 500 companies by analyzing tweets.
2. To apply sophisticated sentiment analysis methods, such as the Loughran-McDonald lexicon, to assess the sentiment of tweets.
3. To create predictive models using logistic regression with Lasso regularization to classify sentiment accurately, evaluate the performance of the model and discuss its implications for stakeholders in the financial markets.
4. To determine the most influential words and phrases that contribute to positive and negative sentiment classifications.

This study contributes to the growing body of research on sentiment analysis in financial contexts. By leveraging Twitter data, it provides real-time insights into public sentiment, which is crucial for making timely investment decisions. The findings can help investors and companies better understand market sentiment and its potential impact on financial performance. Additionally, the study highlights the importance of advanced text mining techniques in extracting valuable information from unstructured data sources like social media.

The significance of this project lies in its potential to bridge the gap between public sentiment and financial decision-making. By providing a detailed analysis of sentiment trends towards Fortune 500 companies, it offers a novel approach to understanding market dynamics.

The next sections will elaborate on the methods used, the results obtained, and the implications of these findings.

Methods:

Data Collection:

The data for this project was sourced from a Kaggle dataset titled “Tweets about the Top Companies from 2015 to 2020” and stored in the variable stockerbot_df. This dataset focuses on tweets mentioning top Fortune 500 companies. For this analysis, the data was filtered to include only tweets written in English. A random subset of 5,000 rows was sampled from stockerbot_df, providing a robust foundation for the analysis. In total, stockerbot_df comprises 28,277 observations across 8 variables: “id”, “text”, “timestamp”, “source”, “symbols”, “company_names”, “url”, and “verified”. The head of the dataset is shown below:

id	text	timestamp	source	symbols	company_ names	url	verified
1017223 6020249 10800	$AABA put play. …	Thu Jul 12 01:46:20 +0000 2018	iampatelrp	AABA	Altaba Inc.	—	False
1017244 8744694 66100	Sgenyou betterbuy somesoo …	Thu Jul 12 03:10:52 +0000 2018	stophahs	PEG	Public Service Enterprise Group Incorporated	https://t witter.c om/i/w eb/stat us/101 724487 446946 6113	False
1019427 5626180 07600	Invest your crypto on …	Wed Jul 18 03:44:05 +0000 2018	jesse_ vallejo	DKS	Dick's Sporting Goods	http://bi nance. com/?r ef=100 78236	False

The preprocessing steps involved cleaning and transforming the tweet text to ensure that it was suitable for analysis. The following steps were taken:

- *Text Conversion:* All tweet text was converted to lowercase to maintain uniformity.
- *Stopword Removal:* Common stopwords, such as "the", "and", "is" as well as additional context-specific stopwords like "rt" (retweet) and "inc" (incorporated), were removed to reduce noise.
- *URL Removal:* URLs were stripped from the tweet text using regular expressions to eliminate irrelevant content.

- *Punctuation and Number Removal:* All punctuation and numbers were removed to focus solely on the words.
- *Whitespace Removal:* Extra whitespace was eliminated to streamline the text.
- *Stemming:* Words were stemmed to their root forms to ensure consistency in analysis.

Variable Creation:

The primary variable of interest was the sentiment score assigned to each tweet. Sentiment scores were calculated using the Loughran-McDonald sentiment lexicon, which is specifically tailored for financial texts. Each tweet was assigned a sentiment score based on the presence of positive and negative words from the lexicon. Tweets with a positive sentiment score greater than 0.1 were classified as "positive", while the rest were classified as "negative".

Analytic Methods:

- *Document-Term Matrix and Sentiment Analysis:* A Document-Term Matrix (DTM) was created from the cleaned tweet text, which formed the basis for further text mining and analysis. The tweets were then tokenized, and the Loughran-McDonald lexicon was applied to categorize words into positive and negative sentiments. The frequencies of these sentiment words were aggregated and analyzed to identify common terms and their impacts.

- *Logistic Regression Model with Lasso Regularization:* A logistic regression model with Lasso regularization was developed to predict the sentiment of tweets. The steps involved were:
    - *Data Splitting:* The dataset was split into training and testing sets using an 80-20 split, stratified by the sentiment rating.
    - *Recipe Preparation:* A recipe was created to preprocess the text data, including tokenization, stopword removal, token filtering, and TF-IDF transformation. The predictors were then normalized.
    - *Model Specification:* A logistic regression model with Lasso regularization was specified, with the penalty parameter set to be tuned.
    - *Hyperparameter Tuning:* A grid of lambda values (penalty) was created, and a cross-validation procedure using bootstraps was conducted to tune the model. Metrics such as ROC AUC (Area Under the Receiver Operating Characteristic Curve), positive predictive value (PPV), and negative predictive value (NPV) were used to evaluate the classification model.

- *Model Evaluation:* The final model was evaluated using the testing set to determine its performance. The following steps were performed:
    - *Model Fitting:* The best model, based on the highest AUC, was selected and fitted to the training data.
    - *Variable Importance:* The importance of different predictors (words) was assessed, and the top 16 most influential terms were visualized.
    - *Final Model Testing:* The finalized model was tested on the testing set to collect metrics such as accuracy, ROC AUC, PPV, and NPV. A confusion matrix was generated to assess the model's classification performance.

- *Visualization:* Several visualizations were generated to aid in the interpretation of the results:

- *Pareto Chart:* A Pareto chart was created to visualize the summed sentiment scores by company symbol, helping to identify which companies had the highest positive or negative sentiments.
- *Word Clouds:* Word clouds for positive and negative words were created to visually depict the most frequent sentiment-bearing terms in the tweets.
- *Bar Charts*: Bar charts were used to show the frequency of positive and negative words, sorted in decreasing order, to identify the most influential words in determining tweet sentiment.

Results:

Word Cloud of Tweet Data:

To gain a better understanding of the data, a word cloud was created to display the most frequently used words in the tweets. This word cloud includes only those words that appeared 100 times or more across all tweets. The resulting word cloud is shown below:

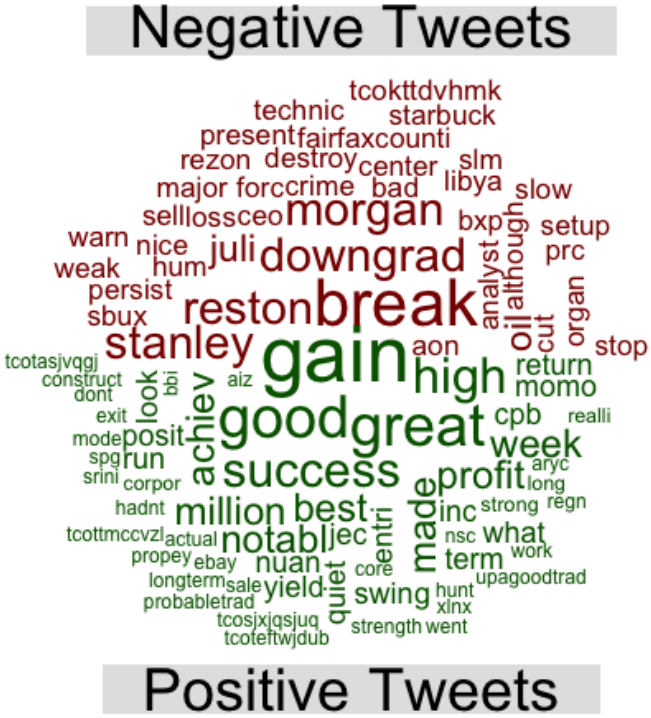
A: Code Snippets[1]

In the word cloud, the most prominent word is "amp", indicating a high frequency of its occurrence in the tweets. Other frequently appearing words include "corpor", "stock", "compani", "analyst", "trade", "price", "earn", "hold", and "share". These words are central to the discussions within the tweets and reflect the primary focus of the dataset, which is on financial and corporate topics related to the top Fortune 500 companies.

The presence of words like "energi", "group", "week", and "alert" also suggests that the tweets are not only discussing specific companies and their stocks but also broader topics such as energy sectors, weekly market movements, and alerts regarding stock performance.

Comparison Cloud of Positive and Negative Tweets:

Next, a comparison cloud of the positive and negative words from the 25 most positive and negative tweets was generated to understand the overall emotional tone and trends in public opinion about the Fortune 500 companies. In both the word cloud and the comparison cloud, the size of each word corresponds to its frequency, with larger words appearing more frequently in the tweets. The comparison cloud is shown below:



A: Code Snippets[2]

Words such as "gain", "good", "great", "success", "profit", and "notable" appear prominently. These terms are indicative of positive sentiment and reflect optimism, achievement, and financial success.

On the other hand, words such as "break", "downgrad", "warn", "slow", and "weak" are among the most prominent in the negative tweets. These terms convey caution, underperformance, and potential risk or loss.

Overall Sentiment Analysis:

The overall sentiment score of the dataset, which is the sum of the sentiment scores for each tweet, was calculated to provide a high-level summary of the sentiment profile of the dataset. This score was derived using the sentiment analysis function from the "sentiment" package, specifically utilizing the Loughran-McDonald sentiment lexicon. The result is shown below:

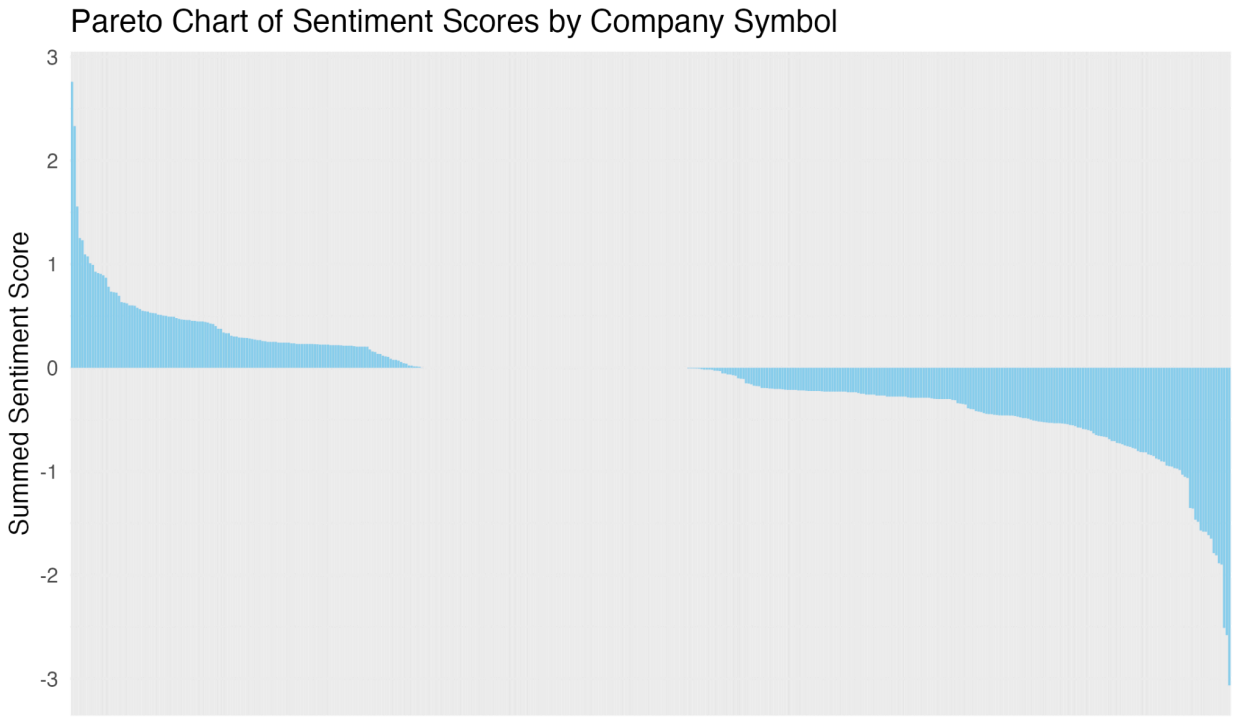
Metric	Score
Overall Sentiment Score	-49.08121

A: Code Snippets[3]

The negative overall sentiment score indicates that the aggregate sentiment of all the tweets in the dataset leans towards negativity. This means that, on balance, the tweets contain more negative sentiment than positive sentiment.

Pareto Chart of Sentiment Scores by Company Symbol:

Additionally, a Pareto chart of the summed sentiment scores for all companies mentioned in the dataset was created to analyze how different companies are perceived based on the sentiment of tweets mentioning them. Each bar represents a company's overall sentiment score. The x-axis labels (company symbols) are intentionally hidden to provide a cleaner view. Detailed properties of each bar can be viewed by hovering over the 3D plot in the project's presentation file.



A: Code Snippets[4]

The majority of the sentiment scores are clustered around zero, indicating a mix of positive and negative sentiments with no extreme biases for most companies.

Top Positively and Negatively Perceived Companies:

The pareto chart above was used to identify the most positively and negatively perceived companies. The results are tabulated below:

Top Positive:

Company	Ticker Symbol	Sentiment Score
WPX Energy	WPX	2.759275
Omisego	OMG	2.332231
DXC Technology Company	DXC	1.555975
International Business Machines Co.	IBM	1.252009
TE Connectivity Ltd.	TEL	1.230723
Assurant	AIZ	1.093717

A: Code Snippets[5]

Top Negative:

Company	Ticker Symbol	Sentiment Score
Cisco Systems	CSCO	-3.064218
McDonald's Co.	MCD	-2.579139
Juniper Networks	JNPR	-2.509452
Mallinckrodt Public Limited Company	MNK	-1.898289
Boston Properties	BXP	-1.885198
Kimberly Clark Co.	KMB	-1.809068

A: Code Snippets[6]

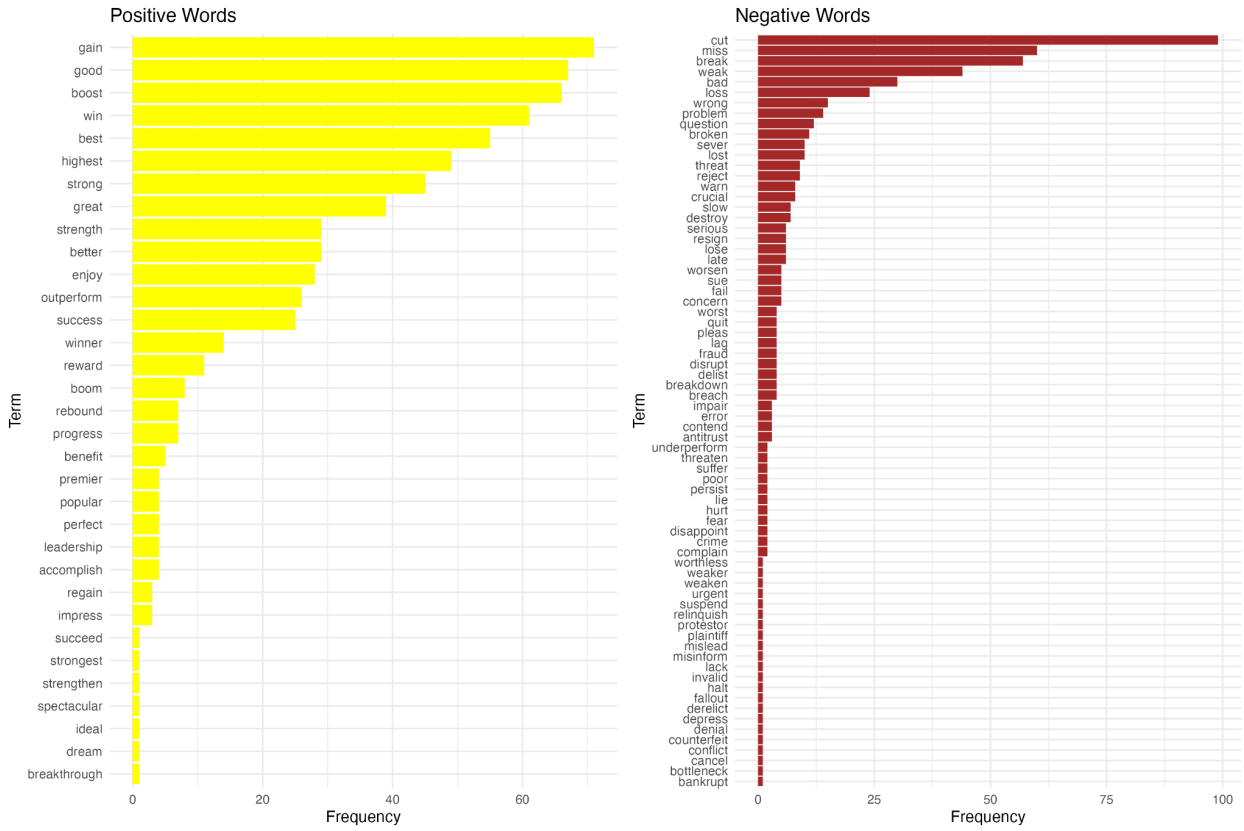
WPX Energy (WPX) leads the positive sentiment with a score of 2.759275, indicating a strong favorable perception among Twitter users. Other companies like Omisego(OMG), DXC Technology Company (DXC), International Business Machines Co. (IBM), TE Connectivity Ltd. (TEL), and Assurant (AIZ) also show significant positive sentiment, suggesting positive discussions and favorable opinions on social media.

On the contrary, Cisco Systems (CSCO) holds the highest negative sentiment score of -3.064218, indicating a strong unfavorable perception. McDonald's Co. (MCD) and Juniper Networks (JNPR) also feature prominently with negative scores of -2.579139 and -2.509452, respectively. Mallinckrodt Public Limited Company (MNK), Boston Properties (BXP), and Kimberly Clark Co. (KMB) round out the list of companies with the most negative perceptions, indicating they are frequently discussed unfavorably on Twitter.

The list of top companies discussed negatively by the public also includes major names like Alphabet Inc. (GOOG), General Motors Co. (GM), and Texas Instruments Inc. (TXN), in addition to McDonald’s Co. (MCD) and Cisco Systems (CSCO).

Frequency Distribution of Positive and Negative Words in Tweets:

Next, a frequency distribution plot of the positive and negative words in tweets was created to gain a deeper understanding of the emotional tone and sentiment patterns of the data. The results are shown below:



A: Code Snippets[7]

The most common positive words include "gain", "good", "boost", "win", and "best". The positive terms generally relate to success, strength, and improvement. Words like "highest", "strong", "great", and "success" emphasize positive outcomes and achievements. Many positive words such as "outperform", "winner", "reward", and "progress" suggest a sense of motivation and accomplishment.

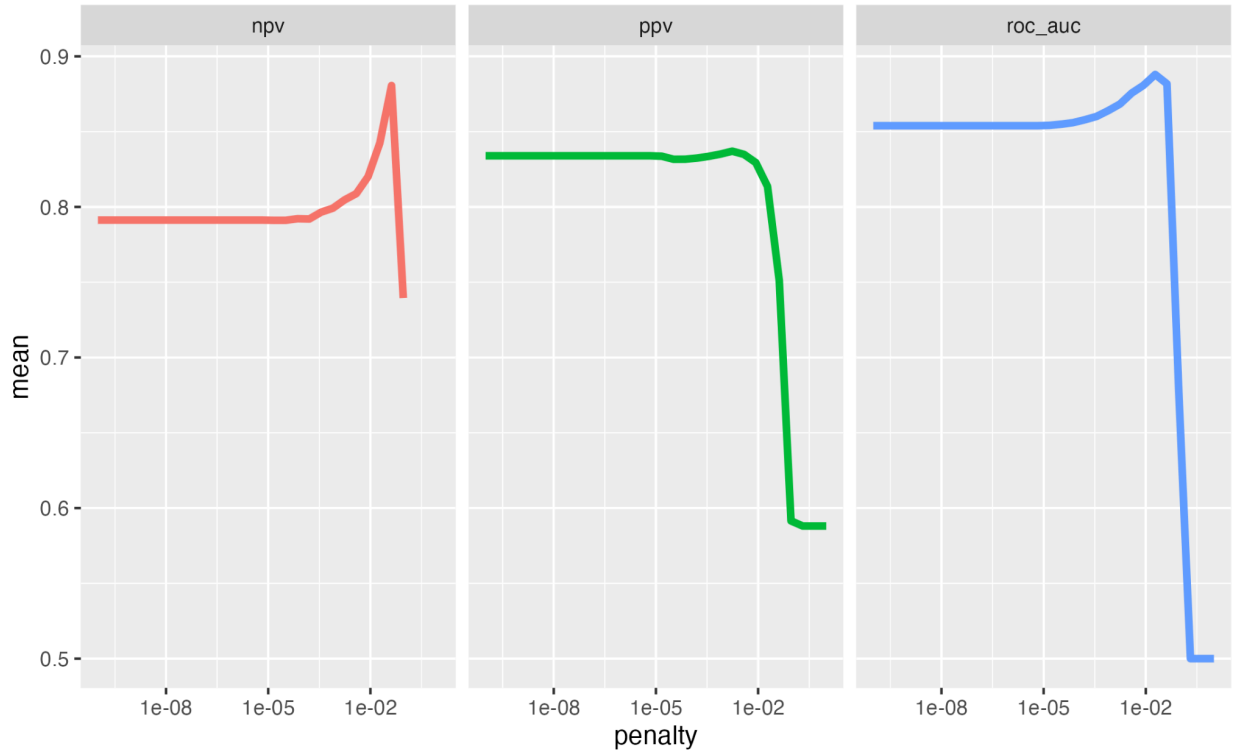
The most frequent negative words are "cut", "miss", "break", "weak", and "bad". Negative terms often reflect issues, problems, and failures. Words like "problem", "question", "loss", and "warn" highlight concerns and challenges. Many negative words such as "fail", "concern", "worst", "fraud", and "disrupt" indicate a critical view and significant issues faced by companies.

The fact that more unique negative words were identified compared to positive words suggests a wider variety of ways in which people express negative sentiments.

Machine Learning Model for Sentiment Analysis using Logistic Regression with Lasso Regularization:

Evaluation of Hyperparameter Tuning:

For the machine learning model for sentiment analysis, the results of hyperparameter tuning for the lasso regression model were first plotted using three evaluation metrics: negative predictive value (NPV), positive predictive value (PPV), and area under receiving operator characteristic curve (ROC AUC). The x-axis represents different penalty values (log scale), and the y-axis represents the mean value of the respective metric.



A: Code Snippets[8]

The NPV remains relatively stable across the penalty values until it reaches 1e−02. At this point, there is a sharp increase in NPV, indicating improved performance in predicting true negatives. However, beyond this peak, NPV drops suddenly, suggesting over-penalization.

The PPV stays consistent across a wide range of penalty values. As the penalty value increases beyond 1e−02, there is a sharp decline in PPV, indicating that higher penalties negatively impact the model's ability to predict true positives.

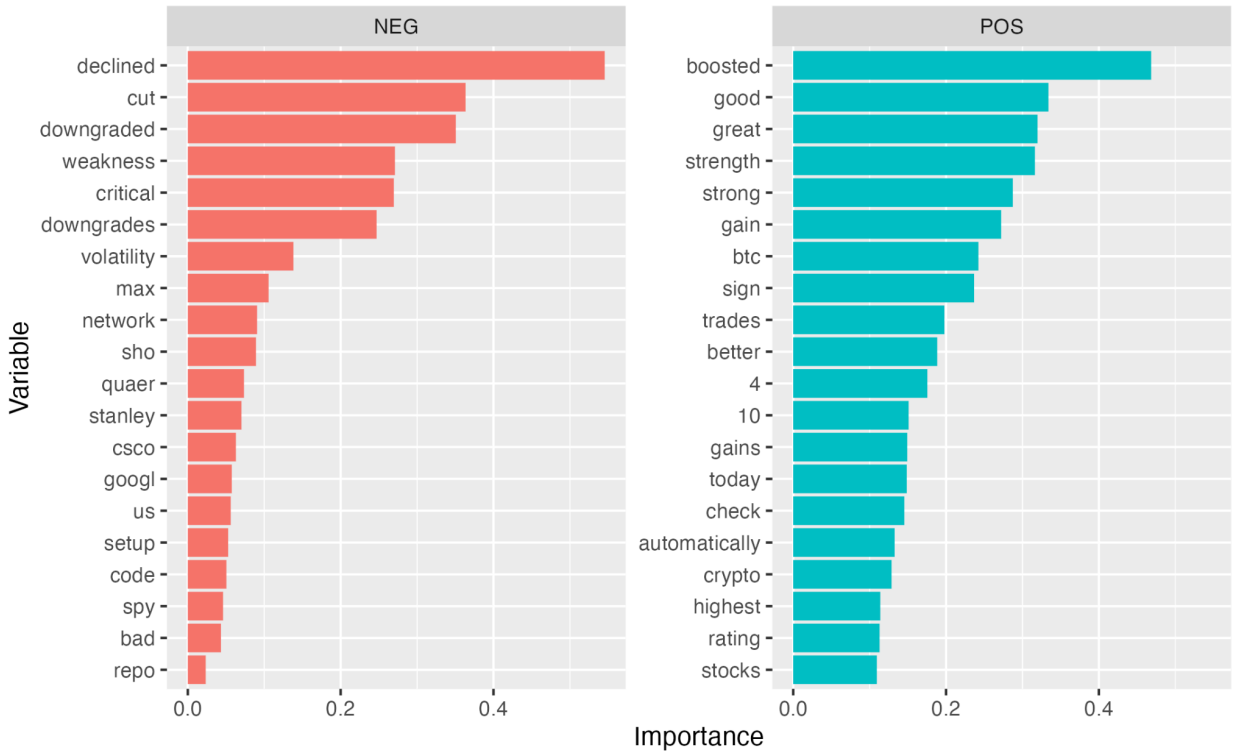
The ROC AUC remains fairly stable initially, with a slight increase as the penalty value reaches 1e−02. The peak performance for ROC AUC is around 1e-02, after which there is a dramatic drop. This suggests that the model performs best with a moderate penalty value, optimizing the balance between sensitivity and specificity.

The best performance across all three metrics is observed around a penalty value of 1e-02. This value balances the tradeoff between regularization and model complexity. High penalty values lead to a significant drop in performance metrics, indicating over-penalization and loss of important features.

The NPV and PPV remain relatively stable over a range of lower penalty values, suggesting that the model is robust to slight changes in regularization strength.

Important Terms for Positive and Negative Sentiment Classifications:

Next, a plot of the top 20 most important terms contributing to positive and negative sentiment classifications, as determined by the model, was generated. This plot helps to explore which terms are contextually significant in sentiment classifications. The result is shown below:



A: Code Snippets[9]

The term “boosted” has the highest importance for positive sentiment, indicating it strongly influences the classification of tweets as positive. The terms “good”, “great”, “strength”, “strong”, and “gain” are also highly influential in predicting positive sentiment, reflecting positive performance or outlook. The term “btc” likely refers to Bitcoin’s stock symbol, suggesting positive discussions involving Bitcoin.

The most important term for negative sentiment classification is “declined”, strongly indicating negative performance or outcomes. The terms “cut”, “downgraded”, and “weakness” likely denote reductions or negative adjustments in financial or stock matters. The term “stanley” likely refers to a company or analyst that is contextually associated with negative sentiment.The terms “csco” and “googl” are company ticker symbols for Cisco Systems and Google, indicating that these companies are subjects of negative sentiment.

Final Model Results:

Estimates of Metrics:

Metric	Estimator	Estimate
accuracy	binary	0.8274336
roc_auc	binary	0.9134026
brier_class	binary	0.1273854

A: Code Snippets[10]

The model's accuracy is approximately 82.74%.

The ROC AUC score is 0.9134. This high value suggests that the model has a strong ability to distinguish between positive and negative sentiments.

The Brier score is 0.1274. The Brier score is a metric used to assess the accuracy of probabilistic predictions. More specifically, the Brier score measures the mean squared difference between the predicted probabilities and their actual binary outcomes. A Brier score of 0 indicates perfect predictive accuracy whereas a score of 1 indicates worst possible predictive accuracy. A score of 0.1274 indicates that the model's probability estimates are reasonably accurate.

Confusion Matrix:

	Truth	
Prediction	negative	positive
negative	127	34
positive	5	60

A: Code Snippets[10]

There are 127 true negatives. This means that the model correctly identified 127 instances of negative sentiment.

There are 34 false positives. This indicates that there are 34 instances where the model incorrectly identified positive sentiment when it was actually negative.

There are 5 false negatives. This means that the model incorrectly identified 5 instances of negative sentiment when it was actually positive.

There are 60 true positives. This indicates that the model correctly identified 60 instances of positive sentiment.

Discussion/Conclusions:

The primary objective of this study was to analyze the sentiment of tweets mentioning Fortune 500 companies, to identify key trends and patterns in public perception, to determine the most influential terms contributing to positive and negative sentiments, and to to develop a robust predictive model using logistic regression with Lasso regularization to classify sentiment accurately. This analysis aimed to provide valuable insights for stakeholders and decision-makers in these companies, helping them understand and manage public sentiment effectively. By employing logistic regression with Lasso regularization, the goal was to identify significant patterns in public perception that could offer critical intelligence and strategic information for finance industry players.

The word cloud generated from the tweet data provided a clear and immediate understanding of the key terms and themes present in the dataset. It highlighted the significant emphasis on stock market activities and corporate analyses. This visual tool effectively summarized the dominant topics, guiding further, more detailed analysis of the tweets. The emphasis on words related to stock market activities, corporate events, and specific companies underscores the public's focus on financial performance and corporate actions. This insight aligns with our objective of identifying key themes in public discussions about Fortune 500 companies. The presence of terms like "stock," "market," "company," and "analysis" suggests that Twitter users are highly engaged with and responsive to corporate performance and news, which can influence market perceptions and investor behavior.

Furthermore, by separating words into positive and negative categories, the comparison cloud highlighted the contrast in sentiment-laden vocabulary used in the tweets. This allowed us to infer the overall sentiment of the dataset and track changes in public perception over time. The balanced view of public opinion, with positive words shown in green and negative words in red, provided a nuanced understanding of the emotional tone of the tweets. This dual perspective is valuable for stakeholders in understanding both praise and criticism related to their companies, directly addressing our objective of identifying sentiment trends. For instance, the prevalence of positive terms like "boosted," "good," and "great" contrasted with negative terms like "declined," "cut," and "downgraded" reveals the emotional dichotomy in public discussions. The ability to visualize these sentiments can help companies tailor their communication strategies to address public concerns more effectively and to capitalize on positive sentiment.

After conducting the overall sentiment analysis, the overall sentiment score indicated a predominantly negative sentiment in the tweets. This could reflect broader economic concerns or specific events affecting multiple companies during the period from 2015 to 2020. However, it's important to note that the overall sentiment score is not a fully accurate representation of public sentiment, as it does not account for nuances in text, such as context and sarcasm. This limitation highlights the need for more sophisticated sentiment analysis techniques to fully capture the complexities of public opinion, which is crucial for achieving our study's objective. The high prevalence of negative sentiment could also be influenced by the nature of social media, where negative experiences and news often garner more attention and engagement than positive ones. This trend necessitates a deeper investigation into the specific events and narratives that drove negative sentiments during the analyzed period.

After understanding the overall sentiment score, the Pareto chart helped identify which companies had strong positive or negative public perceptions. This information is valuable for company stakeholders, investors, and analysts in understanding public sentiment trends over

the period analyzed. By pinpointing companies with extreme sentiments, this chart supports our objective of providing targeted insights for reputation management. Companies like Alphabet Inc. (GOOG), General Motors Co. (GM), and Texas Instruments Inc. (TXN) being frequently mentioned in negative contexts suggests that even large, influential firms are not immune to public scrutiny and criticism. This emphasizes the importance of proactive reputation management and crisis communication strategies to mitigate negative sentiment and enhance public perception. Understanding the reasons behind public criticism can help these companies address issues more effectively, aligning with our objective of offering actionable insights for sentiment management. The fact that companies such as McDonald's Co. (MCD) and Cisco Systems (CSCO) also appeared frequently in negative contexts highlights the diverse range of industries subject to public scrutiny. This diversity suggests that public sentiment is influenced by a variety of factors, including company performance, industry trends, and broader economic conditions.

For the frequency distribution analysis of positive and negative words in tweets, the analysis identified a greater number of unique negative words compared to positive words, indicating a wider variety of expressions for negative sentiments. This suggests that negative experiences are more diverse and multifaceted, requiring a richer vocabulary to describe specific issues. On the other hand, positive sentiments might be more homogeneous. This imbalance highlights the complexity of negative sentiments and the challenges companies face in addressing a broader range of public concerns. The presence of more unique negative words indicates that people have more varied ways to articulate their dissatisfaction, which could imply that negative experiences leave a stronger and more detailed impression than positive ones. This finding is essential for our objective of understanding the nuances in public sentiment. Companies need to be aware of the diverse ways in which negative sentiments can manifest, and prepare to address a wide array of public concerns effectively.

In respect to the machine learning model for sentiment analysis, The evaluation of hyperparameter tuning revealed that the model's performance was robust to slight changes in regularization strength, with the best performance observed around a penalty value of 1e-02. This balance between regularization and model complexity ensured optimal model performance without overfitting or underfitting. This insight is critical for developing effective sentiment analysis models, addressing our objective of utilizing advanced machine learning techniques. The robustness of the model to variations in hyperparameters suggests that it can be reliably used for sentiment analysis across different datasets and conditions, providing consistent insights into public sentiment.

The final model's performance, with a high ROC AUC score of 0.9134 and a reasonable accuracy rate of 0.8274, demonstrated its effectiveness in sentiment analysis tasks. The model's ability to correctly classify both negative and positive sentiments, despite some errors, underscores its utility for real-world applications. The high number of true positives and true negatives indicates strong predictive capabilities, while the presence of false positives and false negatives suggests areas for further refinement. This aligns with our objective of developing a reliable sentiment analysis model. The model's high ROC AUC score indicates that it is particularly effective at distinguishing between positive and negative sentiments, which is crucial for accurate sentiment analysis.

In regards to the limitations of this study, several limitations must be acknowledged:

- The dataset includes tweets only in English and spans a limited time frame (2015-2020), potentially missing out on broader global sentiment.

- While the logistic regression model with Lasso regularization performed well, exploring other models more sophisticated NLP techniques, such as transformer-based models like BERT or GPT which can better understand context and subtle linguistic cues, might yield even better results.
- Expanding the analysis to other social media platforms and integrating additional data sources could provide a more comprehensive view of public sentiment.

Future research should address these limitations by:

- *Expanding the Dataset:* Including tweets in multiple languages and extending the time frame could provide a more comprehensive sentiment analysis.
- *Exploring Advanced Models:* Investigating other machine learning models, such as neural networks, ensemble methods, or context-aware models, could improve predictive accuracy.
- *Refining Sentiment Lexicons:* Developing or incorporating additional, comprehensive sentiment lexicons specifically tailored to capture the nuances of financial language could significantly enhance the accuracy of sentiment classification.

In conclusion, this study successfully demonstrated that sentiment analysis of tweets could provide valuable insights into the public perception of Fortune 500 companies. The logistic regression model with Lasso regularization showed strong predictive performance, and the findings align with existing research, offering new perspectives on market sentiment. By addressing the limitations and exploring future research avenues, this study significantly contributes to the field of financial sentiment analysis, aiding investors and companies in better understanding market dynamics and making informed decisions.

References:

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear." *Procedia-Social and Behavioral Sciences, 26*, 55-62.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35-65.

Silge, J. (2020, May 6). *Sentiment analysis with tidymodels for Animal Crossing user reviews*. YouTube. https://www.youtube.com/watch?v=whE85O1XCkg

Appendix:

Appendix A: Code Snippets

1

```r
# Creating Vector Corpus of tweet text
text_corpus <- Corpus(VectorSource(stockerbot_df$text))
text_corpus <- tm_map(text_corpus, content_transformer(tolower))

# Removing stopwords
my_stopwords <- c(stopwords("en"), "rt", "jul", "inc")
text_corpus <- tm_map(text_corpus, removeWords, my_stopwords)

# Removing urls from tweets
remove_url_http <- function(x) gsub("http[^[:space:]]*", "", x)
text_corpus <- tm_map(text_corpus, content_transformer(remove_url_http))
remove_url_www <- function(x) gsub("www\\S+", "", x)
text_corpus <- tm_map(text_corpus, content_transformer(remove_url_www))

# Removing anything other than english letters and space
remove_cust_punc <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
text_corpus <- tm_map(text_corpus, content_transformer(remove_cust_punc))
text_corpus <- tm_map(text_corpus, removePunctuation, preserve_intra_word_dashes = TRUE)
text_corpus <- tm_map(text_corpus, removeNumbers)
text_corpus <- tm_map(text_corpus, stripWhitespace)

# Stem corpus documents
text_corpus <- tm_map(text_corpus, stemDocument)
```

```r
wc_tweets <- wordcloud(text_corpus, min.freq = 100, colors = brewer.pal(8, "Set2"), random.order = F)
```

2

```r
#tolower
stockerbot_df$text <- tolower(stockerbot_df$text)
#remove alphanumberic words
stockerbot_df$text <- gsub("[^0-9A-Za-z///' ]", "", stockerbot_df$text)
#remove links
stockerbot_df$text <- gsub("http\\w+", "", stockerbot_df$text)
#remove retweet (rt)
stockerbot_df$text <- gsub("rt", "", stockerbot_df$text)
#remove @
stockerbot_df$text <- gsub("@\\w+", "", stockerbot_df$text)

sent_tweets <- sentiment(stockerbot_df$text, polarity_dt = hash_sentiment_loughran_mcdonald)
stockerbot_df$sentiment <- sent_tweets$sentiment
positive_tweets <- head(unique(stockerbot_df[order(sent_tweets$sentiment, decreasing = TRUE), c(2, 9)]), 25)
write.table(positive_tweets$text, file = "/Users/anujprabhu/dat301/Honors Contract/Tweet Sentiment Analysis/tweets/positive_tweets.txt")

negative_tweets <- head(unique(stockerbot_df[order(sent_tweets$sentiment), c(2, 9)]), 25)
write.table(negative_tweets$text, file = "/Users/anujprabhu/dat301/Honors Contract/Tweet Sentiment Analysis/tweets/negative_tweets.txt")

# Combine positive and negative tweets into a single character vector
combined_text <- c(negative_tweets$text, positive_tweets$text)

# Create a corpus from the combined text
pos_neg_corpus <- Corpus(DirSource(directory = "/Users/anujprabhu/dat301/Honors Contract/Tweet Sentiment Analysis/tweets"))

# Preprocess the corpus
pos_neg_corpus <- tm_map(pos_neg_corpus, content_transformer(tolower))
pos_neg_corpus <- tm_map(pos_neg_corpus, removePunctuation)
pos_neg_corpus <- tm_map(pos_neg_corpus, removeNumbers)
pos_neg_corpus <- tm_map(pos_neg_corpus, removeWords, stopwords("en"))
pos_neg_corpus <- tm_map(pos_neg_corpus, stripWhitespace)
pos_neg_corpus <- tm_map(pos_neg_corpus, stemDocument)

corpus_tdm <- TermDocumentMatrix(pos_neg_corpus)

corpus_matrix <- as.matrix(corpus_tdm)
colnames(corpus_matrix) <- c("Negative Tweets", "Positive Tweets")

comparison.cloud(corpus_matrix, max.words = 100, random.order = F, colors = c("darkred", "darkgreen"))
```

3

```r
overall_sentiment <- sum(stockerbot_df$sentiment)
overall_sentiment
```

4

```r
# Use aggregate to calculate the sum of sentiment scores for each symbol
summed_sentiment <- aggregate(stockerbot_df$sentiment, by=list(stockerbot_df$symbols), FUN=sum)

# Rename the columns of the aggregated data frame
colnames(summed_sentiment) <- c("Symbol", "SummedSentiment")

# Display the result
# Sort the summed sentiment data frame in decreasing order of sentiment scores
summed_sentiment <- summed_sentiment[order(summed_sentiment$SummedSentiment, decreasing = TRUE), ]

# Create the Pareto chart using ggplot2
score_by_symbol_plot <- ggplot(summed_sentiment, aes(x = reorder(Symbol, -SummedSentiment), y = SummedSentiment)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  theme_minimal() +
  labs(title = "Pareto Chart of Sentiment Scores by Company Symbol",
       x = "",  # Empty x-axis label
       y = "Summed Sentiment Score") +
  theme(axis.text.x = element_blank())  # Remove x-axis labels
sbsp_plotly <- ggplotly(score_by_symbol_plot)
sbsp_plotly

#ggsave("score_by_symbol_plot.png", score_by_symbol_plot) #width: 12, height: 8
```

5

```r
kable(head(summed_sentiment), format = "html") %>%
  kable_styling(full_width = TRUE, bootstrap_options = "hover")
```

6

```r
kable(tail(summed_sentiment), format = "html") %>%
  kable_styling(full_width = TRUE, bootstrap_options = "hover")
```

7

```{r}
dtm <- DocumentTermMatrix(text_corpus)
text_td <- tidy(dtm)

text_loughran <- text_td %>%
  inner_join(get_sentiments("loughran"), by = c(term = "word"))

# Filter words with sentiment scores
positive_words <- text_loughran %>%
  filter(sentiment == "positive") %>%
  arrange(desc(count))

negative_words <- text_loughran %>%
  filter(sentiment == "negative") %>%
  arrange(desc(count))
```

```{r}
# Use aggregate to calculate the sum of frequency for each positive word
summed_freqs_pos <- aggregate(positive_words$count, by=list(positive_words$term), FUN=sum)

# Rename the columns of the aggregated data frame
colnames(summed_freqs_pos) <- c("Term", "Frequency")

# Display the result
# Sort the summed sentiment data frame in decreasing order of sentiment scores
summed_freqs_pos <- summed_freqs_pos[order(summed_freqs_pos$Frequency, decreasing = TRUE), ]

# Create the Bar chart using ggplot2
freq_by_term_plot_pos <- ggplot(summed_freqs_pos, aes(x = reorder(Term, Frequency), y = Frequency)) +
  geom_bar(stat = "identity", fill = "yellow") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Positive Words",
       x = "Term",
       y = "Frequency")

# Use aggregate to calculate the sum of frequency for each positive word
summed_freqs_neg <- aggregate(negative_words$count, by=list(negative_words$term), FUN=sum)

# Rename the columns of the aggregated data frame
colnames(summed_freqs_neg) <- c("Term", "Frequency")

# Display the result
# Sort the summed sentiment data frame in decreasing order of sentiment scores
summed_freqs_neg <- summed_freqs_neg[order(summed_freqs_neg$Frequency, decreasing = TRUE), ]

# Create the Bar chart using ggplot2
freq_by_term_plot_neg <- ggplot(summed_freqs_neg, aes(x = reorder(Term, Frequency), y = Frequency)) +
  geom_bar(stat = "identity", fill = "brown") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Negative Words",
       x = "Term",
       y = "Frequency")

# Combine positive and negative plots side by side
combined_plot <- grid.arrange(freq_by_term_plot_pos, freq_by_term_plot_neg, ncol = 2)

# Display the combined plot
combined_plot

#ggsave("combined_plot.png", combined_plot, width = 12, height = 8)
```

8

```
```{r}
collected_metrics <- lasso_grid %>%
  collect_metrics()
roc_plot <- collected_metrics %>%
  ggplot(aes(penalty, mean, color = .metric)) +
  geom_line(size = 1.5, show.legend = FALSE) +
  facet_wrap(~.metric) +
  scale_x_log10()
roc_plot

#ggsave("roc_plot.png", roc_plot)
```
```

9

```
```{r}
lasso_importance_plot <- final_lasso %>%
  fit(sent_train) %>%
  pull_workflow_fit() %>%
  vi(lambda = best_model$penalty) %>%
  group_by(Sign) %>%
  arrange(desc(abs(Importance))) %>%
  slice_head(n = 20) %>%
  ungroup() %>%
  mutate(Importance = abs(Importance),
         Variable = str_remove(Variable, "tfidf_text_"),
         Variable = fct_reorder(Variable, Importance)) %>%
  ggplot(aes(x = Importance, y = Variable, fill = Sign)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~Sign, scales = "free_y")
lasso_importance_plot

#ggsave("lasso_importance_plot.png", lasso_importance_plot)
```
```

10

```
```{r}
sent_final <- last_fit(final_lasso, data_split) #final model

sent_final %>%
  collect_metrics()

sent_final %>%
  collect_predictions() %>%
  conf_mat(rating, .pred_class)
```
```