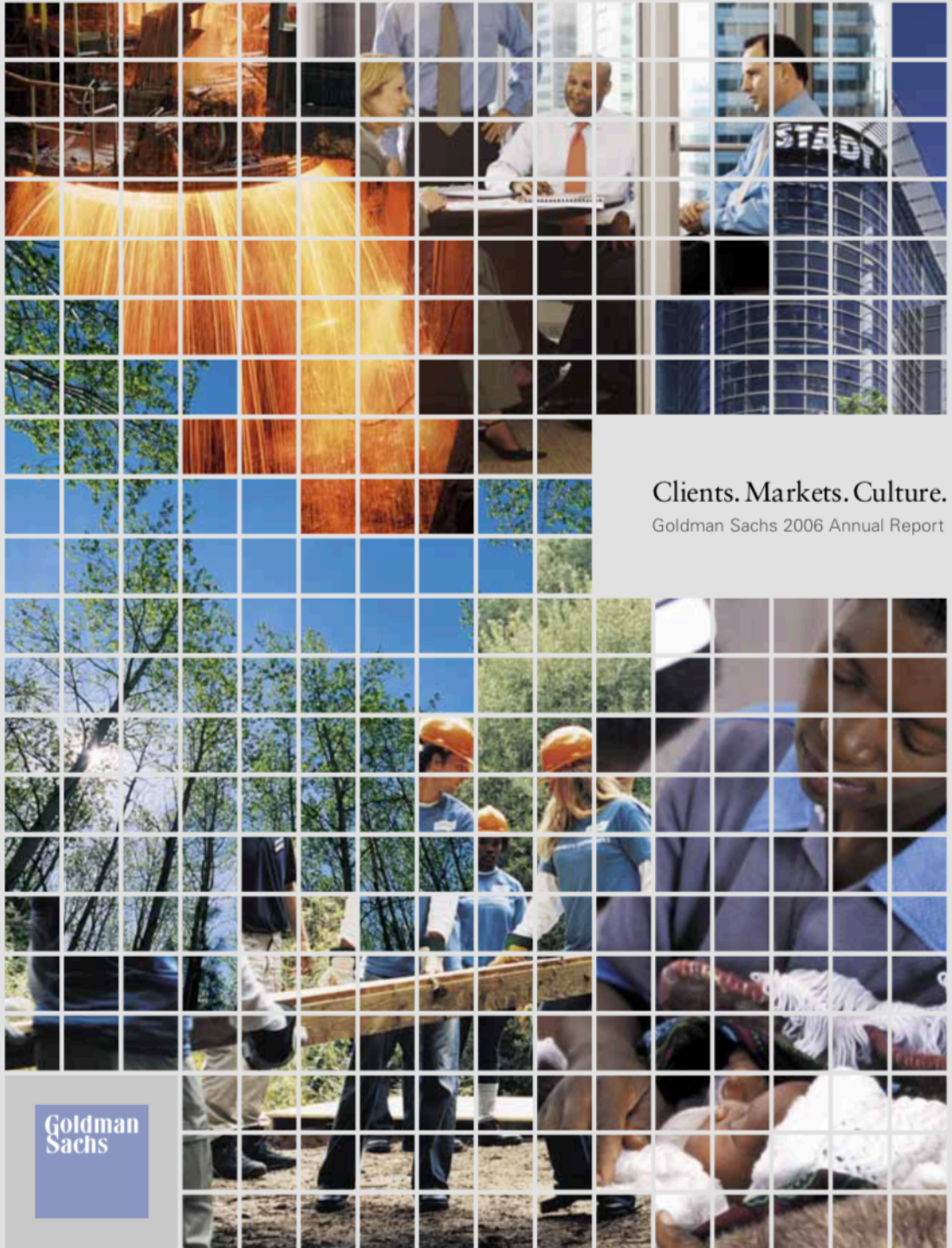


Developing a Document Analysis Tool for the Investment Industry

By: *Samuel Leadley*

Problem Statement

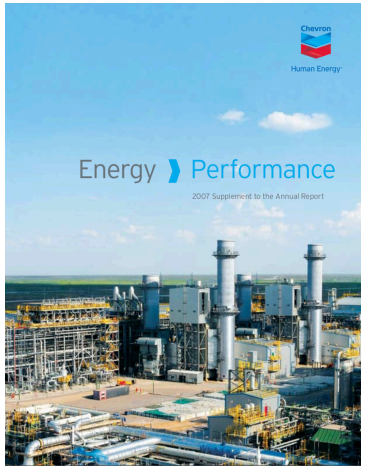
Annual reports, SEC filings, earnings call transcripts, and other text based disclosures are relied heavily on by investment analysts to value and assess companies but they are verbose and jargon heavy.

Can a document analysis tool be created for the investment management industry?

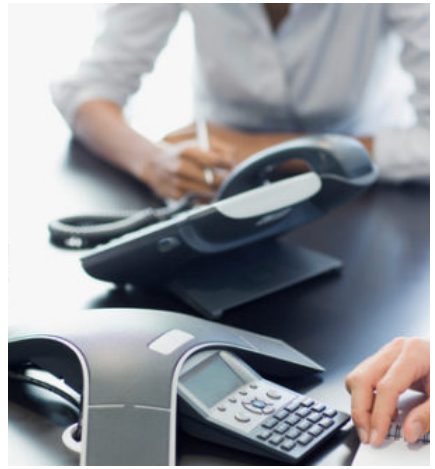
Data Gathering

Possible Data Sources

SEC Filings



Annual Reports



Earnings Call Transcripts

What are Letters to Shareholders?

- A letter written by top executives to shareholders
- Contained in the Annual Report
- Provides overview of firms operations throughout the year

Data Set

Companies	Sectors
Goldman Sachs	Financials
Bank of America	
Chevron	Energy
Haliburton	
Qualcomm	InfoTech
IBM	
Adobe	
Xerox	
Universal Health Service	Health Care
United Health Group	
CVS	

Years	2000 - 2018
Total Documents	166

Creating Target Variable

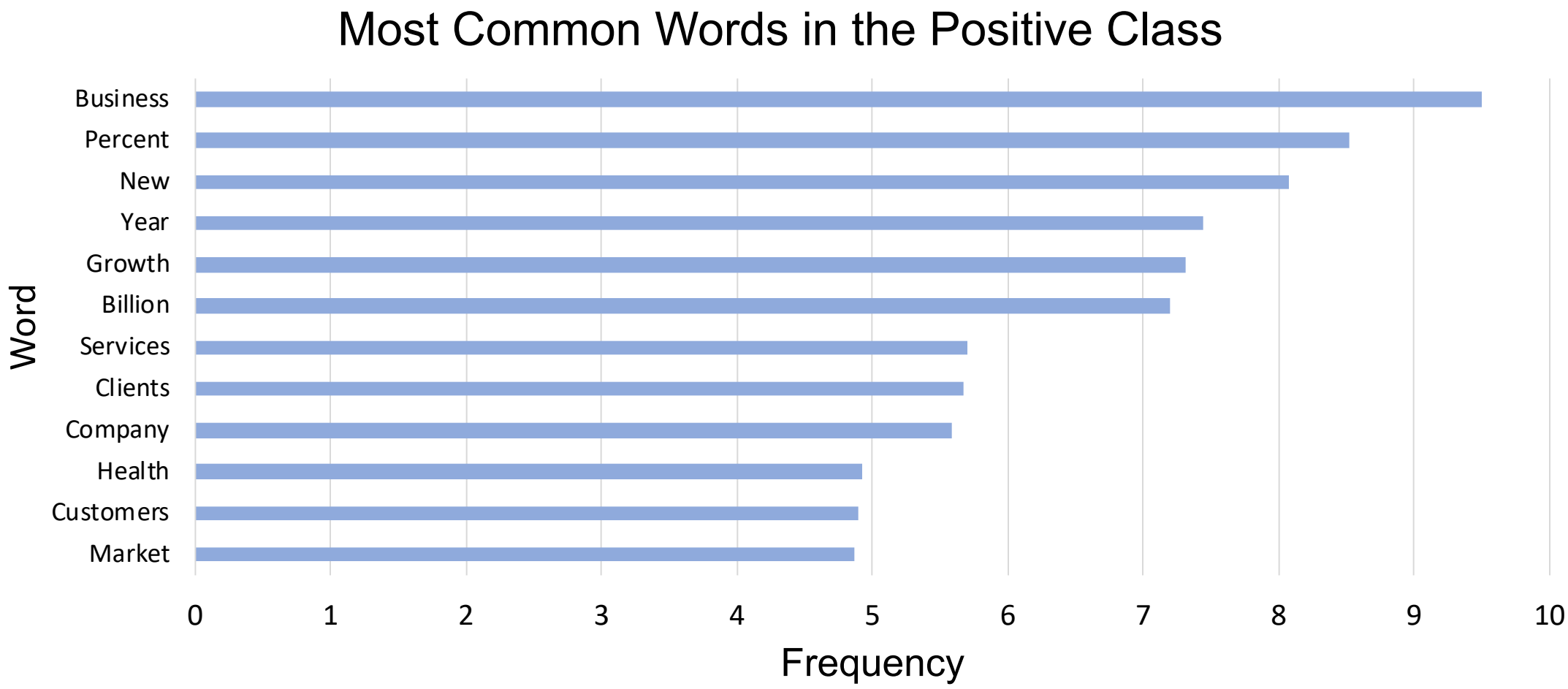
Creating Target Variable

Year-over-year Change in Net Income

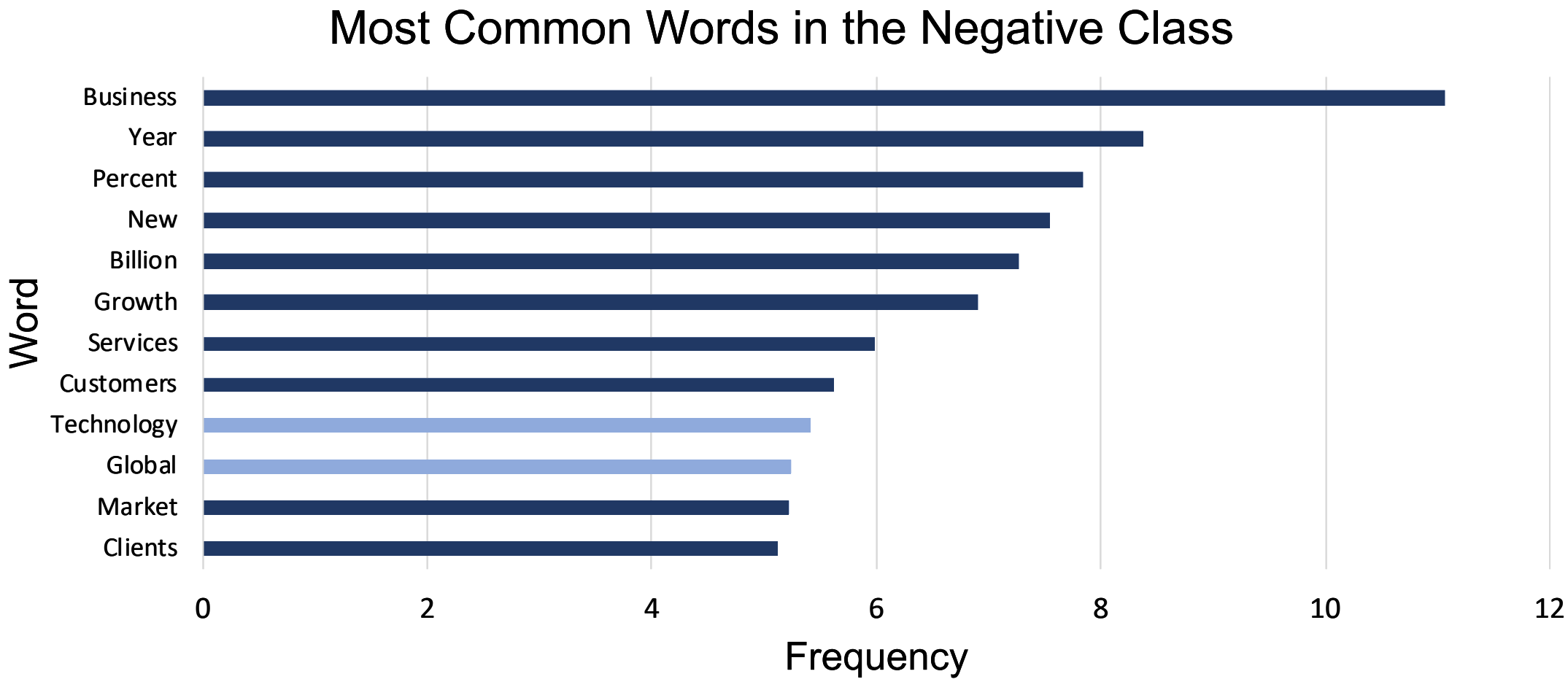
- Revenues – all costs
- Better reflection of earnings

Exploratory Data Analysis

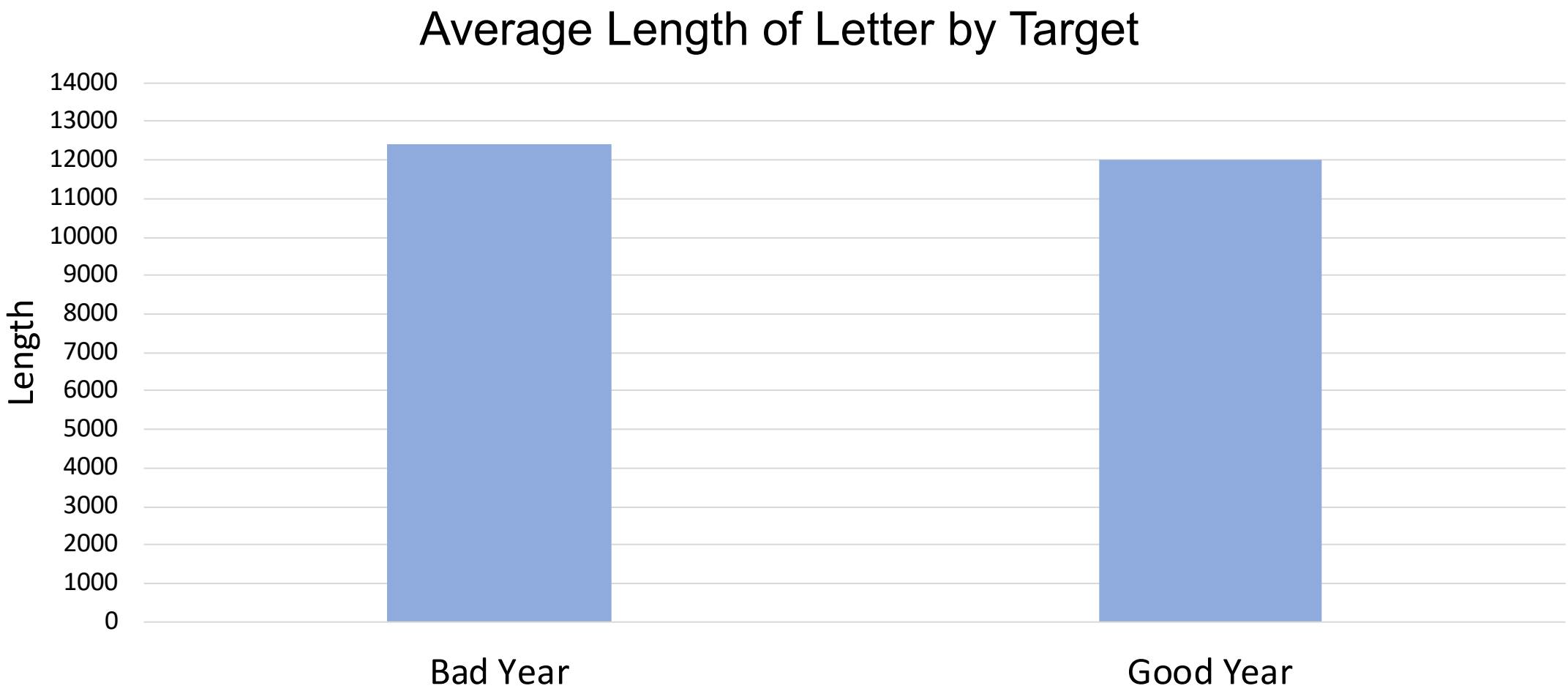
Frequent Words in the Positive Class



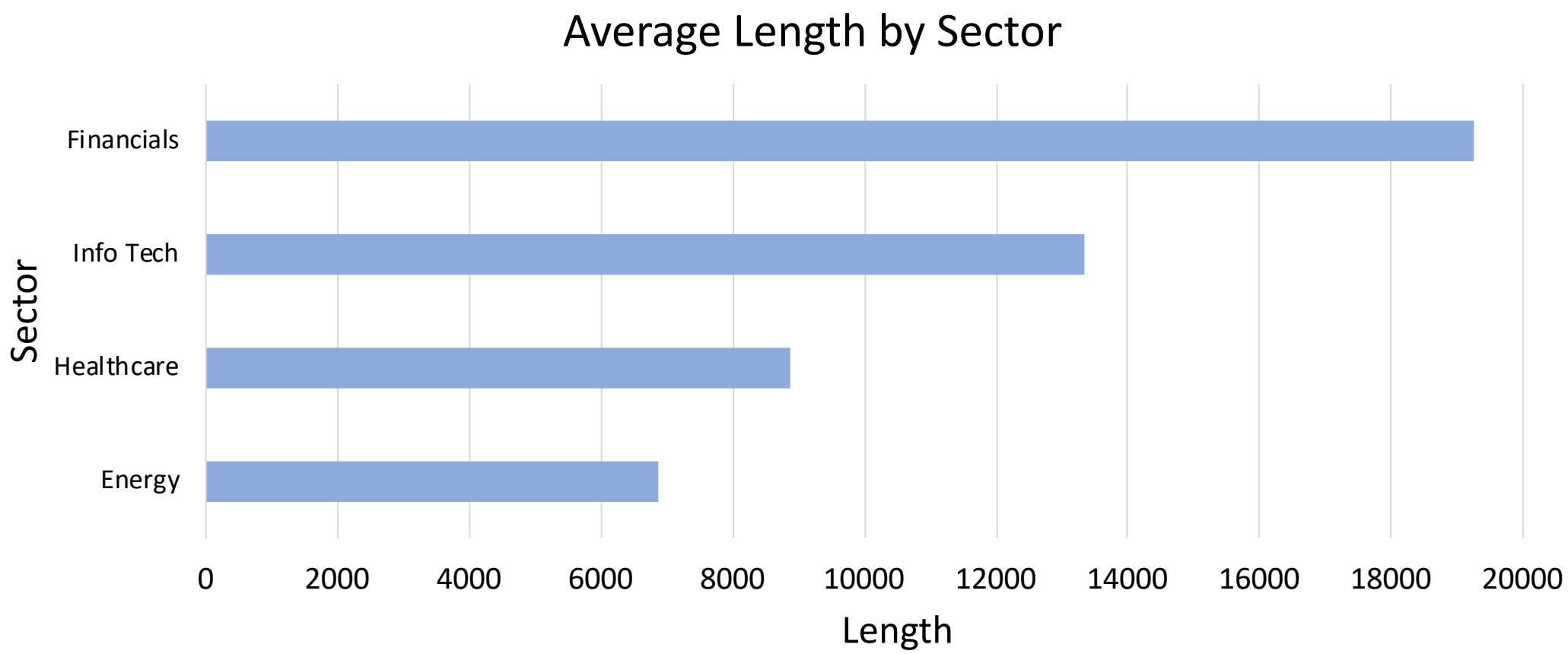
Frequent Words in the Negative Class



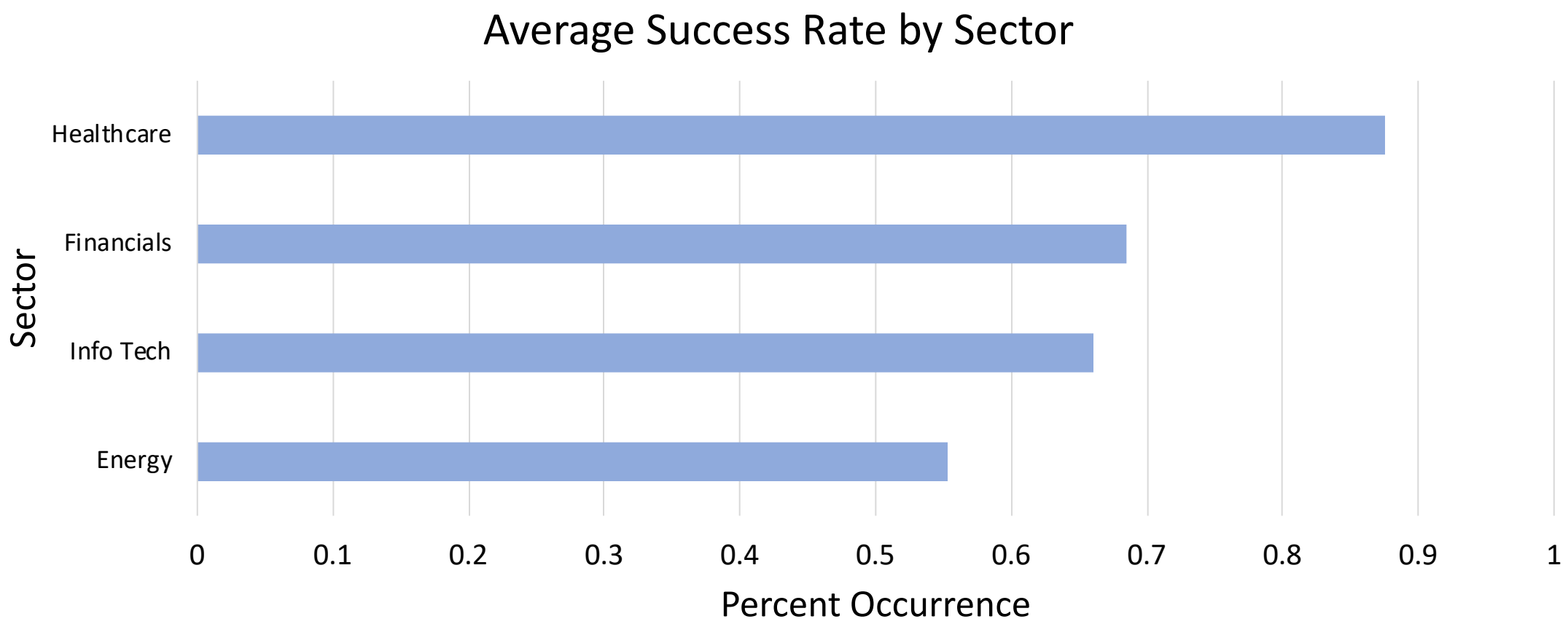
Comparing Lengths of Letters by Target



Comparing Lengths of Letters by Sector



Comparing Sectors by Target



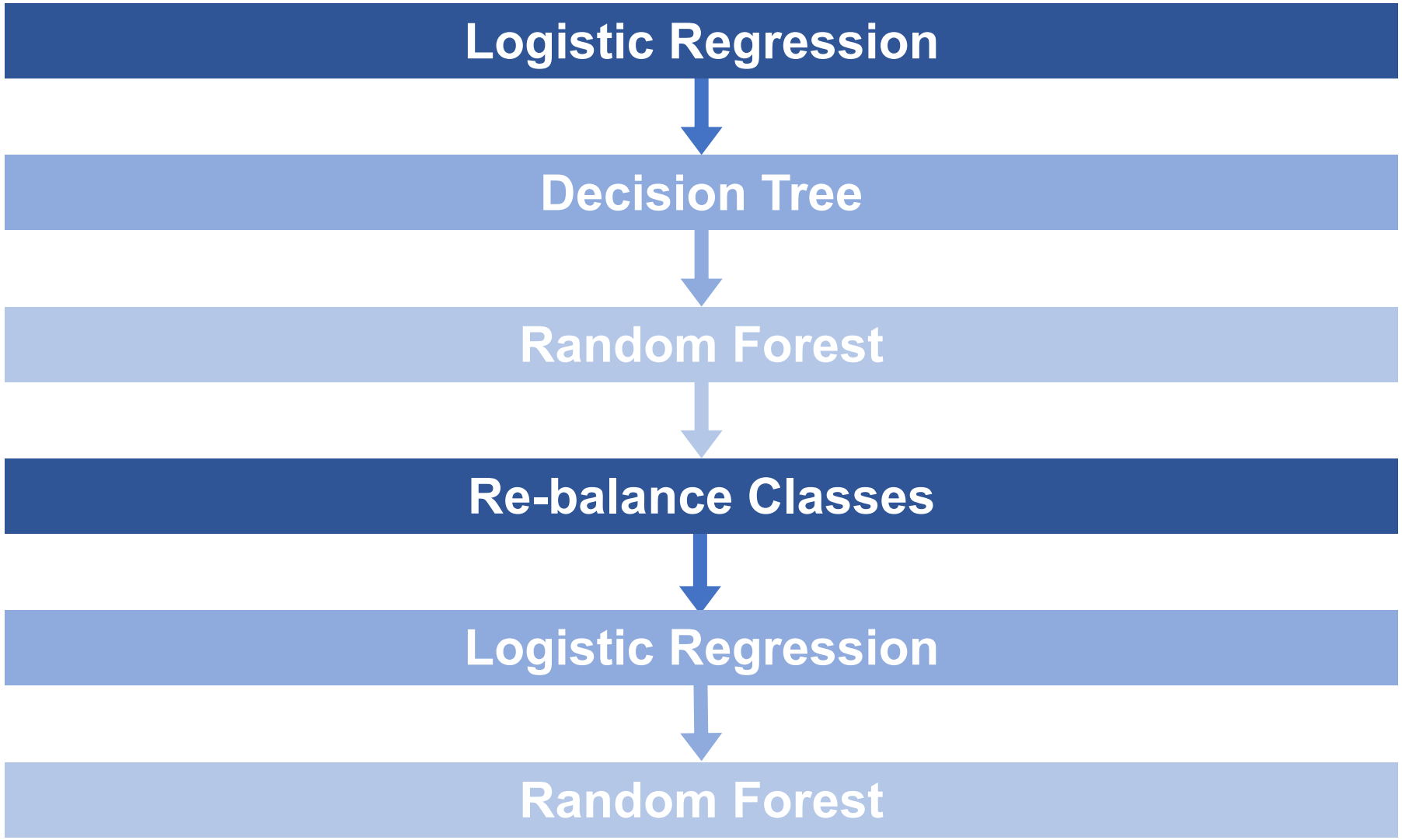
Modeling and Evaluation

Preprocessing

1. Create custom stop words

2. Lemmatize letters and stop words

Modeling


Logistic Regression
Decision Tree
Random Forest
Re-balance Classes
Logistic Regression
Random Forest

Evaluation

Model	Training Accuracy	Testing Accuracy
Logistic Regression	69.35%	69.04%
Random Forest	87.90%	69.04%
Baseline	69.27%	

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	4	10
Actual Positive	0	28

Conclusions and Recommendations

Conclusions

- Shareholder letters are difficult to predict
  - Executives paint a rosy picture
  - They are written similarly year-year
- Too small of a data set

Recommendations

- More data is needed
  - Algorithm to collect data
- Additional data sources

Thank You