10K Text Mining for Predictive Analysis

Loan Le

Introduction

Project Objective

- Evaluate the sentiment of a firm's ***SEC Form 10-K***
- Predictive of the firm's ***stock fluctuation*** – stock price categorically increased or decreased

Dataset

- 3 sections from ten 10Ks:
  - ***Risk Factors*** - outlines potential risk a company expects to face.
  - ***Management Discussion -*** outlines the company's short term & long term goals, future expenses, and most importantly estimated future revenues
  - ***Auditor's Comments -*** an optional segment included in the report to delineate any additional information that could highlight the company's prospects moving forward.
- ***10 observations of 6 variables***
  - Ticker (char)
  - 3 sections mentioned above (char)
  - Stock price pre-10K (numeric)
  - Stock price post-10K (numeric)

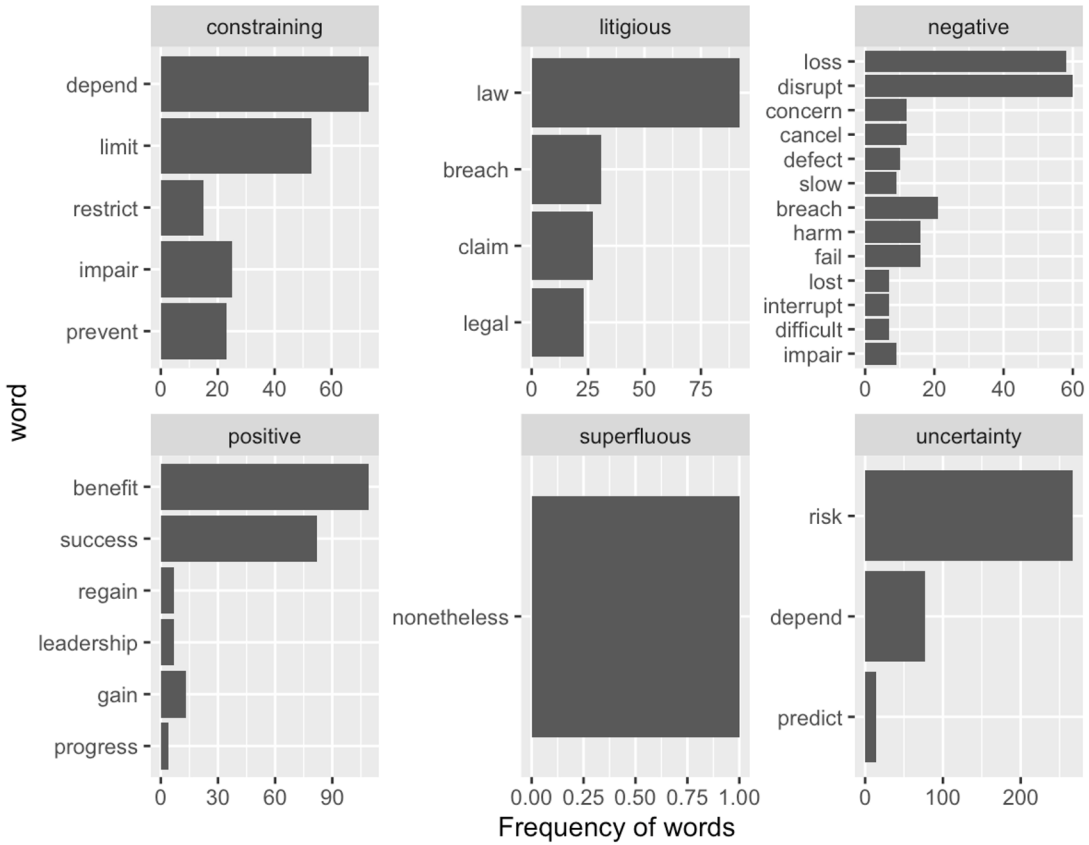
10
Companies

Data Cleaning

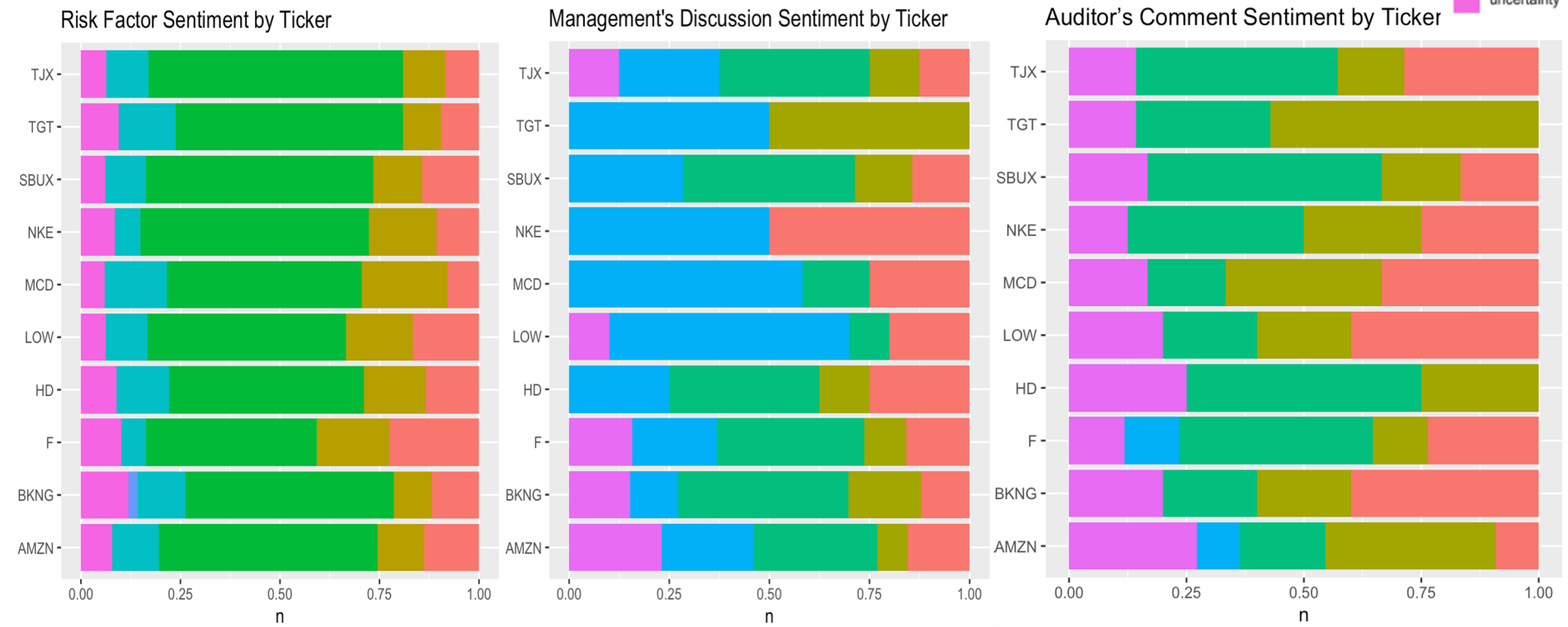
- Added new factor variable - ***Movement***
  - Deliminates if stock increase or decreased from pre 10K price to post 10K price
- Cleaning out ***text variables*** for text mining
  - Convert all words to lowercase
  - Remove punctuation
  - Remove whitespace
  - Remove stop words (e.g. the, a, in)
  - Stem words

Text Mining - Common Words for Each Sentiment

Finance Lexicon (Loughran)



Text Mining -
Finance Lexicon (Loughran)



Predictions

Split Data - 80% Train & 20% Test to predict ***Movement***

Model	Predictor Variables	Train Accuracy %	Test Accuracy %
Logistic Regression	Sentiment + Word Count	65.06%	51.92%
Logistic Regression	Sentiment + Word + Word Count	61.68%	47.77%
Decision Tree	Sentiment + Word Count	52.77%	50.96%
Decision Tree	Sentiment + Word + Word Count	72.77%	46.67%

Results

	predTree2	
	decrease	increase
decrease	150	62
increase	51	152

- Best model is ***Decision tree*** featuring ***Sentiment, Word, Word Count*** variables
    - Had to remove unique words in test but not in train
- Financial library - ***Loughran*** useful to categorize words into sentiments
- ***Larger sample size*** needed for more accurate predictions

Thank you!

Loan Le