Analyzing Filings Sentiment for Applications in Finance

MARSHAL J. WILL

Overview

- Motivation
- Definitions
- Past Research
- Extracting Data
- Extracting Features
- Observations
- Model
- Application

Motivation: Can change of sentiment in Financial Filings be used to make a prediction of a future stock movement?

Definitions

- 10-K

Required annual filings for publicly traded firms

- 10-Q

Required quarterly filings for publicly traded firms

- Edgar

Securities and Exchange Commission (SEC) site of all publicly traded forms

Previous Research

- Loughran and McDonald create word list that is known to have economic meaning for finance and accounting
- Annual reports have become less readable overtime as indicated by Fog Index
- Abnormal rates of return if high negative or positive sentiment in report

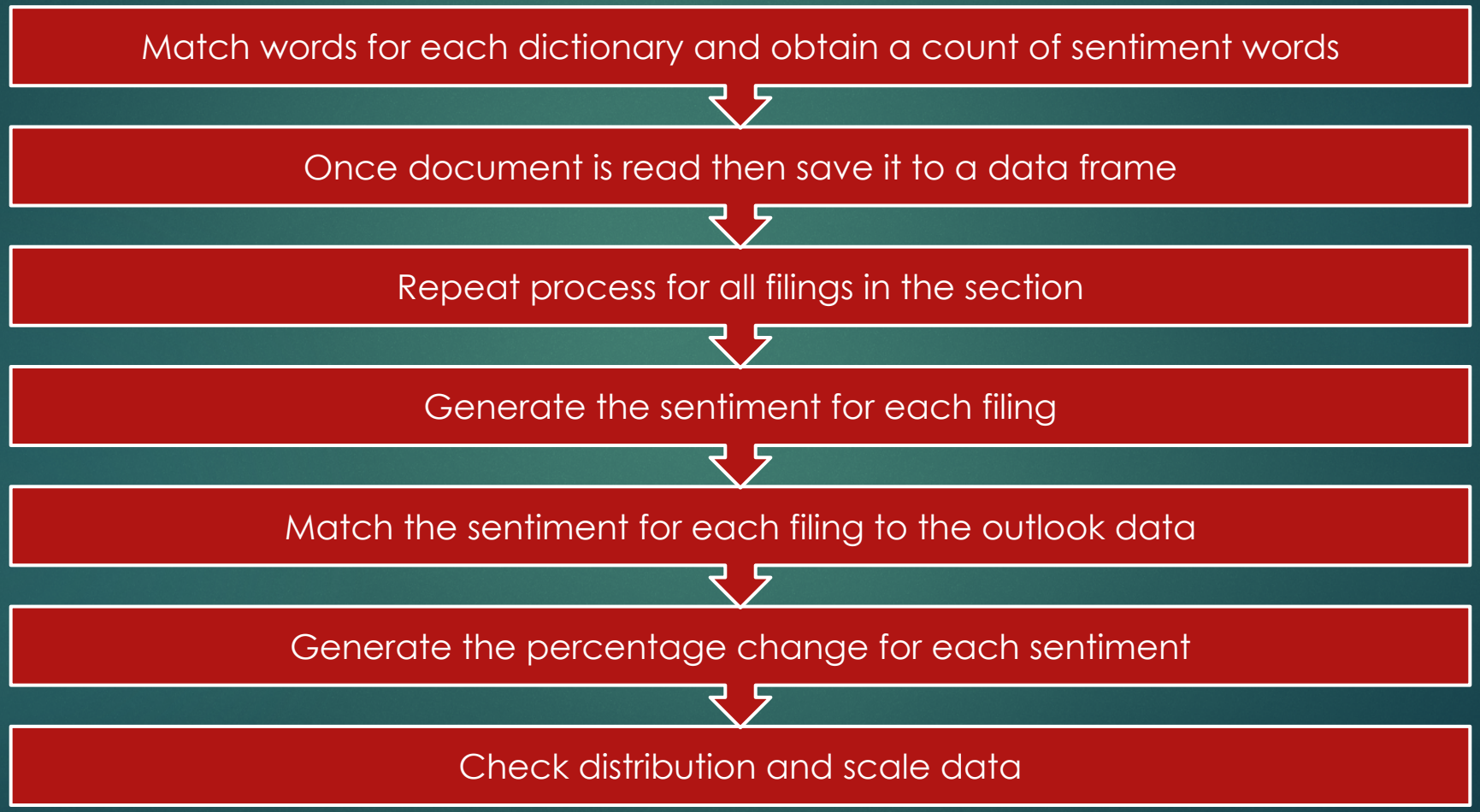
Previous Research

- Firms with high constraining sentiment found to have high cash reserves and pay lower dividends
- Specific Harvard psychology type words have heavy use in certain industries example: cancer, or capital
- Tone in text may capture other information important for investors example: fraud

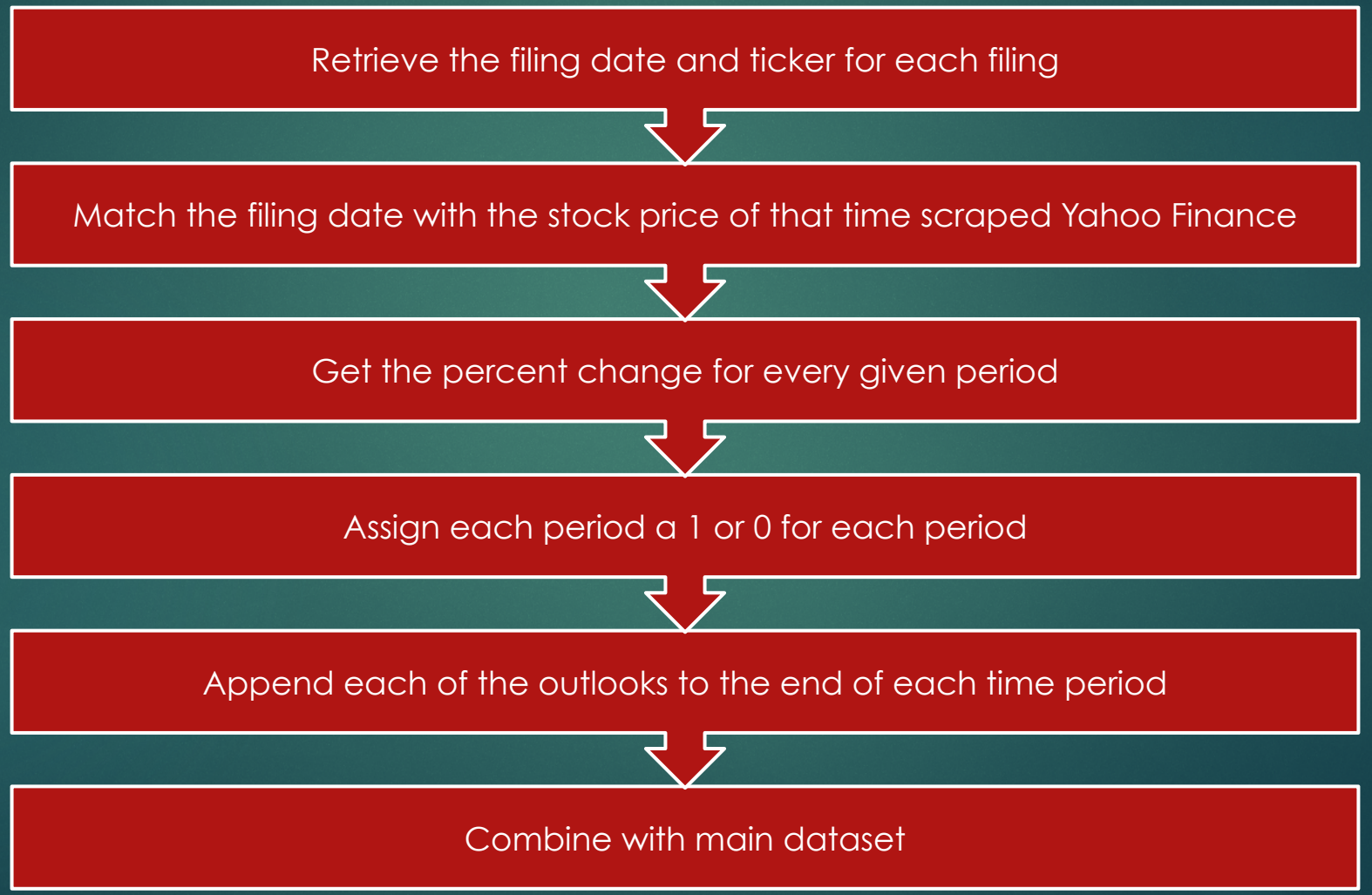
Research Question: Would sentiment of filings be useful in predicting a positive or negative return?

Process of Obtaining Data

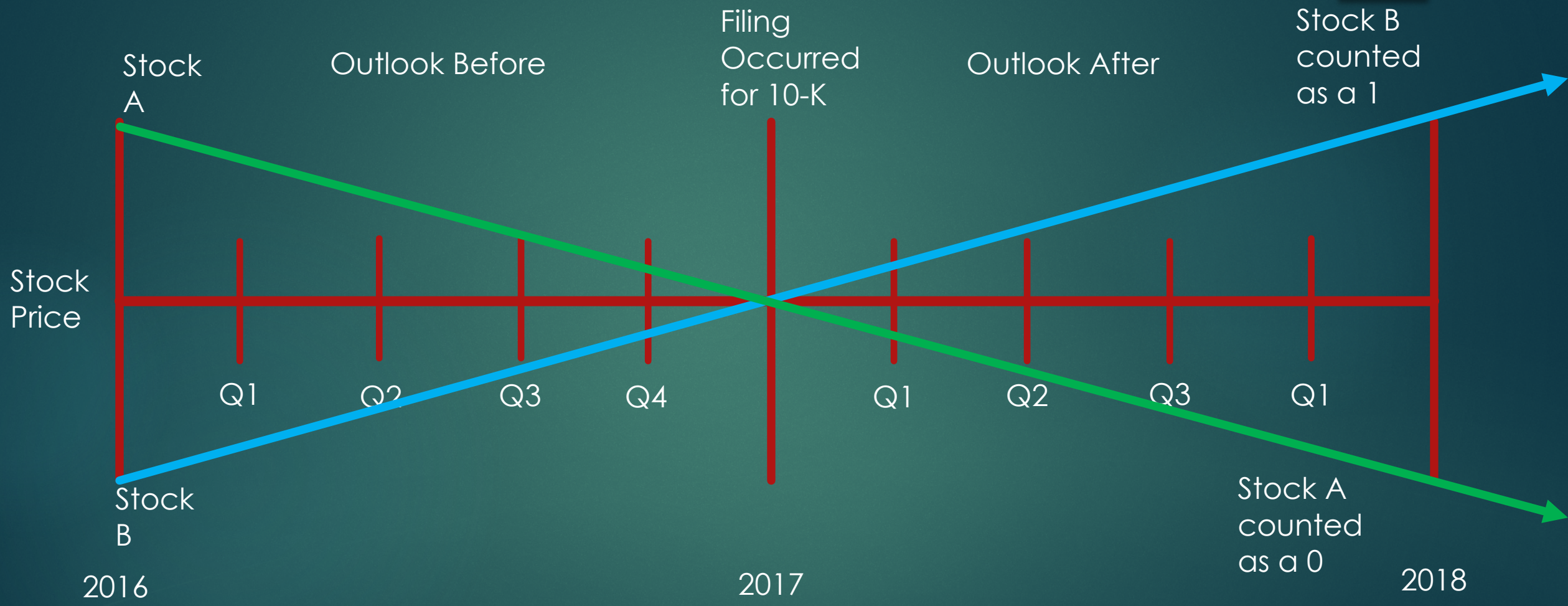
Process of Feature Extraction of Predictors



Process of Feature Extraction Predictors



Predictor Variable



Features Being Used

- Word Sentiment
- Size and Scope of filing
- Percentage Change of Sentiment
- Type of Filing (10-K or 10-Q)
- Farma and Frank Industry Indicator
- Proportion of Sentiment in Filing

Loughran and McDonald Dictionary

- Research of the most financial relevant words based from University of Notre Dame
- Includes Negative, Positive, Uncertain, Litigious, and Constraining word types
- Includes Harvard IV based words that measure psychological based words
- Words are based on importance in financial accounting and business reporting

Types of Sentiment (Based on Loughran and McDonald Dictionary)

Positive
- Able, Enable, Gain

Negative
- Accident, Alleges, Shortfall

Uncertain
- Doubt, Precaution, Risky

Litigious
- Adjourn, Appeals, Lawyer, Lawsuit

Constraining
- Depends, Limits, Require

Negation
- Not, Without, Against

Harvard IV Model Words (Psychology Based Words)

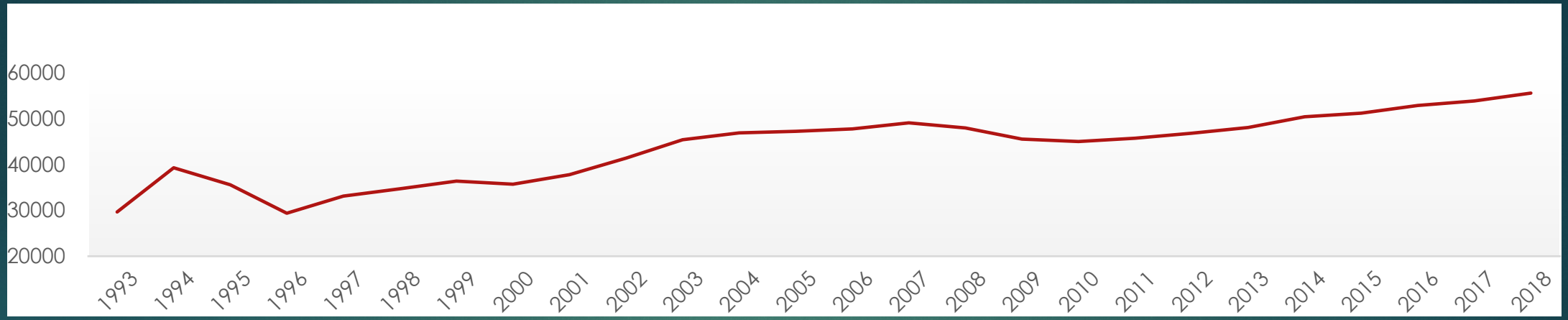
- Strong Model
  - Exe: Best, Clearly, Must
- Moderate Model
  - Exe: Can, Generally, Usually
- Weak Model
  - Exe: Almost, Perhaps, Suggest

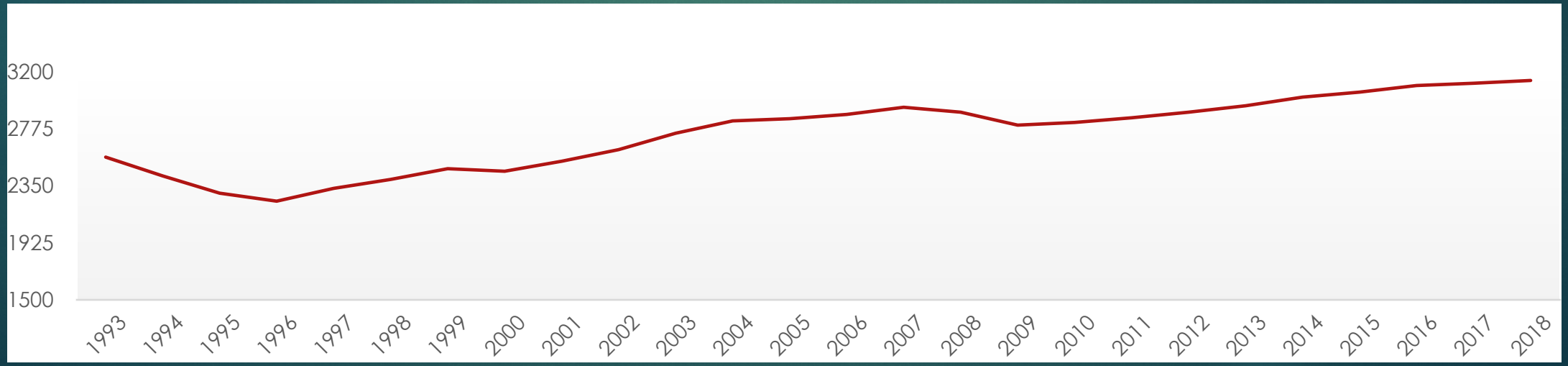
Example of Final Dataset

Ticker	Outlook_bef	Outlook_afte	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change	Perc_change
MMM	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
MMM	1	1	0.64858431	0.18734644	2.28985507	0.60526316	2.28571429	1.15686275	4.85714286	1.61111111	0.375	0.58823529	4	0.00990715	0.53957386
MMM	1	0	-0.0437723	0.01396793	-0.1145374	0.01639344	0.19254658	0.13181818	0.17073171	0.08510638	0.22727273	-0.1666667	-0.2	0.14862446	-0.2229705
MMM	0	1	0.14090439	0.07397959	0.34825871	-0.0725806	0.09895833	-0.0040161	0.02083333	0.09803922	0.33333333	0.2	0.75	0.09424599	0.14238797
MMM	1	1	0.07112009	0.06840855	-0.0479705	0.49565217	-0.0758294	0.02822581	-0.2244898	0.10714286	0.11111111	0	0	0.07785083	0.07305435
MMM	1	1	1.0951421	0.27034237	1.50387597	0.55232558	1.26153846	1.27843137	2.34210526	0.77419355	2.025	1.83333333	3.42857143	0.99079247	0.9761967
MMM	1	1	0.32585281	0.1120056	0.10526316	0.29213483	0.28344671	0.57659208	0.42519685	-0.0272727	0.04958678	0.34640523	0.09677419	1.58403692	0.29704189
MMM	1	0	-0.0394943	-0.0040919	-0.0266106	-0.1478261	0.0229682	-0.1310044	0.14917127	0.08411215	-0.1574803	0.30582524	-0.1764706	2.67948302	-0.038182
MMM	0	1	-0.145486	-0.0496207	-0.152518	-0.0680272	-0.1070812	-0.4032663	-0.4567308	0.06896552	-0.1962617	-0.3159851	-0.25	0.00302784	-0.1278707
MMM	1	1	0.27947175	0.13701363	0.35483871	0.34306569	0.16054159	0.52842105	0.48672566	0.09677419	-0.0348837	0.5	0.33333333	0.17212546	0.2541345

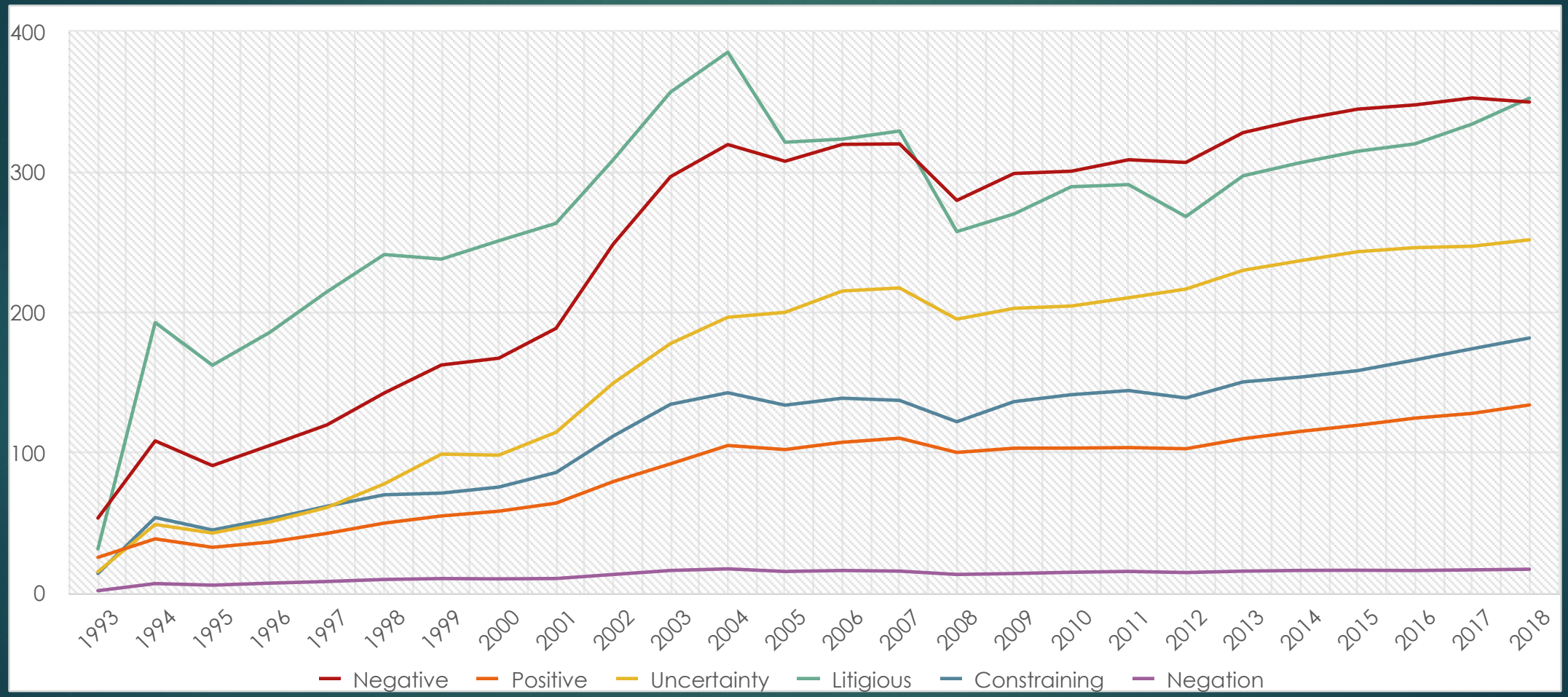
Number of rows:
1026674

Filing Complexity has increased over time

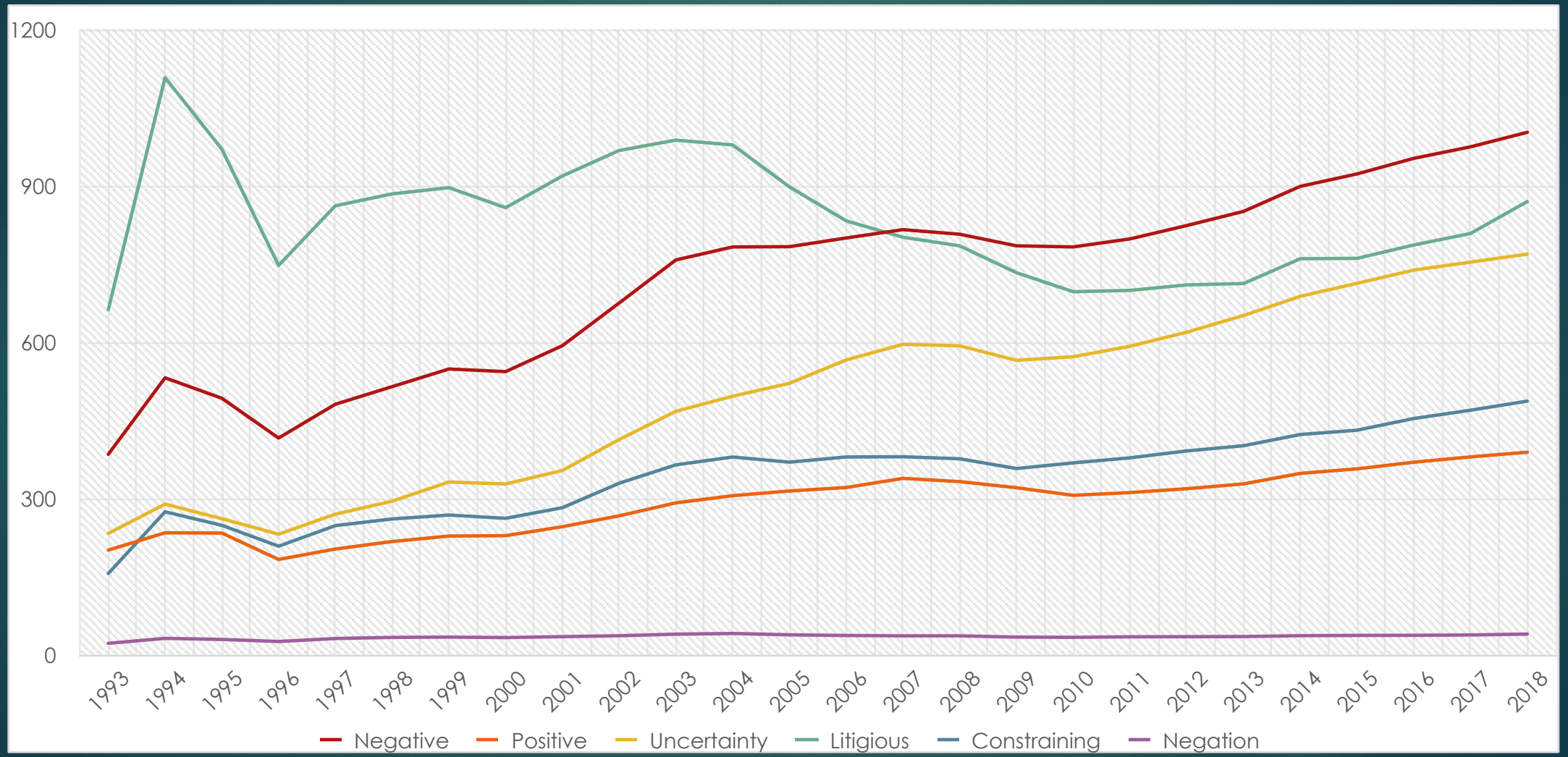




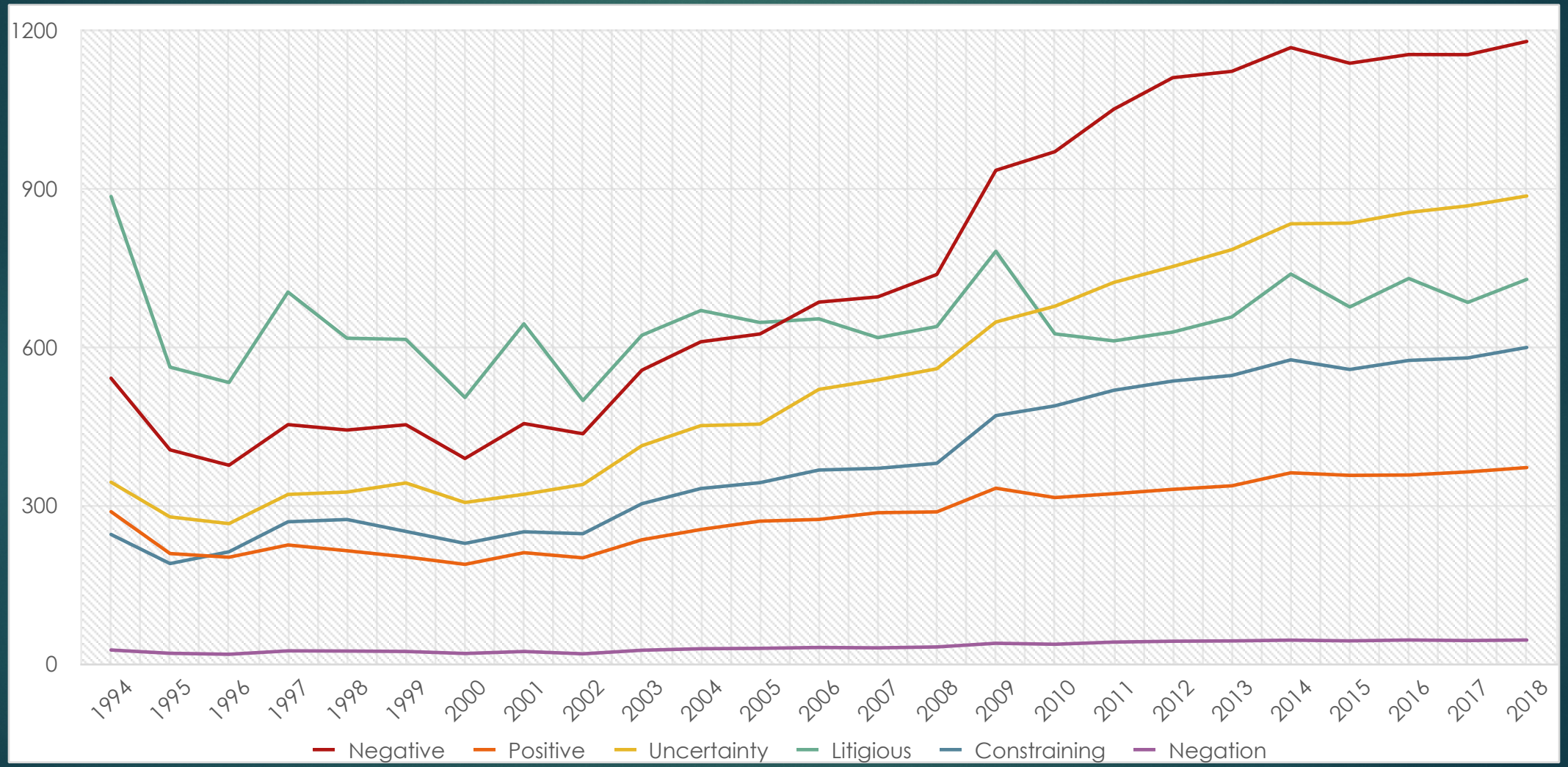
Average Sentiment Over Time for 10-Q Filings



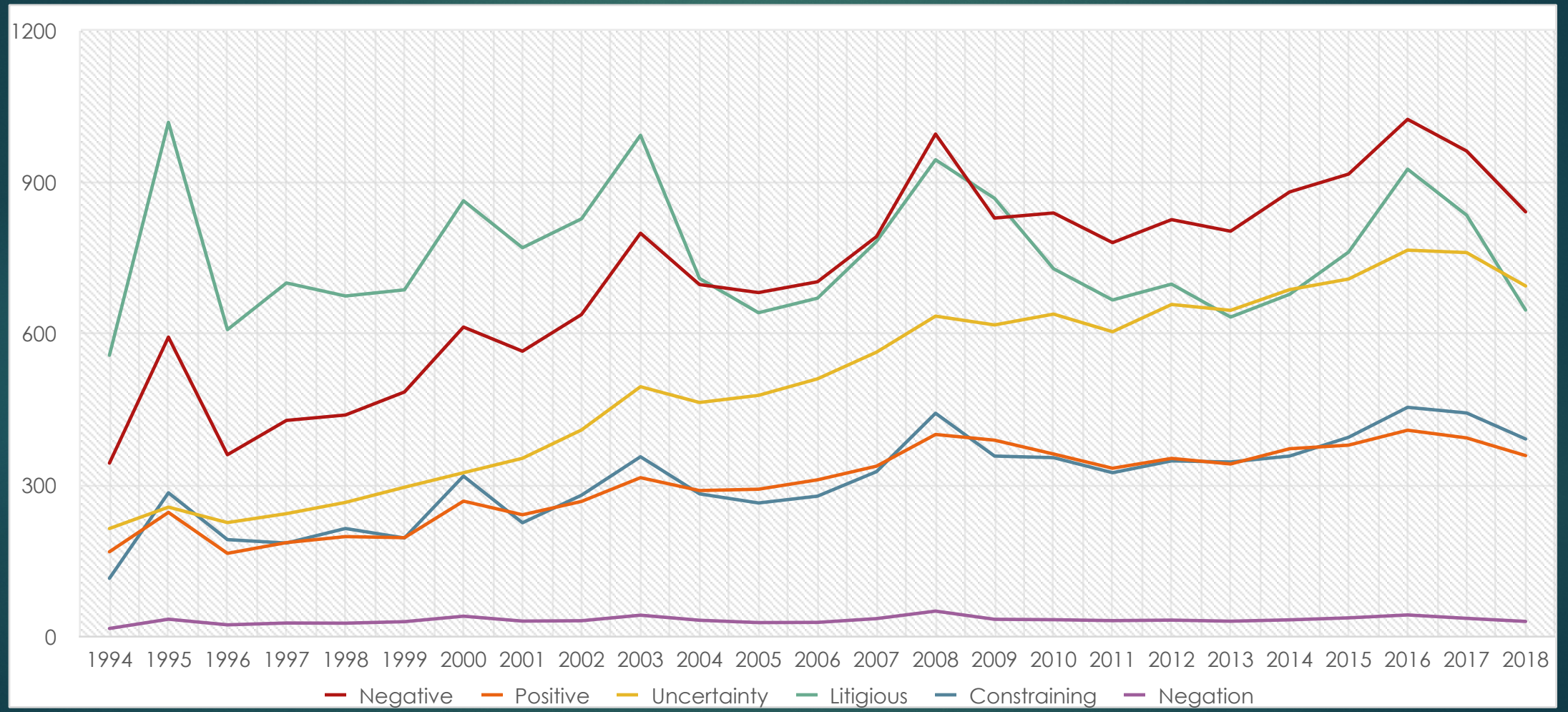
Average Sentiment Over Time for 10-K Filings



Average Sentiment for Banking Sector 10-K's



Average Sentiment for Technology Sector 10-K's



Correlation

Highest Correlations Before Filing	Highest Correlations After Filing
• Gross Filing Size • Unique Word Count • Uncertainty Word Count	• Positive Word Count • Unique Word Count

Models Considered for 10-Q for all years

Models	Accuracy	Precision	Recall	F1 Score	ROC - AUC
Logistic Regression	.54	.54	.95	.69	.55
Gaussian Naïve Bayes	.54	.54	.95	.69	.51
MLP	.55	.57	.67	.61	.56
KNeighbors	.55	.57	.66	.61	.57
Bootstrap Forest	.65	.58	.65	.65	.67

Multilayer Perceptron (MLP) uses 3 layers (100,50,25)

Models Considered for 10-K for all years

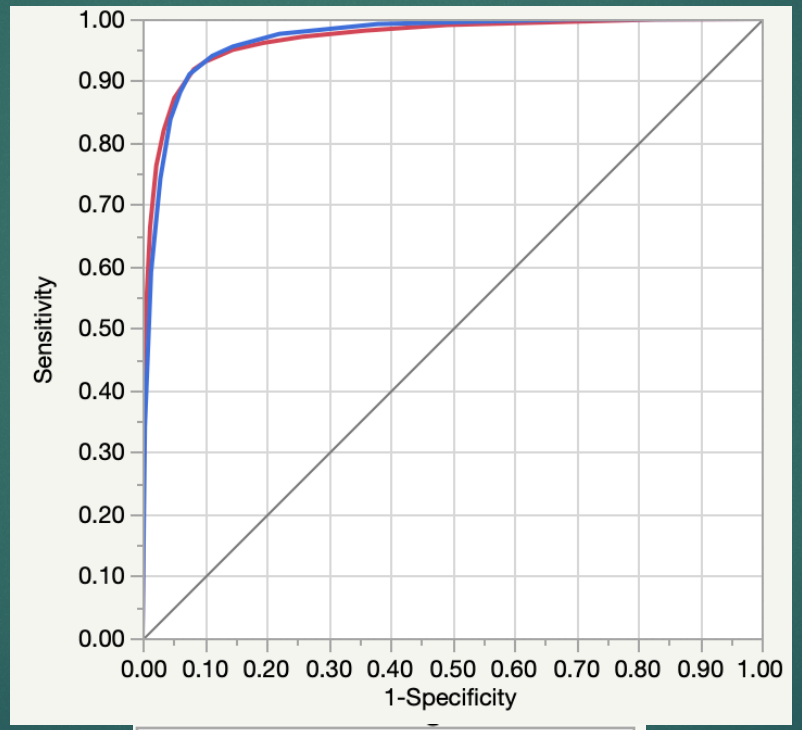
Model	Accuracy	Precision	Recall	F1 Score	ROC - AUC
Logistic Regression	.57	.58	.94	.73	.60
Gaussian Naïve Bayes	.58	.58	.94	.72	.56
MLP	.57	.62	.67	.64	.55
KNeighbors	.59	.62	.76	.68	.60
Bootstrap Forest	.62	.65	.70	.72	.75

Multilayer Perceptron (MLP) uses 3 layers (100,50,25)

Bootstrap Random Forest Contributors All Years

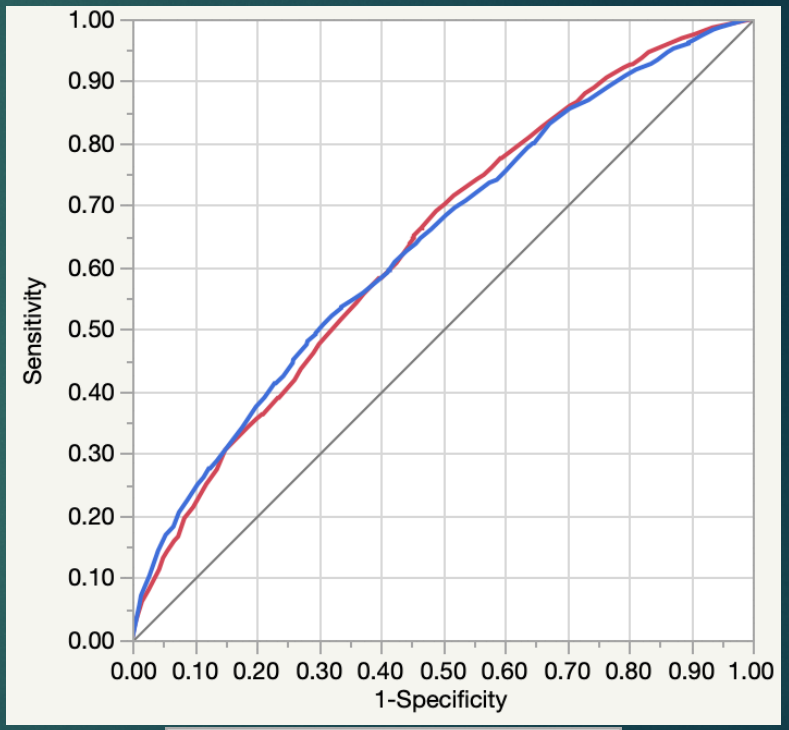
Largest Contributors for 10-K Filing All Years
Percent Change Gross Filing Size
Proportion of Strong Model Words
Proportion of Constraining Words
Proportion of Unique Words

Training



Actual	Predicted Count	
Outlook_after	0	1
0	4787	822
1	287	6244

Validation



Actual	Predicted Count	
Outlook_after	0	1
0	781	1134
1	549	1630

Other Key Findings

- Filing Complexity has increased over time
- Different industries have varying levels of prediction accuracy
- Percent change outside 95% confidence interval was connected to changes in outcomes
- Random Forest Classification overall works the best at predicting outcome
- 10-K have a better overall prediction accuracy than 10-Q's
- 10-K's over the past 5 years have better accuracy
- Second Highest correlation is Negative/Unique words count
- Easy to identify effect, difficult to predict outcome

Application

- Probability generation for Companies, Industries, and Capitalization Size
- Checks for past behavior and may be repetitive
- Sudden changes in sentiment
- Percentage change outside distribution

Predicted Positive Fastenal Probability Next Year
68%

Takeaways

- Better understanding in the challenges in trying to generate Alpha in Finance
- Process of scraping html and cleaning it
- Importance of determining before what you want your features to look like retrieving data
- Analyze previous research and consider their findings in your research

Sources

- Tim Loughran and Bill McDonald, 2011, When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *Journal of Finance*, 66:1, 35-65
- Andriy Bodnaruk, Tim Loughran and Bill McDonald, 2015, Using 10-K Text to Gauge Financial Constraints, *Journal of Financial and Quantitative Analysis*, 50:4, 1-24.
- Tim Loughran and Bill McDonald, 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research*, 54:4,1187-1230.

Questions?