The Determinants of Analyst Forecast Revisions: Evidence From Financial News

Tian Chen

March 2022

1 Introduction

My research studies the value-relevant information embedded in the financial news data using Dow Jones Newswires. I first explore the semantic spaces in the news using word embedding models, which gives us intuition about words most similar to analyst forecast revisions. To answer the research question of how the financial textual news help explain the revisions, I test the explanatory power of sentiments from articles and titles of the news, as well as the tf-idf matrix of texts.

Sentiments of articles are shown to be significantly and positively correlated with analyst forecast revisions in both continuous and binary format, with more statistical power in the binary outcome case. While there are no evidence of significant predictive power for sentiments of titles.

On the other hand, classification problems focus on using financial texts to predict the binary outcome of average revisions, where revisions are aggregated across analysts making revisions on the same announcement day and news are the sum for all the articles one day before the announcement dates. I test classification methods including support vector machine(SVM), logistic, decision trees, K-nearest neighbor, neural network, bagging and random forecast, where SVM outperforms the rest. However, all the methods are unsatisfactory with less than 0.5 of area under the curve (AUC).

2 Literature and Background

Abarbanell and Lehavy (2003) find that analysts' ex post forecast errors are affected by a firm's outstanding stock recommendations, with the firm's incentive to manage earnings being an mediator. If the effects of firm's earnings management cannot be anticipated in the forecasts or analysts deliberately biased the forecasts, there will be an association between firm's earnings management and analysts' forecast errors.

Analyst estimates tend to be optimistic and analyst optimism is greater when estimates are issued early in the forecast period. It is suggested that analyst estimates are upward biased with more than 180 days forecast horizon, unbiased between 91 and 180 days forecast horizon, and negatively biased with 90 days forecast horizon (Cowen et al. (2006)).

Jung et al. (2019) have identified several determinants of analyst forecast revisions, which are prior year's revisions, the level of EPS estimates themselves and future performance signaled by stock splits. They also implement a hedging strategy based on predicted analyst forecast revisions, which earns positive return based on unscaled EPS revisions. Abnormal returns are also found to be associated with firms with lower analyst coverage, which agrees with Healy and Palepu (2001) that firms with lower coverage are subjected to greater undervaluation or overvaluation.

Keskek and Tse(2021) find associations between analyst forecast revisions and upcoming earnings news (the difference between actual earnings and the revised forecast of earnings), which suggests incompleteness of analyst forecast revisions with respect to available information and they show higher post-forecast-revision drift following the incomplete revisions. Moreover, forecast revisions are less complete for industry-wide news than for firm-specific news, and analyst who were optimistic in early period tend to issue less complete forecasts, which generates stronger drifts than others.

Tetlock,Saar-Tsechansky and Macskassy(2008) address the impact of negative words in all Wall Street Journal(WSJ) and Done Jones News Service (DJNS) from 1980 to 2004 in predicting low firm earnings. Moreover, they identify potential profits from using daily trading strategies based on words in DJNS, which is a continuous intraday news source, rather than from based on words updated less frequently (WSJ).

Ettredge, Shane and Smith (1995) focus on the cases where the most recent quarterly earnings report includes an unknown overstatement error, and they find that analysts

effectively ignored 21 percent of the dollar amounts of the overstatements. This suggests analysts using a broader information set, mitigating the effect of earnings overstatements on analysts' earnings forecast revisions.

3 Data

3.1 Textual Data

My data comes from machine text feed and archive database of the Dow Jones Newswires. As previous analysis use texts either from SEC filing or limited news such as the front page *The Wall Street journal*, this database is much more far-ranging in terms of breadth. Following research by Ke et al. (2019), the database has the advantage of being very comprehensive, containing all articles in *The Wall Street journal*, a significantly longer sample than that available from SEC, and include more frequently updated news.

Its articles are time-stamped and tagged with identifiers of firms to which an article pertains.For this exploratory study, I only use data from 2017 and match all the news articles one day before the announcement dates of each analyst forecast revision.

The resulting financial news data are shown in figure 1. We can see that for each revision announcement date there are a substantial amount of news on the day before. And the total amounts of news add up to 1,245,181 for the year of 2017.

Figure 1: overview of the financial news data

anndats	Date	Title	Article	System_GMT_Time	Display_ET_Time	Accession_Number	title_sent	article_sent
2017-01-02	2017-01-01	\nInterbank Foreign Exchange Rates At 23:50 ES...	\n\n \n Latest ...	20170102T045016.304Z	January 01, 2017 23:50 ET (04:50 GMT)	20170101001016	-0.125000	0.071229
	2017-01-01	\nFrench New Car Registrations Up 5.1% in 2016	\n\n \n\n (MORE TO FOLLOW) Dow Jones Newswire...	20170101T131603.377Z	January 01, 2017 08:16 ET (13:16 GMT)	20170101000444	0.068182	0.500000
	2017-01-01	\nRenault New Car Registrations in France Up 0...	\n\n \n\n (MORE TO FOLLOW) Dow Jones Newswire...	20170101T131607.523Z	January 01, 2017 08:16 ET (13:16 GMT)	20170101000444	0.136364	0.500000
	2017-01-01	\nPeugeot New Car Registrations in France Down...	\n\n \n\n (MORE TO FOLLOW) Dow Jones Newswire...	20170101T131611.602Z	January 01, 2017 08:16 ET (13:16 GMT)	20170101000444	-0.009596	0.500000
	2017-01-01	\nTurkish Authorities Hunt for Gunman in Istan...	\n\nBy Margaret Coker and Emre Peker \n\n ...	20170101T132903.071Z	January 01, 2017 08:29 ET (13:29 GMT)	20170101000453	0.000000	0.009984
...	...	...	...	...	...	...	...	...
2017-12-29	2017-12-28	\nUS Chicago Purch Mgmt Adj Dec Index +67.6 vs...	\n\n \n\n (MORE TO FOLLOW) Dow Jones Newswire...	20171228T144546.446Z	December 28, 2017 09:45 ET (14:45 GMT)	20171228002590	0.000000	0.500000
	2017-12-28	\nVP Campbell Surrenders 201 Of Alphabet Inc	\n\n \nSOURCE: Form 4 \n \nISSUER: Alphabet In...	20171228T144556.253Z	December 28, 2017 09:45 ET (14:45 GMT)	20171228002591	0.000000	-0.162500
	2017-12-28	\nZumbrota Hog/Sheep Market For Dec 27	\n\nSource: Central Livestock Association \n \...	20171228T144709.774Z	December 28, 2017 09:47 ET (14:47 GMT)	20171228002592	0.000000	0.100000
	2017-12-28	\nVP Copeland Gifts 3,100 Of Synovus Financial...	\n\n \nSOURCE: Form 4 \n \nISSUER: Synovus Fin...	20171228T143246.937Z	December 28, 2017 09:32 ET (14:32 GMT)	20171228002537	0.000000	0.033333
	2017-12-28	\nPress Release: Cellular Biomedicine Group An...	\n\n Cellular Biomedicine Group Announces Clo...	20171228T123003.016Z	December 28, 2017 07:30 ET (12:30 GMT)	20171228001908	0.000000	0.075608

1245181 rows × 7 columns

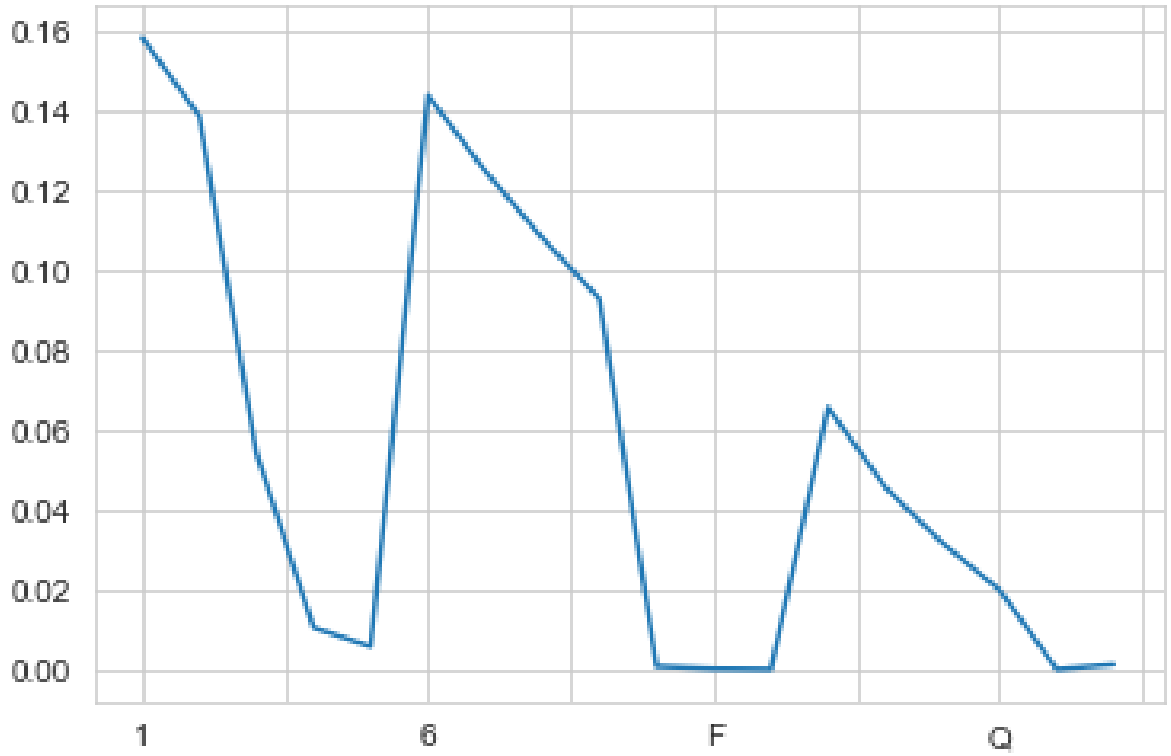
3.2 Analyst Forecast Revision Data

The analyst-level data are collected from Institutional Brokers Estimate System (I/B/E/S). In terms of the variables of interests, *estimator* refers to the sell-side institution or contributor;In the past, I/B/E/S uses 'Broker'. *Analys* refers to the person at the sell-side institution or contributing analyst who makes the forecast. *measure* is a prediction from the analyst about future earnings or another measure for a specific issue/entity and time period, which is earning per share(EPS) in my case. The value of each forecast of the measure is documented in *value*. Announce date(*anndats*) is the date that the forecast/actual was reported.Forecast Period End Date (*fpedats*) is the date to which the estimate applies;For 70% of the companies, an estimate for a particular fiscal year will have an *fpedats* of December 31st of that year. Moreover, because not all companies have the same fiscal year end, Thomson Reuters uses *FPI*(a numeric value or one-character code) to identify estimates for each unique period. Using the reported periods as a base, the period end dates for all estimated periods are easily calculated. For example, if December 2007 is the

last reported annual(assume the calendar year = the fiscal year), the FPI=1, FPI=2 and FPI=3 estimates are for the periods ending December 2008, 2009 and 2010, respectively. Also note FPI=0 is for long term growth.

Figure 2 illustrates the percentage of each forecast period indicator among all firm-analyst-date observations in the full-sized sample from 2013 to 2020. As we can see, observations with FPI=1 make up about 16% of all data, which is substaintial and becomes the major subgroup I will look at.

Figure 2: percentage of observations for each FPI



The resulting analyst forecast data are shown in figure 3.The data are firstly grouped by tickers and analysts, and I only keep the observations with *fpedats* being December 31st of that year and *fpedats* no longer than one year from *anndats*. This results in 615,6128 pieces of revision data, and the number is 35,001 for analyst data in the year of 2017. Furthermore, the forecast revision for a certain analyst targeting at a specific ticker is calculated by taking the difference of his/her *value* for EPS estimates between the latest forecast revision and the last revision the sell-side analyst made for the firm with the

same future period end date.

Figure 3: overview of the analyst forecast revisions data

ticker	analys	cusip	estimator	value	measure	anndats	actdats	anndats_act	fpedats	actual	fpi	year_end	forecast_revision
AA	473.0	01381710	192.0	3.00	EPS	2003-01-08	2003-01-13	2004-01-08	2003-12-31	3.30	1	1	NaN
	476.0	01381710	2383.0	3.24	EPS	2003-10-08	2003-10-08	2004-01-08	2003-12-31	3.30	1	1	0.24
	476.0	01381710	2383.0	3.36	EPS	2003-11-20	2003-11-20	2004-01-08	2003-12-31	3.30	1	1	0.12
	476.0	01381710	2383.0	4.83	EPS	2004-01-09	2004-01-09	2005-01-10	2004-12-31	4.68	1	1	NaN
	476.0	01381710	2383.0	5.85	EPS	2004-03-31	2004-04-01	2005-01-10	2004-12-31	4.68	1	1	1.02
...	...	...	...	...	...	...	...	...	...	...	...	...	...
ZMH	191742.0	98956P10	157.0	4.51	EPS	2020-05-11	2020-05-18	2021-02-05	2020-12-31	5.67	1	1	0.31
	191742.0	98956P10	157.0	4.45	EPS	2020-07-21	2020-07-21	2021-02-05	2020-12-31	5.67	1	1	-0.06
	191742.0	98956P10	157.0	5.00	EPS	2020-08-05	2020-08-17	2021-02-05	2020-12-31	5.67	1	1	0.55
	191742.0	98956P10	157.0	5.05	EPS	2020-10-21	2020-10-21	2021-02-05	2020-12-31	5.67	1	1	0.05
	191742.0	98956P10	157.0	5.57	EPS	2020-11-09	2020-11-09	2021-02-05	2020-12-31	5.67	1	1	0.52

615618 rows × 12 columns

4 Exploring Semantic Spaces

In the data exploration stage, I use the gensim implementation of Word2Vec, with all the articles tokenized and normalized. Taking time constraints into account, I randomly sampled 1% of the textual data I matched with all the revision announcement dates in 2017. Based on the word embeddings of the all the articles in the sample, I could interrogate critical word vectors within my corpus in terms of the most similar words, analogies and other additions and subtractions that reveal the structure of similarity and difference within my semantic space. I also relate the revealing patterns with semantic organization of words in my corpora, bu projecting, and visualize the word embeddings with two separate visualization layout specifications (e.g., TSNE, PCA).

Firstly, there are some interesting word similarity in the corpus related to my research question. According to figuren4, the top similar words for *forecast* include: projection, gdp, downward, consensus,slight, projected, median and forecasts. Interestingly, *forecasts* is similar to *downward* in the given semantic spaces, and this indicates that more downward forecasts are mentioned in the corpora.

Figure 4: most similar words for "forecast"

```
newsW2V.most_similar('forecast')
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_
  """Entry point for launching an IPython kernel.
[('projection', 0.7736994028091431),
 ('gdp', 0.7151694297790527),
 ('downward', 0.7075973153114319),
 ('consensus', 0.6992899179458618),
 ('fy', 0.6907299160957336),
 ('1q', 0.6831955909729004),
 ('slight', 0.6701453924179077),
 ('projected', 0.6678175926208496),
 ('median', 0.6622341275215149),
 ('forecasts', 0.6568032503128052)]
```

Figure 5: most similar words for "revision"

```
newsW2V.most_similar('revision')
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_
  """Entry point for launching an IPython kernel.
[('downward', 0.6949425339698792),
 ('downgrade', 0.692790150642395),
 ('expectation', 0.6820420026779175),
 ('upward', 0.6813136339187622),
 ('indicating', 0.6669085621833801),
 ('negative', 0.6648051142692566),
 ('simulated', 0.6647456288337708),
 ('revised', 0.657730221748352),
 ('assumption', 0.6558386087417603),
 ('stress', 0.6405794024467468)]
```

On the contrary, the most similar words for *revision* includes both *downward* and *upward*, which is shown in figure 5. Since both negative and positive words are closely related to *revision* semantically, more quantitative examination of the relationship between the informativeness of textual news and the revisions are demanded.

Figure 6: sematic relationship between words

```
newsW2V.most_similar(positive=['optimistic', 'forecast'], negative = ['analyst'])
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: DeprecationWarning
  """Entry point for launching an IPython kernel.
[('trend', 0.740863561630249),
 ('somewhat', 0.7294418811798096),
 ('sustained', 0.726111888885498),
 ('slow', 0.723960280418396),
 ('downward', 0.7177110910415649),
 ('slowing', 0.7145942449569702),
 ('seeing', 0.7137243747711182),
 ('modest', 0.7122682332992554),
 ('recession', 0.7080453634262085),
 ('expect', 0.7034046649932861)]
```

Following research by Kozlowski et al. (2019), dimensions produced by word differences such as (positive - negative) in the semantic spaces correspond to dimensions of cultural meaning; also, the projection of words onto these dimensions can reflect widely shared associations.

Figure 6 illustrates the semantic relationship between words" *optimistic + forecast - analyst = downward*. This is to say optimistic to analyst is downward to forecast. This intuition coincide with previous literature, as analysts tend to be overly optimistic in the beginning of the forecast period and revise downward as it approaches to the forecast period end date.

For visualization, we can then extract the vectors and create our own smaller matrix that preserved the distances from the original; then PCA and T-SNE is used to reduce the dimensions (e.g., to 50) and project them down to the two we will visualize. Note note that this is nondeterministic process, and so we can repeat and achieve alternative projections/visualizations of the words.

Figure 7 shows the semantic relationship of top words in the texts. My visualization below puts *u.s.* next to *company* and *forward* near *statements*. *management* is beside *quarter* and *shares*.

Figure 7: Projection of the words

Instead of just looking at just how words embed within in the space, we can look at how the different documents relate to each other within the space using the Doc2Vec technologies. An interesting finding from the word vectors is that: *downward + optimism - negative = revisions*. This indicates that negative to optimism is downward to revisions, which coincides with previous literature that overly optimism in the early stage of forecasts lead to downward revisions in the later forecast periods.

We can plot some words and documents against one another with a heatmap. Figure 8 provides intuitive similarity between words. We can see that there is some significant positive similarity between *upward* and *downward*. Also, *trading* is significantly positively similar to both *upward* and *downward*, but it is negatively associated with *revision*. This gives research ideas of exploring the analyst forecast revisions on substantial stock trading days.

Figure 8: Heatmap of cosine similarity between keyword word vectors

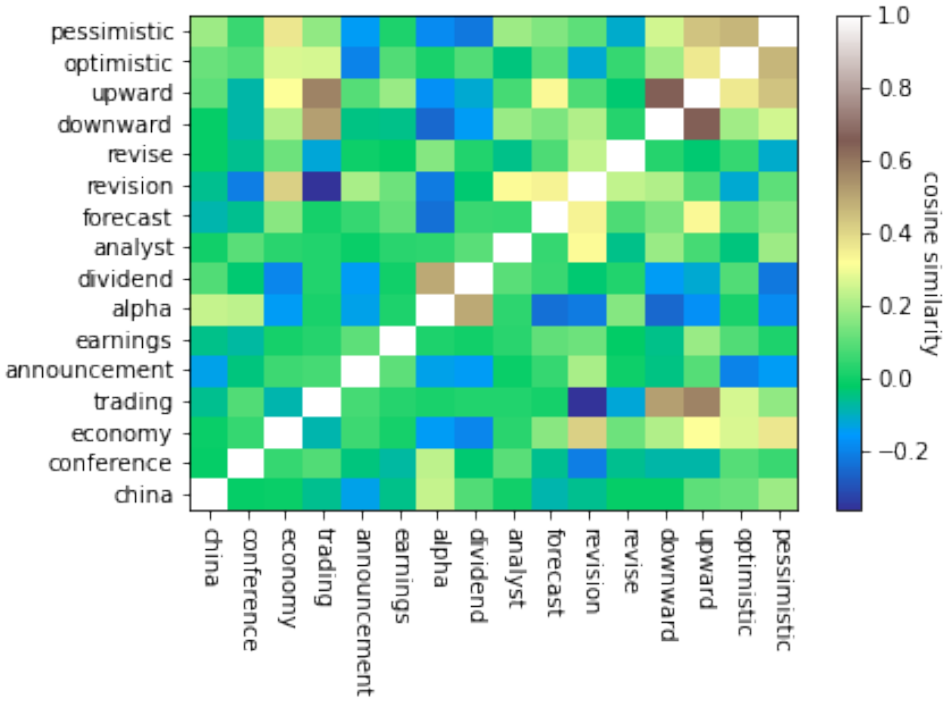
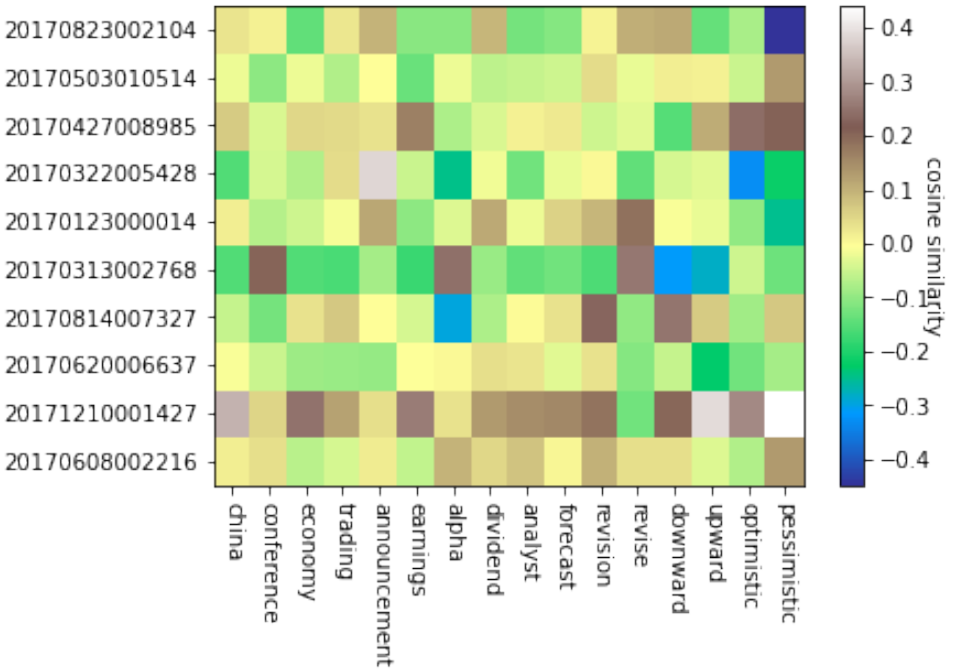


Figure 9 shows the heatmap of similarities between the first ten documents and my keywords selected. We could find a nearly perfectly similar relationship between the *20171210001427* ariticle and the keyword *pessimistic*.

We can examine the text, noting that it mentions words like "under pressure":

Singapore shares edge up Monday helped by strong US jobs data that provide a fresh impetus for risk globally. The benchmark Straits Times Index is up 0.2% at 3432.90 following a 1.1% gain on Friday that finished a choppy week on a strong note. ComfortDelGro is a notable gainer, adding 3.1% after announcing its long-awaited joint venture with Uber late Friday. Singapore Telecommunications and DBS Group are also trading higher. Some real estate and offshore firms, however, are under pressure. Vallianz sheds more than 7% and Sembcorp Marine is down 0.5%.(gaurav.raghuvanshi@wsj.com)

Figure 9: Heatmap of cosine similarity between documents and keywords



Finally, I project word vectors of three groups (earnings, people and actions) onto my dimensions of interests (positivity and certainty). Each dimension is constructed from word vectors of two groups of words with opposing meanings, and I plot each group of words in the two dimensions accordingly. The results are shown in figure 10, figure 11, figure 12 respectively.

Figure 10: Projection of earnings-related words onto the positivity and certainty dimensions

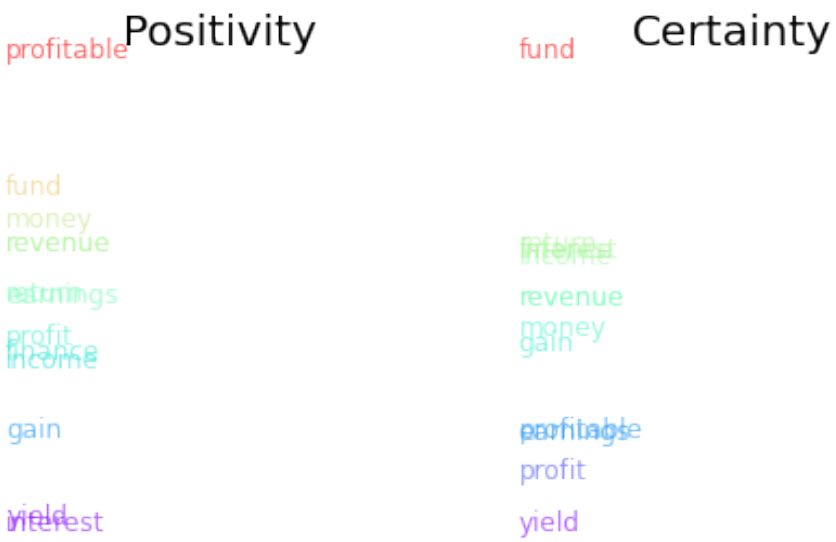


Figure 11: Projection of job-related words onto the positivity and certainty dimensions

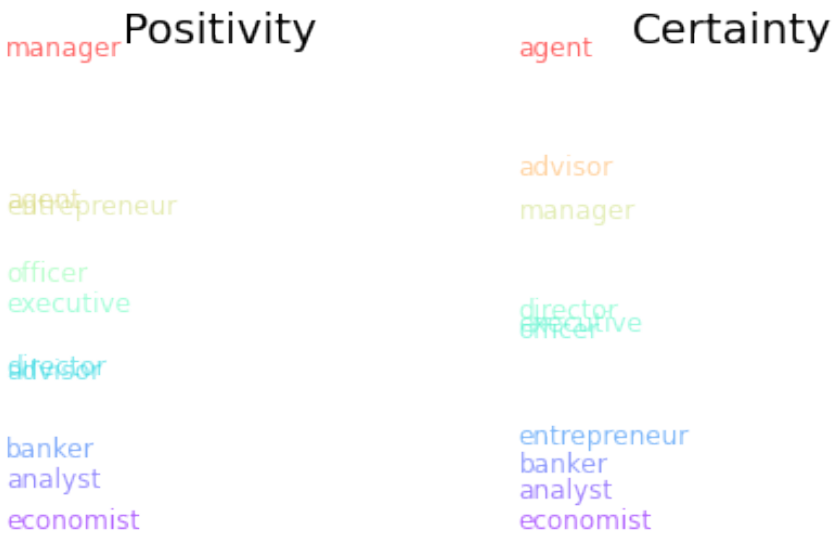
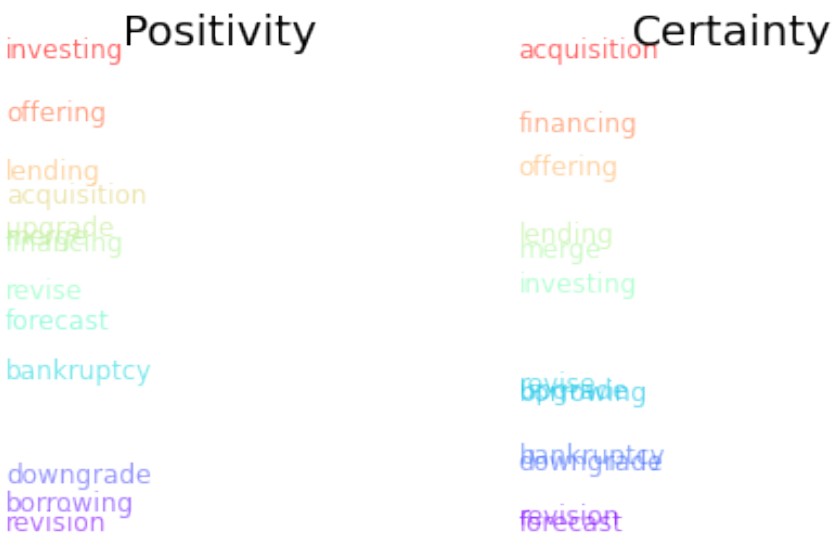


Figure 12: Projection of action-related words onto the positivity and certainty dimensions



Among the first group of words, We can see that *gain*, *interest* and *yield* are among the least positive words and *yield* and *profit* are the least certain words. Interestingly, we find that *analyst* and *economist* and *banker* are among the least positive and least certain words. For the financial-action-related words, it's surprising that *revision* are the most negative and uncertain words, with *forecast* being the least certain words. These projection can provide evidence for previous literature, as well as intuitions that we can further examine in the research.

5 Model and Results of the Empirical Analysis

5.1 Regressions methods

To quantitatively answer the question that how can information from financial news help explain the analyst forecast revisions, linear regressions would be a tremendous start to pin down the associations between low-dimensional information from textual news and analyst forecast revisions.

Sentiment scores are extracted using the package *TextBlob*, which relies on built-in dictionaries of positive and negative words. The sentiment property returns a named tuple of the form Sentiment(polarity, subjectivity). The polarity score is a float within

the range [-1.0, 1.0].

Sentiment scores for every articles and its corresponding titles are calculated using the polarity scores from *TextBlob* sentiment analysis. Since for each of the announcement day of forecast revisions, I use the financial news on the day before the announcements, it remains a question that what are the efficient ways to aggregate all the information from texts on the day before the announcement.

In my analysis, I use a simple and intuitive method - taking mean of all the sentiments on the same dates. This aggregate sentiment on the day before each announcement day represents the value-relevant information incorporated in the texts that may play a significant role in the decision-making process of the analyst forecast revisions.

The results are shown in table1, where I regress the analyst forecast revision on the average sentiment scores from articles, average sentiment scores from titles both separately and jointly. Clustered standard errors are employed, since the independent variable is the same for each announcement date, clustering the standard errors by announcement dates leads to the correct statistics for the linear regression.

Through these three regressions, we could see that only sentiments from articles are significantly positively correlated with the analyst forecast revisions in the single linear regression. For each unit of increase in the average article sentiment scores, the analyst forecast tends to revise upward by 0.7714.

However, when including sentiments from both the articles and titles, article sentiments become insignificant; this may be caused by multicollinearity problem between title sentiments and article sentiments, which leads to a larger standard error and less statistically significant results. Moreover, title sentiment doesn't show significant explanatory power in both single and multiple linear regressions.

Table 1: Main results for continuous forecast revisions

	forecast_revision	forecast_revision	forecast_revision
R-squared	0.0002	0.0001	0.0002
R-squared Adj.	0.0002	0.0001	0.0001
article_sent	0.7714*		0.6646
	(0.4010)		(0.4460)
const	-0.1093**	-0.0414	-0.1122**
	(0.0521)	(0.0321)	(0.0524)
title_sent		1.5963	0.6914
		(1.1497)	(1.2401)
N	35000	35000	35000
R2	0.0002	0.0001	0.0002

Table 2: Main results for binary forecast revisions

	binary_revision	binary_revision	binary_revision
article_sent	1.9460**		2.1162**
	(0.8432)		(0.9400)
const	-0.1630	0.0668	-0.1584
	(0.1117)	(0.0767)	(0.1144)
title_sent		1.7780	-1.1028
		(2.7595)	(2.7736)
N	35000	35000	35000
R2			

Table 2 illustrates the explanatory power of article and title sentiments for the binary analyst forecast revision. The binary outcome equals 1 if the forecast revision is greater than or equal to 0 and 0 otherwise. Surprisingly, article sentiment is significantly and

positively correlated to the binary revision in both single and multiple regressions. In the binary outcome case, the positive association is even more significant than the continuous outcome case, this may because the binary forecast revision can be better pinned down by our sentiment variables.

5.2 Classification methods

Finally, I am interested in the high-dimensional information of financial news text themselves, as extracting one-dimensional sentiments from texts essentially loses substantial information.

Considering the huge computational resources demanded in the current data, I transform the data from firm-analyst-date level to date-level by setting the dependent variable as the means of analyst forecast revisions on each announcement date and converting them to binary outcomes. In terms of the financial news, all the articles are aggregated for the same announcement date of analyst forecast revisions across different analysts and targeting firms. This results in a downsized sample, shown in figure 13. Then the texts are vectorized and converted to tf-idf matrix.

Figure 13: overview of the classification data

	anndats	forecast_revision	Article	binary_revision
0	2017-01-02	-0.776375	\n\n \n\n Items about Japan&aposs markets tha...	0
1	2017-01-03	0.055480	\n\n \n \nThe following is a press release fro...	1
2	2017-01-04	0.116143	\n\nBy Sarah Krouse \n\n Abigail Johnson o...	1
3	2017-01-05	-0.176804	\n\nMorgan Stanley Schedules Quarterly Investo...	0
4	2017-01-06	-0.338777	\n\n \n \nThe following is a press release fro...	0
...	...	...	...	...
315	2017-12-22	0.307982	\n\n \n \n \n ...	1
316	2017-12-26	-0.733250	\n\n \n Sigma Koki Co. (7713.TO) \nPARENT ...	0
317	2017-12-27	0.481429	\n\n The latest Market Talks covering Commod...	1
318	2017-12-28	-0.984071	\n\n \nSOURCE: Form 4 \n \nISSUER: Western All...	0
319	2017-12-29	-0.017500	\n\nSource: ICE \n \nContract Settle \n \nM...	0

320 rows × 4 columns

I test classification methods including support vector machine(SVM), logistic, decision

trees, K-nearest neighbor, neural network, bagging and random forecast, where SVM outperforms the rest. The performances are shown in figure 14 and sorted by AUC scores in descending order.

Figure 14: evaluations of all classifiers

method	Category	Error_Rate	AUC	Precision	Average_Precision	Recall
svm	1	0.484375	0.495074	0.444444	0.450730	0.275862
	0	0.484375	0.495074	0.543478	0.544449	0.714286
logistic	0	0.484375	0.483251	0.537037	0.538724	0.828571
	1	0.484375	0.483251	0.400000	0.445797	0.137931
tree	0	0.531250	0.446305	0.510638	0.522027	0.685714
	1	0.531250	0.446305	0.352941	0.432397	0.206897
knearest	0	0.531250	0.446305	0.510638	0.522027	0.685714
	1	0.531250	0.446305	0.352941	0.432397	0.206897
nn	1	0.546875	0.443842	0.384615	0.429501	0.344828
	0	0.546875	0.443842	0.500000	0.521429	0.542857
bag	0	0.609375	0.377833	0.450000	0.497054	0.514286
	1	0.609375	0.377833	0.291667	0.414152	0.241379
random forest	1	0.656250	0.329064	0.217391	0.412481	0.172414
	0	0.656250	0.329064	0.414634	0.482644	0.485714

I train each classifier with the training data and evaluate the classifier using the test data to get relevant evaluations based on measures like precision, recall, the F-measure, and AUC. Among all classification methods, SVM has the best performance, with error rate of 0.484375 and AUC of 0.495074. Logistic classification performs second best, with a slightly lower AUC score of 0.483251. However, all the methods are unsatisfactory with less than 0.5 of area under the curve (AUC). All the tree-based methods tend to perform badly for my data.

If we take a close look at the ROC curve in figure 15, we will find out that in many times the True positive rate is less than the False positive rate, which is a bad indication for the classification. Moreover, figure 16 visualize the classification results among the

observations, where the observations for category 0 and 1 are still messed up together.

Figure 15: ROC curve for the SVM classifier

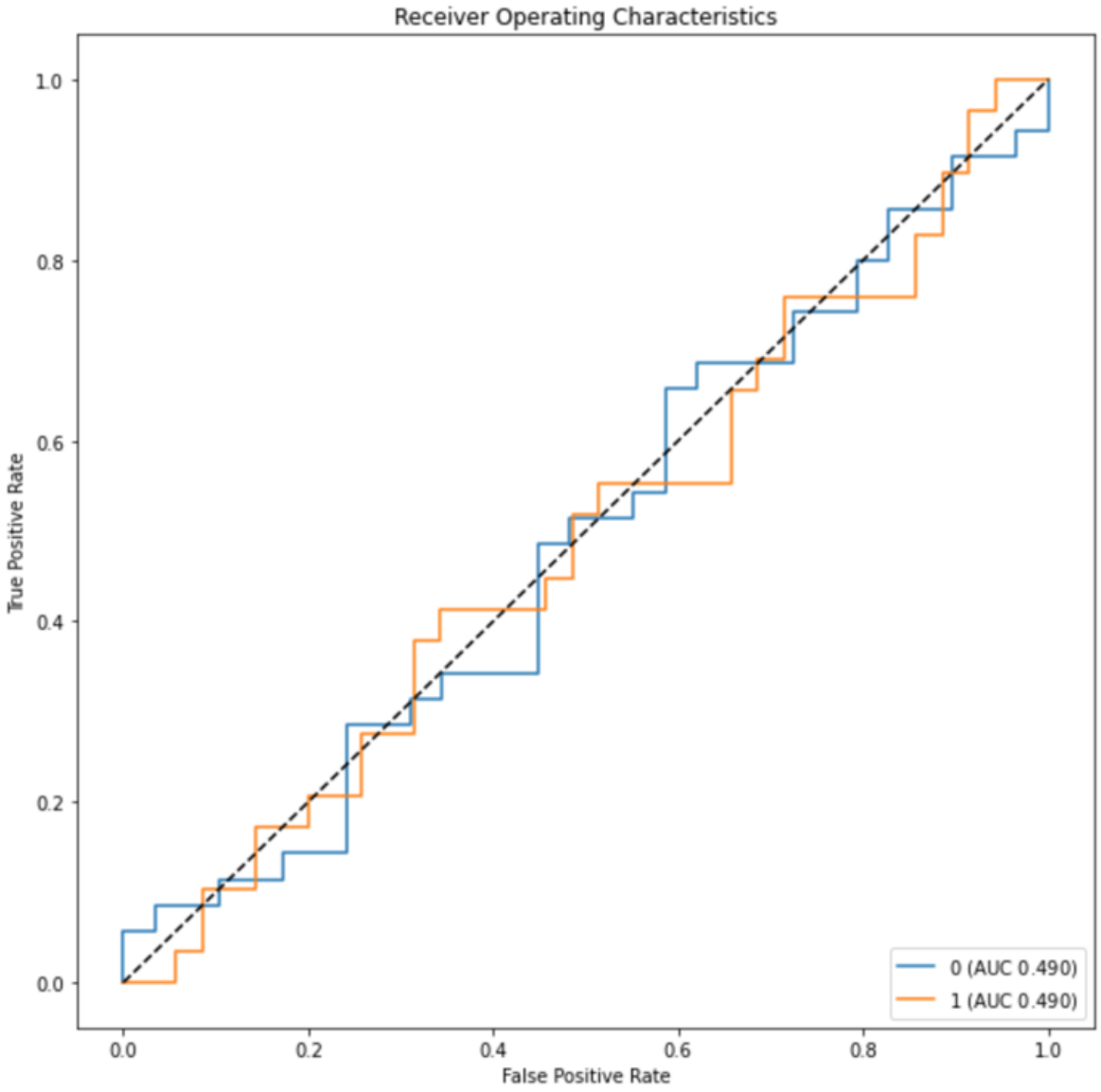
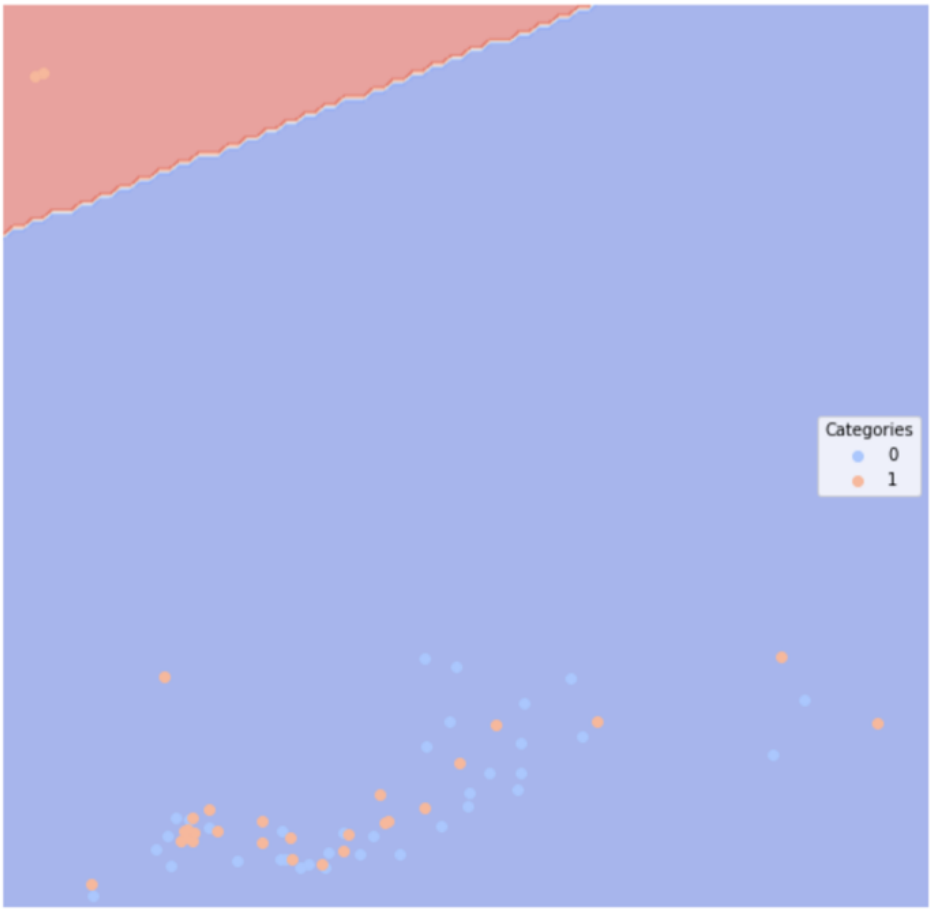


Figure 16: Plotted region for the SVM classifier



Considering the unsatisfactory results, there are many future improvements for the project. I will first extend to full-sized data from 2003 to 2020 using Midway cluster. Secondly, filtering firm-specific news for each target firm would generate more precise value-relevant information during the decision-making process of the analyst forecast revisions. Lastly, the word vectors from embedding models with the classification methods may predict the revisions with more statistical power.

Finally, I want to express my sincere thanks to James, Jacy and Junsol. I have received too many help from you for this course and the project, and I am always grateful about learning standing on the shoulders of giants.

References

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of accounting and economics*, *31*(1-3), 405–440.

Abarbanell, J., & Lehavy, R. (2003). Can stock recommendations predict earnings management and analysts' earnings forecast errors? *Journal of accounting research*, *41*(1), 1–31.

Cowen, A., Groysberg, B., & Healy, P. (2006). Which types of analyst firms are more optimistic? *Journal of Accounting and Economics*, *41*(1-2), 119–146.

Jung, M. J., Keeley, J. H., & Ronen, J. (2019). The predictability of analyst forecast revisions. *Journal of Accounting, Auditing & Finance*, *34*(3), 434–457.

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). *Predicting returns with text data* (tech. rep.). National Bureau of Economic Research.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905–949.