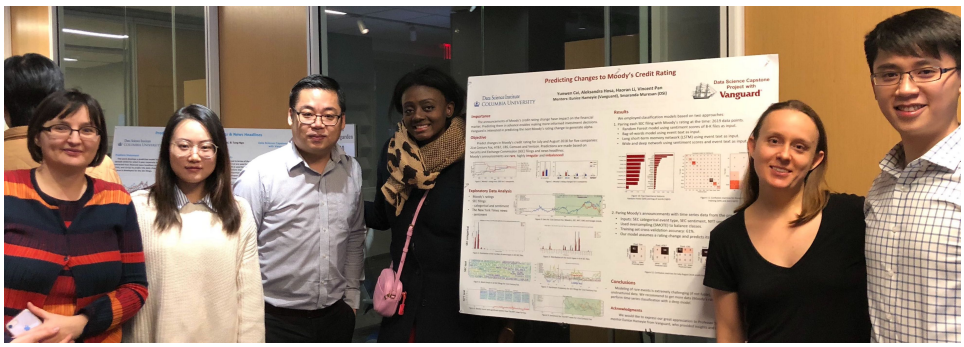
# Capstone Final Report
# Vanguard FY1801
# Team 2

*Yunwen Cai*
*Aleksandra Hosa*
*Haoran Li*
*Vincent Pan*



# Contents

**Abstract**

The announcements of Moody's credit rating have an impact on the financial market. In this project we are partnering with Vanguard to help them make more informed investment decisions by predicting the next Moody's credit rating update type based mainly on the Security and Exchange Commission (SEC) company filings.

Our proof-of-concept model predicts Moody's rating for July and August 2018 for five companies: *21st Century Fox*, *AT&T*, *CBS*, *Comcast* and *Verizon*, based on decades worth of data. Predictions are made based on the SEC filings, news headlines and financial data. We employ models with three different approaches to target-feature matching, covering different time-spans and using different datasets. Best models in each approach reach prediction test accuracy of: 1) 45%, 2) 80%, 3) 60%.

Predicting Moody's credit rating updates is very challenging, since their announcements are rare, highly irregular and imbalanced. We recommend working with data for more companies to improve model performance and to allow for more sophisticated models (e.g. deep models for time series classification).

# 1 Introduction

In this project we are partnered with DSI affiliate Vanguard to help them make more informed investment decisions based mainly on the Security and Exchange Commission (SEC) company filings. The announcements of Moody's credit rating have an impact on the financial market. Predicting them in advance enables making more informed investment decisions. Vanguard is interested in predicting the next Moody's rating change to generate alpha. The goal of this project was to predict changes in the Moody's credit rating for July and August 2018 for five companies: *21st Century Fox*, *AT&T*, *CBS*, *Comcast* and *Verizon*.

Based on our understanding about how Moody's ratings are generated, financial experts' confidence in the firm's future performance is essential in their decision making process. People use news and other forms of text to establish their confidence (Ostertag, 2010). Narrative sections of financial reports have gained more attention in recent research (Kang et al., 2018) for several reasons. First, financial reports can provide investors with critical information, such as managerial opinions or diagnoses of future performance and overall future plans. Second, text is more elastic than numerical data in terms of transferring information. Third, there is no general guideline

regarding processing text data of financial reports. So it's interesting to analyze SEC filings and news headlines in order to assess firm's performance. However, the question must be raised: what can we learn from the text data?

Davis et al. (2010) suggest a positive relationship between firm performance and optimistic tone. After Lopatta et al. (2017) combine traditional bankruptcy prediction models with language features based on Loughran and McDonalds's (2011) word lists and the comprehensive bankruptcy data set to see whether 10-K reports contain useful information to predict company jeopardy, they are able to further show the relationship also holds in the other direction: companies that file for bankruptcy protection ought to use significantly more negative words than their peers with a good performance. Moreover, they also find that firms' 10-Ks filed within a year of their bankruptcy contain more litigious words than healthy businesses. In another study, Karapandza (2016) shows that firms using less future tense in their 10-Ks generate higher returns since they are riskier.

Even though there's lack of studies predicting Moody's rating using SEC filings, the above mentioned studies are critical to our analysis, since Moody's rating is an indicator of the firm's capacity to meet its financial commitments, in other words, how risky it is to invest in the firm. So our underlying goal is essentially similar to predicting bankruptcy and riskiness. Knowing that previous studies show a relationship between sentiment in SEC filings and firm's performance, we can extract sentiment from 10-K, 10-Q and 8-K and generate language features to predict Moody's rating.

In addition to the SEC filings sentiment, we scraped company-related news headlines from The New York Times and calculated corresponding sentiments, as well as obtained financial data. We employed models with three different approaches to target-feature matching, covering different timespans and using different datasets and the results are significantly better than random guessing.

## 2 Data

We explored several datasets including: Moody's credit ratings (prediction target), the SEC filings (text), the news headlines and summaries, financial data such as stock price and financial ratios (Figure 1). Moreover, we used the Loughran and McDonald's Financial Dictionary in order to perform a

lexicon-based sentiment analysis on the text data. In the following sections we elaborate on the acquisition and processing methods, as well as on the exploratory analysis of these datasets. Most of the figures show our data and analyses for one selected company: *21st Century Fox*, figures for the remaining companies can be found in the Appendix (Section A.3).

Since the SEC is our main explanatory dataset, the time spanned by it dictates the time we are interested in for the training purposes, i.e. going back to early 2000s for *21st Century Fox* and *Comcast* and to early 1990s for *AT&T*, *CBS* and *Verizon* (shaded green in Figures 1 and 10). This considerable time span poses a challenge to obtaining additional datasets that would cover that time. Another challenge stems from the sparsity, irregularity and the imbalanced nature of our target: the Moody's credit ratings.

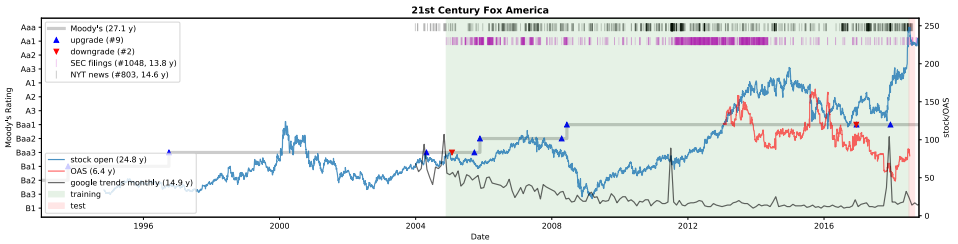



Figure 1: Our datasets for *21st Century Fox* displayed as a time series. Numerical data (stock price, OAS and Google trends) are scaled to allow clearer view in one image. The occurrences of text items are displayed towards the top of the graph: SEC filings - pink, The NYT news - black. Interactive version of the figure available here. Figures for the remaining companies (including links to the interactive plots) available in the Appendix (Figure 10).

## 2.1 Moody's Credit Ratings

The Moody's dataset provided by Vanguard contains 72 data points for all companies combined, going back to the 1990s. Figure 2 illustrates that Moody's updates are **rare** and **irregular**, which poses a challenge in our predictive task.

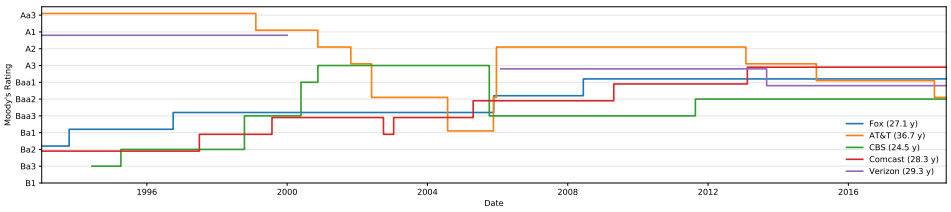



Figure 2: Moody's credit rating since 1993 for all companies of interest.

Since the goal of the project was to predict the *type* of change in Moody's rating (*upgrade*, *downgrade* or *no change*), we derived the target type variable directly from the *Moody's Rating* variable (disregarding the *Watch* variable, details in Report 2). After filtering out all Moody's records dated before the earliest SEC filing for each company 52 records remain. Figure 3 illustrates how **imbalanced** our classes are: *same* or *no change* class is the most frequent for all companies combined as well as for most companies considered individually.

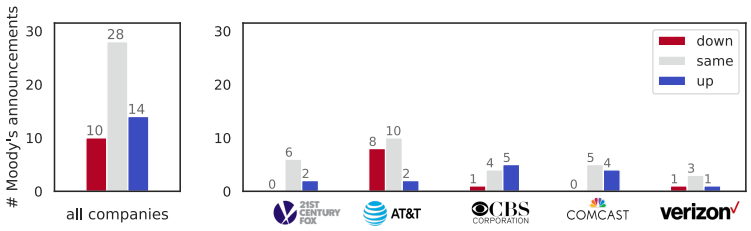



Figure 3: Count of Moody's rating change classes: *upgrade*, *downgrade* and *same* (no change), shown for all companies combined (left) and per company (right).

## 2.2 SEC Filings

### 2.2.1 Processing

The U.S. Securities and Exchange Commission (SEC) is an independent federal government agency responsible for protecting investors, maintaining fair and orderly functioning of securities markets and facilitating capital formation. The SEC filing is a financial statement that public companies are required to submit.

Vanguard provided us with a list of links to 2790 SEC filings, made public on 2506 unique days for any given company, spanning time up to the end

of September 2018. Based on the links, we developed a scraping tool to download all the related SEC filings, cleaned the data by (code on GitHub repo: SEC extract and clean):

- removing all html tags;
- removing structural formatting symbols;
- removing non-English words;
- breaking the text corpus by sentence and paragraphs
- tokenize SEC filings and build uni, bi and tri grams.
- extract event types and the associated languages from 8-Ks
- extract single level indexed tables(financial ratios) from 10-K and 8-Ks
- storing them in a tabular format (>9GB).

#### 2.2.2 Sentiment Score

We downloaded full Loughran and McDonald's Master Dictionary (2016), which contains sentiment classifications, counts across all filings, and other useful information about words. Specifically, the sentiment classes are: *Negative*, *Positive*, *Uncertainty*, *Litigious*, *Constraining*, *Superfluous*, *Interesting*, and *Modal*.

The original dictionary comprises of 85,221 words, however, some of them are classified as none of these sentiment categories. Removal of these words rendered 4009 words that are classified as at least one of these sentiment categories.

For each SEC filing, we mapped words in this filing with the Loughran and McDonald's Master Dictionary and generate sentiment score for each sentiment class as our features. We calculated 9 sentiment scores in total: *Tone_measure*, *Negative_freq*, *Positive_freq*, *Uncertainty_freq*, *Litigious_freq*, *Constraining_freq*, *Superfluous_freq*, *Interesting_freq*, and *Modal_freq*. All frequencies are derived by counting the occurrence of each sentiment in the document, and then dividing it by the total word count in that document. For *Tone_measure*, we subtract the positive word count from the negative word count, and divide it by the total word count.

#### 2.2.3 Event Announcement

As we observed that there is a relationship between material events and Moody's ratings update, we scraped item categories and the raw text of

event announcement section from SEC filings. The reason for not using the entire file instead is that documents are overall very similar in terms of content. We believe the repetitive information is not going to be very helpful to our prediction. We conjecture that the event announcement section is the main section focused on has changed since the last filing, which is generally what people are mostly interested in any SEC filing.

### 2.2.4 Analysis

Figure 4 shows the evolution of the sentiment score with time along with Moody's rating. It illustrates that *Tone_measure* and Moody's rating tend to move in the same direction, while *Litigious_freq* and Moody's rating tend to move in the opposite direction. This effect is clearer for some companies more than for the others (Appendix Figure 11). To access interactive figures click here: Fox, AT&T, CBS, Comcast, Verizon.

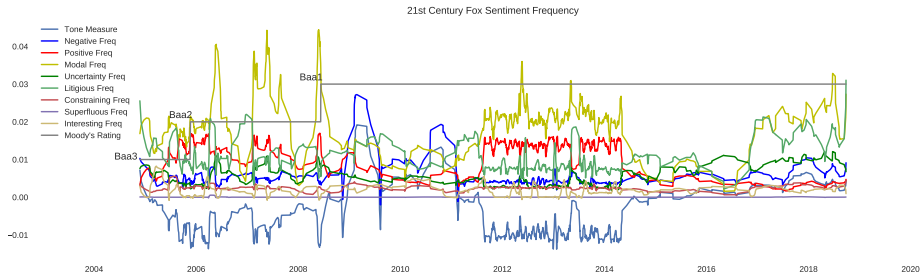



Figure 4: Sentiment frequency of SEC filings for 21st Century Fox

## 2.3 News

To supplement our SEC dataset, we decided on acquiring company-related news headlines from news websites. Scraping news for our project posed an extreme challenge because of the long time-span we were interested in (going back to the 1990s) and the difficulty of getting news from that long ago.

### 2.3.1 Acquisition (Scraping)

We did extensive research and prototyped scraping methods for several news sources (links to our scrapers on GitHub in the Appendix Section A.2.2).

However, we found that the services that do not require a log-in provide only a few days worth of data, while the ones that require setting up an account require agreeing to terms of use that prohibit using any automatic methods to scrape data. This is the case with Yahoo News, The Wall Street Journal, ProQuest (available though Columbia Libraries service) and Seeking Alpha (details in Report 1).

Surprisingly, The New York Times does not require setting up an account and a log-in in order to scrape news headlines and summaries. Therefore, we settled on that service as our source of the news. Moreover, while scraping we took an approach of not doing it too intensely - we introduced a wait time between accessing each new website so as not to overload the The New York Times servers (details in the scraping code on our GitHub: The New York Times scraper, as well as in Report 2).

### 2.3.2 Exploratory Analysis

We obtained a dataset of 8641 news headlines and summaries for our 5 companies of interest. The number of downloaded items varies considerably among the companies (e.g. 754 for *Comcast* vs 4358 for *CBS*, Figure 10), as well as for individual companies over time (Figure 1). Figures 5 and 12 illustrate that the peaks in news count correspond to significant events in the company, such as mergers, acquisitions and scandals (details in Report 2). Interactive figures for each company: Fox, AT&T, CBS, Comcast, Verizon, include hover-on annotations with the headline and a summary for every news item and enable manual exploration of the topics covered by The New York Times around the high count peaks.

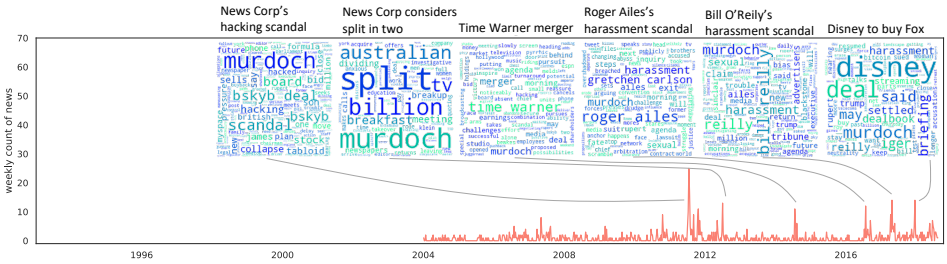



Figure 5: Weekly count of The New York Times news records for *21st Century Fox*. Peaks signify periods of extreme intensity of news coverage and correspond to significant events in the company, as demonstrated by the word clouds and labels. Figure for the remaining companies available in the Appendix (Figure 12).

### 2.3.3 Sentiment Analysis

Since the Loughran & McDonald's Financial Dictionary (L&M) is optimized for sentiment in the financial documents, it might not be the optimal lexicon to analyze the news, as these a considerably different kind of documents. We used both the Vader lexicon and the L&M to perform sentiment analysis on our news dataset. Vader was created for social media text and is one of the sentiment analysis methods available in the *nltk* package. It was previously used in sentiment analysis of news headlines (e.g. on Reddit news).

Figure 6 (and all figures in the Appendix Section A.3.1), show that the positive sentiment scores (both Vader and L&M) tend to be correlated with each other and anti-correlated with negative and litigious scores (details in Report 2).

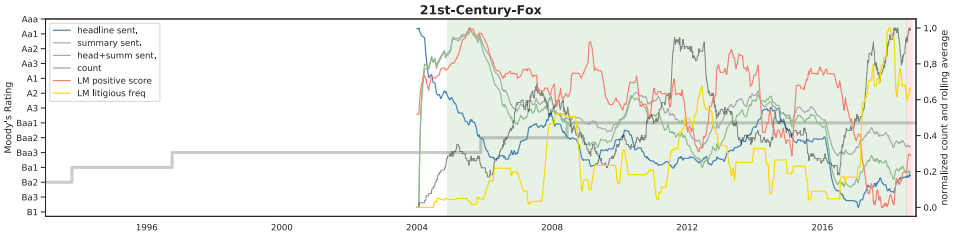



Figure 6: Yearly rolling average of The New York Times sentiment for *21st Century Fox* displayed over time. Selected sentiment scores include Vader compound scores (blue, green and grey) and L&M scores (red and yellow). The count of news records is shown with a thin black line. Moody's ratings are shown with a grey step line. Interactive figure available here. Figures for the remaining companies (including links to interactive plots) available in the Appendix (Figures 13, 14, 15, 16, 17).

## 2.4 Financial Ratios

Based on our discussion with Moodys, we realized that key financial ratios are critical to the determination of Moodys rating. 10-Ks are the annual reports that provide comprehensive summary of companys financial performance. As such, they are less time sensitive and sentiment focused compared to the news or 8-Ks. However, they are great source for financial data. According to *"Moody's Financial Metrics$^{TM}$ Key Ratios by Rating and Industry for Global Non-Financial Corporates: December 2016"*, Moody's uses 11 key financial ratios for global non-financial, non-utility corporations. All the required financial ratios could be derived from either balance sheets or cash flow or income statements, and all the above three are included in 10-

Ks. Initially, we focused on how to extract these three financial tables form 10-Ks. However, Extraction of usable financial tables from 10-Ks is very challenging for the following reasons:

- We need to process 10-Ks back to the 1990s and the earlier 10-Ks are structured differently from the more recent ones: the section names are different, the section sequences are different and the labeling and the numbering format is different.
- Even though the overall structure of 10-Ks is similar, each company files 10-Ks in different format. For example, the financial tables for AT&T are under HTML div tag while for Comcast are under p tag.
- The 10-K filings provided by SEC are sometimes with HTML tags and sometimes as a raw text format. Even in cases we can identify where the financial tables are, it is impossible to reformat most of them to a usable tabular form.

After extensive efforts, we were only able to extract the 15 financial terms that are used for calculating 11 financial ratios from the *AT&T* 10-Ks filed between 2011 to 2018 (GitHub: Extracting financial tables). We used the script from this GitHub repository with some modifications. However, the script doesn't work for other companies and for files before 2010, as discussed previously. Later we learned from the report *"Moody's Approach to Global Standard Adjustments in the Analysis of Financial Statements for Non-Financial Corporations"* that Moody's uses adjusted financial data and ratios. They adjust the financial statement according to their own standard and calculate the ratios based on adjusted statements. Even if we can successfully parse the statements, these ratios would be different from Moody's numbers. In addition, Moody's changes methodology over time and uses different methodology for each industry. Hence, we decided to switch our focus on 8-K files as our main model features.

### 2.5 Other Datasets

We downloaded the daily stock price information, starting in the beginning of 1994, from Yahoo Finance using a Python module *fix_yahoo_finance* (GitHub: EDA). For two of our companies of interest the stock information was not available for the entirety of that period: for Fox starting instead on 1994-11-03 (which is not an issue as we only have SEC data for Fox starting in 2004); and for CBS starting on 2005-12-05. We used the closing

and opening stock price to calculate the daily change for our 3rd approach model (section 3.3)

We also explored other datasets, e.g. Google trends (monthly going back to 2004, weekly going back 5 years) and OAS (the dataset provided by Vanguard, going back to 2014). However, we did not end up using any of these datasets in our models due to their insufficient time span.

# 3 Modeling Approaches and Results

We employed classification models with three approaches regarding the target-feature matching, predicted classes, time span covered and explanatory features (Table 1). In the following sections we elaborate on each of these approaches.

| **A.** | **Target-feature matching** | **Predicted classes** | **Time span** | **Datasets** |
|---|---|---|---|---|
| **1** | pairing each SEC filing with the Moody's rating at the time | the actual Moody's rating (e.g. *Aaa*) | same as the SEC dataset (from early 1990s or 2000s) | sentiment scores of all SEC filings |
| **2** | pairing each Moody's rating announcement with a time series data from the preceding 6 months | Moody's announcement type: *up*, *down*, *same* | same as Approach 1 | same as Approach 1 + SEC 8-K event categorical + The NYT sentiment |
| **3** | pairing each SEC 8-K filing with the Moody's rating at the time | same as Approach 1 | from 2008 to 2018 | 8-K & NYT sentiment scores, 8-K event categorial + number of events per 8-K + stock price change |

Table 1: Comparison of modeling approaches.

## 3.1 Approach 1

Our first approach is to pair each SEC filing with Moody's rating at the time to predict actual rating. There are 2619 data points in the training set and 40 data points in the test set. Details of the implementation can be found in our GitHub repo: Model Approach 1.

### 3.1.1 Random Forest

The random forest model takes sentiment scores of all SEC filings as input. Figure 7 shows the confusion matrices for training and test set. We achieved a training accuracy of 95% and a test accuracy of 45%. Over-fitting is a clear issue for the random forest model. From the feature importance plot, we can see *Tone_measure* has a high importance weight on the predicting result, which corresponds to the relationship we observed between sentiment score and Moody's rating during our exploratory data analysis.

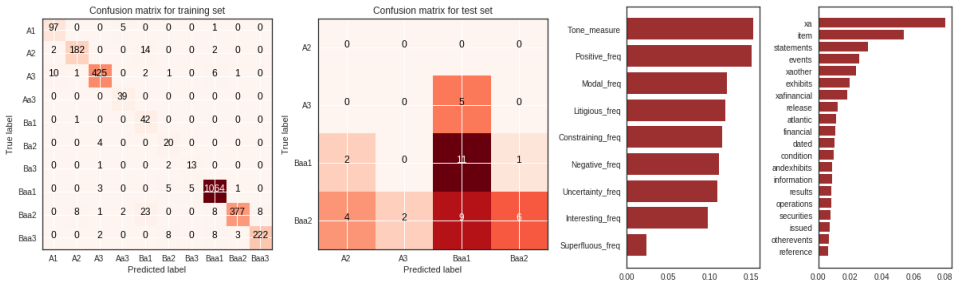



Figure 7: Confusion matrices for random forest training (left) and test (middle-left); feature importance plot for random forest (middle-right) and bag-of-words (right)

### 3.1.2 Bag-of-Words Model

The bag-of-words model takes tokenized event text as input and fits on a random forest classifier. The feature importance plot in figure 7 shows words that have a heavy weight on predicting the Moody's Rating. There are uncleared text such as *xa* that heavily affects the prediction. The bag-of-words model achieves a training accuracy of 71% and test accuracy of 37%, which is slightly lower than the random forest model. See Appendix A.4.1 for prediction results.

### 3.1.3 Long Short-Term Memory Network (LSTM)

LSTM is a type of recurrent neural network unit, which is well-suited to perform prediction on text data, since it can remember previous input for a period of time and thus can deal with lags of unknown duration between important events.

Since our data is financial documents, pre-trained word embeddings (such as word2vec and GloVe) that were trained on standard corpus like Wikipedia might not be a good choice for us. We addressed it by training our own word embedding by adding an embedding layer before the two LSTM layers.

Our LSTM network takes raw event text as input. Since our LSTM model only predicts the majority class, it doesn't really have any predicting power (See Appendix A.4.1 for prediction results and more details). We can use this majority class prediction as a baseline against any other models.

One possible reason for such low training/test accuracy could be that our training data has only 2619 records, which is generally not considered as a large dataset for deep learning neural networks. The amount of training data might have limited the learning ability of LSTM network. Therefore, it's worth trying to apply a pre-trained word embedding, which will be our next step.

### 3.1.4 Wide and Deep Network

A wide and deep network takes both structured and unstructured data as input. It concatenates the output of a deep neural network with the structured data and then jointly trains a wide linear model. Generally, a wide model is good at memorizing interactions with a large number of features, while a deep model is good at generalizing learned interactions. By applying a wide and deep model, we can combine the strengths of both deep and wide models.

Our Wide and deep network takes sentiment scores and event text combined as input. It has a single layer of Gated Recurrent Unit (GRU) with an embedding layer. Its output is combined with structured data and is then used as input for a two-layer dense network. The prediction result (See Appendix A.4.1) is slightly better than a majority class prediction. A major problem with the wide and deep model is that the prediction result varies vastly between runs, which is possibly due the dependency on random initialization.

### 3.2 Approach 2

Approach 2 was inspired by the intuition that Moody's update is a reaction to what happens in a time period before the update. Therefore, the signal recorded in the time interval preceding the Moody's update could help us predict the update *type*. Following on that intuition we paired each Moody's announcement with time series features from the preceding 6 months. We obtained an imbalanced set of 51 examples with 3 classes. The advantage of this approach is that we could use all datasets. To maximize the number of training examples, we chose to only work with the datasets that cover all of our training time span, which includes only the SEC and The NYT datasets. Table 2 summarizes the model characteristics.

| | |
|---|---|
| **Inputs** | SEC sentiment scores, The NYT sentiment scores, SEC 8-K categorical event type |
| **Output classes** | rating change type: *up*, *down*, *same* |
| **Time span of data** | from the earliest SEC filing for each company (i.e. from 1990s or 2000s) |
| **Size of training set** | 51 - number of Moody's updates in our dataset (oversampled to 81) |
| **Size of test set** | 62 - number of days in July and August |
| **Training accuracy** | cross-validation: 61% |
| **Prediction result** | 4/5 companies classified correctly (majority vote of all days in July-August) |

Table 2: Approach 2 model summary.

Feature engineering for this model involved obtaining Moody's announcement (*upgrade*, *downgrade*, *same*) from the provided dataset (details in Report 2). For all the explanatory time series we followed these pre-processing steps:

1. Resample each time series daily, using *mean* for collisions (i.e. when there is more than one event on the same day) and *forward*-filling missing values.
2. For each date in the Moody's dataset extract time interval of 6 months preceding the Moody's date for all features.
3. Resample the extracted time series using the *mean* to limit the number of features, e.g. resample monthly to obtain 6 model features per each original variable.

The test set was obtained by repeating the above procedure for every day in July and August 2018. This resulted in 62 examples in the test set, all of which are labelled as *same* (since there were no Moody's announcements during these two months for any company of interest).

Due to the fact that the number of features (hundreds) is significantly greater than the number of training examples (51), we ran our models on a subset of the features (selected using *SelectKBest* from *sklearn*). We handled the imbalance of the training set by oversampling the minority classes using the *SMOTE* algorithm, which resulted in 81 training examples.

We ran several classification algorithms (Linear SVC, K Neighbors, Logistic Regression, Kernel SVM, Random Forest) and found that XGBoost gave the best cross-validation accuracy. We used this model to predict Moody's rating update type for all days in July and August.

In the prediction stage we effectively assumed that Moody's rating announcement occurs on each day in July and August and predicted the *type* of the announcement. Figure 8 shows the confusion matrices for the prediction for each company. These results might be interpreted as the probability of Moody's update type in the 2-month time period: 98% chance of *same* for *21st Century Fox*, 94% chance of *downgrade* for *AT&T*, 73% chance of *same* for *CBS*, 90% chance of *same* for *Comcast* and 76% chance of *same* for *Verizon*.

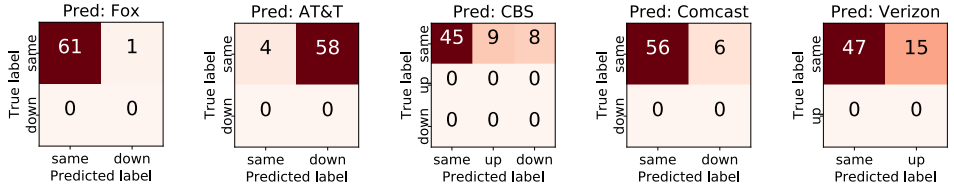



Figure 8: Confusion matrices for July-August 2018 of rating change prediction using XGBoost algorithm.

80% of our companies of interest are classified correctly with high probability, i.e. they are found as *no change* in Moody's rating (there was no Moody's announcements during these two months for any company of interest). The only exception is *AT&T*, which is found a *downgrade*. This can be explained by the fact that Moody's downgraded *AT&T* on 15 June, 2018 - only two weeks before the start of our test period. Since our model features are based on a time period of 6 months, there is a significant overlap between all of our test features and features corresponding to this recent

*AT&T* downgrade, which is in the training set. The mis-classification for *AT&T* is a direct result of that and is in fact expected.

More details of the implementation can be found in our GitHub repo: Model Approach 2.

### 3.3 Approach 3

Approach 3 is to pare each 8-K filing with Moody's rating at the time to predict actual rating by adding more features.

| | |
|---|---|
| **Inputs** | 8-K categorical event type and number of events per 8-K, 8-K sentiment scores, The NYT sentiment scores, 90 days average sentiment scores, stock price change |
| **Output classes** | the actual Moody's rating (e.g. *Aaa*) |
| **Time span of data** | from 2008-01-01 to 2018-08-31 |
| **Size of training set** | 1,243 - number of 8-K filings from 2008-01-01 to 2018-06-31 |
| **Size of test set** | 18 - number of 8-K filings from 2018-07-01 to 2018-08-31 |
| **Training accuracy** | 100% |
| **Prediction result** | 3/5 companies classified correctly (majority vote of the actual rating prediction for each company) |

Table 3: Approach 3 model summary.

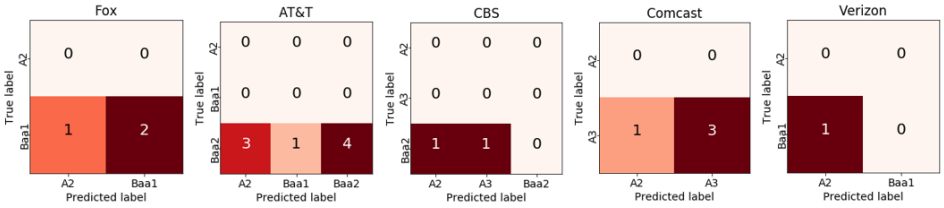



Figure 9: Confusion matrices for July-August 2018 of actual rating prediction using XGBoost model.

Based on the majority vote, *Fox*, *AT&T*, and *Comcast* will remain the same rating, while *CBS* and *Verizon* will be upgraded.

Detailed implementation can be found in our GitHub repo: Model Approach 3.

# 4 Conclusions

We developed models to predict Moody's credit rating change *type* for 5 companies: *21st Century Fox*, *AT&T*, *CBS*, *Comcast* and *Verizon*. We made our prediction based on the SEC filings (categorical even-based and sentiment), The New York Times headlines and summaries (sentiment), financial ratios and stock price. Our models follow three different approaches regarding the target-feature matching, time span covered and datasets used and. We found that a tree-based ensemble model was the best performer in each approach.

Modeling of rare events is extremely challenging (if not futile), especially with limited unstructured data. Moody's rating announcements are rare, highly irregular and result in imbalanced training sets:

- size of the dataset: only 51 Moody's announcements in the time overlapping with our SEC dataset;
- irregularity: announcements are not scheduled and can happen at any time - in our dataset we observe time intervals between consecutive announcements between a few days and 8 years;
- imbalance of classes: most of the announcements are *no change* (with a watch for an upgrade or downgrade).

Further difficulty in modeling stems from the fact that Moody's changes its rating methodology from time to time (e.g. it changed its methodology for the telecommunication industry multiple times since the 1990s), so a set of circumstances that led to an upgrade in 1990s might not necessarily lead to an upgrade in the present time. Moreover, Moody's employs a different methodology for each industry type, and our companies span several industries according to Moody's. Additionally, we are dealing with changes in explanatory data sources over time, e.g. the SEC changed its 8-K events definitions in 2004.

Nevertheless, our models pick up on the correlations between our explanatory variables and Moody's ratings, as our models result in test prediction

accuracy for July-August 2018 that is above the random guess baseline. In particular, our model in Approach 2 predicts Moody's update type correctly for 4 out of 5 companies and the mis-classification for *AT&T* can be explained by a recent Moody's downgrade for that company (June 15, 2018).

As far as next steps are concerned, our recommendation is to acquire more data (Moody's ratings for more companies) and develop deep models for time series classification (as mentioned in Report 2).

## 5 Acknowledgements

We would like to express our great appreciation to Professor Smaranda Muresan as well as our mentor Eunice Hameyie from Vanguard, who provided insights and NLP expertise to this project.

## 6 References

Cai, Y., Hosa, A., Li, H. and Pan, V. (2018) Capstone First Progress Report, Vanguard FY1801, Team 2, *Columbia University Data Science Institute Capstone Progress Report*, `https://github.com/lihaoranharry/Capstone_Vanguard_NLP_Prediction/blob/master/Report%201.pdf`

Cai, Y., Hosa, A., Li, H. and Pan, V. (2018) Capstone Second Progress Report, Vanguard FY1801, Team 2, *Columbia University Data Science Institute Capstone Progress Report*, `https://github.com/lihaoranharry/Capstone_Vanguard_NLP_Prediction/blob/master/Report%202.pdf`

Ostertag, S. (2010). Establishing news confidence: A qualitative study of how people use the news media to know the news-world. Media, Culture Society, 32(4), 597-614. doi:10.1177/0163443710367694

Kang, T., Park, D., & Han, I. (2018). Beyond the numbers: The effect of 10-K tone on firms performance predictions using text analytics. Telematics and Informatics, 35(2), 370-381. doi:10.1016/j.tele.2017.12.014

Ben-David, I., Graham, J., & Harvey, C. (2010). Managerial Miscalibration. doi:10.3386/w16215

Lopatta, K., Gloger, M. A., & Jaeschke, R. (2017). Can Language Predict Bankruptcy? The Explanatory Power of Tone in 10-K Filings. Accounting Perspectives, 16(4), 315-343. doi:10.1111/1911-3838.12150

Karapandza, R. (2016). Stock returns and future tense language in 10-K reports. Journal of Banking Finance, 71, 50-61. doi:10.1016/j.jbankfin.2016.04.025

Hutto, C.J. and Gilbert, E. (2014) VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, *Association for the Advancement of Artificial Intelligence*, `http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf`

Sentiments Analysis on Reddit News Headlines with Pythons Natural Language by Toolkit (NLTK), `https://www.learndatasci.com/tutorials/sentiment-analysis-reddit-headlines-pythons-nltk/`, accessed Dec 15, 2018

Documentation for the Loughran & McDonald Master Dictionary, `https://www3.nd.edu/~mcdonald/Word_Lists_files/Documentation/Documentation_LoughranMcDonald_MasterDictionary.pdf`, accessed Dec 16, 2018

Moody's Approach to Global Standard Adjustments in the Analysis of Financial Statements for Non-Financial Corporations, `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1831433`

# A Appendix

## A.1 Team Contributions

- Yunwen Cai: **Report 1**: collected and cleaned Loughran and McDonald's dictionary; initialized sentiment analysis; researched extracting text features from SEC filings; write-up on literature review; write-up and code on scraping Yahoo News. **Report 2**: modeling - constructed and merged features; created training and test sets; explored and evaluated 3 models. **Poster**: modeling via the first approach (Result 1); figures 6, 7, 10, 11. **Final Report**: Data: SEC Sentiment Score (Section 2.2.2), SEC Event Announcement (Section 2.2.3), SEC Exploratory Analysis (Section 2.2.4); Modeling Approach 1 (Section 3.1).

- Aleksandra Hosa: **Report 1**: initial EDA on numerical/categorical data (Figure 1); updated and tested WSJ scraping code; prototyped ProQuest scraping, researched Yahoo News, Seeking Alpha and NYT scraping, looked into legality of scraping; collected stock price and Google Trends; write-up on: Moody's, SEC (partial), OAS, stock and Google Trends, news scraping, initial EDA, fast.ai and transfer learning and the next steps strategy (partial). **Report 2**: The New York Times dataset - wrote the scraping code and executed the data download, performed EDA on the news, performed sentiment analysis, analyzed possible biases; Moody's dataset - derived the *Type* variable, filtered the records, performed intervals analysis; researched the idea for a deep model for multivariate time series classification. **Poster**: Moody's dataset analysis and visualization; visualization of all datasets through time, The NYT dataset sentiment analysis, EDA and visualization; modelling via Approach 2; on the poster: Importance and Objective (partial); EDA (partial); Modeling - Approach 2; Figures 1, 2, 3, 8, 9, 12. **Final Report**: Abstract (partial); Intro (Section 1) (partial); Data: Intro (Section 2), Moody's (Section 2.1), News (Section 2.3), Other (Section 2.5) (partial); Modeling: Intro (Section 3), Approach 2 (Section 3.2); Conclusions (Section 4) (most); Appendix: Project Repo (Section A.2) (most), Supplementary EDA Figures (10, 12), Sentiment on The NYT (Section A.3.1).

- Haoran Li: **Report 1**: collected and cleaned SEC filings; obtained language related features from SEC filings; EDA on SEC (partial); set up multiple meetings with current Moody's staff to discuss the

proper use of data, insights, project potentials and recommendations, form and quantify project problem statement and next step strategies. **Report 2**: processed the SEC filings - obtained, constructed and engineered features from 8-Ks and 10-Ks (events/item categories and associated text from 8-Ks; tokenized and obtained uni-, bi-, and tri-grams from 8-Ks and 10-Ks; single level indexed financial tables from 8-Ks and 10-Ks, multi-level indexed financial tables (not successful)). **Poster**: extracted, engineered and visualized SEC 8-K filings events and distribution; Figures 4, 5. **Final Report**: Abstract (partial); Intro (Section 1) (partial); Data: SEC Filings (Section 2.2) (partial); Conclusions (Section 4) (partial).

- Vincent Pan: **Report 1**: identified financial terms that were used to calculate 11 key financial ratios as model features. **Report 2**: worked on extracting financial tables from 10-Ks (not successful). **Poster**: Abstract and project description. **Final Report**: Intro (Section 1) (partial); Data: Financial Ratios (Section 2.4); Modeling Approach 3 (Section 3.3).

## A.2 Project Repository

Please find our project repository following this link: Capstone Vanguard NLP Prediction Git Repo.

### A.2.1 Reports

- Report 1
- Report 2
- Poster

### A.2.2 IPython Notebooks

- EDA: overall EDA, Moody's analysis
- SEC Processing: SEC extract and clean
- News scraping: Yahoo news, Wall Street Journal, ProQuest, The New York Times
- News sentiment analysis: The New York Times
- Modeling: Approach 1 + Sentiment analysis on SEC filings, Approach 2, Approach 3

### A.2.3 Data

- The New York Times news scraped raw data
- The New York Times news processed data with sentiment

### A.2.4 Figures

- The New York Times analysis figures
- Sentiment analysis on SEC filings and Modeling Approach 1

### A.2.5 Interactive Plots

- Overall EDA plots: Fox, AT&T, CBS, Comcast, Verizon.
- The New York Times news weekly count and sentiment with **hover-on** headlines and summaries: Fox, AT&T, CBS, Comcast, Verizon.
- The New York Times news sentiment - yearly rolling average: Fox, AT&T, CBS, Comcast, Verizon
- The SEC filings sentiment with smoothing: Fox, AT&T, CBS, Comcast, Verizon

## A.3 Supplementary EDA Figures

1. Datasets overview for all companies: Figure 10.
2. Sentiment on SEC for all companies: Figure 11.
3. Weekly count of The NYT news with wordclouds: Figure 12.
4. Sentiment on The NYT for all companies: Section A.3.1 with Figures 13, 14, 15, 16, 17.

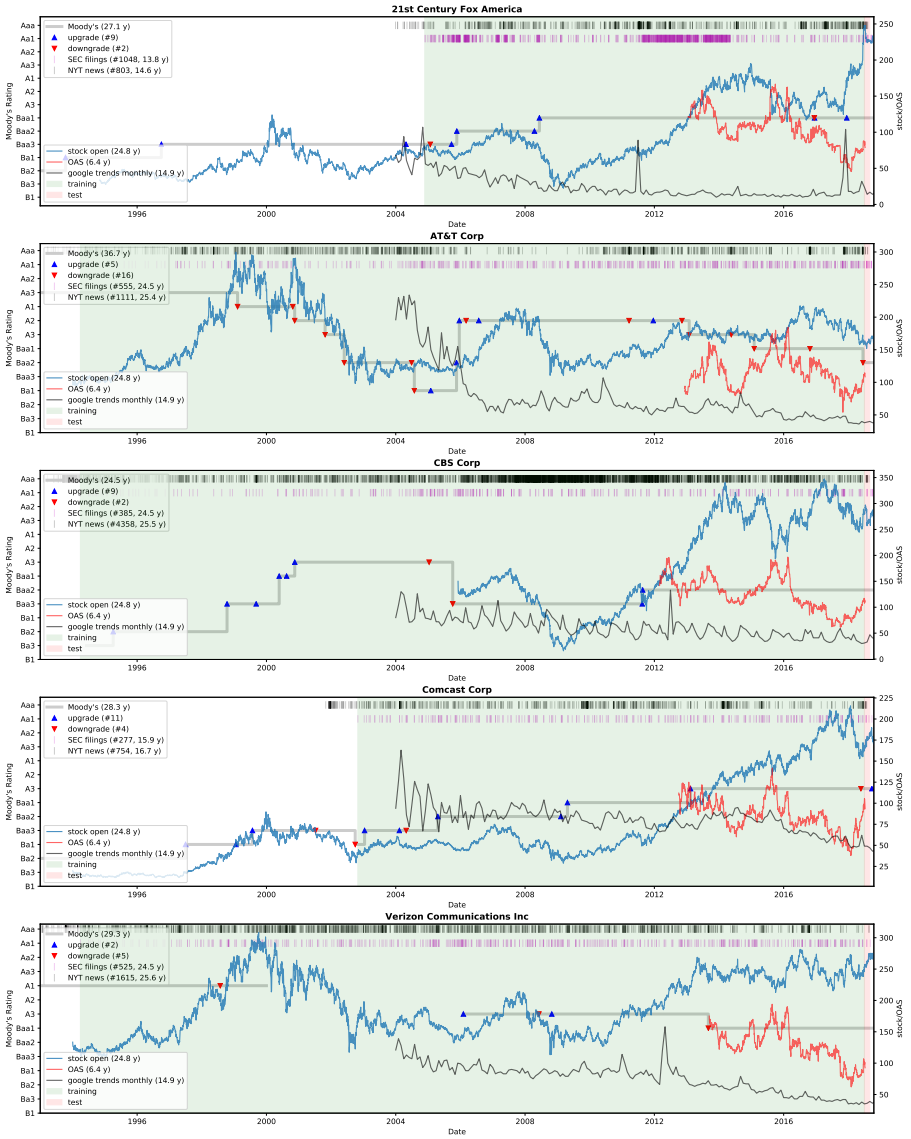



Figure 10: Our datasets displayed as a time series. Numerical data are scaled to allow clearer view in one image. To access interactive figures click here: Fox, AT&T, CBS, Comcast, Verizon. Moody's ratings are shown with a grey step line with upgrade or watch for upgrade (blue) and downgrade or watch for downgrade (red) events; the occurrences of text data points (SEC filings - pink, The NYT news - black) are displayed towards the top of the graph; stock price, OAS and monthly Google trends are shown as blue, red and black lines, respectively. The area shaded in green shows the span of time from the first SEC filing until the end of June 2018, which is our training period; red shaded area (July and August 2018) is out testing time period.

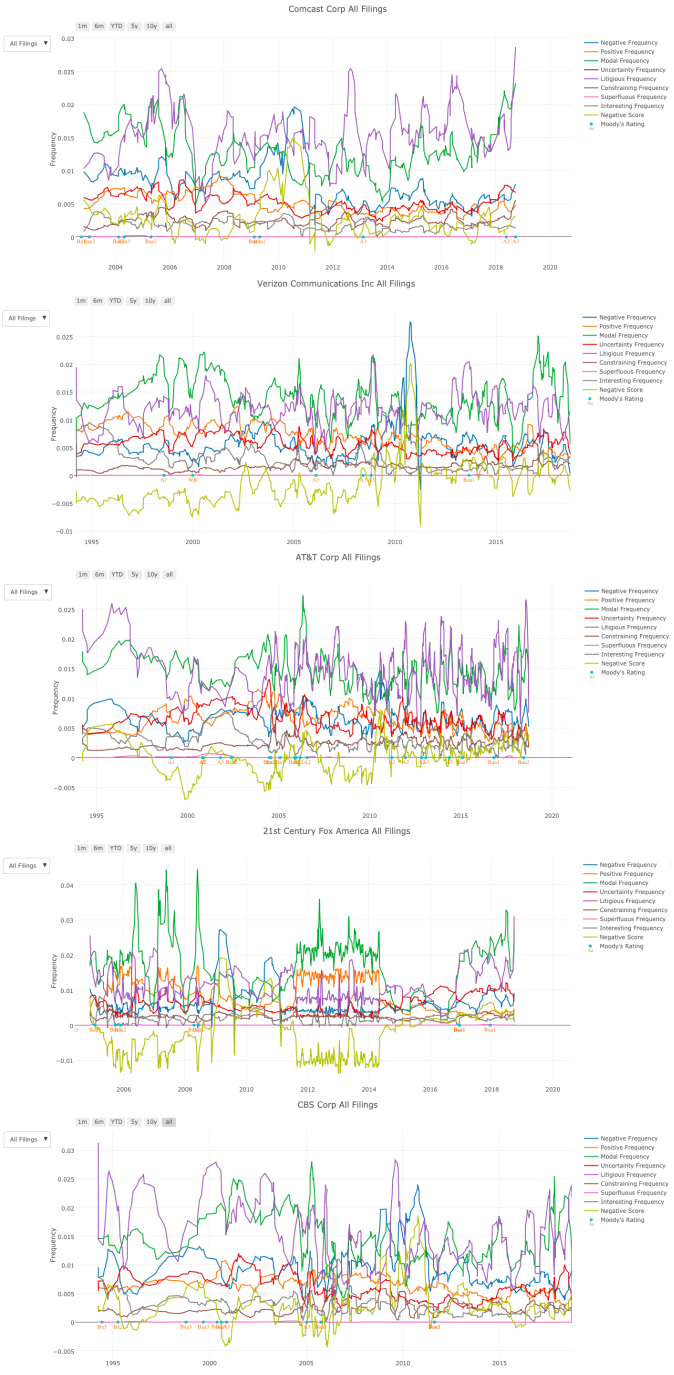



Figure 11: Sentiment frequency of SEC filings


Figure 12: Weekly count of The NYT news records for all five companies. Peaks signify periods of extreme intensity of news coverage and correspond to significant events in the company, as demonstrated by the word clouds and labels.

### A.3.1 The New York Times Sentiment Analysis

The series of figures below (Figures 13, 14, 15, 16, 17) show the The New York Times sentiment analysis for all five companies of interest, including:

a) Correlation of all computed sentiment scores - only values of correlation with absolute value greater than 0.1 are labelled.

b) Selected sentiment scores shown for individual news records over time: Vader compound score for concatenated headline and summary (green dots), L&M positive score for summary (red) and L&M litigious frequency for summary (yellow). Moody's rating is represented with a thick grey step line. The area shaded in green indicates the span of time from the first SEC filing until the end of June 2018 (our training period); red shaded area (July and August 2018) indicates our testing time period.

c) Yearly rolling average of selected sentiment scores and records count shown over time: Vader compound score for concatenated headline and summary (green) headline (blue) and summary (grey), L&M scores for summaries: positive score (red) and litigious frequency (yellow). Rolling average of the count of news records is shown with a thin black line. Moody's rating is represented with a thick grey step line. The area shaded in green indicates the span of time from the first SEC filing until the end of June 2018 (our training period); red shaded area (July and August 2018) indicates our testing time period.

Interactive versions of plots b) and c) are available under the following links:

- Individual news records plots (b): Fox, AT&T, CBS, Comcast, Verizon.

  These plots include hover-on labels with headline and summary text for each data point.

- Rolling average plots (c): Fox, AT&T, CBS, Comcast, Verizon.

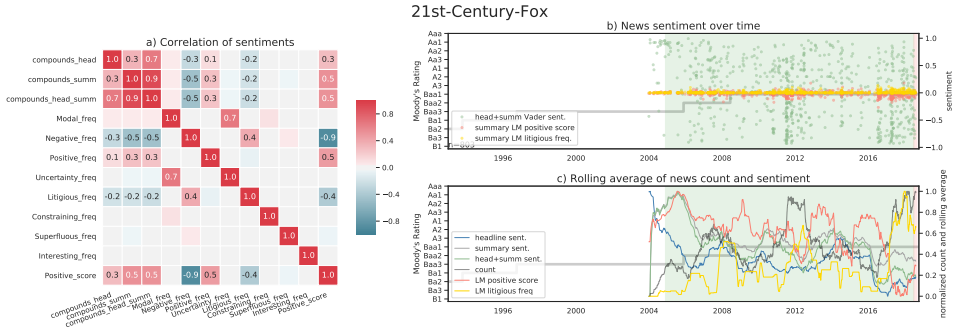



Figure 13: EDA on The NYT sentiment for 21st Century Fox.

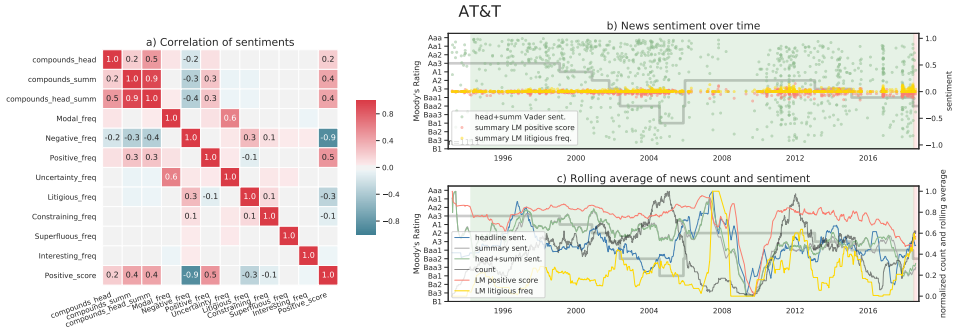



Figure 14: EDA on The NYT sentiment for AT&T.

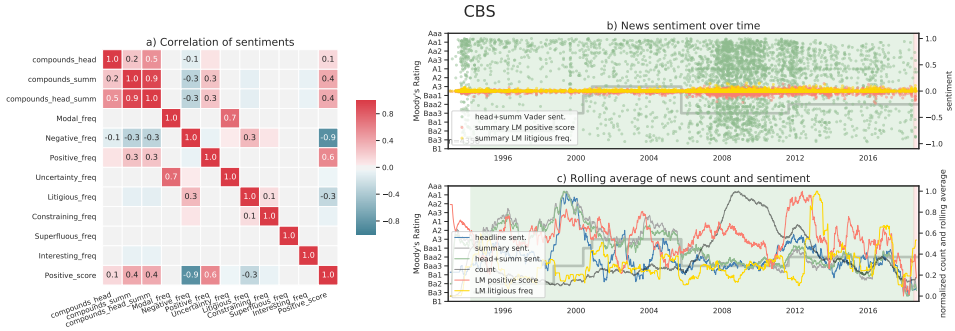



Figure 15: EDA on The NYT sentiment for CBS.

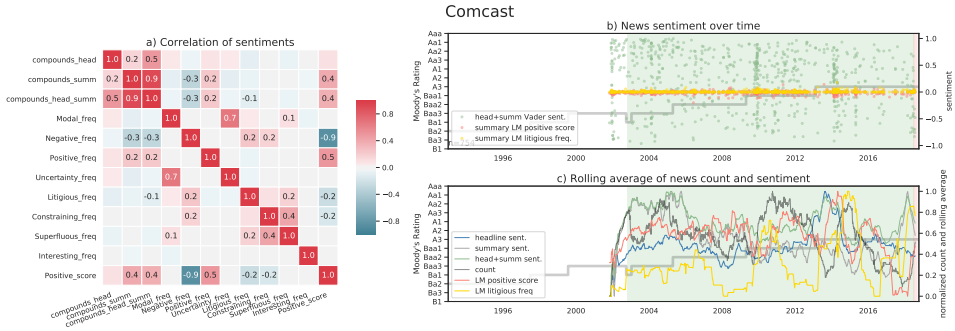



Figure 16: EDA on The NYT sentiment for Comcast.

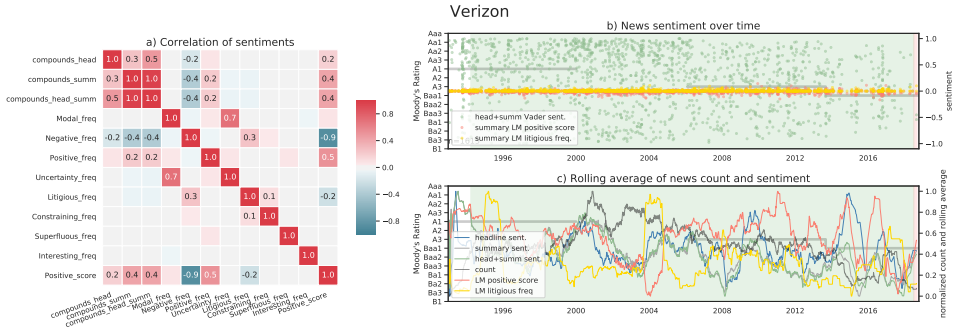



Figure 17: EDA on The NYT sentiment for Verizon.

## A.4 Modeling

### A.4.1 Approach 1 Details

- Bag-of-words model

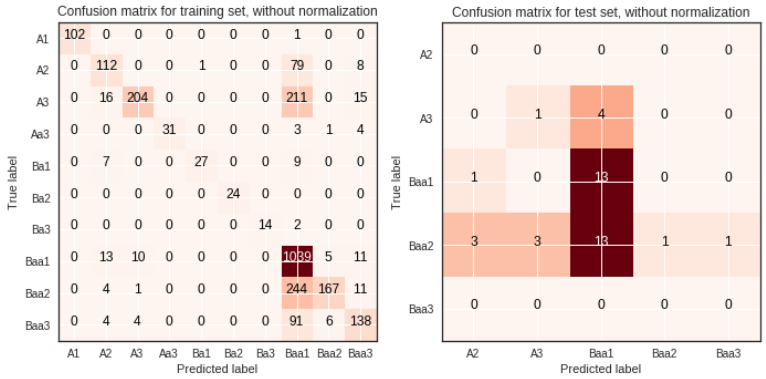



Figure 18: Confusion matrices for bag-of-words model training (left) and test (right)

- LSTM

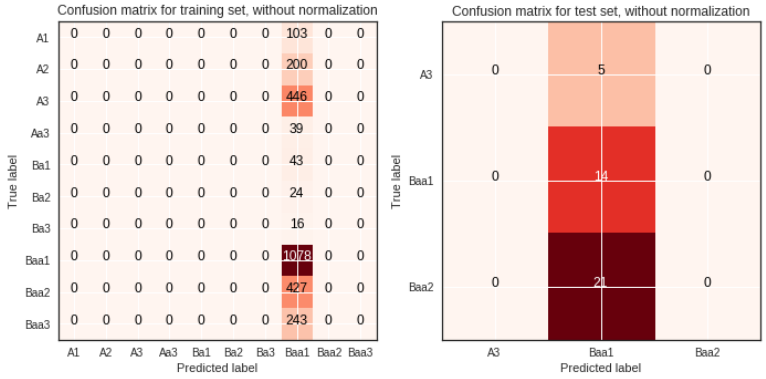



Figure 19: Confusion matrices for LSTM training (left) and test (right)

- Wide and deep model

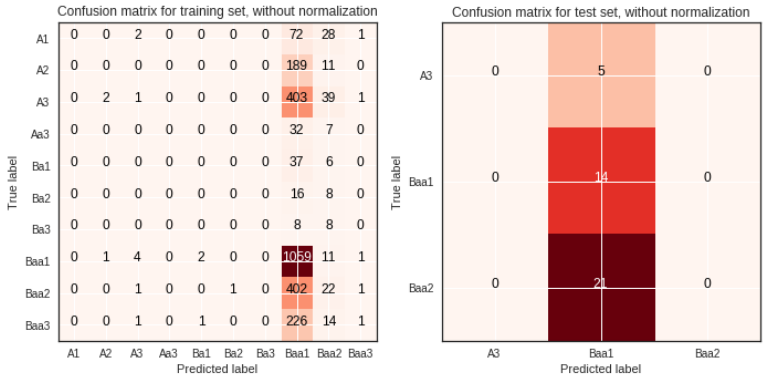



Figure 20: Confusion matrices for wide and deep network training (left) and test (right)