Capstone First Progress Report
Vanguard FY1801
Team 2

Yunwen Cai
Aleksandra Hosa
Haoran Li
Vincent Pan

Contents

1 Team Contributions & Project Repository

- Yunwen Cai: collect and clean Loughran and McDonald's dictionary, initialize sentiment analysis, research extracting text features from SEC filings, write-up on literature review, write-up and code on scraping Yahoo News.
- Aleksandra Hosa: 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).
- Haoran Li: collect and clean SEC filings, obtain 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.
- Vincent Pan: extracting financial numbers from 10-Q and 10-K SEC filings, calculating financial ratios as features.

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

2 Problem Statement

In this project we are tasked with finding the correlation between company filings submitted to the U.S. Securities and Exchange Commission (SEC) and Moody's rating scores. We are going to utilize all existing types of SEC filings: 8-K, 10-K and 10-Q. We are also going to use supplementary data, such as news headlines (and possibly full text), company daily option-adjusted spread (OAS), stock price etc. Our focus is the following five telecommunication/mass media companies:

1. 21st Century Fox America
2. AT&T Corp
3. CBS Corp
4. Comcast Corp
5. Verizon Communications Inc.

A critical challenge that goes along with our project is to find the time relationship between the SEC fillings and the Moody's rating score updates. We have a set of Moody's rating scores and a set of SEC filings, but the matching function between filings and rating scores is unclear, i.e. we don't know if we should consider SEC fillings and other explanatory variables just for the past one month to predict Moody's ratings action for a given company, or from the past two months etc. Further, if any such matching function exists, we don't know if it is consistent across all companies.

3 Data Collection, Description & EDA

3.1 Moody's Credit Rating

Our prediction objective is Moody's credit ratings - provided by Vanguard from the Bloomberg portal. Our dataset comprises of 72 data-points (Moody's ratings updates) in total for all companies. Table 1 shows that Moody's updates are rare events, with between 0.3 to 0.7 events per year per company on average. Further, Figure 1 shows that these events are highly irregular, with higher intensity of updates possibly occurring within a year or two (e.g. Fox circa 2005) and nothing happening for nearly a decade (e.g. Fox 2009-2017). This sparsity of data and highly irregular nature of it poses a challenge for our modeling.

Company name	# Moody's events	Initial Moody date	Avg # Moody's events per year
21st Century Fox America	13	1991-11-27	0.5
AT&T Inc	25	1982-04-26	0.7
CBS Corp	11	1994-06-08	0.6
Comcast Corp	15	1990-09-04	0.5
Verizon Communications Inc	8	1989-08-16	0.3

Table 1: Data summary for provided Moody's ratings.

Our data comprises of 14 unique values of Moody's rating, from *Aaa* (highest) to *B1* (lowest). We also have one occurrence of *WR*, recorded for Verizon on 2000-01-01 (then known as Bell Atlantic Corp). *WR* indicates that the rating given at this date was later withdrawn, which could be due to e.g. liquidation. For the purposes of our modeling, we assume that the previous rating of *A1* held for Verizon until the next update on 2006-02-03 (which is reflected in Figure 1).

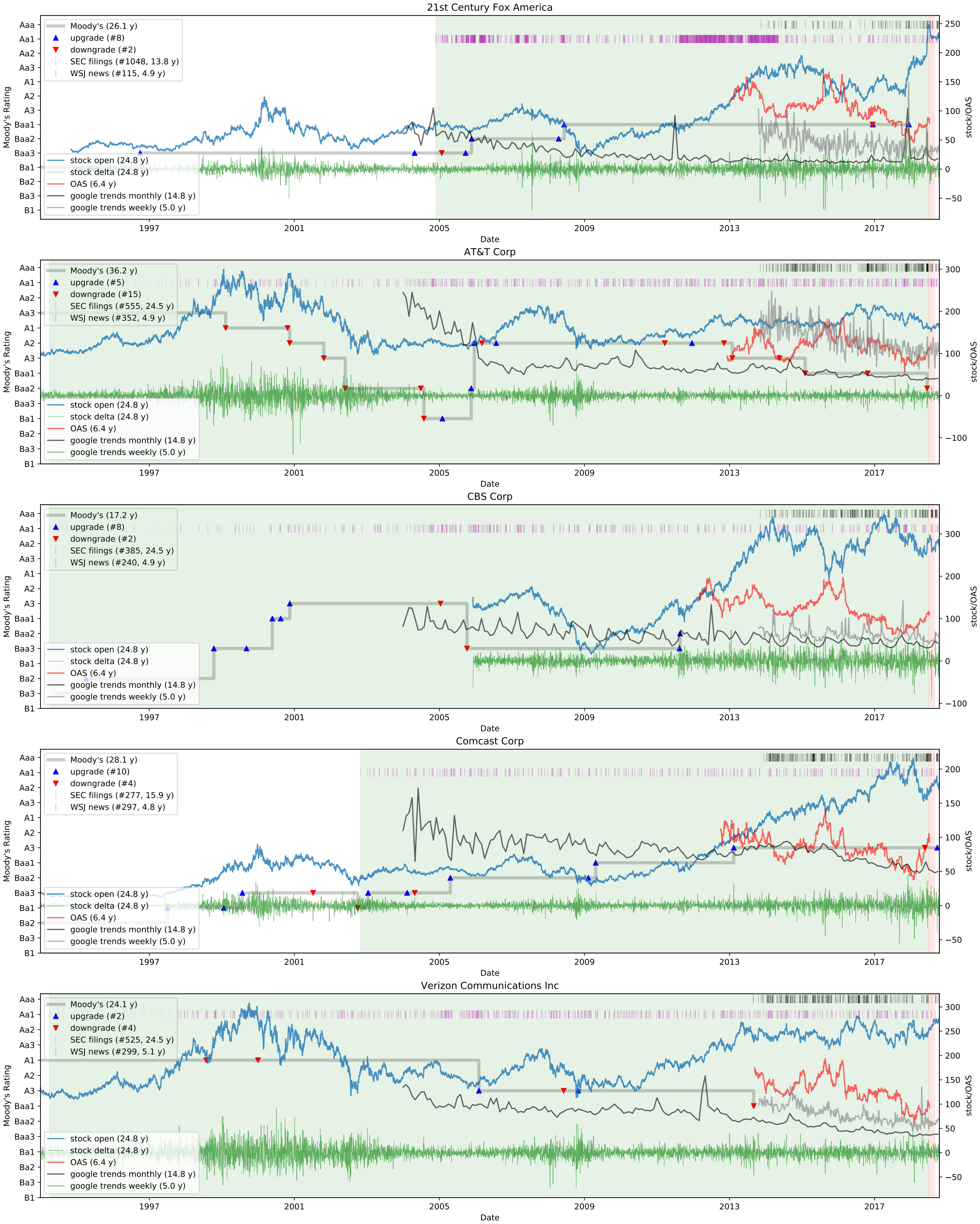
Besides the rating itself, we were also provided with two other features: *Watch* ('+', '-', or none) and *Type* ('up', 'down', or 'initial'). The reason why in Figure 1 we can observe an 'upgrade' (blue triangle) even when the rating is not changed is because the *Watch* value is '+' for that particular Moody's rating action. Pending confirmation from Vanguard whether the *Watch* feature is provided by Moody's, or whether it's a Bloomberg feature, we'll make a decision whether to model the pure Moody's rating, or the rating concatenated with the *Watch* modifier.

3.2 SEC Filings (10-K, 10-Q and 8-K)

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 (*findexdate* feature) 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;


21st Century Fox America

AT&T Corp

CBS Corp

Comcast Corp

Verizon Communications Inc

Figure 1: 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 (blue) and downgrade (red) events; the occurrences of text data points (SEC filings - pink, WSJ news - black) are displayed towards the top of the graph; stock price and daily delta are shown in blue and green, respectively; OAS is shown as a red line; monthly Google trends scores are shown in black and weekly in grey. 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.

- removing structural formatting symbols;
- removing non-English words;
- breaking the text corpus by sentence and paragraphs
- storing them in a tabular format (>9GB).

We observe that the time of the earliest available filing differs for the companies of interest, as shown in Table 2. In all cases the time-span covered by the SEC dataset is shorter than the time-span covered by the Moody's dataset. Again, Figure 1 shows that SEC filings can occur with varying frequency and it is depending on the company, e.g. Comcast can be characterized by stable frequency of SEC filings since 2002, while Fox undergoes periods of sparse filing (circa 2010) and very frequent filing (circa 2013).

Company name	# SEC filings	earliest SEC filing	Avg # SEC filings per year
21st Century Fox America	1048	2004-11-24	75.7
AT&T Inc	555	1994-03-15	22.7
CBS Corp	385	1994-01-12	15.7
Comcast Corp	277	2002-10-30	17.4
Verizon Communications Inc	525	1994-01-21	21.5

Table 2: Data frequency summary for provided SEC filings.

The types of SEC filings we work on include:

- 10-K, which is a comprehensive analysis of the company and contains more information than an annual report, e.g. the financial statements are more detailed. Companies have to submit 10-K annually within 90 days of the end of their fiscal year.
- 10-Q, which is a truncated version of 10-K. 10-Q is to be filed 3 times a year, within 45 days of the end of each of the first three quarters of the company's fiscal year. Its focus is the company's latest developments and a preview of the next steps. Generally, it is less detailed than 10-K and may include un-audited financial statements.
- 8-K, which is an unscheduled filing designed to address specific events that didn't make it in time for 10-Q or 10-K but which must be communicated to the investors. Such events may include bankruptcy, management departures, acquisitions etc.

The above rules prescribing filing frequency for each form type are reflected in our dataset, as it is dominated by the 8-K filings (Figure 2).

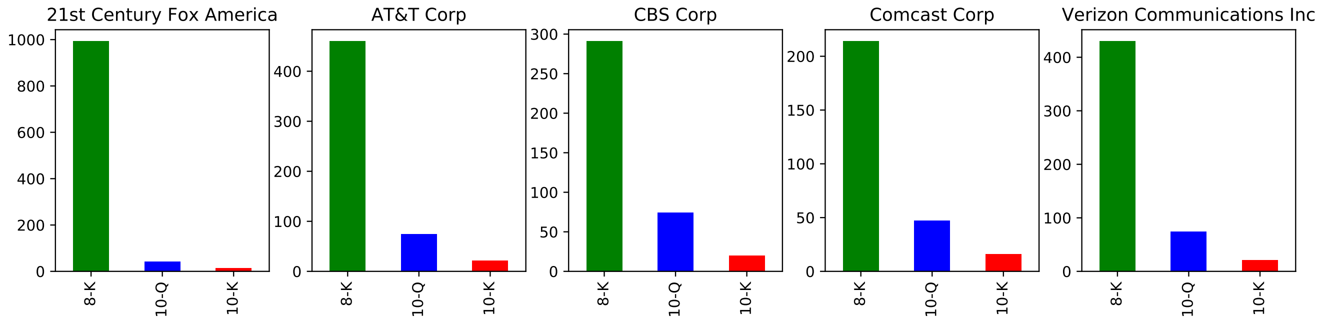


Figure 2: SEC filing type count.

3.3 Option-Adjusted Spread (OAS)

The daily OAS dataset was provided by Vanguard. It has a significantly shorter time-span than either the Moody's or the SEC datasets, as it only covers around 2013 until the present (Figure 1). Since it covers so little of our time-span of interest, at this time it is not clear if we will be able to use it as one of the explanatory features it in our model.

3.4 Loughran and McDonald's Financial Dictionary

We downloaded full Loughran and McDonald's Master Dictionary (2016), which is used to determine which tokens (collections of characters) are classified as words. It also contains sentiment classifications, counts across all filings, and other useful information about each word. For each SEC filing, we will map words in this filing with the dictionary and generate sentiment score for each sentiment category as our features. Specifically, we plan to use *Negative*, *Positive*, *Uncertainty*, *Litigious*, *Constraining*, *Superfluous*, *Interesting*, and *Modal* as our sentiment categories.

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. Table 3 shows sentiment categories we are going to use with several examples from the dictionary.

Sentiment Category	Example
Negative	abnormal, bankrupt, cyberattack
Positive	creative, enhance, outperform
Uncertainty	pending, perhaps, reconsider
Litigious	regulate, restitution, subclause
Constraining	unavailable, strict, restrain
Superfluous	proscribe, ubiquitous, wheresoever
Interesting	aggressive, death, diseased
Modal	usually, would, rarely

Table 3: Sentiment categories in Loughran and McDonald's Financial Dictionary.

3.5 News

We would like to scrape company-related news headlines and full text from news websites as an additional source for text analysis. However, scraping news poses an extreme challenge because of the long time-span we are interested in (25 years for some companies) and the difficulty of legally getting data from that long ago. Additionally, while scraping we have to be careful to use the correct company name at any given time, as these were changing though time (Table 4).

	Company name	Time range (from SEC) from	to
1	News Corp	2004-11-24	2013-06-17
	21st Century Fox America	2013-07-05	now
2	Southwestern Bell Corp	1994-03-15	1994-11-08
	SBC Communications Corp	1995-05-09	2005-11-05
	AT&T Inc	2005-11-21	now
3	Viacom Inc	1994-01-12	2006-01-05
	CBS Corp	2006-01-09	now
4	AT&T Comcast Corp	2002-10-30	2002-10-30
	Comcast Corp	2003-01-10	now
5	Bell Atlantic Corp	1994-01-21	2000-09-08
	Verizon Communications Inc	2000-11-14	now

Table 4: Company names through time.

We have done extensive research and prototyped scraping methods for several news sources. In summary, we found that services that don't require a log-in provide only a few days worth of data, while the services that require setting up an account and logging-in require agreeing to terms of use that prohibit using any automatic methods to scrape data. This is the case with Yahoo News, Wall Street Journal, ProQuest and Seeking Alpha. Our research regarding these services is detailed in the following sections. At the moment our best prospect is The New York Times annotated corpus and archives (more in Section 3.5.5).

3.5.1 Yahoo News

We wrote code to scrape the news from Yahoo News (GitHub: Yahoo scraper), which doesn't require logging-in to the service. However, we found that a search for 'AT&T' only renders 13 pages of results, which amounts to only 3 days of news. Although so far we haven't checked the log-in Yahoo service and we don't know what kind of time-range that would give us access to, we expect that the log-in service requires agreeing not to scrape when setting up an account.

3.5.2 Wall Street Journal

We were able to update the deprecated Wall Street Journal (WSJ) scraping code by Jason Weinreb (GitHub repo). Our modified code is on our GitHub repo: WSJ scraping. We managed to establish that using this method we would be able to obtain between 115 to 352 articles for each company, amounting to a total of 1303 articles. Only data for the past 5 years is available using this service (going back to fall 2013). Figure 1 shows the time-span covered by this dataset in the context of our other data. It also shows that coverage frequency tends to vary through time, which is most probably correlated with significant events that warrant news coverage. The collected dataset would include a headline, sub-header and the full text for each news item.

However, in order to use this service we have to *'agree not to display, post, frame, or scrape the Content for use on another website, app, blog, product or service, except as otherwise expressly permitted by this Agreement'* (from WSJ Subscriber Agreement and Terms of Use, section 8.3.1.).

3.5.3 All news - ProQuest via Columbia Library Services

We also managed to develop and test scraping code for ProQuest via Columbia Library Services (GitHub: ProQuest scraping), which includes news from many sources (e.g. WJS, The New York Times, Financial Times) dating back to 1994 and beyond. Manual search of the database revealed that there are tens of thousands of relevant articles covering all of our time-span of interest, i.e. time-span of our SEC dataset.

However, we have to agree not to scrape in order to use that service: *'Customer and its Authorized Users shall not: [...] Perform automated searches against ProQuests systems (except for non-burdensome federated search services), including automated bots, link checkers or other scripts'* (from ProQuest's Terms and Conditions Section 10e).

3.5.4 Seeking Alpha

Similarly to the above sources of news, Seeking Alpha (seekingalpha.com) gives access to news over limited time-range when you use it without logging-in: our searches for 'Verizon' and 'AT&T' only produced 10 pages of results going back to mid 2017. After signing-up and logging-in we were able to establish that at least for these two companies we would be able to access news or news summaries as far back as 2005-2006.

Again, singing-up to the service requires agreeing to terms and conditions, which include: *'you agree not to: [...] Use any robot, spider, site search/retrieval application, or other manual or automatic device or process to download, retrieve, index, "data mine", "scrape", "harvest" or in any way reproduce or circumvent the navigational structure or presentation of the Site or its contents'*.

3.5.5 The New York Times

At the moment our best bet seems to be The New York Times: via Columbia we might have access to the The New York Times Annotated Corpus, which covers January 1, 1987 to June 19, 2007. To cover time from 2007 up to the present we could explore using the collection of *free to read articles* available at spiderbites.nytimes.com - these articles seem to be available free of charge and accessing them doesn't require a log-in and agreeing to any terms and conditions.

3.6 Financial Ratios

According to Moody's Financial Metrics$^{\text{TM}}$ Key Ratios by Rating and Industry for Global NonFinancial Corporates: December 2016, Moody's uses eleven key financial ratios for global non-financial, non-utility corporations. We will use these ratios as part of our model features and extract relevant financial numbers from 10-K and 10-Q filings. Following is the list of eleven ratios and their definitions:

- EBITA / Average Assets
  Earnings Before Interest, Tax, Depreciation and Amortization / Average of Current and Previous Year Assets
- EBITA / Interest Expense
  Earnings Before Interest, Tax, Depreciation and Amortization / Interest Expense
- EBITA Margin
  Earnings Before Interest, Tax, Depreciation and Amortization / Net Revenue

- Operating Margin
  Operating Profit / Net Revenue
- (FFO + Interest Expense) / interest Expense
  (Funds From Operations + Interest Expense) / Interest Expense
- FFO / Debt
  Funds From Operations / (Short-Term Debt + Long-Term Debt)
- RCF / Net Debt
  (FFO Preferred Dividends Common Dividends Minority Dividends) / (Short-Term Debt + Long-Term Debt, gross - Cash and Cash Equivalents)
- Debt / EBITDA
  (Short-Term Debt + Long-Term Debt) / Earnings Before Interest, Tax, Depreciation and Amortization
- Debt / Book Capitalization
  (Short-Term Debt + Long-Term Debt) / (Short-Term Debt + Long-Term Debt + Deferred Taxes + Minority Interest + Book Equity)
- CAPEX / Depreciation Expense
  Capital Expenditure / Depreciation Expense
- Revenue Volatility
  Standard Deviation of Trailing Five Years of Net Revenue Growth

3.7 Stock price

We have 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 also have used the opening stock price to calculate the daily change (delta) and both quantities are plotted in Figure 1.

3.8 Google trends

Similarly, we have downloaded the monthly (going back to 2004) and weekly (last 5 years) Google Trends information using a Python module *pytrends* (GitHub: EDA). We have downloaded the trends data for all company names and added the results before plotting it on Figure 1, e.g. results for 'Bell Atlantic' and 'Verizon' were added.

3.9 Differences in time-spans and frequencies

Figure 1 shows that the time-spans of our datasets are very different. Since our objective is the prediction of the Moody's rating, ideally we would have all other datasets covering the same time-range, which is challenging, as Moody's goes back to the 1980s in some cases. SEC filings provided by Vanguard are the main dataset to be used for prediction, so we have defined the training period for each company as the period starting with the earliest SEC filing in our dataset and finishing on 2018-06-30 (shaded green in Figure 1). Ideally, all our additional data would cover the same time-span as the SEC dataset. However, this is still very challenging, as SEC goes to back to 1994 in most cases (2002 for Comcast and 2004 for Fox). Stock prices mostly fulfill that condition, and Google Trends gets relatively close - the monthly variety going back to 2004. The real challenge is posed by the OAS and the news - our current OAS dataset only goes back to 2013 and we are still unclear if we will be able to obtain any news dataset going sufficiently far back. It remains to be seen if the 'short' datasets will be of any use as explanatory variables in our model.

Of less concern are the differences in frequency of datapoints: stock price and OAS are daily and regular, Google Trends are either monthly or weekly and regular, SEC filings are fairly irregular, especially in case of the forms 8-K, which can be issued at the discretion of the company and quite often, news items are also fairly irregular, Moody's ratings are not only very sparse, but also highly irregular (as mentioned in Section 3.1).

3.10 The timing of Moody's rating actions in context

Figure 1 (in particular the interactive versions: Fox, AT&T, CBS, Comcast, Verizon) is designed to give us an initial insight whether there is any correlation between Moody's announcements and events that might have caused the decision. Especially, we wanted to establish if there is any correlation between the timing of a SEC filing and a Moody's event. Our analysis has revealed that 18 out of the total of 72 Moody's updates (25%) happened on the same day as a SEC filing by a given company. Further 19% of Moody's updates happened within 4 days of the last SEC filing. This could suggest that Moody's responds very fast to significant events revealed in SEC filings. However, our sources at Moody's have informed us that they do not operate on such short-time scales and that they typically take weeks/months to come to a conclusion about a rating action (more on that in Section 5).

We can also use Figure 1 to visually investigate if there is any correlation between significant events (e.g. exhibited by peaks in Google Trends or dramatic changes in stock price or OAS value) and Moody's rating updates. However, our investigation in that regard remains inconclusive. For instance, we would expect the high peak in Google Trends in July 2011 for 21st Century Fox (then News Corp), which happened due to the News International (a subsidiary of News Corp) phone hacking scandal (Wiki), would be followed by a Moody's score adjustment. This scandal resulted in a lot of bad press and management changes in the company. Interestingly, for almost 3 years following that event, until summer 2013 - which coincides with News Corp's change to 21st Century Fox, News Corp practiced extremely frequent SEC filing and enjoyed fast and steady stock price increase. Despite all that turmoil, there was no rating change from Moody's. In fact, Moody's rating was not updated until December 2016, 8 years after the previous update.

4 Literature Review

4.1 NLP on SEC filings

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 trasnferring 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 sentiments from 10-K, 10-Q and 8-K and generate language features to build model that predicts companies' Moody's rating.

Besides language features, traditional financial ratios are being used widely as numerical features. Lopatta et al. (2017) incorporate numerical financial features, specifically bankruptcy-indicating financial ratios such as profitability, liquidity, market value of equity, variance inflation factor (VIF), cash, and market-to-book ratio, into logit models. Kang et al. (2018) also adopt an earnings persistence ordinary least square (OLS) regression model containing financial indicators to measure the usefulness of sentiment. There are overlaps between Kang's choice of financial ratios and Lopatta's choice of financial ratios. Besides ratios in Lopatta's model, Kang's model also incporparates earning volume and age of the firm. Inspired by these studies, we are going to collect financial ratios for companies of interest, and combine these ratios with language features to develop our model.

4.2 fast.ai and Transfer Learning in NLP

We would like to explore the new ML library for deep learning developed by fast.ai (`http://www.fast.ai`). fast.ai is an open source outfit with a mission to make ML simple and user-friendly and in that way to make it available to broad audiences (ZDNet article Oct 2, 2018). fast.ai is built on top of the PyTorch library and it's 1.0 version was released in the beginning of October 2018. Already it has gained support of Google Cloud, AWS and Microsoft's Azure cloud service. Howard and Ruder (2018) were able to apply fast.ai in a transfer learning NLP classification task to achieve a significant accuracy gain over other methods, even though they only used a relatively low number of labelled examples (NLP in fast.ai). We would like to test their methodology in our text classification task.

5 Goals and Next Steps

In order to better determine the next steps, the team had three meetings/discussions with current Moody's staff. Below is a summary of the discussion that helped us shape our goals:

- Moody's internally has two major groups: analytics and quantitative research. The analyst group consists of industry experts who analyze the ratings based on press, news, fillings, and most importantly - on their business experience. The quantitative group consists of researchers who focus on predicting the Moody's rating change (same prediction objectives as this project) through statistical and data science methods.

- For each industry, Moody's has an internal scorecard, which details the rules between sub-ratings and sub-categories (including all the ratios listed in Section 3.6). For example, for a given industry, if *EBITA/Average Assets* ratio falls between a certain range, it is an A grade for this sub-category. Based on discussions on sub-ratings between the quantitative and the analyst groups, Moody's analyst group deterimines the final rating for the company of interest.
- Internally, with the industry-specific score card, the quantitative group is able to achieve around 60% accuracy of their prediction.
- While the quantitative group focuses solely on financial numerical data (including but extending far beyond 10-K) to model the score change, the analyst group focuses both on the numerical data, as well as the news, press, and industry expertise.
- Moody's quantitative group focuses primarily on 10-K because it contains more detailed financial data, however, 8-Ks and 10-Qs can reflect the information/news the analyst group also focuses on.

Based on the above insights, from a modeling perspective, there is room for improvement by incorporating NLP techniques on the news and SEC filings because that is what's currently missing from the Moody's quantitative group's process. However, we acknowledge that numerical financial data are extremely important for a successful prediction.

The combined insights from our EDA and from Moody's made us formulate a big picture plan to mimic the rating process of Moody's analytics and quantitative groups. We will use both financial data from 10-Ks and the sentiment from the news as our major explanatory variables to predict Moody's ratings, and we will use the key words from 8-Ks and 10-Qs as an extra signal to enhance the prediction. Blow is our detailed plan:

- Find and extract all relevant financial data from all available 10-Ks.
- Find and extract sensitive key words that signify a major event (e.g. merger, acquisition, litigation, etc.) from all SEC filings (10-K, 10-Q and 8-K).
- Scrape the news from any available sources, ideally as far back as the earliest SEC filing for each company: establish if and how we can use The New York Times archives and explore any other potential sources (paying attention to the legality of use).
- Utilize the financial sentiment map to build a sentiment score for the news and SEC filings (explore using fast.ai and transfer learning for NLP).
- Structure all features and Moody's ratings in a per Moody's score base (heavy data engineering).
- Utilize all the above as features and build models to predict the Moody's rating and, thus, predict the change of Moody's rating ('up', 'down', 'stable').

  In order to find the 'right' matching between SEC fillings and Moody's score, we will define the number of months of information used as another feature, and perform grid-search to find the 'best' matching function (likely to be different across the five companies).

6 Appendix

6.1 Supplementary Data and Materials

- Our GitHub Project: Capstone Vanguard NLP Prediction (`https://github.com/lihaoranharry/Capstone_Vanguard_NLP_Prediction`), including: EDA, SEC extract and clean, Yahoo scraping, WSJ scraping, ProQuest scraping.
- Interactive EDA plots (Plotly): Fox, AT&T, CBS, Comcast, Verizon.
- Jason Weinreb's WSJ scraping GitHub Repository: `https://github.com/jweinreb/python-wsj/blob/master/wsj-parser.py`
- Loughran and McDonald's Financial Dictionary: `https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary`
- Moody's Financial Metrics^{TM} Key Ratios by Rating and Industry for Global NonFinancial Corporates: December 2016: `https://www.researchpool.com/download/?report_id=1537315`
- Free to read NYT articles (1851-2018): `http://spiderbites.nytimes.com`
- The New York Times Annotated Corpus (1987-2007): `https://catalog.ldc.upenn.edu/ldc2008t19`

6.2 References

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

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

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

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

Howard, J. and Ruder, S. (2018) Universal Language Model Fine-tuning for Text Classification, ACL `https://arxiv.org/abs/1801.06146`

Text Classification with fast.ai: `http://nlp.fast.ai`

ZDNet article 'Fast.ai's software could radically democratize AI' (Oct 2, 2018) `https://www.zdnet.com/article/fast-ais-new-software-could-radically-democratize-ai/`

Wikipedia article on News International phone hacking scandal: `https://en.wikipedia.org/wiki/News_International_phone_hacking_scandal`