Capstone Second Progress Report
Vanguard FY1801
Team 2

Yunwen Cai
Aleksandra Hosa
Haoran Li
Vincent Pan

Contents

1 Team Contributions & Project Repository

- Yunwen Cai: modeling - constructed and merged features, created training and test sets, explored 3 models (Sections 5 and A.4).
- Aleksandra Hosa: 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 (Sections 3 and A.3); Moody's dataset - derived the *Type* variable, filtered the records, performed intervals analysis (Section 4); researched the idea for a deep model for multivariate time series classification (Section 5.4.2).
- Haoran Li: 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)) (Sections 2.1 and 2.2).
- Vincent Pan: extracted financial tables from 10-Ks (Section 2.2.2).

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

2 Processing the SEC Filings

2.1 Summary Update

In the previous phase, we were able to scrape all the raw SEC filings from the SEC website. Further, we were able to remove all strings that were not formatted as English words. We also obtained uni-grams, bi-grams and tri-grams from the SEC filings.

Based on our discussions with Moody's, we understood that their internal rating methodology is based upon **key financial ratios** and **key events**. Since the 8-K is an event driven filing and the 10-K is an annual financial report, we believe it is critical to obtain the events from the 8-Ks and the balance sheets, cash flow tables and income statements from the 10-Ks. Therefore, during this phase of the project, we focused on:

1. How to identify and extract the key events from these events;
2. How to extract the financial tables from 10-Ks and use the data within the table.

So far we managed to obtain the following new features from the SEC filings:

- identify and extract ***'events'/'categories'*** from all 8-Ks;
- identify, extract and clean ***'sentences'/'paragraphs'*** that describe the respective ***'events'/'categories'*** from all 8-Ks;
- identify and extract all ***financial tables*** from all SEC filings;
- ***clean and tabulate*** some of the financial tables from 8-Ks.

2.2 Approaches and Challenges

2.2.1 8-K Item Categories and Related Language

According to the SEC's Investor Bulletin, **effective August 23, 2004**, all 8-K filings contain one or several of the following items:

- Item 1
  - Entry into a Material Definitive Agreement
  - Termination of a Material Definitive Agreement
  - Bankruptcy or Receivership
- Item 2
  - Completion of Acquisition or Disposition of Assets
  - Results of Operations and Financial Condition
  - Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant
  - Triggering Events That Accelerate or Increase a Direct Financial Obligation or an Obligation under an Off-Balance Sheet
  - Costs Associated with Exit or Disposal Activities Arrangement
  - Material Impairments
- Item 3
  - Notice of Delisting or Failure to Satisfy a Continued Listing Rule or Standard; Transfer of Listing
  - Unregistered Sales of Equity Securities
  - Material Modification to Rights of Security Holders
- Item 4
  - Changes in Registrants Certifying Accountant
  - Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review
- Item 5
  - Changes in Control of Registrant
  - Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory Arrangements of Certain Officers
  - Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year
  - Amendments to the Registrants Code of Ethics, or Waiver of a Provision of the Code of Ethics
  - Submission of Matters to a Vote of Security Holders
- Item 6
  - No documentation on Item 6 on the SEC's Investor Bulletin
- Item 7
  - Regulation FD
- Item 8

  - Other Events
- Item 9
  - Financial Statements and Exhibits

The above **Items** naturally categorize the events that are being reported in any 8-K. We used the same categories. However, it is very challenging to extract the events along with their associated paragraphs from the 8-K fillings for the following reasons:

- Since we need to use SEC filings back to the 1990s, a big portion of the filings are before 2004. These earlier filings have to be categorized and formatted differently. Figure 1 shows two files formatted differently: Filing to the left and Filing to the right.

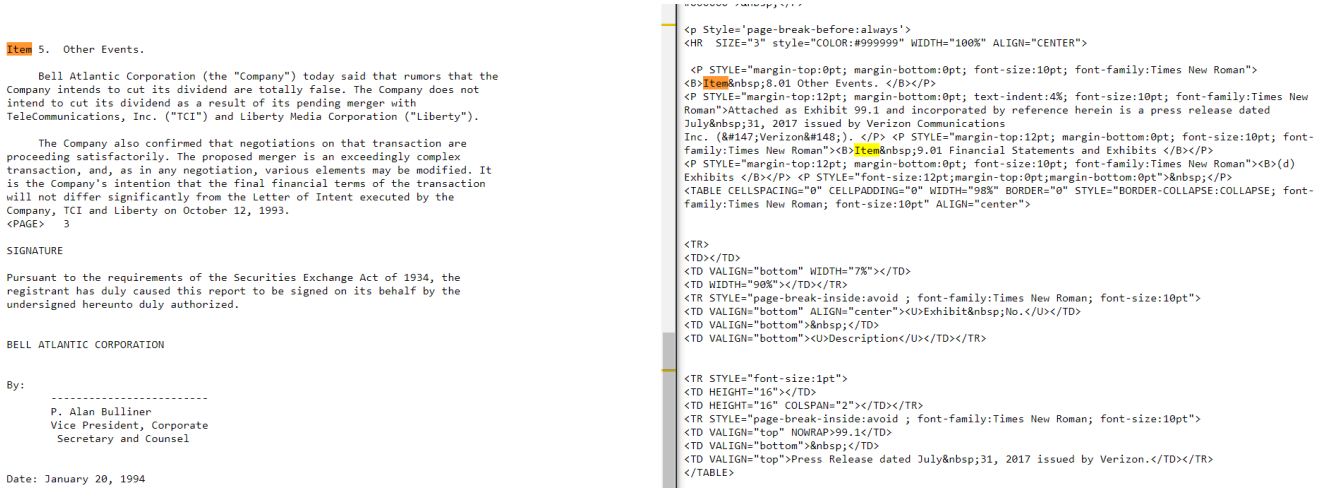


Figure 1: Different format used in 8-Ks.

- Filings from before August 23, 2004 do not share the same item category/item number as filings made after August 23, 2004 (e.g. *Item 1. Financial Statements* vs *Item 1. Legal Proceedings* vs *Item 1.01 Entry into a Material Definitive Agreement* vs *Item 1. Financial Statements of this Quarterly Report* all appeared in the SEC filings with the same Item number).
- We are interested in analyzing 5 companies and they use different formats when filing to SEC. Consequently, event though SEC provides all filings with a .txt link (click for example), the filings are formatted differently.
- Company might add different descriptive language for same item category (e.g. *Item 9.01. Exhibits* vs *Item 9.01. Exhibits* vs *Item 9.01. Financial* vs *Item 9.01. Financial Statements and Exhibits* all concern the same item category).
- Every 8-K could have one or more topics resulting in one or more item categories in any 8-K filing.

Extensive effort was put into extracting the desired categories and the associated language out of the 8-K filings. We created multiple parsers to process the filings, using both the HTML-tag and the regular expression approach.

We noticed that all filings made before August 23, 2004 have single digit Item number (e.g. *Item 5*) followed by a descriptive language, while all filings made after August 23, 2004 have two Item numbers with two decimal places (e.g. *Item 5.01*) followed by a descriptive language. Further, while processing

using a regular expression, we noticed that even though the descriptive language may be different, each within time period, Items with same Item number essentially concern the same category (e.g. *Item 9.01 Financial Statements and Exhibits* and *Item 9.01 Financial Statement* and *Item 9.01 Exhibits* all include *Item 9.01* but have different descriptive language; nevertheless, they can be merged into one category). Therefore, we decided to preserve only the Item numbers and treat same item number as same category.

```
print(all_data_8k['updated_item_category_splited'].unique())
print('Unique item category before cleaning: ', len(all_data_8k['updated_item_category_splited'].unique()))
```

```
 'Item 2.03.    Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Regist
rant'
 'Item 4.01 section (b) of the form 8-K as it does not relate to our Firm.  Very truly yours,  /s/ BDO Seidman, LLP  BDO Seid
man, LLP'
 'Item 5.02.    Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensat
ory Arrangements of Certain Officers'
 'Item 5.03.    Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year'
 'Item 3.03.    Material Modifications to Rights of Security Holders'
 'Item 1.02.    Termination of a Material Definitive Agreement'
 'Item 3.01.    Notice of Delisting or Failure to Satisfy a Continued Listing Rule or Standard; Transfer of Listing'
 'Item 3.03' 'Item 2.06 Material Impairments.' 'Item 3.03.'
 'Item 1.01 of this Current Report on Form 8-K is incorporated herein by reference.'
 'Item 9.01 Financial Statements and Exhibits.  (d)'
 'Item 2.01Completion of Acquisition or Disposition of Assets.'
 'Item 9.01Financial Statements and Exhibits.'
 'Item 2.03. Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet'
 'Item 7.01 is not intended to, and does not, constitute a representation that such furnishing is'
 'Item 5.02 Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory
Arrangements of Certain Officers.']
Unique item category before cleaning:  491
```

Figure 2: Before cleaning/merging *Item* categories: a total of 491 unique categories.

```
print(all_data_8k['updated_item_category_splited_v2'].unique())
print('Unique item category after cleaning: ', len(all_data_8k['updated_item_category_splited_v2'].unique()))
```

```
['Item 5.' 'Item 7.' 'Item 6.' 'Item 1.' 'Item 2.' 'Item 4.' 'Item 9.'
 'Item 8.01' 'Item 2.02' 'Item 1.01' 'Item 9.01' 'Item 5.02' 'Item 5.03'
 'Item 2.01' 'Item 2.04' 'Item 7.01' 'Item 3.01' 'Item 2.03' 'Item 5.07'
 'Item 1.02' 'Item 4.01' 'Item 5.05' 'Item 3.03' 'Item 3.02' 'Item 8.'
 'Item 5.04' 'Item 5.01' 'Item 3.' 'Item 1.03' 'Item 2.06']
Unique item category after cleaning:  30
```

Figure 3: After cleaning/merging *Item* categories: a total of 491 unique categories.

We also examined the distribution of the number of items within any 8-K filing. The hypothesis is that there might be a correlation between the number of Item categories a company is filing in a given 8-K filing and the instability of this company at the time of filing. Since the Moody's score is determined by the credit risk of a company, more Item categories could lead to a higher probability of a change in the Moody's score.

Figure 4 shows that there is always at least one Item category in any 8-K regardless of company and time period. However, the distribution varies across companies (e.g. 94% of 21st Century Fox's 8-Ks has 2 categories and there is only one 9 Item categories 8-K filing throughout the whole time period while AT&T's 8-Ks have more spread-out number of Item categories being filed: 40% 1 event, 52% 2 events and a very long right tail.) Figure in Section also demonstrates that *AT&T's* Moody's ratings change much more frequently than *21st Century Fox's*.

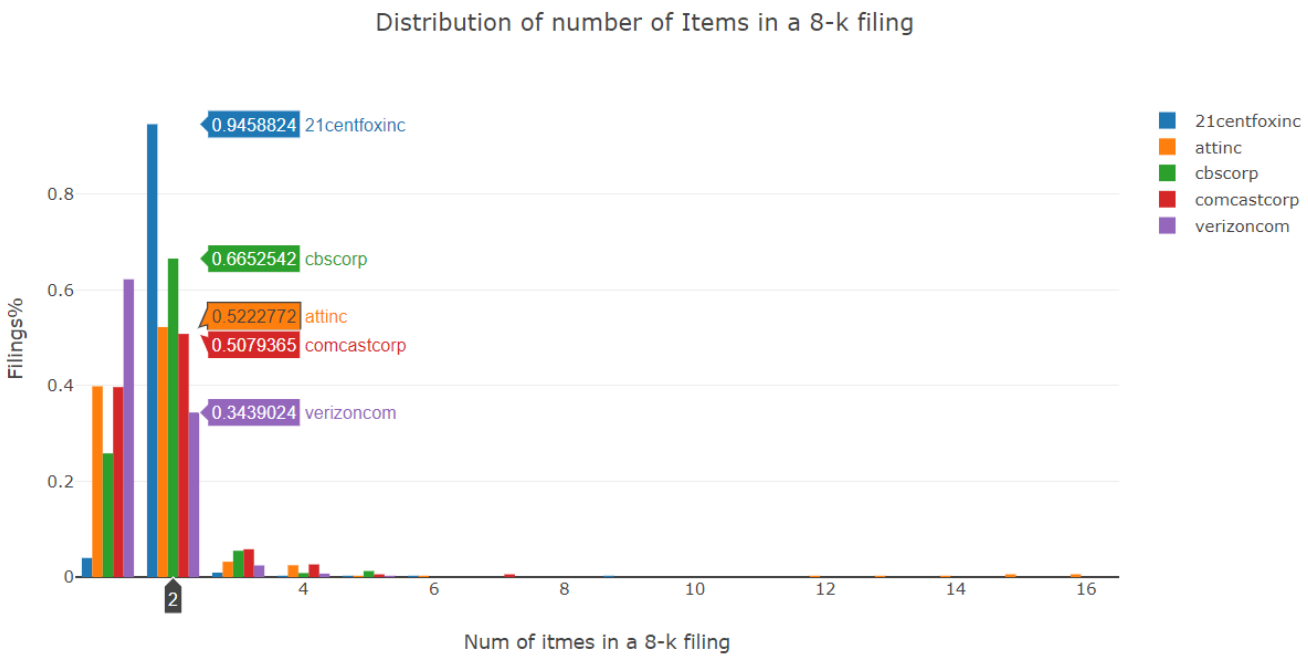


Figure 4: Distribution of the number of *Items*.

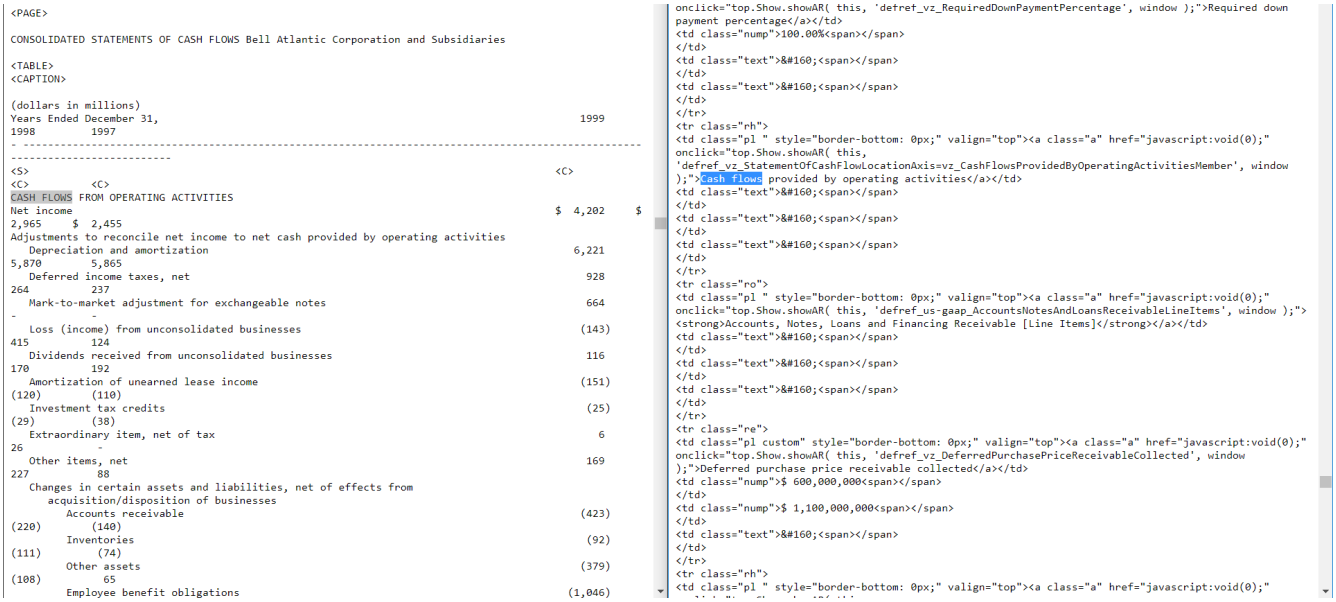
Moreover, we extracted text that directly explain the Item category. The hypothesis is that this language is more focused and centered towards the events being filed in any given 8-K. Therefore, if we perform sentiment analysis on them, we could achieve a higher prediction score rather than on the entire text. Once we identify all the Item categories, we need to extract all text in-between the Item categories. This is accomplished with basic cleaning techniques such as removing stop words, non-English words and tokenization.

2.2.2 10-K Financial Tables

Based on our discussion with Moody's, we realized that key financial ratios are critical to the determination of Moody's rating. 10-Ks are the annual reports that provide comprehensive summary of company's 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. Upon research, all the required financial ratios (listed in our first report) could be derived from either balance sheets or cash flow or income statements, and all the above three are included in 10-Ks. Therefore, we focused on how to extract these three financial tables form 10-Ks. 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.
- Similar to the 8-Ks' challenges described in the previous sections, event though the overall structure of 10-Ks is similar, each company files 10-Ks in a slightly different way. These differences make identifying the financial tables very challenging.
- 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. Figure 5 shows an example screenshot of a 1990s *Verizon* 10-K filing (left) and a 2018 *Verizon* 10-K filing. Full forms are also available following the links: 1990s *Verizon* 10-K and 2018 *Verizon* 10-K.



```
<PAGE>

CONSOLIDATED STATEMENTS OF CASH FLOWS Bell Atlantic Corporation and Subsidiaries

<TABLE>
<CAPTION>

(dollars in millions)
Years Ended December 31,                                                    1999
1998         1997
- ------------------------------------------------------------------------------------------------
-------------------------
<S>                                                                        <C>
<C>          <C>
CASH FLOWS FROM OPERATING ACTIVITIES
Net income                                                                $  4,202     $
2,965     $  2,455
Adjustments to reconcile net income to net cash provided by operating activities
   Depreciation and amortization                                            6,221
5,870        5,865
   Deferred income taxes, net                                               928
264          237
   Mark-to-market adjustment for exchangeable notes                         664
-            -
   Loss (income) from unconsolidated businesses                            (143)
415          124
   Dividends received from unconsolidated businesses                        116
170          192
   Amortization of unearned lease income                                   (151)
(120)        (110)
   Investment tax credits                                                   (25)
(29)         (38)
   Extraordinary item, net of tax                                             6
26           -
   Other items, net                                                         169
227          88
   Changes in certain assets and liabilities, net of effects from
      acquisition/disposition of businesses
         Accounts receivable                                               (423)
(220)        (140)
         Inventories                                                        (92)
(111)        (74)
         Other assets                                                      (379)
(108)        65
         Employee benefit obligations                                    (1,046)
```

```
onclick="top.Show.showAR( this, 'defref_vz_RequiredDownPaymentPercentage', window );">Required down
payment percentage</a></td>
<td class="numP">100.00%<span></span>
</td>
<td class="text"> <span></span>
</td>
<td class="text"> <span></span>
</td>
</tr>
<tr class="rh">
<td class="pl " style="border-bottom: 0px;" valign="top"><a class="a" href="javascript:void(0);"
onclick="top.Show.showAR( this,
'defref_vz_StatementOfCashFlowLocationAxis=vz_CashFlowsProvidedByOperatingActivitiesMember', window
);">Cash flows provided by operating activities</a></td>
<td class="text"> <span></span>
</td>
<td class="text"> <span></span>
</td>
<td class="text"> <span></span>
</td>
</tr>
<tr class="ro">
<td class="pl " style="border-bottom: 0px;" valign="top"><a class="a" href="javascript:void(0);"
onclick="top.Show.showAR( this, 'defref_us-gaap_AccountsNotesAndLoansReceivableLineItems', window );">
<strong>Accounts, Notes, Loans and Financing Receivable [Line Items]</strong></a></td>
<td class="text"> <span></span>
</td>
<td class="text"> <span></span>
</td>
<td class="text"> <span></span>
</td>
</tr>
<tr class="re">
<td class="pl custom" style="border-bottom: 0px;" valign="top"><a class="a" href="javascript:void(0);"
onclick="top.Show.showAR( this, 'defref_vz_DeferredPurchasePriceReceivableCollected', window
);">Deferred purchase price receivable collected</a></td>
<td class="numP">$ 600,000,000<span></span>
</td>
<td class="numP">$ 1,100,000,000<span></span>
</td>
<td class="text"> <span></span>
</td>
</tr>
<tr class="rh">
<td class="pl " style="border-bottom: 0px;" valign="top"><a class="a" href="javascript:void(0);"
```

Figure 5: Different 10-K filing format: a 1990s *Verizon* filing (left) and a 2018 *Verizon* filing (right).

- The formatting of strings and characters in 10-Ks varies between companies and across documents for the same company. There is a unifying way of recording cash flow, balance sheet and income statement. Figure 6 shows an example of a financial table we extracted from the 10-K in raw string format. Even though the table visually looks fine, note the several problems with this table, including:
  - multi-level column;
  - row headers are broken into multiple lines;
  - stars and dashes.

```
1  working_file['raw_table'][2]
```

```
[<table>
 <caption>
                                                Three months ended                  Year ended
                                                   December 31                      December 31
                                                ------------------              ------------------
                                                1993          1992*             1993          1992*

 <s> <c> <c> <c> <c>
 OPERATING REVENUES
  Communications and
   Related Services                          $3,194.5      $3,117.9          $12,534.8     $12,164.6
  Financial, Real Estate,
   and Other Services                           122.7         156.3              455.4         553.8

      Total operating revenues                3,317.2       3,274.2           12,990.2      12,718.4

 OPERATING EXPENSES
  Employee costs, including
    benefits and taxes                        1,019.0       1,004.3            4,027.6       3,941.5
  Depreciation and
```

Figure 6: Raw multi-level index financial table.

Figure shows what it looks like after our attempt to re-construct it into a tabular format. Note it is not quite right.

```
[[['', ''],
  ['',
   'Three months endedYear endedDecember 31December 31------------------------------------19931992*19931992*'],
  ['',
   'OPERATING REVENUES Communications andRelated Services$3,194.5$3,117.9$12,534.8$12,164.6 Financial, Real Estate,and Other
Services122.7156.3455.4553.8'],
  ['', 'Total operating revenues3,317.23,274.212,990.212,718.4'],
  ['OPERATING EXPENSES Employee costs, including',
   'benefits and taxes1,019.01,004.34,027.63,941.5 Depreciation andamortization649.8597.22,545.12,417.4 Other1,041.21,154.63,
619.93,853.3'],
  ['', 'Total operating expenses2,710.02,756.110,192.610,212.2'],
  ['OPERATING INCOME', '607.2518.12,797.62,506.2'],
  ['Other income and',
   'expense, net29.142.388.1214.4Interest expense, excludingFinancial Services141.3162.3612.1694.9Income before provision for
incometaxes, extraordinary item, andcumulative effect of changes inaccounting principles495.0398.12,273.62,025.7Provision for
income taxes157.870.6792.0643.5Income before extraordinary itemand cumulative effect of changesin accounting principles337.23
27.51,481.61,382.2Extraordinary item -- earlyextinguishment of debt,net of tax(4.1)(10.2)(58.4)(41.6)Cumulative effect of cha
nges inaccounting principles:Income taxes----65.2--Postemployment benefits,net of tax----(85.0)--Total cumulative effectof ch
anges in accountingprinciples----(19.8)--'],
  ['NET INCOME', '$333.1$317.3$1,403.4$1,340.6']]
```

Figure 7: Multi-level financial table after an attempted reconstruction.

So far we were 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 (Figure 8). 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.

```
2009-02-25 :Can't extract tables
2010-02-25 :Can't extract tables
2011-03-01
                                        0        1        2        3
10                Total operating revenues  124,280  122,513  123,443
15           Depreciation and amortization   19,379   19,515   19,673
20                       Interest expense    2,994    3,368    3,369
10                  Deferred income taxes    1,170    1,247      NaN
12                   Total current assets   19,951   25,187      NaN
13    Property, Plant and Equipment - Net  103,196   99,519      NaN
29              Total current liabilities   33,951   36,951      NaN
30                          Long-Term Debt   58,971   64,720      NaN
32                  Deferred income taxes   22,070   23,579      NaN
43                Noncontrolling interest      303      425      NaN
7            Depreciation and amortization   19,379   19,515   19,673
39                          Dividends paid    9,916    9,670    9,507
44  Cash and Cash Equivalents End of Year    1,437    3,741    1,727
```

Figure 8: Parsing result for an *AT&T's* 10-K file.

Overall, in order for us to extract the financial terms for all 5 companies, we would need to create at least 10 scripts, which is very time consuming. Additional to it, 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. As mentioned in our first report, Moody's changes methodology over time and uses different methodology for each industry. Hence, we switched our focus to 8-K files. We have not yet put any financial ratios into any of our models, but we will include some ratios as we believe there is a positive correlation between Moody's ratings and the financial ratios.

Action-wise, we are looking for Vanguard to provide the financial ratios as well as OAS going as far back as possible (preferably all the way to the 1990s).

3 Scraping and Sentiment Analysis of The New York Times

The New York Times makes it possible to scrape headlines with short summaries for all search results without the need to log in to their online service. Logging in would be necessary in order to download the full text of the news, which we are not going to pursue for the purposes of this project due to time constraints.

3.1 Scraping Procedure and Challenges

After manually typing a search term in The NYT search tool (start and end date optional) and hitting enter, a website with search results appears with the number of search results showing in the top left corner. Only about 10 results are shown in this initial page and to see more, we need to scroll down and click the *Show more* button. The scraping code starts by constructing the url that would result form a manual search and it proceeds by iteratively automatically clicking the *Show more* button to reveal additional 10 search results, until all search results are revealed. The scraping function takes as inputs the search term (e.g. "CBS") as well as the start and end dates, and it returns a dataframe with a date

of publication, headline, summary and a link to the full text, for each news record. For the details of the implementation please see the code on our GitHub: NYT scraping.

We have encountered the following challenges while scraping:

- **Multiple company names**

  We need to keep track of all company names (search terms) and time periods, resulting in multiple search terms used for each company, with their associated search start and end dates - this is taken care of in the scraping code.

- **Dependence of search result on search query**

  We found that the number of search results depends on the search term, this dependence is illustrated in Table 2 in the Appendix - note the high number of search results for *"AT&T"* in 2005-2018 vs 1000 times smaller number of results for *"AT&T Inc"*.

- **The NYT search tool failures/malfunctions**

  The scraping procedure could theoretically work very fast, with a couple of second per each *Show more* click. However, in practice we found that such high speed of scraping tends to break The NYT search tool - an error page appears an no more scraping is possible. Through experimentation we found that a waiting time of 30-40 seconds between each *Show more* click limited the number of failures considerably. In a few cases where failures persisted, we modified the search time period by e.g. splitting the original time in two, which lead to different amount of downloaded files for different search terms (see in GitHub: The NYT scraped raw data).

  Another way the search tool malfunctioned as the scarping proceeded was observed: after some time has passed since the scraping began many of the search results tended to repeat (the code ignored the repeated records and did not download them).

- **Gaps in news coverage**

  The scraped dataset on our GitHub represents our final attempt after experimenting with various search terms. In particular, we were trying to fill in the news gap for *AT&T* between 2006-2010. This gap appeared in our initial dataset, obtained by using *"AT&T Inc"* as the search term. Since this query returns a very small amount of search results (about 650) and *"AT&T"* returns to many to be feasible for us to scrape (about 670,000), we decided to use an alternative: *at&t inc*, as returned by the prompter in The NYT search tool, which produced about 330,000 results (a compromise).

- **Irrelevant search results**

  We found that most of the search results for any given search query do not concern the company of interest: there is no mention of the company either in the headline or in the summary (Figure 9 shows that the number of records drops dramatically if we only keep the news records that contain the company name). Moreover, a manual spot-check of the full text for some search results revealed that often there is no mention of the company of interest in the full text of the article. It is really unfortunate that The NYT search tool returns so many irrelevant results, as we have spent a lot of time scraping this data, most ow which turns out to be irrelevant.

3.2 Downloaded News Dataset

Table 1 and Figure 9 show the number or records obtained using the final choice of search terms for each company and each time period. In all cases but one a search term was used as shown in Table 1, including

the double quotation marks, for the corresponding time span (shown in Table in the Appendix). In the case of *AT&T* for the time period 2005-2018, the search term was used as prompted by The NYT search tool, *at&t inc*, as mentioned in the previous section.

Company	Search Terms	Search Results	Downloaded	De-duped	Filtered
21st-Century-Fox	"News Corp", "21st Century Fox"	1971	1970	1970	803
AT&T	"Southwestern Bell", "SBC", at&t inc*	33345	32930	30899	1111
CBS	"Viacom", "CBS"	28468	24950	23730	4358
Comcast	"Comcast"	8618	4107	4026	754
Verizon	"Bell Atlantic", "Verizon"	9692	7108	7096	1615

Table 1: Number of records: as given by The NYT search, actually downloaded, de-duplicated and filtered with respect to key words (company names).

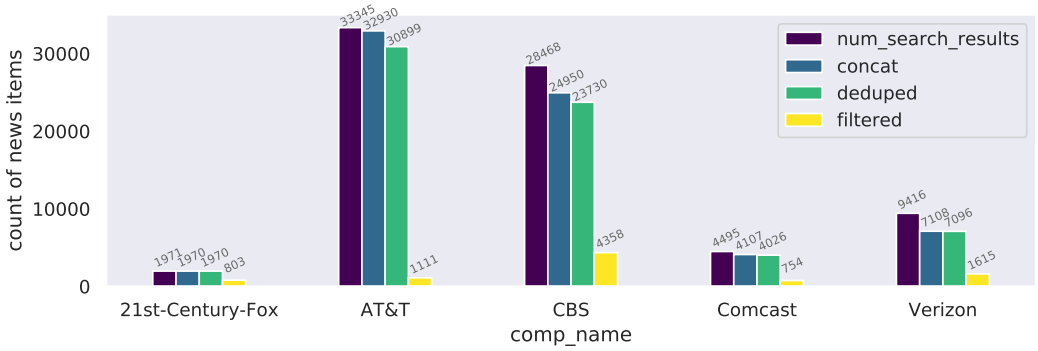


Figure 9: Count of news records: number of search results as shown in the NYT search tool (dark blue), number of successfully scraped/downloaded news items (light blue), number after de-duplication by unique link (green) and after filtering out items that don't include key words (company name) in the headline or the summary (yellow).

In some cases (*CBS*, *Verizon*) the number of downloaded news is considerably lower than the number of search results as indicated by The NYT search tool. This is perhaps because the search tool does not indicate the correct number of results to begin with, or perhaps some of the results are lost in the process of the augmentation of the search results (continuous clicking of the *Show more* button) - as the server might struggle displaying so many results. The number of records decreases even more after de-duplication of the records (done by keeping the records with unique links), which points out another malfunction of The NYT search tool in displaying the same record multiple times with a different link.

Finally, the records for each company are filtered to keep the records that contain key words in either the headline or the summary. The key words user were: 'fox' and 'news corp' for *21st Century Fox*; 'at&t', 'sbc' and 'southwestern' for *AT&T*; 'cbs' and 'viacom' for *CBS*; 'comcast' for *Comcast*; and 'verizon' and 'bell atlantic' for *Verizon*. Filtering decreased the number of records dramatically, suggesting that most of the search results provided by the The NYT's search tool is irrelevant to the search query.

Figure 10 shows the weekly count of news records that remained after data processing (key-word filtering). The intensity of news coverage varies between companies and also for each company over time. Of note is

the gap in coverage for *AT&T* between 2006-2010, as well as the period of prolonged intensive coverage for *CBS* between 2008-2010. The spikes in intensity correspond to significant events in the company, such as mergers, acquisitions and scandals (e.g. data hacking, sexual harassment).

3.3 Sentiment Analysis

Lexicon-based sentiment analysis was performed using two lexicons:

1. Vader (a build-in *nltk* function): sentiment scores were calculated for: headlines, summaries, concatenated headlines and summaries. We used the compound of the positive score.
2. LM financial lexicon (Yunwen's function): sentiment was calculated on summaries to derive multiple metrics, such as positive score, litigious frequency etc.

The code implementing the sentiment calculation can be found on our GitHub: The NYT Sentiment code, along with the output sentiment features for each company: The NYT Sentiment data.

Figure 12a shows a correlation matrix for all sentiment scores calculated for *AT&T* (correlations for other companies look very similar and the figures can be found on our GitHub: The NYT Figures). Vader sentiments (compound positive scores) tend to be correlated with each other and with the LM positive score and frequency. This result can be considered a successful 'sanity check' - it is expected that the two lexicons give correlated results, even though one is a general language lexicon (Vader) and the other is optimized for financial texts (LM). Also of note is the correlation of litigious frequency and negative frequency, which are anti-correlated with the positive scores. These relationships can also be seen in Figure 12c - as positive scores move together over time and in the opposite direction to the litigious frequency. These results are not surprising, as intuitively higher frequency of litigious language should correspond to more negative (less positive) overall sentiment. This trend can be observed for all companies (Figure 17 in the Appendix and the associated interactive plots for each company: Fox, AT&T, CBS, Comcast, Verizon). Of interest is also the correlation of modal frequency with uncertainty frequency.

Figure 12b shows several sentiment scores for each news record over time. Note the differences in the scale: Vader score is distributed between -1 and 1, while the LM scores are concentrated closer to 0. LM positive score is skewed towards the positive values, while LM litigious frequency is skewed towards the negative. This is true for all companies, as demonstrated in Figure 18 in the Appendix and the associated interactive plots for each company: Fox, AT&T, CBS, Comcast, Verizon. The interactive plots have hover-on labels including headline and summary text displayed for each data point, which is quite useful for spot-checking the accuracy of sentiment scored or for investigating the topics covered around the spikes of high intensity news coverage.

Figure 10: Weekly count of The NYT news records for all 5 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.

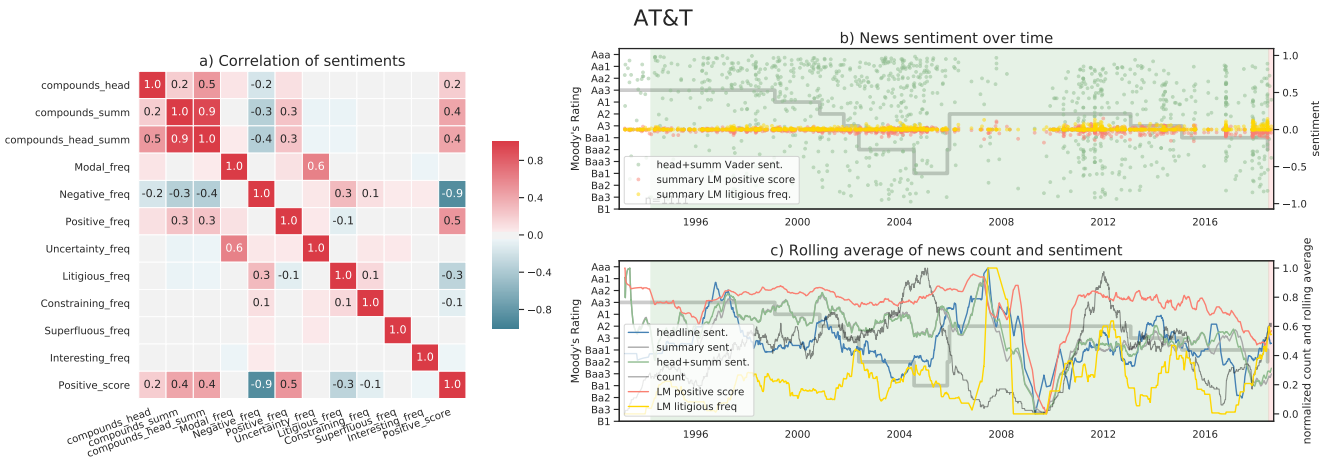


Figure 11: The NYT sentiment analysis for *AT&T*: a) correlation of all computed sentiment scores - only values of correlation with absolute value greater than 0.1 are labelled; b) selected sentiment scores shown for individual news records over time, Moody's rating is represented with a thick grey step line; c) yearly rolling average of selected sentiment scores and records count shown over time, Moody's rating is represented with a thick grey step line. Interactive versions of plots b) and c) can be accessed using these links: individual records plot, rolling average plot; the individual records plot includes hover-on labels with headline and summary text for each data point. Figures 17 and 18 in the Appendix, and the associated interactive plots, show The NYT Sentiment over time for all companies.

3.4 Possible Sources of Bias

The bias in our sentiment analysis of The NYT news may stem from several sources:

- **Bias of The New York Times**

  It possible that our data source itself is biased: The NYT may cover different companies/industries in a different way, e.g. they might dislike or have some political agenda against *AT&T* and tend to write often in a negative tone about that company. Since The NYT is our only source of news it we are not able to check if this bias takes place.

- **Variability of news coverage over time**

  We do observe variability in the news coverage intensity over time for most companies (e.g. 'gaps' or periods of very intense coverage) and it is unclear whether it is due to a fault in our scraping procedure, a malfunction of The NYT search tool, or perhaps our scraped dataset a true reflection of what The NYT wrote about each company. For instance, consider the coverage gap for *AT&T* in 2006-2010 (Figure 12b and c): in this case lower news coverage led to more extreme rolling average of sentiment and low rolling average of positive sentiment during that time - it is possible that if we had better coverage, the sentiment would be more balanced.

- **Impurity of search results**

  It is a fact that some search terms used in our scraping are more popular than others, e.g. we managed to obtain 4358 records for *CBS* but only 803 for *21st Century Fox*. It should be noted that *CBS* is not only a multimedia company (*CBS Corporation*) but also a very popular network TV station. Using "CBS" as a search term in The NYT search tool captures both articles about the multimedia company as well as articles related to the TV station (e.g. articles about the developments in some CBS reality show).

We decided to broaden our search to "CBS" instead of "CBS Corp" due to small amount of search results for the latter and our intuition that the multimedia company might sometimes be referred to as "CBS', especially in the headlines and summaries, which value brevity. However, we have to accept the fact that a subset of our dataset relates only to the TV station and it is unclear if a negative sentiment in these articles reflects is a good proxy for what happens at the multimedia company.

In contrast, we expect that *21st Century Fox* search results do not have this impurity problem, as there are no other entities with that name. Therefore, our news dataset varies in the impurity level between companies, which introduces bias.

- **Lexicon inadequacy**

  Financial sentiment lexicon (LM) might not be the best indicator of the sentiment in news headlines or summaries, as these are not financial documents.

3.5 Possible Further Work

Given more time, we would like to:

- Use transfer learning to classify the news items into good and bad, using *fast.ai* and transfer learning (as mentioned in the first Report). This would require manually labelling a subset of news records for each company (e.g. 100, but the more the better) and then using a deep network pre-trained on a general corpus to perform the classification task. Prediction derived this way could potentially be a better indicator of the sentiment, rather than a lexicon-based analysis.
- Download full text of the news records using the scraped links and perform sentiment analysis done on the full text. It is not clear if this would improve our prediction, as full text can potentially have more noise rather than a headline or a summary, but it could be worth trying. However, even though we have the code that would allow us to scrape full text, the legality of doing that is somewhat questionable, as The NYT only allows to use the downloaded materials for personal reasons in the terms and conditions of their online service.

4 Moody's Ratings

4.1 Moody's *Type* Derivation and Record Filtering

The originally provided Moody's dataset contains 72 data points for 5 companies combined. The goal of the project is to predict the **change** in Moody's rating, however, the provided *Type* column cannot be used, as it takes into account the *Watch* variable and ends up a binary column only including 'up' and 'down'. This is because even when the *Moody's Rating* remains unchanged, the *Watch* is either '+' or '-' and *Type* becomes 'up' or 'down'. We needed to compute our own *Type* variable - a direct derivative of the *Moody's Rating*, that includes a 'no change' class. The code performing this and the following analyses is on our GitHub: Moody's Analysis.

Moreover, we have filtered out all Moody's records dated before the earliest SEC filing in our SEC dataset for any given company. Additionally, we removed the one record where Moody's score was withdrawn for *AT&T*. Finally, any 'initial' records were removed for all companies, since we could not derive the Type

(change in rating) variable for such records. 52 records remained at the end of this procedure. Figure 12 shows that the 'same' or 'no change' class is the most frequent for all companies combined as well as for most companies considered individually. Of note is that some companies are missing one of the classes: e.g. *21st Century Fox* and *Comcast* are missing 'downgrade'. Our target dataset is **small** and **imbalanced**, which makes our modeling very challenging.

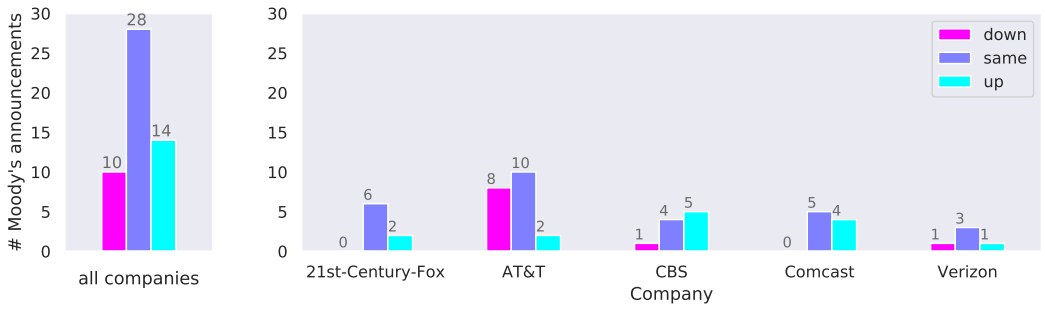


Figure 12: Count of Moody's rating change classes: *upgrade*, *downgrade* and *same* (no change), shown for all companies combined (left) and per company (right).

4.2 Analysis of Intervals Between Moody's Events

For the purposes of modelling we want to investigate periods of time of set length before any Moody's announcement. Figure 13 shows cumulative distribution of the intervals between consecutive Moody's events in our filtered dataset. The intervals are distributed from very short (days) to very long (over 9 years). This analysis informs our choice of the length of the time period to be taken into consideration, e.g. if we choose 6 months, we will have to deal with 17 (33%) of our time series overlapping; if we choose 2 months, only 7 time series will be overlapping, but it remains to be seen if such a short time period would provide better performance when used in modeling.

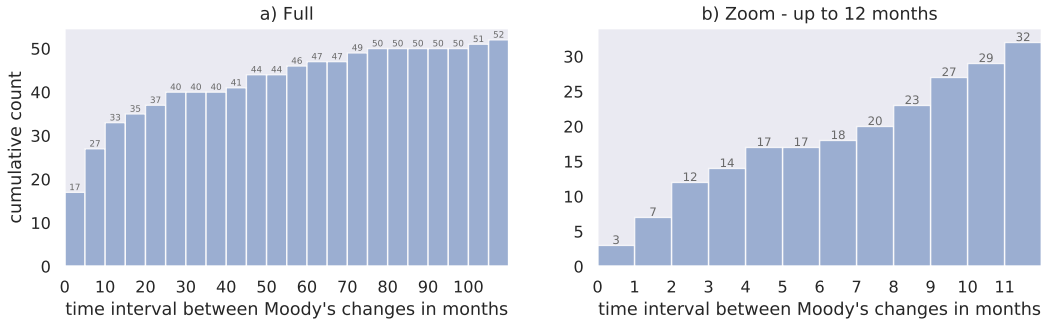


Figure 13: Cumulative distribution of time **intervals** between consecutive Moody's rating announcements (the announcements where no change was made are included): a) entire filtered Moody's dataset, b) up to 12 months.

5 Modeling

5.1 Features and EDA

The dataset used as features in the models presented in this report can be divided into a structured part and an unstructured part.

The structured part includes 9 sentiment scores: *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.

The unstructured part includes the raw text of the **event** announcement extracted from the SEC filings. The reason we only use event announcement instead of the entire document are two-fold. Firstly, 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. Secondly, there is a lot of noise in the original documents, such as in forms containing financial ratios. We try to avoid noise in our text as well.

Additionally, we calculated the sentiment score exclusively on the event text, however, the results of these models were not as good as using sentiment score based on the entire document.

Figure 19 in the Appendix shows the evolution of the sentiment score with time along with Moody's rating (indicated by blue dots). It illustrates that *Tone_measure* (Negative_score in plot) 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.

5.2 Target

For the purposes of the models presented in this report we used the value of Moody's rating at the time of every SEC filing, and we associated (matched) with each SEC filing. Note that we removed all the data (features) from the time period when Verizon withdrew from Moody's (rating *WR*: 01/01/2000 - 02/03/2006).

5.3 Explored Models

Since we have two parts of information, we can build models using only the structured data, only the unstructured data or both.

5.3.1 Random Forest

The random forest model takes only structured data (sentiment scores) as input/features.

Figure 14 shows the actual and predicted training and test result of random forest. The training accuracy is 95%, which is much higher than the test accuracy 40%. We conclude that this model overfits our dataset. Figure 14 also shows the feature importance of our model. As expected, positive and negative frequency are the most influential factors among all features.

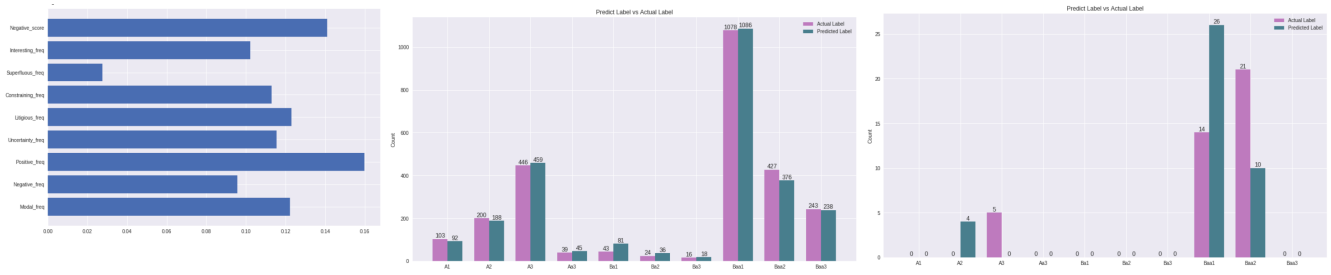


Figure 14: Random forest feature importance (left), training result (middle), and test result (right).

5.3.2 Long Short-Term Memory (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. Our LSTM network takes only unstructured data (event text) as input.

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.

Figure 15 shows the actual and predicted training and test results of our LSTM network. The training accuracy is 41%, which is similar to the test accuracy of 35%. However, since the model predicts the majority class, it doesn't really have any predicting power. We can use this majority class prediction as a baseline for predicting accuracy or 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. Therefofr, it's worth trying to apply a pre-trained word embedding, which will be our next step.

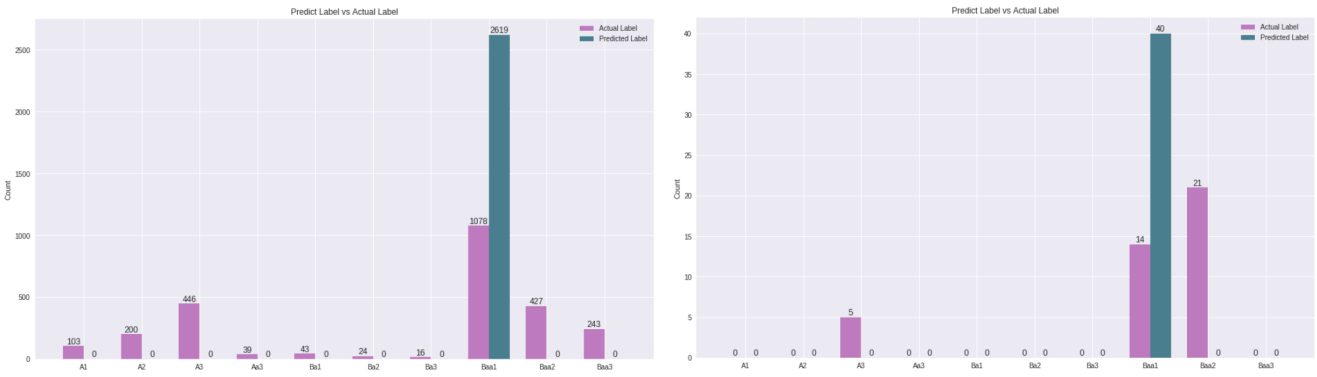


Figure 15: LSTM model training (left) and test (right) results.

5.3.3 Wide and Deep Model

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.

The deep model we used is a network with 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. Figure 16 shows the actual and predicted training and test results of our wide and deep network. The training accuracy is 40.2%, and the test accuracy is 35%. It does slightly batter 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.

A possible reason that the wide and deep model does not give a significant improvement on the performance compared to the random forest and LSTM network is that its major successful applications are for generic large-scale regression and classification problems with **sparse** inputs, such as recommender systems, while our input is not sparse.

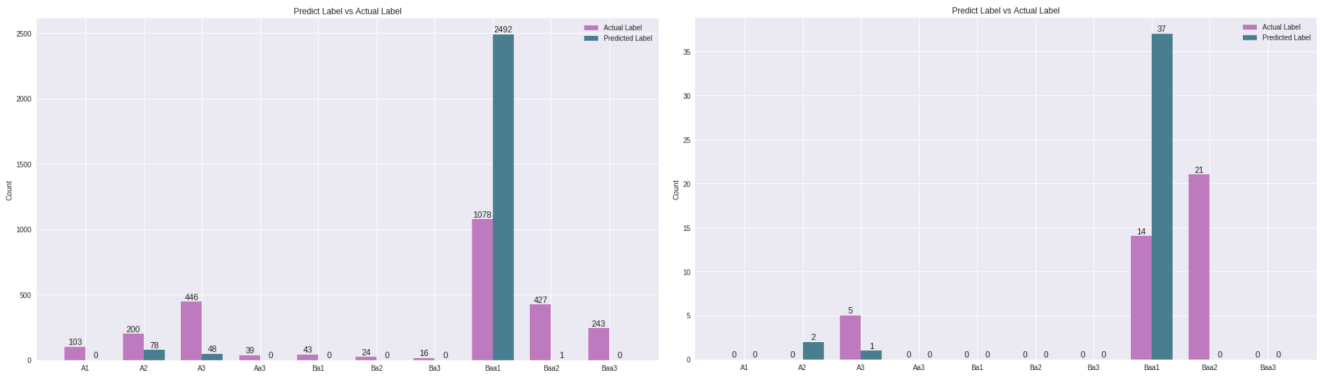


Figure 16: Wide and deep model training (left) and test (right) results.

5.4 Plan for Further Models

5.4.1 Feature Reconstruction

We were advised by Professor Muresan that using actual sentiment count instead of frequency might improve the result, so we will try using sentiment count. Also, we will try transfer learning on word embeddings, i.e. using one of the pre-trained word embeddings instead of training by ourselves.

Moreover, we will find a way to incorporate event types of 8-Ks into our features.

5.4.2 Deep Model for Time Series Classification

We would like to explore a deep-learning model for classification of multivariate time series. Recurrent or convolutional layers have been shown to be successful in this kind of tasks. This was illustrated by e.g. *Strodthoff and Strodthoff (2018)* with their convolutional model detecting abnormal heart activity based on ECG signal input. Also recently, *Laptev et al. (2017)* from Uber AI Labs built an end-to-end forecast model for multivariate time series, consisting of an autoencoder/feature extractor and a forecaster, both using LSTM layers. Laptev et al. report major gains in accuracy of prediction of rare events (e.g. the demand for Uber rides during holidays) compared to a previously used in-house model. We believe a similar approach could work in our application.

Once we obtain the financial data from Vanguard, we will have the following time series:

1. multiple sentiment scores on the SEC filings,
2. multiple sentiment scores on The NYT news,
3. even-driven categorical variables from the SEC filings,
4. financial data (from Vanguard),
5. possibly other, e.g. stock price, Google trends.

The idea would be to match each of the 52 Moody's rating type data points (target) with all of the above time series from a time period preceding every given Moody's rating (features). The model would be fed the multivariate time series as input, and would learn to recognize the *time patterns* that lead to Moody's announcements, hopefully learning to classify if the pattern leads to an *upgrade*, a *downgrade* or *no change* in Moody's rating.

We would also like to explore a wide-and-deep version of this model, where the input to the *deep* part would be the multivariate time series and the input to the *wide* part would be some non-temporal variables, such as e.g. categorical variables identifying a company (so that the model can distinguish between e.g. *AT&T* vs *CBS*) or an industry type (e.g. multimedia vs telcomm).

A Appendix

A.1 Supplementary Data and Materials

- Rajat Agrawal's Edgar Company Filings Web-crapping GitHub Repository: https://github.com/ragraw26/Edgar-COMPANY-FILINGS-Web-Scrapping-Data-Analysis
- Moody's Approach to Global Standard Adjustments in the Analysis of Financial Statements for Non-Financial Corporations: http://web.dpworld.com/wp-content/uploads/2014/05/Moodys-Approach-to-Global-Standard-Adjustments-in-the-Analysis-of-Financial-Statements-for-Non-Financial-Corporations-1012.pdf

A.2 References

1. Strodthoff, N. and Strodthoff, C. (2018) Detecting and interpreting myocardial infarctions using fully convolutional neural networks,

A.3 NYT Data Tables and Figures

Search Term	Company	Start Date	End Date	Number of Results
News Corporation	21st-Century-Fox	2003-11-24	2013-07-05	4463
News Corp	21st-Century-Fox	2003-11-24	2013-07-05	1104
21st Century Fox America Inc	21st-Century-Fox	2013-06-17	2018-08-31	0
21st Century Fox America	21st-Century-Fox	2013-06-17	2018-08-31	0
21st Century Fox	21st-Century-Fox	2013-06-17	2018-08-31	867
Southwestern Bell Corp	AT&T	1993-03-15	1995-05-09	29
Southwestern Bell	AT&T	1993-03-15	1995-05-09	166
SBC Communications Corp	AT&T	1994-11-08	2005-11-21	0
SBC Communications	AT&T	1994-11-08	2005-11-21	1157
SBC	AT&T	1994-11-08	2005-11-21	1576
AT&T Inc	AT&T	2005-11-05	2018-08-31	654
AT&T	AT&T	2005-11-05	2018-08-31	671546
Viacom Inc	CBS	1993-01-12	2006-01-09	1157
Viacom	CBS	1993-01-12	2006-01-09	3148
CBS Corp	CBS	2006-01-05	2018-08-31	854
CBS	CBS	2006-01-05	2018-08-31	25320
AT&T Comcast Corp	Comcast	2001-10-30	2003-01-10	0
AT&T Comcast	Comcast	2001-10-30	2003-01-10	45
AT&T	Comcast	2001-10-30	2003-01-10	46166
Comcast	Comcast	2001-10-30	2003-01-10	186
Comcast Corp	Comcast	2002-10-30	2018-08-31	470
Comcast	Comcast	2002-10-30	2018-08-31	4309
Bell Atlantic Corp	Verizon	1993-01-21	2000-11-14	447
Bell Atlantic	Verizon	1993-01-21	2000-11-14	1852
Verizon Communications Inc	Verizon	2000-09-08	2018-08-31	695
Verizon Communications	Verizon	2000-09-08	2018-08-31	1711
Verizon	Verizon	2000-09-08	2018-08-31	7840

Table 2: Dependence of the number of search results on the search query.

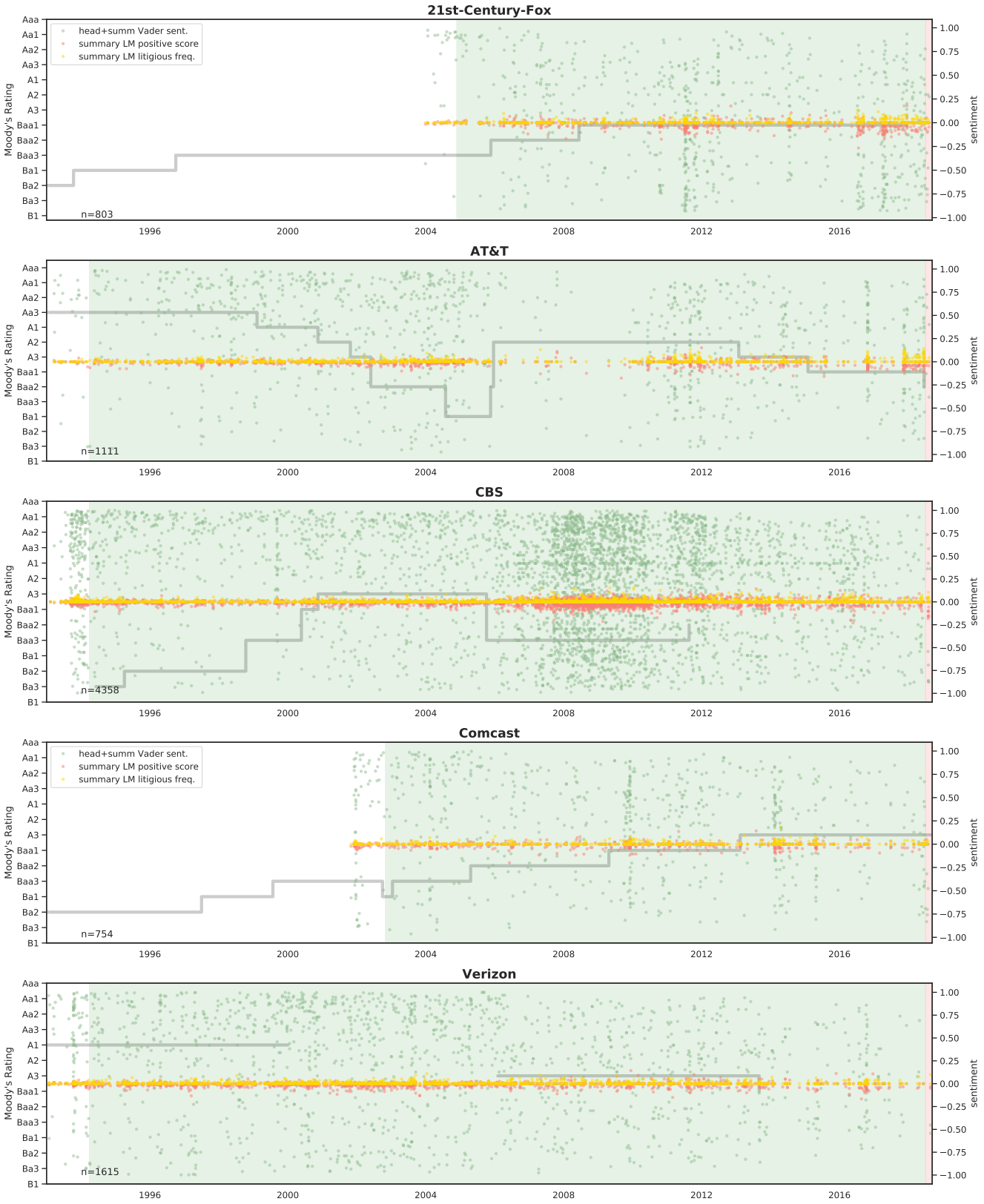


Figure 17: The NYT sentiment over time for all 5 companies: Vader compound score for concatenated headline and summary (green dots), LM positive score for summary (red) and LM litigious frequency for summary (yellow). To access interactive figures click here: Fox, AT&T, CBS, Comcast, Verizon. Moody's ratings are shown with a 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.

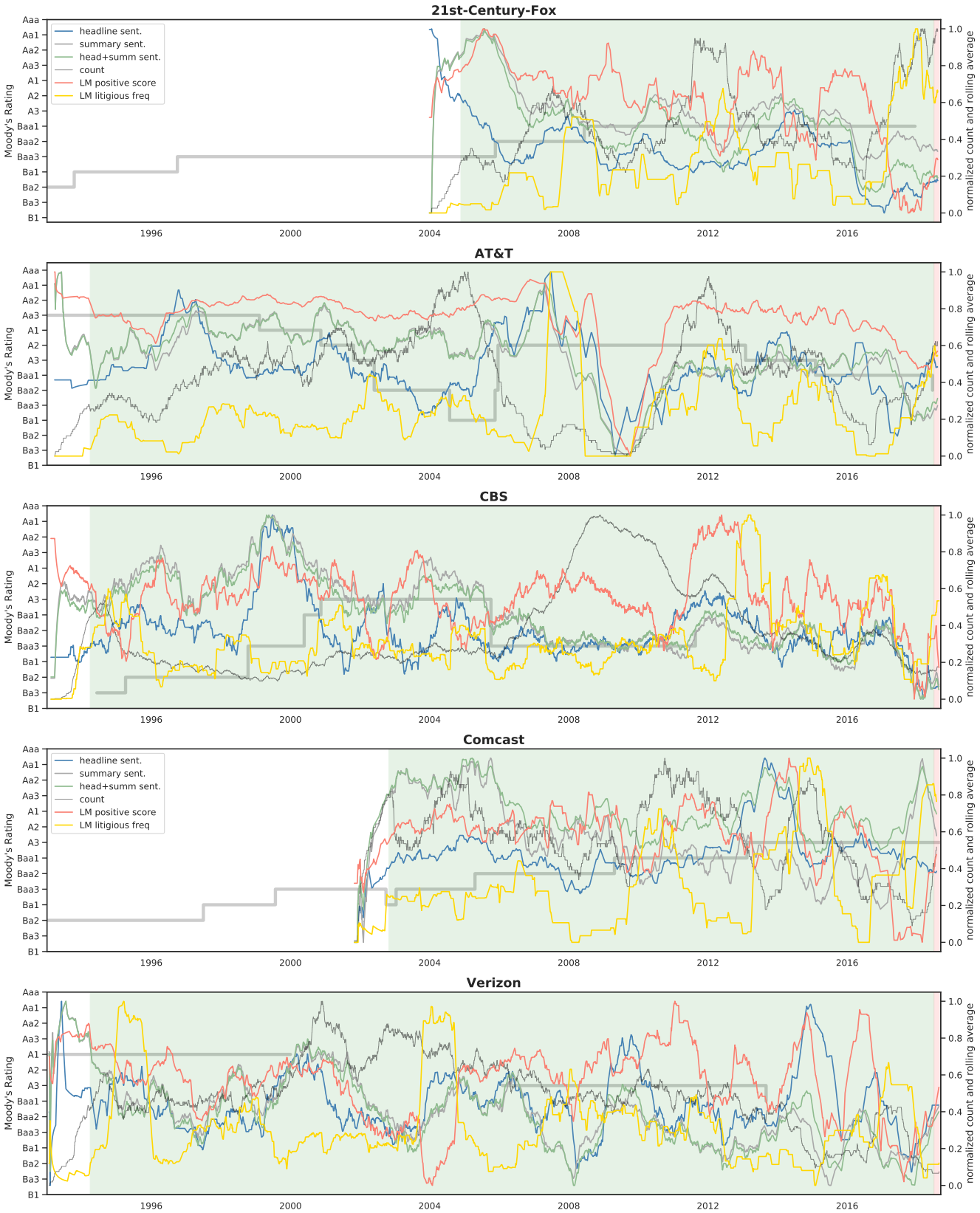


Figure 18: Rolling average (over 1 year) of The NYT sentiment displayed over time: Vader compound score for concatenated headline and summary (green) headline (blue) and summary (grey), LM scores for summaries: positive score (red) and litigious frequency (yellow). Rolling average of the count of news records is shown with a thin black line. To access interactive figures click here: Fox, AT&T, CBS, Comcast, Verizon. Moody's ratings are shown with a 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.

A.4 Figures of Sentiment on SEC Files

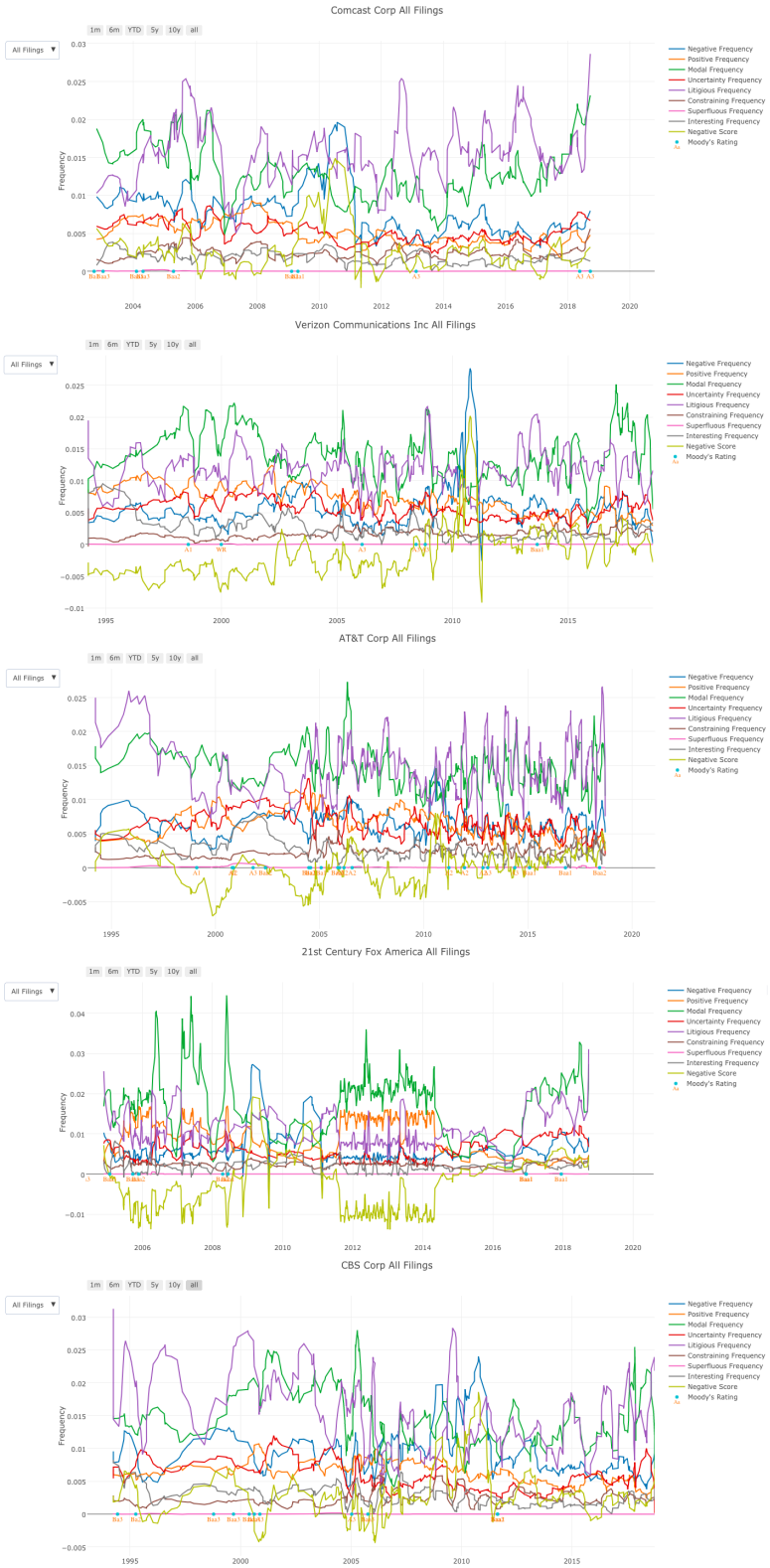


Figure 19: Sentiment score for all SEC filings. Savitzky-Golay filter was applied to data points to smoothen the lines without greatly distorting the signal.