Julex Final Presentation

August 7th

Yuxuan Zhang, Yuting Wang,

Ziyu Fan, Hasan Allahyarov

Brandeis | INTERNATIONAL BUSINESS SCHOOL

Sample and Research Design

Samples covers period from 2018 through 2022. (totally 69676 data)

We start with 35,395 10K and 24,526 10-Q filings.(2018-2022 QT1 &QT4). 6610 10-Q and 935 10-k(2022 QT2 &QT3)

Algorithm design: download text data > Regex to search occurrence > Beautifulsoup to Parse > Extract > Tokenize sentences as batches > put in Finbert > sentiment analysis

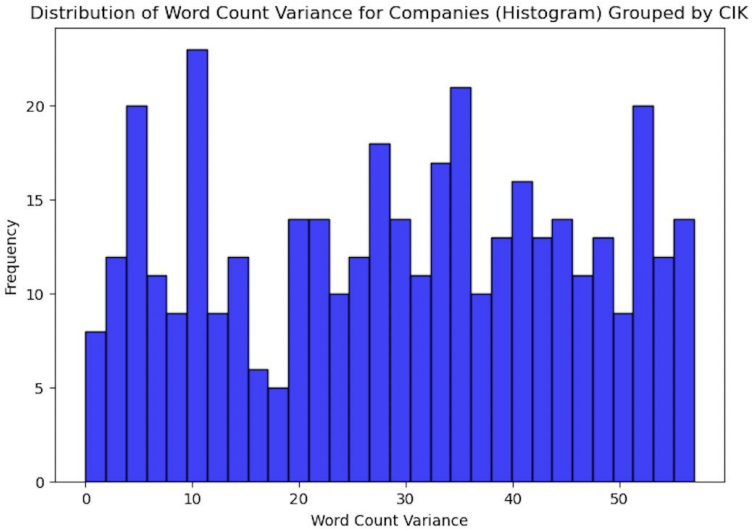
```
count     59921.000000
mean       9771.429516
std       12134.926999
min           1.000000
25%        1693.000000
50%        5784.000000
75%       12214.000000
max      130458.000000
Name: num_words, dtype: float64
```

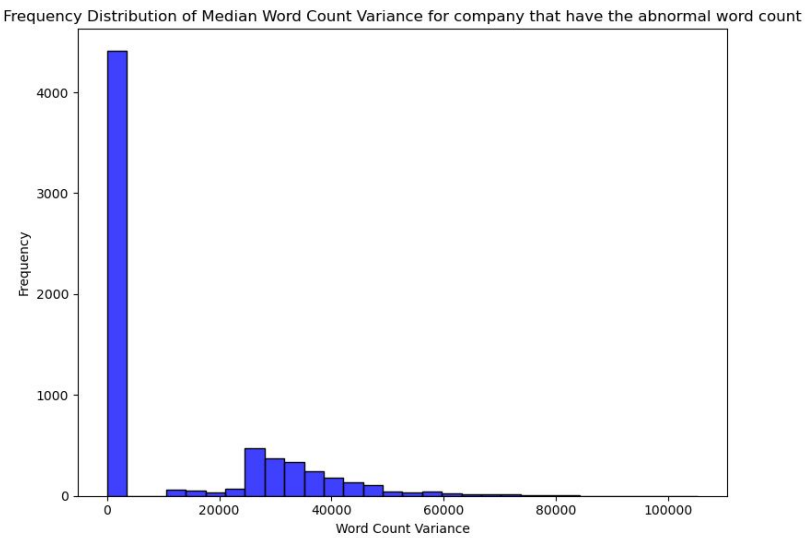
2018-2022 Q1/Q4

```
(                Q2             Q3
 count    7545.000000    2209.000000
 mean     2601.304175    2353.822091
 std      5855.292572    5488.728273
 min         0.000000       0.000000
 25%        10.000000      10.000000
 50%        26.000000      12.000000
 75%      3014.000000    2849.000000
 max    111193.000000  100231.000000,
 1.8353719295195232,
 0.06652868836385127,
 0.09860834990059641,
 0.10140334993209597)
```

2022 Q2/Q3

Companies with Median word counts vs with low word counts

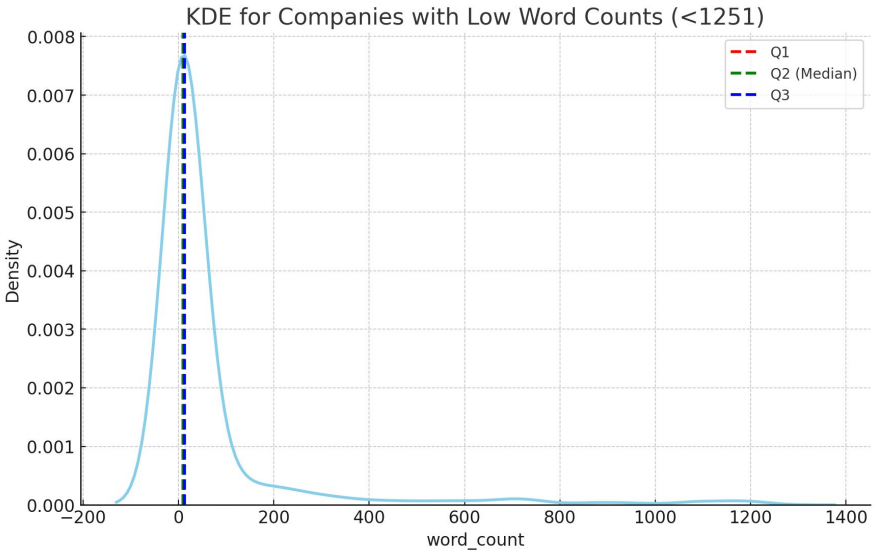


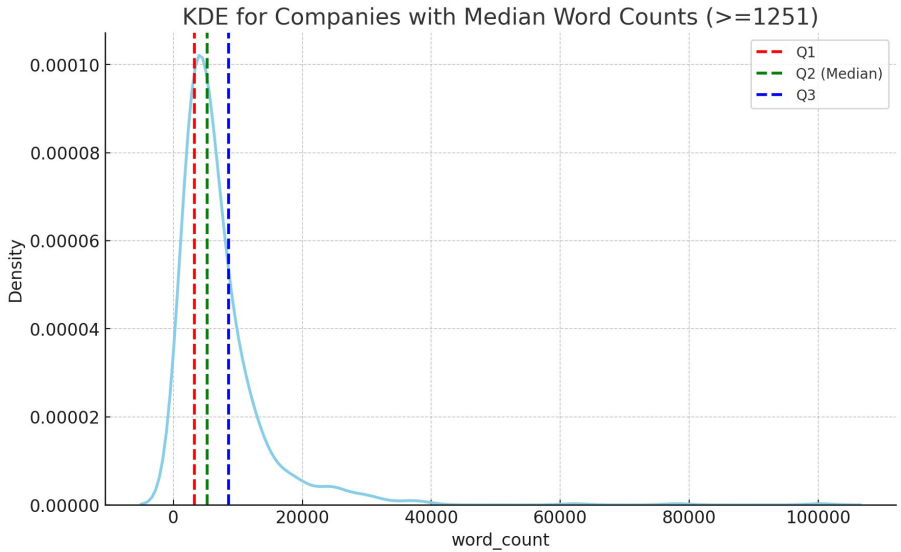


434 companies are situated close to the median count, falling within a 1% upper and lower bound. For companies exhibiting abnormal word counts (<1251 or >24052), the range for word count variance is significantly larger in comparison to those companies that have word counts more proximate to the median.

In the subsequent graph, the results consistently show a low word count variance. However, it's noteworthy that the majority of instances with low word count variance also coincide with a low quantity of Management Discussion and Analysis (MDA) extractions. This suggests that the algorithm's performance is suboptimal across multiple reports for specific companies

Word counts Variance





There are 24 companies whose word counts fall within 5% of the median word count. This means that these companies' word counts are closely aligned with the median word count across all companies in the dataset.

Strategy to evaluate algorithm accuracy on extraction: word count

We referred to “the information content of 10-K narratives” and multiple other documents…(research paper has 28712 10-k reports)

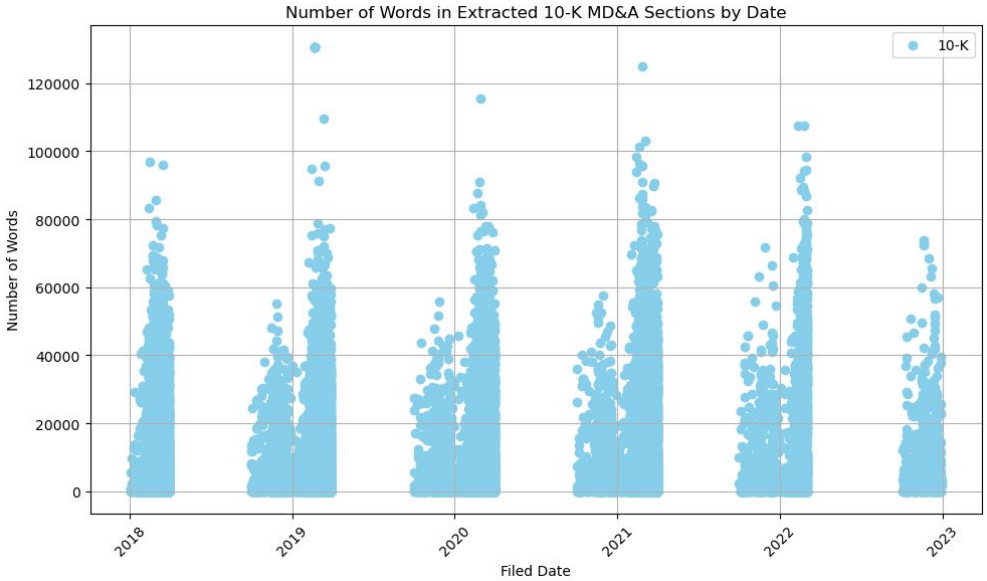
- Based on 20 years of 10-K extraction, Mean MDA word count length is 7470
- Consider flagging outlier >99% confidence level as abnormal
  - 1560/4886 for 2018,19,20,22 all 10-k
- Consider flagging reports with less than 1251 words as
  - There are repetitive subheadings
  - Very limited useful information
  - Inadequate extraction
- Usage extraction rate after is

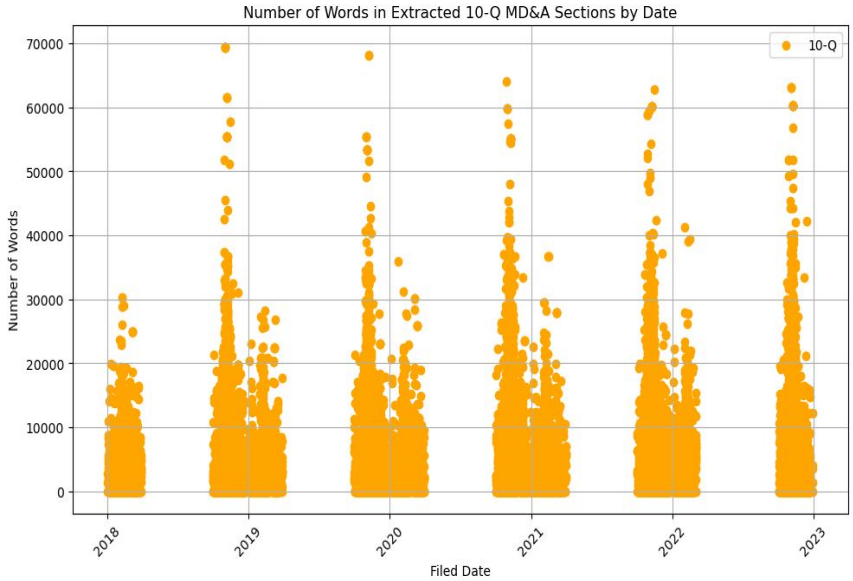
However, we have to check on excel sheet +EDGAR filing website to confirm the accuracy of the extraction

Panel A: Disclosure Variables

Disclosure Variables	Mean	St. Dev.	1%	10%	25%	50%	75%	90%	99%
Footnote Length (Words)	8,832	5,186	1,377	3,748	5,226	7,640	11,117	15,355	27,577
MD&A Length (Words)	7,470	4,722	1,251	2,645	3,984	6,459	9,764	13,569	24,052
Footnote Change	0.06	0.07	0.01	0.02	0.03	0.04	0.07	0.11	0.41
MD&A Change	0.08	0.07	0.01	0.02	0.04	0.06	0.10	0.17	0.36
Footnote Within-Industry Sim.	0.67	0.10	0.34	0.54	0.62	0.69	0.73	0.76	0.81
MD&A Within Industry Sim.	0.55	0.10	0.32	0.43	0.49	0.55	0.61	0.68	0.78
MD&A-Footnote Similarity	0.64	0.13	0.30	0.47	0.56	0.65	0.74	0.80	0.90
Footnote Sentiment									
Positive	0.95%	0.33%	0.30%	0.57%	0.73%	0.92%	1.14%	1.38%	1.95%
Negative	2.40%	0.96%	0.62%	1.29%	1.74%	2.28%	2.93%	3.69%	5.25%
"Net Positive"	-1.45%	0.98%	-4.31%	-2.75%	-1.99%	-1.33%	-0.78%	-0.31%	0.45%
MD&A Sentiment									
Positive	1.43%	0.55%	0.47%	0.81%	1.05%	1.35%	1.73%	2.15%	3.10%
Negative	2.74%	1.20%	0.58%	1.35%	1.88%	2.58%	3.45%	4.39%	6.06%
"Net Positive"	-1.31%	1.23%	-4.56%	-2.94%	-2.07%	-1.22%	-0.47%	0.18%	1.28%
Net Positive: MD&A - FN	0.14%	1.12%	-2.63%	-1.27%	-0.55%	0.15%	0.83%	1.49%	2.94%

Word Count Visualization After Flagging





10-Q abnormal rate vs normal: 0.289294430481186

10-K abnormal rate vs normal: **1.1451937374267471**

Current Algorithm Evaluation/ Challenges

a. Our general Algorithm is incapable to pick up headings listed below(and more…), which caused either under or over extracting in MDA session

Management's Discussion of Material Instance of Noncompliance and Steps Taken to Remedy the Material Instance of Noncompliance

Management's Report on Internal Control Over Financial Reporting

Item 7. Combined Management's Discussion and Analysis of Financial Condition and Results of Operations

Item 2. Management's Narrative Analysis

Current Algorithm Evaluation/ Challenges

```
pattern = re.compile(r'\bDiscussion\s+and\s+Analysis\s+of\s+Financial\s+Condition[s]?\b', re.IGNORECASE | re.DOTALL
matches = re.finditer(pattern, file_content)
indices = [match.start() for match in matches]

if len(indices) >= 1:
    start_index_1 = indices[0]
    if len(indices) >= 2:
        start_index_2 = indices[1]
    else:
        start_index_2 = -1
else:
    # If no matches found for the plain text pattern, check the bold HTML format
    pattern_html = re.compile(r'<b>\s*Management\s+Discussion\s+and\s+Analysis\s*</b>', re.IGNORECASE | re.DOTALL)
    matches_html = re.finditer(pattern_html, file_content)
    indices_html = [match.start() for match in matches_html]

    if len(indices_html) >= 1:
        start_index_1 = indices_html[0]
        if len(indices_html) >= 2:
            start_index_2 = indices_html[1]
        else:
            start_index_2 = -1
    else:
        # Handle the case when no occurrence is found for both patterns
        start_index_1 = -1
        start_index_2 = -1

pattern_end = re.compile(r"Disclosure[s]? About Market Risk", re.IGNORECASE)
```

```
if end_index_2==-1 and start_index_2==-1 and start_index_1<end_index_1:
    extracted_text = file_content[start_index_1:end_index_1]
elif start_index_2<end_index_2:
    extracted_text = file_content[start_index_2:end_index_2]
else:
    extracted_text=""
if start_index_1==-1:
    extracted_text=''
```

Reasons why it stops before the next item's actually heading if end_index_2<start_index_2.

Index list only store 2 occurrence which can either leads to over or under extraction

Moving forward: Strategy to improve Algorithm

- Polishing algorithm so it can detect headings are highlighted/bolded in HTML format more accurately (instead of using inside regex/ try in beautifulsoup)-failed
  - **Will significantly improve extraction accuracy based on manually checking**
  - **Particularly useful for identifying titles and headings, which often follow certain linguistic patterns.**
- Considering different formats of title
  - Include the frequently used expressions of title in our code ‘if-else’s
- Improve the algorithm by training on a diverse set of financial documents
  - Allows the algorithm to learn from a variety of formats and styles
  - Increases the algorithm's robustness and adaptability.

Regression

Get the return of a stock price of the filed date and the next day's price

- Challenge: No Ticker = No price

The effect of number of words on the return of the stock:

- Firm Fixed effects: Filed Date, Ticker, Document Type
- **Percentage_Diff = β0 + β1 * num_of_words + β2 * day_1 + ... + βn * day_n + βn+1 * ticker_1 +... + βm * ticker_m + βm+1 * document_type_1 + βm+2 * document_type_2**

Overfitting

- Regularization (Lasso or Ridge)
- Cross Validation

```
Mean Squared Error: 1.0614494786830598e+18
R-squared: -50806498393773.38
Coefficients:
num_words                   2.718531e-04
day_2018-01-02 00:00:00    -4.789800e+09
day_2018-01-03 00:00:00    -4.789800e+09
day_2018-01-04 00:00:00    -4.789800e+09
day_2018-01-05 00:00:00    -4.789800e+09
                                    ...
ticker_ZVSA                -4.162015e+00
ticker_ZWS                  2.653531e+00
ticker_ZYME                -2.106211e-01
ticker_ZYXI                 3.474707e+00
document_type_10-Q          2.661219e+01
Length: 5932, dtype: float64
Intercept:
4789799964.97034
```

Moving forward: Sentiment analysis method

- Use tokenized
- Slice the paragraph into sentences (sent_tokenize)
- Check whether the sentence exceeds 512
- Take only sentences with more than 6 words.

- FLS prediction
- Record the percentage of FLS sentences ('yiyanghkust/finbert-fls' model)
- Record the resulting sentiment for 'FLS' sentences ('ProsusAI/finbert' model)

Moving forward: Suggestion to Implement Sec API

- Sec Api’s advantages:
  - Provide correct extraction for all reports
  - No need to download reports at all. Possible to save only MDA extraction
  - Very user-friendly
  - Time and memory saver
  - Flexible Querying: filter and search data based on various parameters
- Sec Api’s disadvantages:
  - Need to clean HTML part (Tables, Subheadings and etc.)
  - Hard to find reports of non-existing companies (or without tickers)
  - Rate Limits: if requests a large amount of files in a short time, the user will encounter limitations

THANK YOU

Feel free to ask any questions

Brandeis Field Project Team

Brandeis | INTERNATIONAL BUSINESS SCHOOL