Fintech and AI in Finance: Coding Project Report

Mert Ergin

1982539

18.05.2025

Abstract

This report presents the coding project conducted for the course Fintech&AI and discusses the resulting findings. In this project, I apply the methodology developed by Cohen et al. (2020) to a more recent dataset.

Cohen et al. (2020) analyze financial filings of U.S. companies from 1995 to 2014 and demonstrate that textual changes in these filings have implications for firms' future returns. Specifically, they find that smaller changes in the filings compared to the previous year predict higher future returns. Moreover, a trading strategy that goes long on firms with minimal changes and short on those with substantial changes generates significant abnormal returns.

I apply their methodology to a more recent dataset—10-K filings from 2018 to 2024—to assess whether their findings remain valid in the current period. When using the Jaccard similarity measure to assess textual changes, I am able to closely reproduce the authors' results—both in terms of the performance of the long-short portfolio and the outcomes of the regression analysis. Interestingly, when employing cosine similarity as the measure of textual change, the long-short strategy yields returns that are more than nine times higher than those reported in the original study. However, this effect becomes milder when I restrict the sample to stocks with average monthly returns below 200 percent, bringing the results in line with those of the original paper. This suggests that the initially observed excess returns are likely driven by outliers. Therefore, I conclude that while the exaggerated performance using cosine similarity is unsystematic, the core findings of the original study can still be replicated using this alternative similarity measure.

The remainder of the report is structured as follows: the next section provides a brief summary of the original study; Section 3 outlines my implementation of the methodology; and the final section presents and discusses the results of my analysis.

1 Summary

Cohen et al. (2020) begin their study with a case analysis of Baxter International Inc., whose stock fell by over 20% within two weeks following April 23, 2010. They argue that this decline had already been signaled in the company's 10-K filing from February 23, 2010, for the 2009 fiscal year, but the market response was delayed due to investor inattention. The authors show that the 2009 filing was notably less similar to the previous year's and contained a higher frequency of concerning terms—for instance, FDA appeared 48 times (vs. 28 in 2008), Recall 30 times (vs. 20), and Colleague Pump nearly tripled in frequency.

Motivated by this case study, the authors conduct a comprehensive analysis of all 10-K and 10-Q filings from 1995 to 2014 to examine whether textual changes between consecutive financial reports are predictive of future stock returns. They further investigate whether this relationship can be exploited through a trading strategy that takes long positions in "non-changers" (firms with minimal textual changes) and short positions in "changers" (firms with substantial changes).

After collecting 10-K and 10-Q filings from the SEC's EDGAR database, the authors preprocess the texts following the approach of Loughran and McDonald (2011), removing tables, HTML tags, and similar non-textual elements. To quantify textual changes between reports, they apply four similarity metrics: (i) cosine similarity, (ii) Jaccard similarity, (iii) minimum edit distance, and (iv) simple similarity.[1] Monthly stock return data is obtained from CRSP. Similarity scores are measured relative to the corresponding report from the previous year—for instance, 2005 Q1 10-Qs are compared to 2004 Q1 10-Qs, and 2005 10-Ks to 2004 10-Ks.

The authors use 86,965 10-K filings and 258,271 10-Q filings in their analysis, resulting in a total of 327,130 similarity observations. For cosine similarity, they report a mean of 0.8721 and a standard deviation of 0.1910. For Jaccard similarity,

[1] For the mathematical definitions of these metrics, see the paper.

the mean is 0.3949 with a standard deviation of 0.1906. Additionally, they find a correlation of 0.6049 between the two similarity measures.

To conduct the portfolio analysis, it is assumed that stocks enter portfolios in the month following the public release of the report and are held for three months. For example, if a report is filed on 24.05.2002, the corresponding stock enters the portfolio on 01.06.1998 and remains there until 01.09.1998.

To classify firms as "changers" or "non-changers," the authors sort each filing into similarity quintiles based on the cross-sectional distribution of similarity scores within the same filing month and year. Quintile 1 (Q1) represents firms with the largest textual changes, while Quintile 5 (Q5) includes those with the least.

Using cosine similarity, the authors report average monthly returns of 0.63%, 0.72%, 0.72%, 0.85%, and 0.92% for quintiles Q1 through Q5, respectively. A long-short portfolio (Q5–Q1), which buys high-similarity firms and sells low-similarity ones, yields an average return of 0.31%. This suggests that larger textual changes in financial reports are associated with lower future returns. A similar pattern emerges with the Jaccard similarity measure: returns across quintiles are 0.59%, 0.67%, 0.69%, 0.82%, and 0.98%, with the Q5–Q1 strategy earning 0.38% per month.

To investigate whether these results are robust to different holding periods, the authors examine cumulative abnormal returns over a six-month horizon. They find that the cumulative return of the Q5 portfolio consistently exceeds that of the Q1 portfolio for each of the six months following the public release of the financial report. This suggests that the return differential between "nonchangers" and "changers" persists beyond the initial three-month holding period.

Lastly, the authors conduct Fama-MacBeth cross-sectional regressions of future firm-level stock returns on established return predictors and their similarity measures. They report that each similarity measure is a positive and statistically significant predictor of future returns, even after controlling for known factors. Without controls, the coefficients are 0.45 for cosine similarity and 0.31 for Jaccard similarity, indicating that higher textual similarity is associated with better subsequent performance.

2 My Implementation

2.1 Data

Due to the memory and computational constraints of my personal computer, I restrict my analysis to a subset of financial reports. I choose to focus on 10-K filings submitted during the first quarter of each year, as they are typically longer and contain more comprehensive information than 10-Q filings Cohen et al. (2020). The majority of firms file their 10-Ks in February or March of the following year, although some with "off-cycle" fiscal year-ends submit them outside the first quarter. The exclusion of 10-Q filings, however, constitutes a limitation of my study.

I obtain 10-K filings for the periods 2008–2012 (used to revisit the Baxter case) and 2018–2024 (used for the main analysis) from the dataset provided by Bill McDonald.[2] The documents in this dataset are already parsed according to the standards of Loughran and McDonald (2011) and are well suited for textual analysis.

I use monthly stock return data from CRSP, following the approach of the original authors. The data is manually downloaded from the WRDS platform.[3] Since CRSP identifies securities by ticker symbols, whereas SEC filings are indexed by Central Index Keys (CIKs), I use the official ticker-to-CIK mapping provided by the SEC[4] to link firms'

[2] `https://sraf.nd.edu/data/stage-one-10-x-parse-data/`

[3] `https://wrds-www.wharton.upenn.edu/pages/get-data/center-research-security-prices-crsp/annual-update/stock-security-files/monthly-stock-file/`

[4] Available at `https://www.sec.gov/include/ticker.txt`.

financial reports with their corresponding return data.[5]

Initially, my dataset includes 36,135 10-K filings. However, after matching these filings with stock return data from, the sample is reduced to 23,130 observations, as some firms do not have corresponding stock data available in CRSP.

Although the authors do not explicitly specify which rate they use as the risk-free return, I follow standard practice in the finance literature (e.g., Fama and French (1993)) and use the one-month U.S. Treasury bill rate. This data I obtain from FRED platform.[6] I compute excess returns by subtracting the risk-free rate from the monthly stock returns.

2.2 The Case of Baxter

I begin my analysis by revisiting the Baxter case study. I download the 10-K filings for Baxter from 2008 to 2012 and compute the Jaccard similarity between successive reports, following the methodology used in the original study. For the Jaccard measure, I convert each 10-K into a set of unique words and calculate similarity as the ratio of the intersection to the union of the word sets from consecutive years.

Surprisingly, I am unable to replicate the findings presented in Figure 2 of the original paper, which reports a notably low similarity between the 2009 and 2010 filings. In contrast, my results show relatively stable similarity scores: 0.6472 (2008–2009), 0.6499 (2009–2010), and 0.6933 (2010–2011).

I also fail to reproduce the keyword frequencies reported by the authors for the 2009 10-K filing. Notably, my counts for FDA, Recall, and Colleague Pump in the 2007 and 2008 reports exactly match those reported in the paper—33 and 28 for FDA, 16 and 20 for Recall, and 29 and 28 for Colleague Pump, respectively. However, I find substantial discrepancies for the 2009 filing: I count only 26 occurrences of FDA, 14 of Recall, and 27 of Colleague Pump, whereas the paper reports 48, 30, and 79, respectively. To ensure the robustness of my results, I manually retrieved the 2009 10-K filing from the EDGAR database and verified the word counts using a text editor. The manual counts confirmed my Python-based results.

2.3 Main Analysis

After loading the 10-K filings and merging them with stock price data as previously described, I compute textual similarity between consecutive filings. While the original study employs four similarity measures, I exclude minimum edit distance and simple similarity due to their high computational cost. Specifically, calculating Levenshtein distance for minimum edit distance and identifying additions, deletions, and substitutions for simple similarity both require significant processing time.

For Jaccard similarity, I follow the same method I described in the Baxter case study. For cosine similarity, I use the *CountVectorizer* function from the *sklearn.feature_extraction.text* library to convert each document into a vector of word counts. I then compute the cosine similarity between vectors using the *cosine_similarity* function from *sklearn.metrics.pairwise*.

After calculating the similarity scores, I assign each filing to a quintile based on its relative similarity within the same filing month and year, as described in Section 1. This classification allows me to distinguish between changers and non-changers.

For each 10-K filing, I extract the stock price data for the month following the filing date and calculate cumulative

[5] Additionally, I exclude all months in which the CRSP dataset reports a stock price of zero or a negative value. A negative price indicates that the closing price is unavailable and has been replaced with a bid/ask average, while a zero price suggests that neither a closing price nor a bid/ask average is available.

[6] `https://fred.stlouisfed.org/series/DGS1MO`

returns over a five-month horizon.[7] I also compute cumulative returns for the risk-free asset over the same period and use these to derive excess cumulative returns. These excess returns are then used to calculate the average performance of the Q1 to Q5 portfolios and to conduct a regression analysis assessing the relationship between report similarity and future stock returns.

2.3.1 Results and Discussion

Summary statistics for cosine and Jaccard similarities are presented in Figure 1. In contrast to the original study, I find higher average similarity scores—0.98 for cosine and 0.6798 for Jaccard—as well as lower standard deviations (0.0192 for cosine and 0.1077 for Jaccard). This may indicate that firms have become more consistent in their financial reporting, possibly in anticipation of increasingly sophisticated textual analysis by investors. Such behavior would align with the feedback effects in corporate disclosure in response to technological advancements, as discussed by Cao et al. (2023).

Despite the lower dispersion in similarity measures, I am able to reproduce the portfolio-level findings of the original study. Specifically, for cosine similarity, I find that the Q1 portfolio yields an average monthly return of 0.16%, while the Q5 portfolio earns 2.96%, resulting in a Q5–Q1 long-short return of approximately 2.79% per month. For Jaccard similarity, the Q1 and Q5 portfolios yield 3.3% and 3.5% average monthly returns, respectively, producing a Q5–Q1 spread of 0.28%, which is also consistent with the original study's findings. The average monthly returns for all quintiles, along with corresponding t-test results, are presented in Figures 2, 3 and 4.

As a robustness check, I examine excess cumulative returns over the five-month period following the month after each 10-K filing. For cosine similarity, Quintile 5 consistently outperforms Quintile 1 in each of the five months. In the case of Jaccard similarity, Q1 slightly outperforms Q5 in the first and fifth months; however, the overall trend remains consistent with the original findings. Detailed results are presented in Figures 5 and 6. In this analysis, I use excess cumulative returns rather than abnormal returns, as I was unable to obtain the data required to calculate risk-adjusted returns—such as book-to-market ratios—for implementing the Fama-French three- or five-factor models.

Due to similar data limitations, as well as my limited theoretical background on Fama-MacBeth regressions and their methodological complexity, I do not fully implement the regression component of the original study. Instead, I perform a simplified analysis by regressing average three-month excess returns on the similarity measures. As shown in Figures 7 and 8, both similarity measures are positively and statistically significantly associated with future stock returns.

2.4 Exclusion of Outliers

Given the relatively high long-short return observed for cosine similarity—2.79% compared to 0.34% in the original study—I conduct an additional robustness check by excluding stocks with exceptionally high excess returns over the three-month holding period. Specifically, I remove observations above the 99th percentile of the excess cumulative return distribution, which corresponds to a return of 233%, indicating that these stocks more than doubled in value within three months.

After excluding these extreme outliers, the long-short return for cosine similarity decreases to 0.55%, while it increases to 1.20% for Jaccard similarity. Importantly, even after this adjustment, the regression analysis continues to show a positive and statistically significant relationship between both similarity measures and future stock returns. Detailed results are presented in Figures 9, 10, 11 and 12.

[7] Although the main results are based on a three-month holding period, I extend the analysis to five months to test the robustness of the findings, following the approach of the original authors.

References

CAO, S., W. JIANG, B. YANG, AND A. L ZHANG (2023): "How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI," *The Review of Financial Studies*, 36, 3603–3642.

COHEN, L., C. MALLOY, AND Q. NGUYEN (2020): "Lazy Prices," *Journal of Finance*, 75, 1371–1415.

FAMA, E. F. AND K. R. FRENCH (1993): "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3–56.

LOUGHRAN, T. AND B. MCDONALD (2011): "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *Journal of Finance*, 66, 35–65.

```
count    20339.000000
mean         0.985971
std          0.019233
min          0.560317
25%          0.984945
50%          0.993176
75%          0.996199
max          0.999956
Name: cosine_sim, dtype: float64
----------------------------------------
count    20339.000000
mean         0.679752
std          0.107681
min          0.094046
25%          0.622482
50%          0.692267
75%          0.755680
max          0.977281
Name: jaccard_sim, dtype: float64
----------------------------------------
----------------------------------------
Correlation(cosine, jaccard):  0.6561514181061432
```

Figure 1: Summary Statistics of Similarity Measures

```
Monthly Average Excess Returns of the Portfolios:
Quintile Portfolios for Cosine Measure
------------------------------------------------------------
Quintile 1: 0.0017
Quintile 2: 0.0325
Quintile 3: 0.0438
Quintile 4: 0.0217
Quintile 5: 0.0296
------------------------------------------------------------
Q5 - Q1 Average Return Spread: 2.79%
```

Figure 2: Main Results—Calendar-Time Portfolio Returns: Cosine Similarity

```
Monthly Average Excess Returns of the Portfolios:
Quintile Portfolios for Jaccard Measure
------------------------------------------------------------
Quintile 1: 0.0331
Quintile 2: 0.0192
Quintile 3: 0.0237
Quintile 4: 0.0178
Quintile 5: 0.0360
------------------------------------------------------------
Q5 - Q1 Average Return Spread: 0.28%
```

Figure 3: Main Results—Calendar-Time Portfolio Returns: Jaccard Similarity

T-Test Results: Monthly Average Excess Returns
Quintile Portfolios Based on Cosine Similarity

Quintile	T-Stat	P-Value
1	0.358	0.7203
2	4.106	0.0000
3	3.449	0.0006
4	4.099	0.0000
5	2.886	0.0039

T-Test Results: Monthly Average Excess Returns
Quintile Portfolios Based on Jaccard Similarity

Quintile	T-Stat	P-Value
1	2.651	0.0081
2	2.996	0.0028
3	3.312	0.0009
4	4.152	0.0000
5	3.342	0.0008

Figure 4: Main Results—Calendar-Time Portfolio Returns: T-Tests

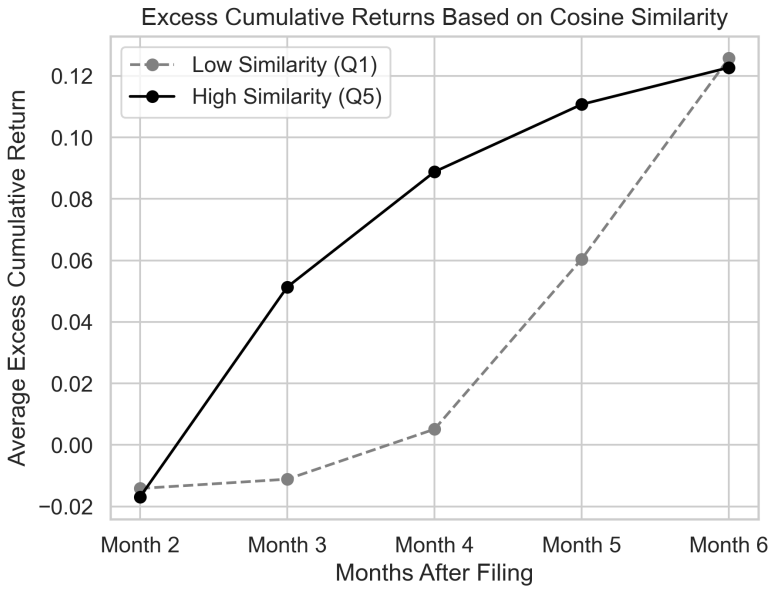

Figure 5: Monthly Cumulative Excess Return: Cosine Similarity

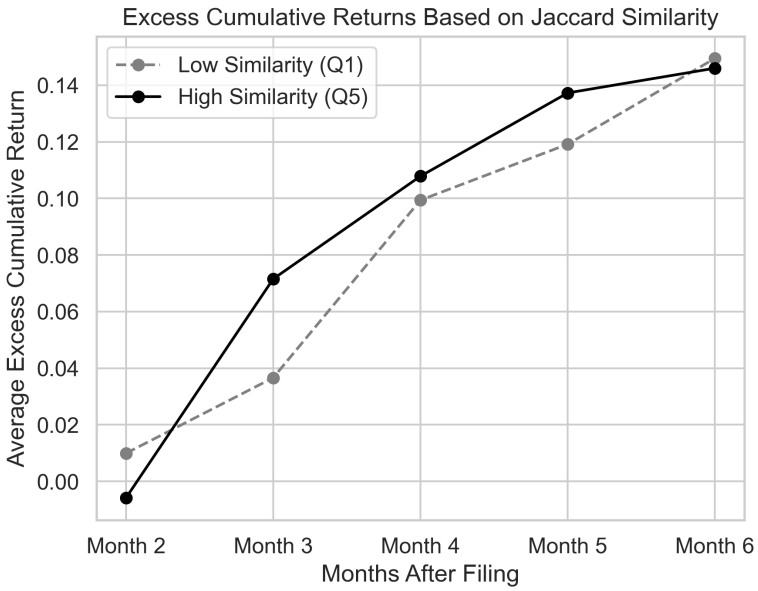

Figure 6: Monthly Cumulative Excess Return: Jaccard Similarity

OLS Regression Results

Dep. Variable:	Three-Months Excess Return (Monthly Average)	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	9.872
Date:	Sun, 18 May 2025	Prob (F-statistic):	0.00168
Time:	10:44:46	Log-Likelihood:	-15823.
No. Observations:	19014	AIC:	3.165e+04
Df Residuals:	19012	BIC:	3.166e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.6172	0.205	-3.015	0.003	-1.019	-0.216
cosine_sim	0.6523	0.208	3.142	0.002	0.245	1.059

Omnibus:	50651.060	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1699016276.735
Skew:	32.014	Prob(JB):	0.00
Kurtosis:	1466.026	Cond. No.	102.

Figure 7: Main Results—Regression Analysis: Cosine Similarity

OLS Regression Results

Dep. Variable:	Three-Months Excess Return (Monthly Average)	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	3.892
Date:	Sun, 18 May 2025	Prob (F-statistic):	0.0485
Time:	10:46:31	Log-Likelihood:	-15826.
No. Observations:	19014	AIC:	3.166e+04
Df Residuals:	19012	BIC:	3.167e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0254	0.026	-0.965	0.334	-0.077	0.026
jaccard_sim	0.0754	0.038	1.973	0.049	0.000	0.150

Omnibus:	50651.913	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1698493547.288
Skew:	32.016	Prob(JB):	0.00
Kurtosis:	1465.801	Cond. No.	13.9

Figure 8: Main Results—Regression Analysis: Jaccard Similarity

```
Monthly Average Excess Returns of the Portfolios (Outliers Excluded):
Quintile Portfolios for Cosine Measure
------------------------------------------------------------
Quintile 1: -0.0124
Quintile 2: -0.0113
Quintile 3: -0.0028
Quintile 4: -0.0069
Quintile 5: -0.0069
------------------------------------------------------------
Q5 - Q1 Average Return Spread: 0.55%
```

Figure 9: Portfolio Returns after Exclusion of Outliers: Cosine Similarity

```
Monthly Average Excess Returns of the Portfolios (Outliers Excluded):
Quintile Portfolios for Jaccard Measure
------------------------------------------------------------
Quintile 1: -0.0161
Quintile 2: -0.0105
Quintile 3: -0.0060
Quintile 4: -0.0037
Quintile 5: -0.0040
------------------------------------------------------------
Q5 - Q1 Average Return Spread: 1.20%
```

Figure 10: Portfolio Returns after Exclusion of Outliers: Jaccard Similarity

OLS Regression Results

Dep. Variable:	Three-Months Excess Return (Monthly Average)	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	21.54
Date:	Sun, 18 May 2025	Prob (F-statistic):	3.49e-06
Time:	10:56:31	Log-Likelihood:	16761.
No. Observations:	18823	AIC:	-3.352e+04
Df Residuals:	18821	BIC:	-3.350e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1785	0.037	-4.859	0.000	-0.250	-0.106
cosine_sim	0.1729	0.037	4.641	0.000	0.100	0.246

Omnibus:	7303.657	Durbin-Watson:	1.788
Prob(Omnibus):	0.000	Jarque-Bera (JB):	74129.308
Skew:	1.578	Prob(JB):	0.00
Kurtosis:	12.196	Cond. No.	101.

Figure 11: Regression Analysis after Exclusion of Outliers: Cosine Similarity

OLS Regression Results

Dep. Variable:	Three-Months Excess Return (Monthly Average)	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	49.01
Date:	Sun, 18 May 2025	Prob (F-statistic):	2.63e-12
Time:	10:57:22	Log-Likelihood:	-3904.7
No. Observations:	18823	AIC:	7813.
Df Residuals:	18821	BIC:	7829.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1220	0.014	-8.622	0.000	-0.150	-0.094
jaccard_sim	0.1439	0.021	7.001	0.000	0.104	0.184

Omnibus:	7396.944	Durbin-Watson:	1.792
Prob(Omnibus):	0.000	Jarque-Bera (JB):	75253.668
Skew:	1.602	Prob(JB):	0.00
Kurtosis:	12.256	Cond. No.	13.9

Figure 12: Regression Analysis after Exclusion of Outliers: Jaccard Similarity