Beyond Cosine Similarity: Predicting Stock Declines from Financial Disclosures with LLM Sentiment and Deep Learning

Anonymous EMNLP submission

Abstract

This paper analyzes quarterly SEC corporate disclosures for S&P 500 companies from January 2000 to December 2019 demonstrating how large language models (LLMs) and Concatenated Deep Learning are able to detect which companies under perform. This research finds that by comparing two quarterly corporate disclosures combined with the reasoning capabilities of the Claude2 large language model, negative excess returns of -11% over a 180 day period (-22% annualized) can be avoided. The paper introduces two novel approaches: (A) Concatenating Deep Learning architectures comparing quarterly filings, and (B) Summarization methods using Claude2 to extract sentiment signals related to major business risks, profitability, legal, market pressures, etc. Together, these techniques demonstrate new ways of expanding beyond rudimentary natural language processing approaches, such as lexicons and cosine similarity, to answer fundamental questions related to firm performance.

1 Introduction and Related Work

Cohen et al. (2020) is perhaps the most prominent research that has shown that textual changes in corporate disclosures (measured by cosine similarity in 10-Qs) can be predictive of stock returns, particularly if changes reflect underlying changes in risk or operational performance. However, such similarity-based methods do not capture deeper semantic meanings or hidden cues about the tone of the management.

The primary contribution of this paper is to combine (1) advanced deep learning architectures and (2) large language model summarization to detect changes in management's discussion and analysis (MD&A) sections that may indicate future firm performance. We propose novel deep learning architectures to concatenate from the 10-10-10-Q of the current quarter and the 10-Q of the previous quarter, feeding RNN or max embedding pipelines, then classifying whether the firm's excess(note that excess is defined as the individual stock return minus the overall market return, i.e. S&P 500 in this case) will be positive or negative over the next 90 days. In parallel, we apply generative AI using Anthropic Claude2 to produce condensed summaries that highlight business risks, profitability, and other changes, assigning a sentiment strongly correlated with future negative returns.

In contrast, the traditional approach of using cosine similarity scores alone can miss subtle shifts in context, tone, or emphasis. With modern deep learning, CNN and RNN-based architectures can process text more holistically, learning local and sequential patterns, respectively (Goldberg, 2016). Recent LLMs such as ChatGPT (Schulman et al., 2023) or Claude2 (Bai et al., 2022) can reason about entire documents, identify meaningful changes, and produce high-level summaries. As Cao et al. (2022) shows, companies may even attempt to obfuscate negativity in disclosures to game naive natural language processing techniques. This indicates the importance of advanced semantic methods or generative AI to detect deeper signals.

Overall, the empirical results show that these deep learning approaches outperform the naive cosine similarity in detecting negative future returns by a factor of 3% per year. Moreover, Claude2 summarization reveals that negative sentiment signals after major business changes can be associated with a -11% average excess return over 180 days, indicating a powerful ability to avoid underperforming firms. In general, the study demonstrates that deep learning and LLM can predict when companies perform poorly using corporate disclosure data that simpler methods may miss.

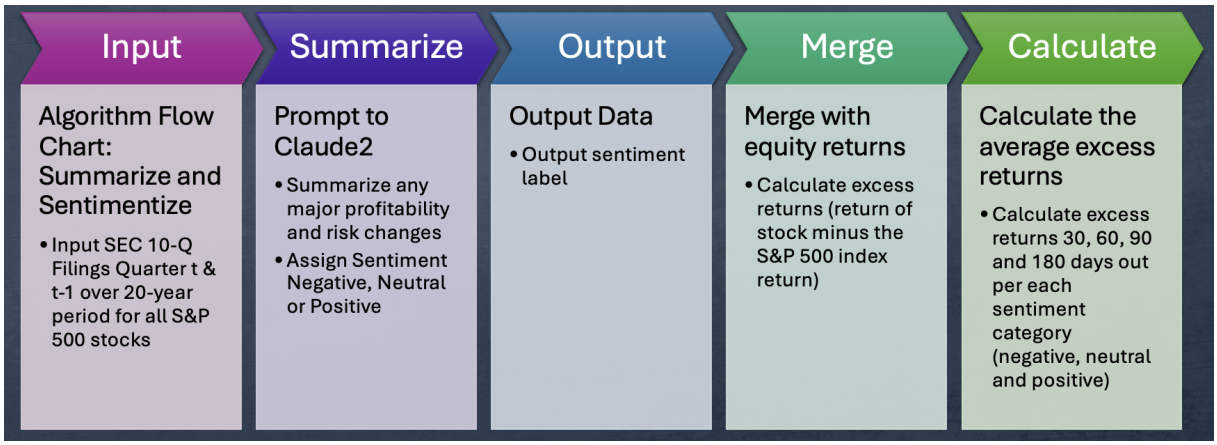

Figure 1: Algorithm Flow Chart for Summarize and Sentimentize

2 Data

2.1 Data Collection

This research collects SEC 10-Q filings for firms in the S&P 500 from January 2000 until December 2019, ensuring historical index compositions are considered to reduce survivor bias. The raw text data was sourced from Wharton Research Data Services (WRDS), which provides structured tables, Central Index Keys (CIKs), filing publication dates, and 10-Q texts. The price data for these firms was also obtained from WRDS, adjusting for splits and generating excess returns by subtracting the performance of the S&P 500 index.

In total, we obtained 28,669 10-Q documents spanning the sample. After removing corrupt or empty files (file size zero), adjusting for extreme outliers (e.g., unusually long or short MD&A text), and filtering for realistically parsed Management Discussion and Analysis sections, the final dataset had 22,002 records. We used 17,863 for training and validation (Jan 2000 to Dec 2012) and 4,139 for out-of-sample testing (Jan 2013 to Dec 2019).

2.2 Management's Discussion and Analysis Section

Each 10-Q can contain numerous boilerplate sections, such as "Controls and Procedures" or disclaimers on forward-looking statements. We focus on the "Management's Discussion and Analysis of Financial Condition and Results of Operations" (MD&A), where managers discuss performance drivers, risks, accounting changes, and critical assumptions. This narrative should capture material changes from quarter to quarter.

3 Concatenation Methodology Overview

Merity (2016) introduced a model that sums or takes the maximum over GloVe word embeddings for each text, then merges them and feeds into fully connected layers. We adopt a similar approach, extending it by concatenating two separate neural pipelines (one for the current quarter 10-Q, one for the previous quarter 10-Q).

We want to use the current 10-Q compared against the previous quarters 10-Q to predict whether the excess return of a stock will be positive or negative over a 90-day horizon. Defined as follows:

$$f(\text{10-Q}_t, \text{10-Q}_{t-1}) \rightarrow \begin{cases} 0 & \text{(negative return)} \\ 1 & \text{(positive return)} \end{cases}$$

Where we use utilize Cosine Similarity via a Logistic Regression as a benchmark and the following Deep Learning Concatenation architectures: Max of Embeddings, CNN Concatenation and Bidirectional LSTM Concatenation.

For Cosine Similarity Logistic Regression, we simply take the cosine similarity score between quarter t and $t-1$ and feed it into a logistic regression classifier that predicts positive/negative future returns. This is the simplest baseline, akin to earlier work by Cohen et al. (2020).

For Max of Embeddings, mirroring the Quora question matching architecture (Merity, 2016) to 10-Q pairs. We embed each MD&A using pretrained GloVe vectors, apply a dimension-wise max pooling across tokens, then merge the two quarter embeddings via concatenation. A deep MLP with

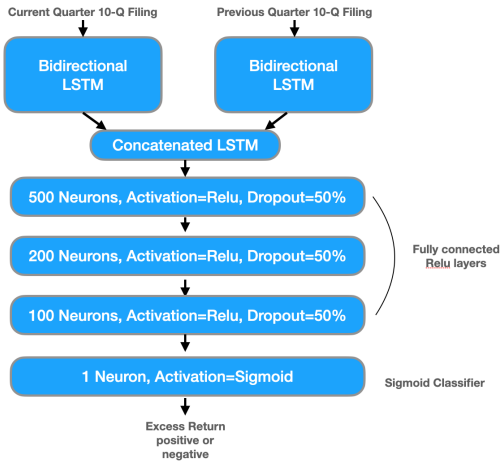

Figure 2: Concatenated BiLSTM architecture for current vs. previous 10-Q.

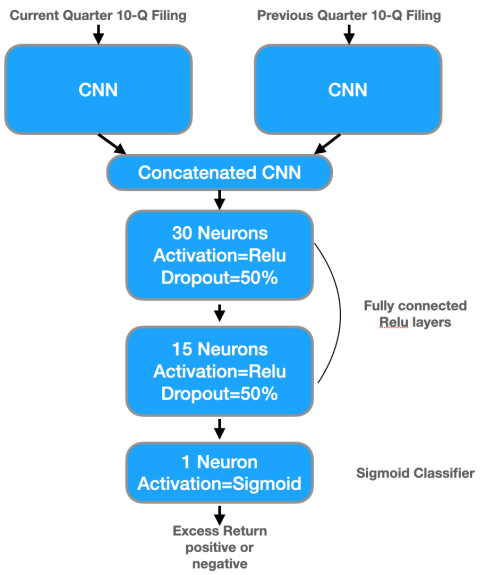

Figure 3: Concatenated CNN architecture for current vs. previous 10-Q.

four 200-neuron layers predicts a binary label (positive/negative).

The Concatenated CNN approach, we build a CNN for each quarter's MD&A, each with an embedding layer (300-dim GloVe), a 1D convolution, and max pooling. The outputs flatten, then are concatenated and pass through two fully connected layers before a final sigmoid output. See Figure 3. Dropout of 50% is used, with Adam for optimization and binary cross-entropy loss.

Similarly, we build a BiLSTM subnetwork for each 10-Q, each embedding up to 500 tokens (to mitigate gradient issues). Outputs are concatenated, feeding fully connected layers. Figure 2 shows the architecture. Again, we use 50% dropout, binary cross-entropy, and Adam.

4 Using Claude2 for reasoning

We use Claude2 (Bai et al., 2022) for reasoning as at the time it could process up to 100k tokens context window in its API. Due to its large context window, Claude2 can accommodate entire MD&A sections (median 10k words) in one pass, making it well suited for analyzing 10-Q pairs.

4.1 Zero-shot LLM 10-Q Summaries and Sentiment Analysis

4.1.1 Methodology

Instead of just measuring textual overlap, we prompt Claude2 to read both quarters' MD&A, summarize major changes regarding business risks, profitability, legal and market pressures, and then assign a sentiment tag (*positive, neutral, negative*). Algorithm 1 details how we supply the prompt and parse its output. This "signal compression" step harnesses Claude2's generative reasoning to highlight the core differences in narrative that might be relevant to future performance. We then examine the subsequent excess returns of each label.

Algorithm 1 Signal Compression with Sentiment

Input: 10-Q text of quarter t and t-1

1: **Prompt to Claude2:**
"Please respond with one word [0..1] indicating change magnitude (0 = max, 1 = none). Then summarize in 3-4 sentences any significant changes impacting underlying business profitability. If no major changes, say 'no changes'. Then label sentiment as 'positive', 'neutral', or 'negative' in one word."
2: Output ← Claude2(Prompt)
3: Parse Output → {change score, summary text, sentiment}
4: **return** {sentiment}

5 Results

5.1 Model Results

5.1.1 Concatenation Deep Learning Models

Table 1 compares the F1 scores for predicting negative or positive 90-day returns. All concate-

nated deep learning models outperform the cosine-similarity logistic benchmark in capturing negative vs. positive classes.

Model	F1 (Neg)	F1 (Pos)
Cosine Sim + Logistic	0.31	0.63
CNN Concatenation	0.48	0.53
Max of GloVe Emb.	0.49	0.54
LSTM Concatenation	0.48	0.53

Table 1: F1 Scores by Class (Test Set)

5.2 Excess Return Results

5.2.1 Concatenation Deep Learning

Table 2 shows average 180-day excess returns for each predicted label. Cosine similarity logistic incorrectly assigns a negative label to instances that yield strong negative returns. Meanwhile, CNN and max-embedding approaches show better discrimination, avoiding the large losses.

Model	Neg	Neutral	Pos
Cosine + LogReg	0.03	0.01	-0.074
CNN Concat	-0.053	0.012	0.003
Max Glove Emb	-0.037	-0.004	0.047
LSTM Concat	-0.03	-0.03	-0.002

Table 2: Avg 180-Day Excess Returns by Predicted Label

5.2.2 Zero-shot LLM 10-Q Summaries and Sentiment

Tables 3 and 4 present out-of-sample average (and median) excess returns over horizons from 30 to 180 days, grouped by the sentiment assigned by Claude2. Notably, "negative" sentiment leads to strongly negative average performance (-5% to -12%) over 180 days, i.e., -22% annualized. "Positive" sentiment yields small positive returns. These results suggest that LLM-based summarization can effectively flag downward risk.

5.2.3 Bringing it all together

Table 5 consolidates the models over a 180-day horizon. The zero-shot LLM approach yields the clearest signal, with a -11% average return for negative-labeled disclosures. By contrast, the concatenated CNN or max-embedding approach is more balanced in capturing both positive and negative sides. In practice, an investment strategy might combine them.

Sentiment	30d	60d	90d	180d
Negative	-0.064	-0.097	-0.088	-0.119
Neutral	0.000	-0.010	0.030	0.015
Positive	0.014	0.020	0.011	0.008

Table 3: Zero-Shot LLM Summaries: Avg Excess Returns by Sentiment

Sentiment	30d	60d	90d	180d
Negative	-0.041	-0.046	-0.018	-0.051
Neutral	0.003	-0.004	0.007	0.005
Positive	0.020	0.024	0.030	0.022

Table 4: Zero-Shot LLM Summaries: Median Excess Returns by Sentiment

Model	Neg	Neutral	Pos
Cosine + LogReg	0.030	0.010	-0.074
CNN Concat	-0.053	0.012	0.003
Max GloVe Emb	-0.037	-0.004	0.047
LSTM Concat	-0.030	-0.030	-0.002
LLM Zero-Shot	-0.220	0.030	0.018

Table 5: Annualized Avg 180-Day Excess Returns Across Approaches

6 Conclusion

This research demonstrates how advanced NLP techniques—concatenated deep learning architectures and LLM summarization—can uncover subtle signals in corporate 10-Q disclosures. We move beyond naive similarity or dictionary-based counts to actual semantic reasoning about business changes and profitability. Experimental results show that large language models like Claude2, when asked to identify sentiment around material changes, can avoid large negative returns. Meanwhile, concatenated CNN or LSTM networks also outperform simple cosine similarity in classification tasks.

7 Limitations

While our approach demonstrates strong predictive performance in identifying negative sentiment in SEC 10-Q filings, several limitations must be acknowledged. First, generalization to future market conditions remains uncertain, as financial markets are dynamic, and shifts in regulatory policies, macroeconomic conditions, or firm-specific strategies may impact the effectiveness of our model. Second, our LLM-based summarization approach relies on Claude2, which, while powerful, may introduce biases or inconsistencies in sentiment classification, particularly if management deliberately obfuscates negative disclosures. Finally, while deep learning architectures such as CNNs and BiLSTMs improve upon naive cosine similarity, they remain limited in interpretability, making it challenging to directly attribute predictions to specific textual features. Future work should explore more robust explainability techniques and test the approach in real-time financial decision-making contexts.

References

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. 2022. A general language assistant as a laboratory for alignment. Anthropic.

S. Cao, W. Jiang, B. Yang, and A. L. Zhang. 2022. How to talk when a machine is listening?: Corporate disclosure in the age of ai. Working paper at NATIONAL BUREAU OF ECONOMIC RESEARCH.

L. Cohen, C. Malloy, and Q. Nguyen. 2020. Lazy prices. The Journal of Finance, vol 75(3):1371–1415.

Y. Goldberg. 2016. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research.

S. Merity. 2016. Keras SNLI baseline example.

J. Schulman, B. Zoph, C. Kim, J. Hilton, J. Menick, J. Weng, J. F. Ceron Uribe, and et al. 2023. Introducing ChatGPT. Open AI Blog.