Trading based on earnings call sentiments: An equal-weighted decile portfolio approach

Ulrik T. Sjøli, Mert Ülgüner, Zhuofu Zhou

FIN-407, EPFL, June 2024

1 Introduction

In this report, we aim to answer the following research question:

Can one elaborate a trading strategy going long the stocks with the most positive earnings calls and short the ones with the most negative?

We first download a sample of earnings call transcripts from between 2017 and 2023. After preprocessing these, we assign a set of positive-negative sentiment scores to each earnings call transcript. To investigate the relationship between returns and sentiment scores, we then, for each month, sort the transcripts by sentiment score to build equally weighted decile portfolios that are held for the following month. Based on these, we describe a long-short trading strategy and investigate its summary statistics, and also interpret the factors driving the sentiment scores by considering word counts and topic analysis.

1.1 Literature review

We build our study on research that uses textual analysis to gauge sentiment in financial data. Loughran and McDonald (2010) [4] extend previous work to create word (or unigram) lists (also called dictionaries) that classify certain words as positive, which we call LM-pos-unigrams, or negative (LM-neg-unigrams) based on

manually selecting words reaching certain criteria in 10-K reports between 1994 to 2008. To avoid endogenizing sentiment and future returns, the authors do not create short sets of tonal words which managers can simply avoid using, instead opting to construct exhaustive word lists. Importantly for our analysis on earnings calls, they argue that the LM-neg-unigrams list likely is generalizable to other financial documents.

Druz, Petzev, Wagner and Zeckhauser (2019) [1] use the Loughran-McDonald dictionaries to count $negative\ words_{jt}$ and $positive\ words_{jt}$ for a firm j's earnings call held in quarter t to calculate a sentiment score

$$negativity_{jt} = \frac{negative\ words_{jt} - positive\ words_{jt}}{negative\ words_{jt} + positive\ words_{jt} + 1}.$$

They show that bleak quarter-to-quarter ($negativity_{jt+1} - negativity_{jt}$) changes in sentiment significantly predict downturns in future earnings, and that both analysts and the market underestimate by how much. This result holds after controlling for information in the earnings press release and other factors. They estimate Fama-MacBeth regressions with monthly return as response, and earnings surprise and negativity change as added predictors, and state that "calendar-time strategies [...] reap significant profits". To improve interpretability in the relationship between $sentiment_{jt}$ and firm j's next month cumulative return, our method does not consider the change in sentiment for firm j from t to $t+1$ and it does not control for firm- or market-specific factors. The findings by Druz et al. do, however, indicate that given sufficient methods for sentiment scoring, we might find a trading strategy affirming the research question.

Garcia, Hu and Rohrer (2022) [2] construct alternative sentiment dictionaries using a technique they call robust multinomial inverse regression. Using text data from 10-Ks and earnings calls, they identify unigrams and bigrams (combinations of two words) that are correlated with stock price reactions. We call the dictionaries they find GHR-pos/neg-bigrams and GHR-pos/neg-unigrams. In contrast to the Loughran-McDonald (LM) dictionaries, which are selected using psychological and financial domain knowledge, the "GHR" dictionaries are selected to correlate with returns. Using TF-IDF weighting, the authors find that their dictionaries signifi-

cantly outperform the LM dictionaries in out-of-sample return prediction. We will use the LM and GHR dictionaries to construct our sentiment scores in section 2.

Inspiring our part on topic analysis is the the paper "A Greenwashing Index" by Gourier and Mathurin (2024) [3]. The authors quantify greenwashing by constructing an index from Wall Street Journal articles from 1986 to 2022. To better understand the nature of greenwashing over time, the authors perform topic analysis by running Latent Dirichlet Allocation (LDA) to investigate the topics contained in the articles.

1.2 Dataset and preprocessing

To conduct our analysis we download the "Scraped Motley Fool Earnings Call Transcripts" dataset from Kaggle [1]. The raw data consists of 18755 earnings call transcripts with different dates and stock tickers from 2017 to 2023. In total, data on 2876 different stocks is included. 99.2% of the transcripts are public firms listed on either the NYSE or the NASDAQ. We remove duplicate rows and remove data points with missing date values. Then we merge the dataset with the corresponding returns extracted from WRDS, only keeping the rows where stock returns exist for a sufficient period after the earnings call is held. Last, for simplicity we drop the time of day for each transcript, only considering the date. In the end, we are left with a dataset consisting of 15804 transcripts and their respective dates and tickers, as well as a time series of daily returns for each ticker for the period after the call.

1.3 Exploratory data analysis

In this part of the report we report some results of our initial data analysis before we construct the sentiment scores in section 2.

1.3.1 Frequency of transcripts

Figure 1 shows that the dataset contains more transcripts in 2021 compared to the other years. We speculate that this is due to data collection methods at Motley Fool or choices made during scraping. When calculating the frequency of the words

[1] https://www.kaggle.com/datasets/tpotterer/motley-fool-scraped-earnings-call-transcripts

in the Loughran-McDonald dictionary in section 1.3.2, we must therefore normalize the frequency of the desired words by considering the amount of monthly transcripts. These large discrepancies in the monthly transcript count, which we take into consideration in section 3.1 when constructing our monthly-rebalanced trading strategy, suggest to us that we should not trade in some months. For 2017 and 2018, there are very few transcripts, so we do not trade in those years.

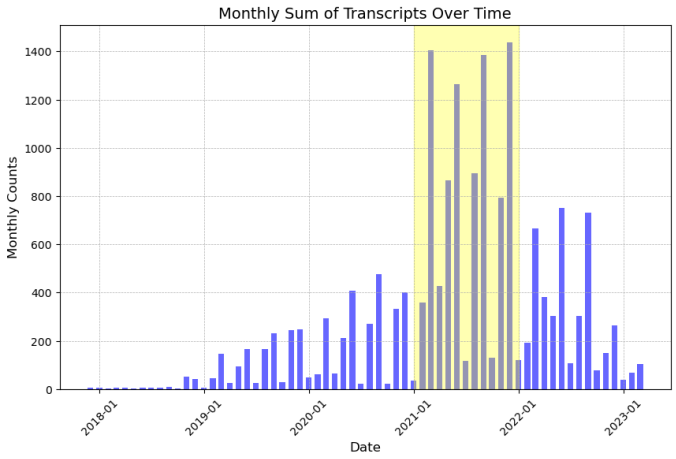


Figure 1: Monthly frequency of transcripts. 2021 highlighted in yellow.

1.3.2 The relative frequency of positive and negative words

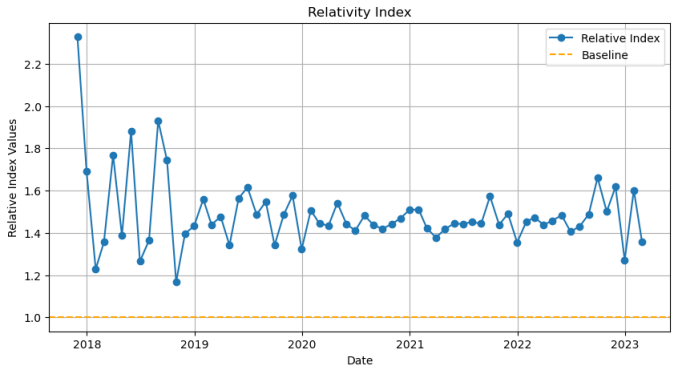


Figure 2: Monthly mean relative frequency of positive and negative words.

First we calculate the frequency of each word in the Loughran-Mcdonald dictionary over the whole corpus of transcripts, to find that there are 1047338 LM-positive words and 776621 LM-negative words. To get an overview of the overall tonality, we plot the normalized mean *positive count/negative count* ratio for each month

in figure 2. For the years with the fewest data points, the ratio fluctuates excessively, another reason to limit trading in 2017 and 2018. We note that there are weak signs of the ratio dropping leading up to and during crisis periods such as the beginning of the coronavirus pandemic and the Russo-Ukrainian war, but overall the plot is inconclusive. The crude ratio measure nevertheless indicates that there are noticeable fluctuations in monthly sentiment, though we seemingly need more elaborate measures of sentiment to draw any concrete inferences on its relationship with market return.

1.3.3 Relationship between word frequency and market returns

In this section we select the top three positive and top three negative words from the Loughran-McDonald dictionary with very high frequency in the transcripts, and analyze their relationship of normalized word frequencies with monthly market returns. With the above caveat regarding 2017 and 2018, this helps determine the drivers of our sentiment scores. These are based on frequencies across transcripts with respect to selected dictionaries. The "BoW" score we construct in section 2.1 is directly related to the word counts in this section, as it uses the same machinery of simply counting LM words.

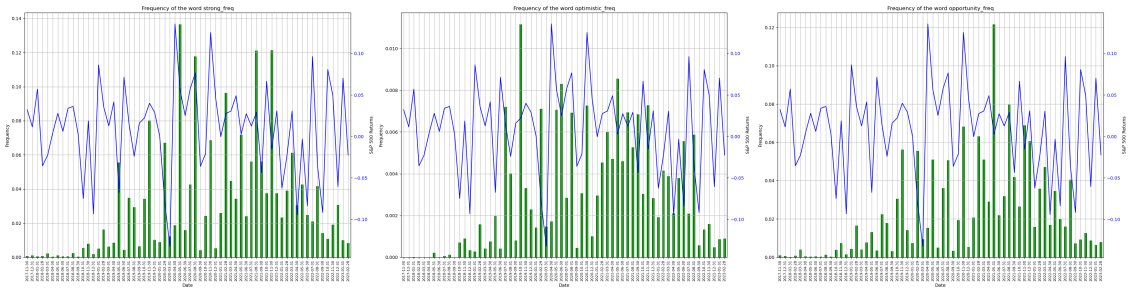


Figure 3: Normalized frequency of the positive LM words strong, optimistic and opportunity over time, with monthly market return superimposed.

In figure 3, we show the normalized word frequencies over time for "strong", "optimistic" and "opportunity" together with the monthly market return. An underlying assumption is that earnings call word choice is stationary (i.e. not changing over time), only reflecting market conditions, which we attempt to justify in the following. Considering the leftmost plot, we observe that there is a spike in re-

turns after the coronavirus crash of February 2020. During the month following this increase, we see an increase in the frequency of "strong". For "optimistic", the frequency somewhat decreases after the crash, before recovering in 2021, only to decrease again in the lead-up to the Russo-Ukrainian war. For "opportunity" there is no discernible trend, showing that these single-word measures probably are weak estimators of market sentiment.

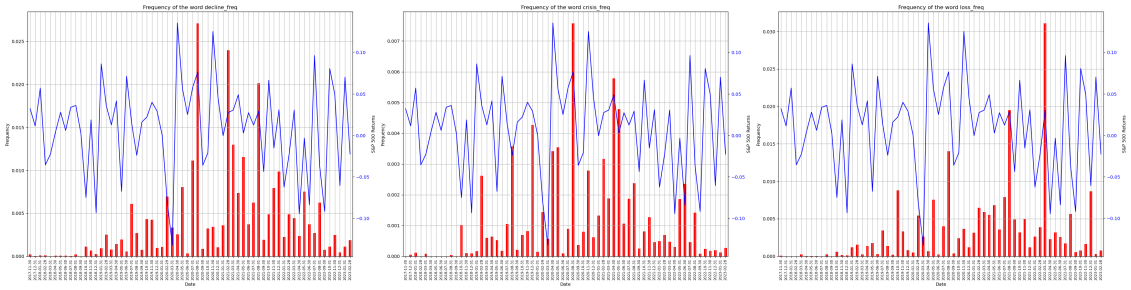


Figure 4: Normalized frequency of the negative LM words decline, crisis and loss over time, with monthly market return superimposed.

In figure 4, we perform the same analysis for negative words. We observe that the frequency of "decline" tends to increase after weak monthly returns. As expected, there is a weak increase in the frequency of "crisis" during the early coronavirus pandemic, tapering off in the latter half of 2021. There is a large increase the frequency of "loss" right after the start of the Russo-Ukrainian war, when returns are somewhat positive, though its frequency seems to lag downturns in return throughout the rest of the period. This shows another limitation of relying on counting single words: They may also be used in a many other contexts.

Summing up, we can see that the majority of the selected words' frequencies are related to the changes in monthly market returns. The above analysis suggests that there is often no simple causal relationship between previous return and word frequency, but that there sometimes is weak correlation between market return and word choice.

In figure 5, we show the Pearson correlation coefficient of selected positive ("gain", "opportunity", "strong", "improvement", "progress", and "optimistic") and negative words ("crisis", "challenge", "decline", "difficult", "loss", and "slowdown") with the monthly market return to investigate whether they co-move. We

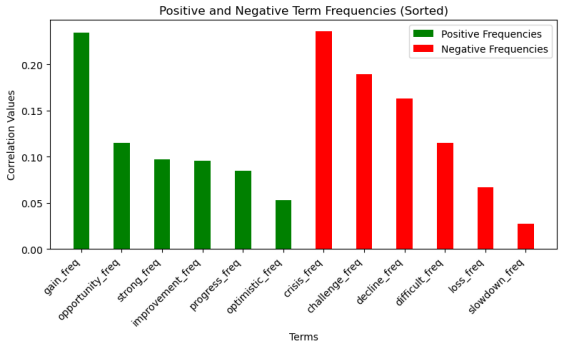


Figure 5: Correlation between selected word frequencies and monthly market returns.

observe that the correlation levels for negative words is higher compared to the levels for positive words. For this limited sample, the negative word frequency explain the monthly return change more accurately than positive word frequency. That negative sentiment is a stronger indicator is supported by Tetlock (2007) [5] who states that negative sentiments are noisy predictors for downward market pressure. Between 2017 and 2023 we identify two such time frames, the early coronavirus pandemic and the the economic tightening starting at the end of 2021. This is a supporting factor which might explain some of the high correlation levels for the negative word frequencies. In addition, Loughran and McDonald (2010) [4] argue that many positive words are easily compromised by managers choosing to qualify positive words instead of using easily detectable negative words when communicating, which also intuitively leads to lower correlation.

2 Sentiment scores

In this section we develop the sentiment scores we use for our trading strategy. For each scoring method, we show the (normalized) score distribution and plot a time series of mean monthly scores superimposed on the monthly market return during the period of interest.

2.1 Bag-of-words

We first preprocess each transcript by tokenizing and lowercasing. Then, we follow the method in Druz et al. (2019), and calculate a score

$$sentiment_j = \frac{positive\ score_j - negative\ score_j}{positive\ score_j + negative\ score_j + 1} \tag{1}$$

for an earnings call j. Here, we let $positive\ score_j$ be the count of Loughran-Mcdonald positive unigrams contained in each transcript, and $negative\ score_j$ be the count of negative unigrams. Like Druz et al., we correct for negation by not counting positive words if they are preceeded by a negation word ("no", "not", "none", "neither", "never", "nobody") or end with "n't". By construction, each score is in the interval [-1, 1]. To be consistent with the methods below, however, we use a scikit-learn MinMaxScaler with range [-1, 1] to normalize the entire set of scores, after computing each transcript's $sentiment_j$ score. The distribution of bag-of-words (BoW) scores and the mean sentiment score time series is given in figure 6.

The time series shows that the BoW-score captures the coronavirus crash well, but the scores otherwise seem fairly uncorrelated with the market return. The score distribution is centered around 0.25, showing that the BoW relative sentiment is positive.

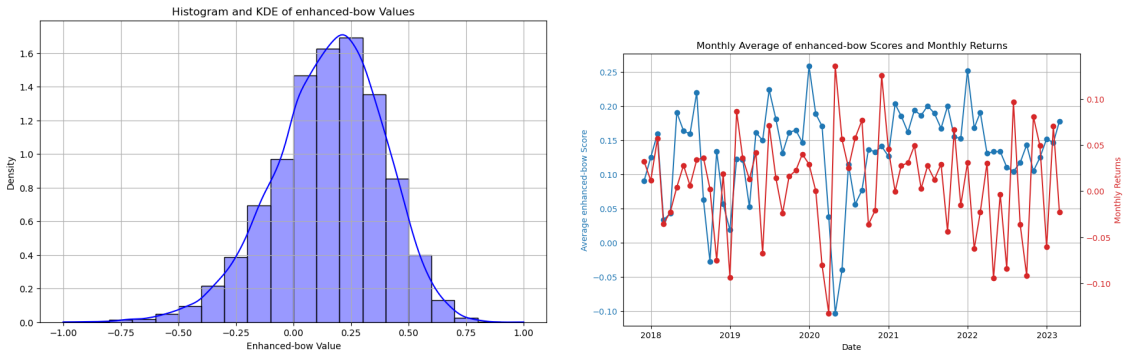


Figure 6: Distribution of BoW scores and time series of mean monthly sentiment.

2.2 VADER

VADER is calibrated to analyze social media post positive-negative sentiment, but we are interested in comparing its performance in portfolio selection to standard methods used in finance literature, such as the bag-of-words approach using the LM dictionary. For each transcript, VADER outputs a positive score, a negative score

and a composite score. Instead of using the composite score, for consistency we calculate the sentiment using (1), letting $positive\ score_j$ be the VADER positive score of earnings call j. Finally, we normalize the set of scores as in subsection 2.1. The distribution of scores and time series are shown in figure 7. The time series show that VADER tracks the coronavirus crash well and the distribution that the overall sentiment is relatively positive. Importantly, the VADER and BoW (figure 6) scores closely co-move, showing that they pick up on similar sentiment cues in the earnings calls. One difference is that the VADER scores are less volatile, which might explain their relative outperformance in figure 12.

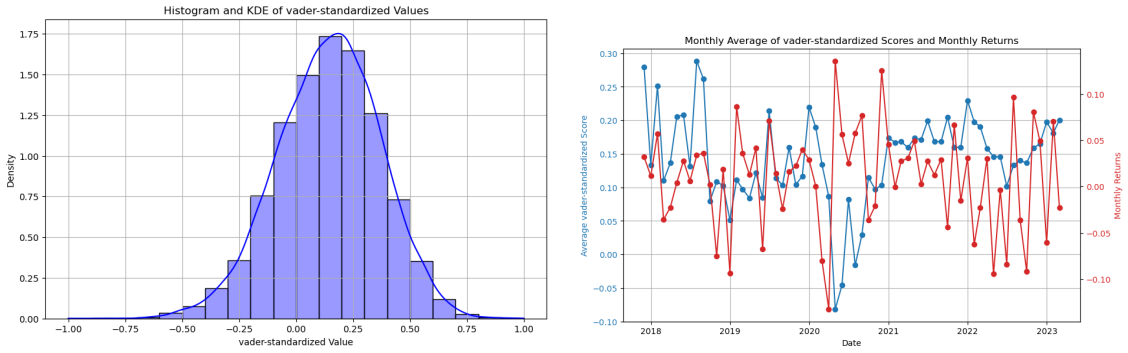


Figure 7: Distribution of VADER scores and time series of mean monthly sentiment.

2.3 GHR-bigram

Following Garcia et al. (2022), we let our list of dictionaries be

$$D = (GHR{-}pos/neg{-}bigrams,\ GHR{-}pos/neg{-}unigrams,\ LM{-}pos/neg{-}unigrams),$$

and decompose each D_i into positive and negative sets of n-gram, $D_{i,\,pos}$ and $D_{i,\,neg}$, for $i \in \{1, 2, 3\}$. Then, let V^{pos} be a set of positive n-grams and V^{neg} be a set of negative n-grams from some union U of the elements D_i in D.

For U and each $k \in \{pos,\ neg\}$ separately, we construct a document-term matrix (DTM) for the entire corpus of earnings call transcripts using a scikit-learn TF-IDF-vectorizer with vocabulary V^k. Each column of the DTM earnings call j represented as a sparse vector

$$tf_j = [tf_{1j}, ..., tf_{pj}],$$

where p is the length of V^k. Next, we calculate the transcript's sentiment score,

$$S_j^k = \sum_{i \in V^k} \frac{tf_{ij}}{N_j},$$

where N_j is the the length of the transcript. We calculate its composite sentiment score using (1), setting $positive\ score_j = S_j^{pos}$ and $negative\ score_j = S_j^{neg}$. Finally, we normalize the scores as in subsection 2.1.

We apply the above to D_1, setting $V^{pos} = D_{1,\,pos}$, the positive GHR bigrams and $V^{neg} = D_{1,\,neg}$, the negative GHR bigrams, which yields the GHR-bigram scores shown in figure 8.

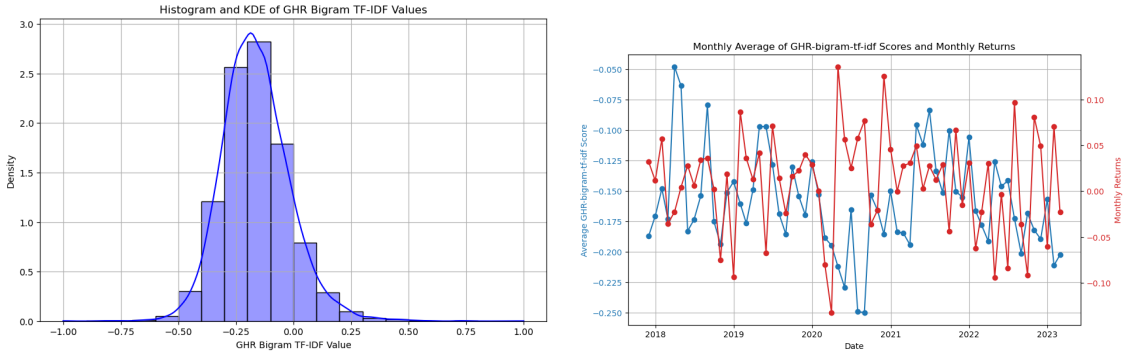


Figure 8: Distribution of GHR-bigram scores and time series of mean monthly sentiment.

2.4 GHR-bigram-unigram

We apply the method developed in section 2.3, finding $positive\ score_j$ by setting $V^{pos} = (D_{1,\,pos} \cup D_{2,\,pos})$ and $negative\ score_j$ by setting $V^{neg} = (D_{1,\,neg} \cup D_{2,\,neg})$, where $(a \cup b)$ is the union of the n-gram lists a and b. That is, the DTM rows are the union of GHR bigrams and unigrams. The resulting sentiment score distribution and time series is shown in figure 9.

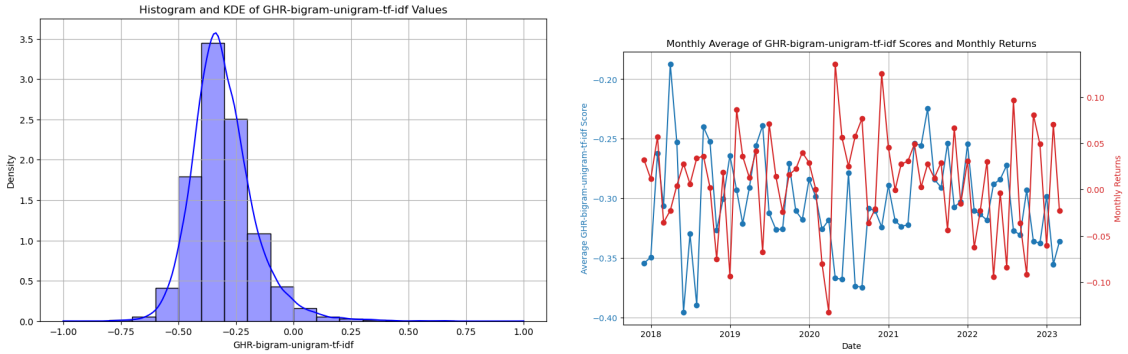


Figure 9: Distribution of GHR-bigram-unigram scores and time series of mean monthly sentiment.

2.5 GHR-bigram-LM-unigram

We replace the GHR-unigrams with LM-unigrams ($i = 3$) and follow the method outlined in subsection 2.4. The results are shown in figure 10.

As they are similar, we interpret the TF-IDF scores from this section and 2.3 and 2.4 together here. Their distributions are narrower than the ones in figures 6 and 7, indicating less variance within the scores. For all the TF-IDF scores, the mean sentiment is relatively negative, indicating that there are some positive transcripts in the right tail, making the normalization "squish" the rest of the transcripts together. Finally, the time series plot shows greater correlation between the TF-IDF scores and the market returns in the period of interest between 2019 and 2023.

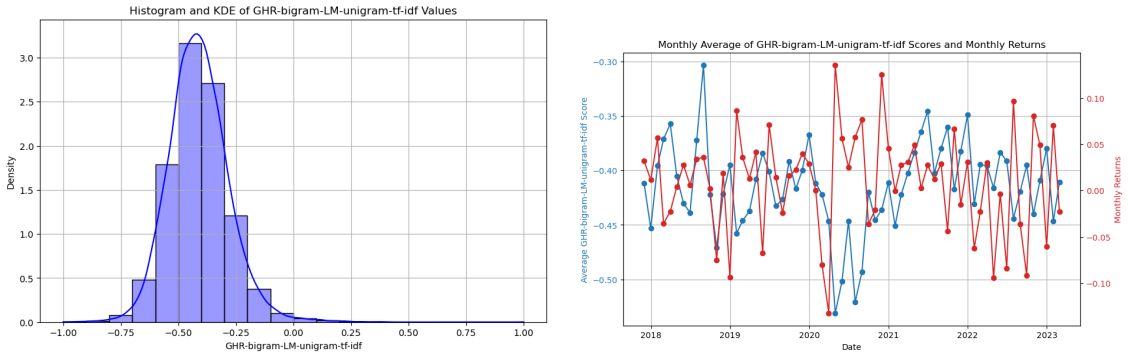


Figure 10: Distribution of GHR-bigram-LM-unigram scores and time series of mean monthly sentiment.

2.6 Correlation heatmap of sentiment scores

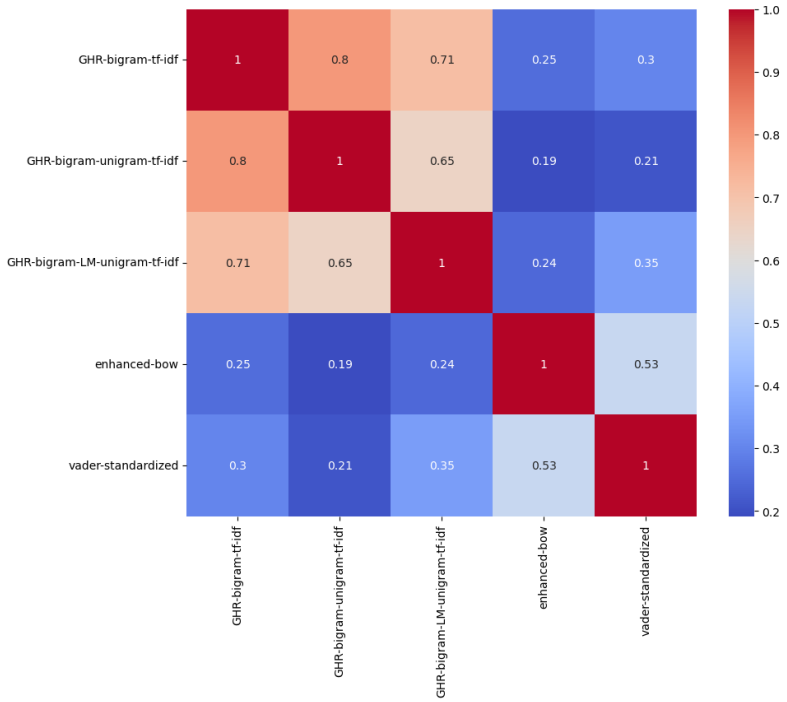


Figure 11: Correlation levels between the different sentiment scores.

In figure 11, we include a heatmap of the correlation levels between the sentiment scores. We confirm the conjecture in section 2.5 that the TF-IDF sentiment scores have high Pearson correlation, which is expected as their calculation methodologies are similar. In addition, we observe a reasonably strong correlation between the VADER and BoW unigram approaches. The plot explain the potential similarities between the trading strategies, and suggests that they will have different cumulative return patterns.

2.7 Sentiment score caveats

For each earnings call transcript, we now have a set of sentiment scores (BoW, VADER, GHR-bigram, GHR-bigram-unigram, GHR-bigram-LM-unigram). Before constructing the sentiment score decile portfolios in section 3, we comment on some shortfalls of the scoring methods that are detailed above, as there are two problems that (potentially) introduce look-ahead bias.

First, the MinMaxScaler we use to normalize the different scoring methods

makes the most positive transcript in the whole corpus of earnings call transcripts have score 1, and the most negative one have score -1. All other scores are made relative to these, so that a transcript released in 2023 impacts the score of a transcript released in 2017. Hence they are no longer directly interpretable as a negative or positive score, instead being scored relative to the rest of the transcripts. The scaling of the list of scores is done after first scoring the transcripts independently, and as ordering is preserved by the scaling, this is not a problem for our below strategies, as we are only interested in the intra-month ordering of transcripts.

Second, the TF-IDF scoring methods in sections 2.3, 2.4, 2.5 are calculated ex-ante over the whole corpus, so that the inverse document frequency (IDF) part of a 2017 transcript's TF-IDF score is influenced by n-gram frequency in 2023. There is clearly no reason why we should consider the IDF of future documents to be known at the start of 2017. As we saw in section 1.3, we can not assume that the word distribution is constant over the entire corpus, so this introduces look-ahead bias, making our TF-IDF-based strategies infeasible for out of sample prediction. Regardless, it is difficult to see the mechanism linking the look-ahead information we gain and future returns, so the cumulative returns in section 3.2 should still be approximately correct.

2.7.1 Other methods for sentiment score calculations

Due to the look-ahead issues discussed above, an improvement might have been to update the TF-IDF matrix monthly instead of ex-ante, for example by only considering the previous month's transcripts to limit time complexity. This implementation represented too big a break with the rest of the project structure, however, so we decided to keep the simple approach. Another admission is that our conclusions might have been strengthened by considering other combinations of TF-IDF scores, such as only LM unigrams, GHR unigrams, and the combination of LM and GHR uni- and bigrams.

Before deciding on the above scoring methods, we also considered using FinBERT,[2] which we ended up not doing due to computing constraints. We also considered replacing the VADER scores (which, we repeat, result from a model trained

[2]https://huggingface.co/ProsusAI/finbert

on social media) with FinVADER[3] scores, which include finance-related unigrams in the scoring lexicon, but we preferred to use standard methods.

3 Trading strategy

In the next sections of the report we will construct trading strategies based on the sentiment scores constructed in section 2. We assume no transaction costs and no price impact. First, we introduce our decile sorting algorithm to give a brief overview on the basic trading strategy. Then, we construct long-short portfolios based on the decile sorting algorithm, and compare them against different sentiment scoring methods to find which method performs the best.

3.1 Decile-based portfolio

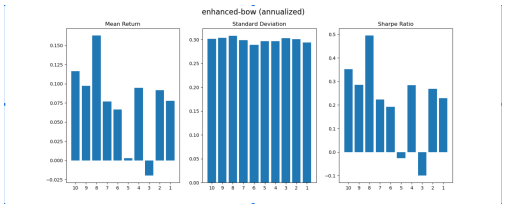
After labelling each earnings call with sentiment scores, we categorize them by their month of release. As discussed in section 1.3.1, we only consider months with more than 100 earnings calls.

For each month t, we gather all the earnings calls and sort them by a given sentiment scoring method, and allocate the firms they correspond to into 10 equal-weighted portfolios based on their sentiment score decile. We hold the decile portfolios from the first day of the next month $(t + 1)$. If more than 100 earnings calls are released in month $t + 1$ we reconstruct the decile portfolios at month-end. If not, we keep the decile portfolios until a month with more than 100 calls arrives or we reach the end of the period where we have earnings calls.

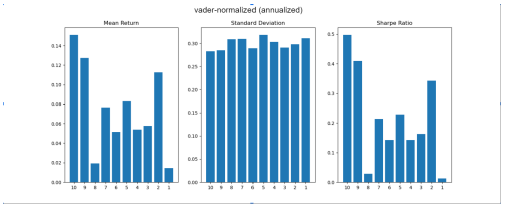
Letting 10 denote the decile with the highest *sentiment* and 1 the smallest, we plot the annual return, standard deviation and Sharpe ratio for all the scoring methods by running the above algorithm, and show the result in figure 12.

We observe that the standard deviations of different deciles remain similar for all methods, but that the annualized returns vary. For all of the methods, there is no linear relationship between returns, but importantly, the top two decile portfolios of all five scoring methods successfully outperform the bottom two deciles. This

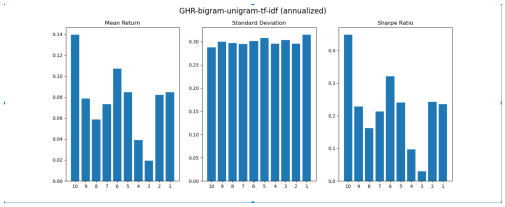
[3]https://github.com/PetrKorab/FinVADER



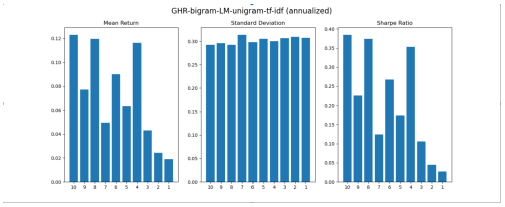
(a) Bag-of-words



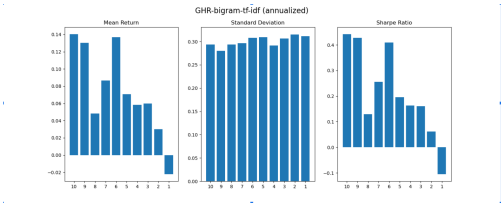
(b) VADER



(c) GHR-bigram-unigram-tf-idf



(d) GHR-bigram-LM-unigram-tf-idf



(e) GHR-bigram-tf-idf

Figure 12: (a) through (e) show decile portfolio performance for given scoring methods.

indicates that all the methods can more or less extract somewhat profitable signals from the earning calls purely using the sentiment scores.

3.2 Long-short Portfolio

Next, we construct long-short portfolios based on the decile-based portfolios. For each sentiment scoring method, we go long the top-two decile portfolios (10 and 9) and short the bottom-two decile portfolios (1 and 2), forming zero-cost long-short portfolios. As a benchmark, we go long the S&P 500 and short US treasury bonds. The time series of long-short portfolio cumulative returns is shown in figure 13.

The VADER and GHR-bigram-unigram methods underperform the benchmark in the given time frame, and the scoring method with the best performing portfolio is GHR-bigram-tf-idf. In figure 12e we see that in this time sample the method performs well at identifying the worst performing stocks, making the long-short portfolio formed by GHR-bigram-tf-idf the best performing trading strategy.

We also display the portfolio performance in figure 14. We again see that GHR-

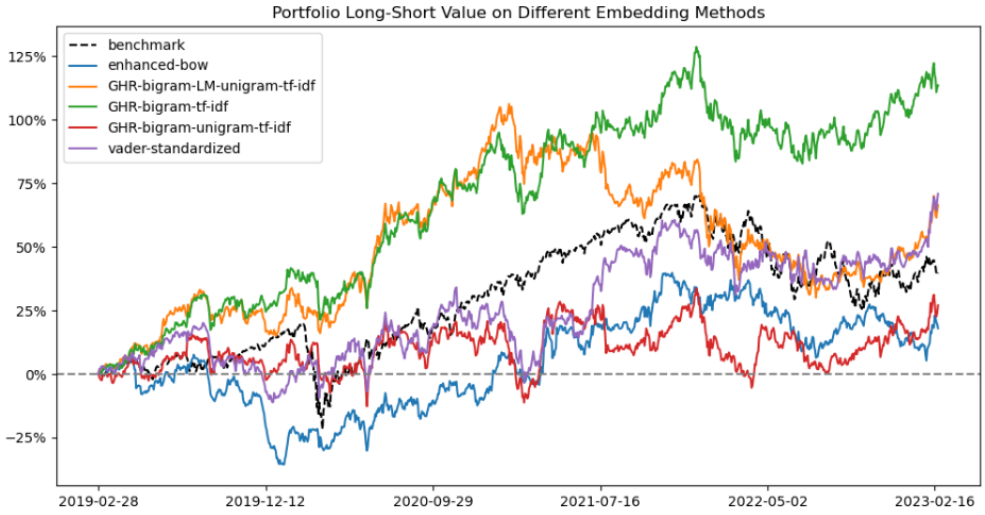


Figure 13: Time series of cumulative long-short portfolio returns compared to benchmark (dotted in black).

bigram-tf-idf best, outperforming the others in annual return, Sharpe ratio, annual cumulative abnormal return CAR_P, tracking error TE_P and information ratio IR_P. The last three are calculated as

$$CAR_P = R_P - R_{Benchmark}, \quad TE_P = \sigma_{CAR_P}, \quad IR_P = \frac{CAR_P}{TE_P},$$

with σ denoting standard deviation.

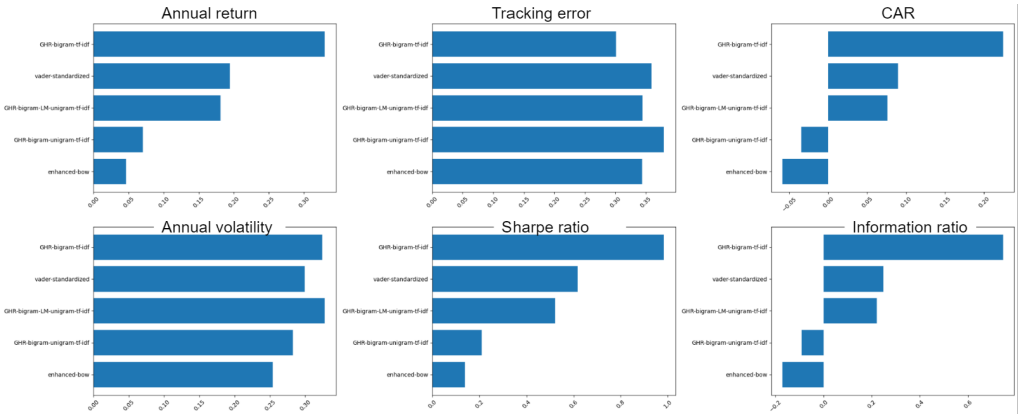


Figure 14: Portfolio performance overview showing return, TE, CAR, SD, SR and IR for each long-short portfolio.

3.3 Economic interpretation

The underlying assumption of our decile portfolio construction algorithm described in section 3.1 is that the earnings calls with greater *sentiment* indicate higher next-month returns. Based on the results above, this assumption is not falsified. We repeat that the the simple bigram embedding method performs better than the bigram-unigram and only-unigram models. This finding is fairly consistent with Garcia et al. (2023)[2], whose bigram portfolios also perform well in regression studies compared to less parsimonious models. One potential explanation is that the bigram embedding method contains higher information about the intrinsic sentiment in each earnings call, that is, many individual bigrams are strong indicators of sentiment. When adding the less informative unigrams, their signals are distorted, diluting the sentiment score. This is despite the bigram-only signals being relatively weaker as there are few bigrams compared to unigrams to count in each transcript. Intuitively, this should mean that they should be more noisy estimators of sentiment, leading to more randomness in the scoring, but our results show that the strength of the signals contained in the bigrams outweighs this concern. Another caveat is that our results may suffer from sampling bias, making them nongeneralizable to other time periods, for example figure 13 shows that the top performing strategies largely avoid the coronavirus crash, explaining their relative outperformance.

In the next section, we will focus on the GHR-bigram portfolio, and try to identify what topics in the earning calls drive the sentiments.

4 Latent Dirichlet Allocation (LDA)

4.1 A brief introduction to LDA

Latent Dirichlet Allocation is a statistical model that extracts topics from desired documents. The model works under specific assumptions, first that the documents consist of a mixture of topics and second that the topics are defined as a distribution over a vocabulary. The topic distribution for each document follows a Dirichlet distribution with parameter "α", and the word distribution in each topic follows

a Dirichlet distribution with parameter "η". Given the distributions, the model determines topics and the words associated with those topics.

4.2 LDA implementation and topic interpretation

To understand what topics drive the sentiments in the top two-decile portfolios and bottom two-decile portfolios of the GHR-bigram strategy, we apply LDA. The discussion in section 3.3 suggests that the bigram strategy most accurately captures sentiments in earnings calls, so determining the topics gives a link between month-ahead return and sentiment. We first gather all the transcripts that were in the top-two decile portfolios and all the transcripts in the bottom two-decile portfolios over the whole period in two separate lists. Before training a gensim LDA model on the transcript lists, we lemmatize the transcripts they contain and remove stop words. We also extend the list of stop words with non-indicative words such as names and roles to get more informative topics. We reduce topic overlap by tuning the η hyperparameter close to zero as it represents the a-priori belief on topic-word distribution. To simplify the interpretation of the output we set the number of topics to three We configure other training parameters, setting 'passes' to 15 and 'chunksize' to 50, to optimize the run-time of the model.

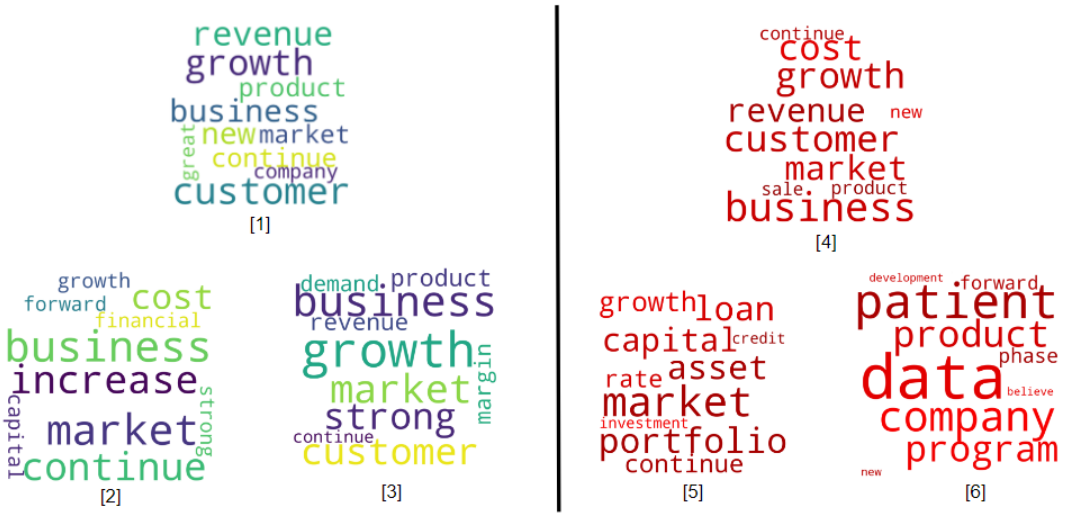


Figure 15: Word clouds for the top-two decile portfolios and bottom-two decile portfolios.

In figure 15 we plot the word clouds of the LDA-identified topics for the top two-

portfolios and bottom two-portfolios for the GHR-bigram strategy. The left part shows the topics in the most positive transcript,s and the right part the most negative. We then manually interpret each topic, and come to the following conclusions: The topic visualized in word cloud [1] is about the future outlook of the company, as the management discusses new opportunities on increasing the company's revenue growth. Word cloud [2] is about corporate finance as management talks about the outcome of the decisions they have taken given the market conditions. The outcome of the decisions are mentioned through the financials of the company. Word cloud [3] is about market dynamics as the management talks about the demand for their products and customer preferences.

The topic represented by negative word cloud [4], revolves around the same words as word cloud [1], switching out "great" with "cost", maybe indicating negative sentiment. Therefore we also label this word cloud's topic as future outlook. Word cloud [5] is similar to cloud [2], disucssing investment decisions given the market conditions. Different to [2], management here focuses on the firms liabilities, including words such as; "loan", "credit", and "rate". Word cloud [6] is about company development, discussing the phase the firm is going through and the actions to get through it.

In conclusion, the majority of the topics mentioned by companies who perform strongly and poorly are similar. The main difference between the topics is that when a company performs poorly, the discussion is more about own development and the current situation, whereas the companies who perform strongly are more forward-looking and reflect their opinions on the market dynamics, conveying confidence.

5 Conclusion

By constructing decile portfolios based on different sentiment scores, this project has shown that it is possible to elaborate a profitable trading strategy going long stocks with the most positive earnings calls and short the ones with the most negative earnings calls. We have used intepretable methods for sentiment scoring, throughout the report emphasizing the economics driving the scores, which in turn drive the performance of our trading strategies.

References

[1] Marina Druz et al. “When Managers Change Their Tone, Analysts and Investors Change Their Tune”. In: *Financial Analysts Journal* 76.2 (2020), pp. 47–69.

[2] Diego Garcia, Xiaowen Hu, and Maximilian Rohrer. “The colour of finance words”. In: *Journal of Financial Economics* 147.3 (2023), pp. 525–549.

[3] Elise Gourier and Hélène Mathurin. “A Greenwashing Index”. In: *Journal of Financial Economics* (2024). ESSEC Business School and CEPR, February 2024.

[4] Tim Loughran and Bill McDonald. “When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks”. In: *Journal of Finance* (2010). Forthcoming in the Journal of Finance.

[5] Paul C. Tetlock. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market”. In: *Journal of Finance* 62.3 (2007), pp. 1139–1168.