# Sentiment Metrics in Finance

Research Internship Report

Riccardo Djordjević[1]

[1] r.djordjevic@students.uu.nl - 0866504

# Table of Contents

## Introduction

Firms and researchers constantly face problems related to investor sentiment. These might be caused by the mispricing that originates in various asset classes, or by the difficulties of measuring investor sentiment. Therefore, firms and researchers use surveys or proxies to measure investor sentiment (Baker and Wurgler, 2007). However, the FEARS index proposed by Da et al. (2015) provides an innovative alternative to already existing metrics of investor sentiment.

In this paper is possible to find a background section explaining the theory behind the script that builds the FEARS index; an observations section to compare the results obtained during the various runs of the script with the expected results of the paper; a contributing section to shed light on the further developments that are needed to sharpen the future functioning of the scripts. This project was inaugurated as part of the Honours Programme of the Utrecht School of Economics (USE). The main goal is to replicate the same results achieved in 'The Sum of all FEARS Investor Sentiment and Asset Prices' (Da et al. 2015) and to extend them until present, since their time series is limited to the 2004-2011 period. The purpose is to engineer an algorithm that can gather the FEARS index automatically daily. This would expand the set of available tools to researchers and firms to measure investor and household sentiment.

Is imperative knowledge that this report has been written before the complete set of information necessary for the index has been gathered. Consequently, the observations section highlights the remarks found throughout the testing of the FEARS index but does not perform any test on the actual FEARS index for the period of 2004–2011.

## Background

This paper is based on the seminal work 'The Sum of all FEARS investor sentiment and Asset Prices' (Da et al. 2015). The authors built the FEARS index for quantifying investor sentiment by directly measuring the behaviour of households. Every other measure of investor sentiment is, or a proxy of it like the Volatility Index (VIX) and the number of Initial Public Offerings (IPOs), which do not directly capture sentiment, or are based on surveys which, unfortunately, are not possible to gather daily.

The index formulation initiates with the gathering of the list of primitive keywords from the Harvard IV-4 Dictionary and the Lasswell Value Dictionary. These keywords have been selected with an Excel Pivot Table when they are marked with '@Econ' or 'ECON', and additionally are also marked with "Ngtv," "Negativ," "Positiv," or "Pstv". This process aims to gather keywords related to economics with a meaning related to a positive or negative sentiment. Subsequently, the first 10 'Top' related queries[2] for each primitive keyword are extracted from Google Trends. The latter step is fundamental to finding what kind of queries Google users ask to the search engine.

After gathering the list of possible keywords that might form the FEARS index, the Pytrends Python library, functioning as Google Trends unofficial API, is employed to group the interest over time for each related query. Since Google Trends does not return as output the absolute search volume, but a relative interest over time for the selected period and

[2] Trending (or rising) searches are those that are accelerating the fastest, while top-searched, or "most-searched," are the most-searched queries in a specific time frame. Trending searches are useful for seeing how things have changed.

doesn't return daily data if the period is too long[3], an alternative strategy was necessitated. This entailed gathering daily interest over time in monthly batches, and then normalising it by the monthly data, collected all at once. This process ensures the comparability of the interest over time of different months and creates a time series.

The following stage consists of performing a series of adjustments on the time series to improve its statistical robustness. This entails removing the related queries that have less than 1000 daily observations. Then exclude keywords that are not related to economics and finance (i.e. if there are the words 'economic depression' and 'postpartum depression', only the former will be kept). Then performing a log difference[4] of the daily change in search term *j*.

$$\Delta SV_j = \log(SV_{j,t}) - \log(SV_{j,t-1}) \quad (1)$$

To mitigate any further disturbance in the data that would make it difficult to proceed with statistical calculations, the time series is first winsorized the log-difference time series at 5% (2.5% in each tail. Then the winsorized time series is regressed on weekday and month dummies and the residuals are kept, removing seasonality. Each time series is subsequently standardised by dividing the residuals by their own standard deviation. The latter is the last stage of statistical adjustment on the time series to engineer the 'adjusted (winsorized, deseasonalized, and standardized) daily change in search volume, $\Delta ASVI_t$'. The following calculations are performed to select the relevant keywords that will build the FEARS index.

[3] Daily results can be obtained by using a timeframe of 269 days or less. Weekly results can be obtained by using a timeframe of between 270 days and 269 weeks. Monthly results can be obtained by using a timeframe of 270 weeks or more.

[4] To avoid the error of $ln(0)$ the value of $1 \times 10^{-10}$ has been assigned to all the zero or 'NaN' observations.

The process reinitiates by merging the $\Delta ASVI_t$ the data frame and the S&P500 daily returns. This procedure leads to the loss of some rows of the data frame since there are no data points for market return during weekends. A backwards expanding rolling regression of the $\Delta ASVI_t$ of each keyword on the S&P500 daily returns is then executed. This regression is 'backwards expanding' because the time frame of the regression gets shorter each time. It is also 'rolling' because some periods of the timeframe are used again at each new regression.

The backward expanding rolling regression functioning is clearly explained in this paragraph. The initial rolling window of the regression equals the complete time frame of the analysed data, without the last six months (in this case from January 1st, 2004, to June 30th, 2011). Then 30 keywords with the smallest t-statistics originating by the regression are selected for the six-month period that follows the regression window (in this example for the first window we select these keywords for the period from July 1st, 2011, to December 31st, 2011). For this six-month window, the FEARS index is given by the average of the $\Delta ASVI_t$ for the 30 selected keywords. The following regression window gets six months shorter while keeping the same start date (from January 1st, 2004, to December 31st, 2010) and from that, the 30[5] most relevant keywords (given by the smallest t-statistics) are selected for the six-month period immediately following the regression window (from January 1st, 2011, to June 30th, 2011). This process is repeated until the regression window is reduced to six months. Since an initial window of six months is needed to perform the regression, the FEARS index starts from July 2004.

## Observations

[5] A further comparison analysis has been performed I the following section to verify the effects on the FEARS index of changing this variable to 25 or 35.

The objective of this section is to compare the results yielded during the construction of the FEARS index scripts with those of the original paper. Also, a few technical difficulties that have been encountered during the construction of the program, are explained for the benefit of future research endeavours in this domain. Further details on future refinements to the algorithm can be found in the 'Contributing' section.

The initial observations regard the number of primitive keywords. Since the list of primitive keywords or the exact dictionary used in the research could not be traced back, the same dictionary has been searched on Google and has been used to group a list of 149 primitive keywords, the same number as in the original paper. Therefore, given the parity, it is logical to assume that the algorithm-grouped list is highly like the original one.

Nevertheless, a major issue has been encountered in the number of keywords after gathering the 'Top' related queries. After removing duplicates, 1399 related queries to the primitive keywords list[6] have been found, a 12.4% variation from the 1245 keywords of the original study. The results suggest that Google changed its data aggregation methods. Google operates many different search engines such as YouTube, Google Shopping, Google Scholar, and others. It also pays or owns other browsers to make them use Google as its primary search engine such as Safari, Opera, Firefox, and others. Additionally, many websites use Google as a search engine to find specific items or pages inside the website itself (i.e. a university website might use Google search engine when a student is trying to look for 'fees' inside their website). Therefore, it is possible that Google Trends transformed the methodology of how it weights different browsers and search engines, leading to different

[6] The initial number of 'Top' related queries is 1490, since we take the first 10 related queries for each of the 149 primitive keywords. Some related queries were not available because there was not enough data available, and Google Trends was not returning any output.

results for the ‘Top’ related queries. Unfortunately, it has not been possible to retrieve information on how Google might have changed its parameters since this is mostly not publicly available information. Yet, this should not cause a major issue to the construction of the FEARS index since the keywords that have the highest influence on the stock market should have been searched enough that even a change in the Google Trends data methodology could not remove them from the list of related queries. In other words, even if at this stage there is a markedly superior number of keywords, for each six-month period there should be the same queries as in the original research.

The tables presented below compare the FEARS index values derived from the original dataset by Da et al. (2015) with those generated by our algorithm, across three different configurations: Table 1, Table 2, and Table 3 (Google Trends, 2024). These configurations differ by the number of keywords selected for each rolling regression window—30, 25, and 35 most negative t-statistics, respectively. Table 2 and 3 ca be found in appendix A. The researchers in the original paper chose to analyze the most negative 25, 30, and 35 keywords, instead of only 30, to examine how varying the number of sentiment-indicating keywords affects the robustness and sensitivity of the FEARS index to capturing investor sentiment. The analysis should be taken with caution as it has not been possible to require information on all keywords, yet.

For each table, the compare dataset shows a lower standard deviation and a tighter minimum-maximum range, than the original dataset. This suggests that the sentiment shifts are more concentrated and less volatile given the smaller sample of keywords. Additionally, the reduced skewness and kurtosis point to fewer extreme outliers.

This paragraph aims to compare the results of the original paper across the three tables. Across the tables, the mean is consistently 0.00 indicating a long-run baseline. As the number of selected keywords increases, the standard deviation gets lower and the min-max

range gets narrower, indicating a smoothing of sentiment signals. Furthermore, selecting more keywords leads to fewer outliers in the FEARS index, shown by the lower kurtosis in Table 3; it also leads to a more positive skewness, which is given by the fact that the additional keywords that are selected in Table 2 and 3 have a less negative t-statistics than those selected for Table 1. Across every table is possible to observe a remarkably low correlation coefficient, which is also represented in the Figure 1 and 2. Figure 1 has created made more comprehensible by taking the 120 rolling average of the FEARS 30 index.

**Table 1**

*Comparative statistics between the results of Da et al. (2015) and the partially gathered FEARS index using the 30 keywords with the most negative t-statistics*

| **Statistic** | **Original Dataset (fears30)** | **Compare Dataset (row_average30)** |
|---|---|---|
| Mean | 0.00 | -0.01 |
| Std | 0.35 | 0.24 |
| Min | -2.55 | -1.17 |
| 25% | -0.15 | -0.16 |
| 50% | -0.02 | -0.04 |
| 75% | 0.13 | 0.12 |
| Max | 3.19 | 1.27 |
| Skewness | 1.87 | 0.58 |
| Kurtosis | 18.31 | 1.54 |
| Correlation | 0.03 | |

**Figure 1**

*ORIGINAL Vs NEW FEARS 30 index - 120 Rolling Average*

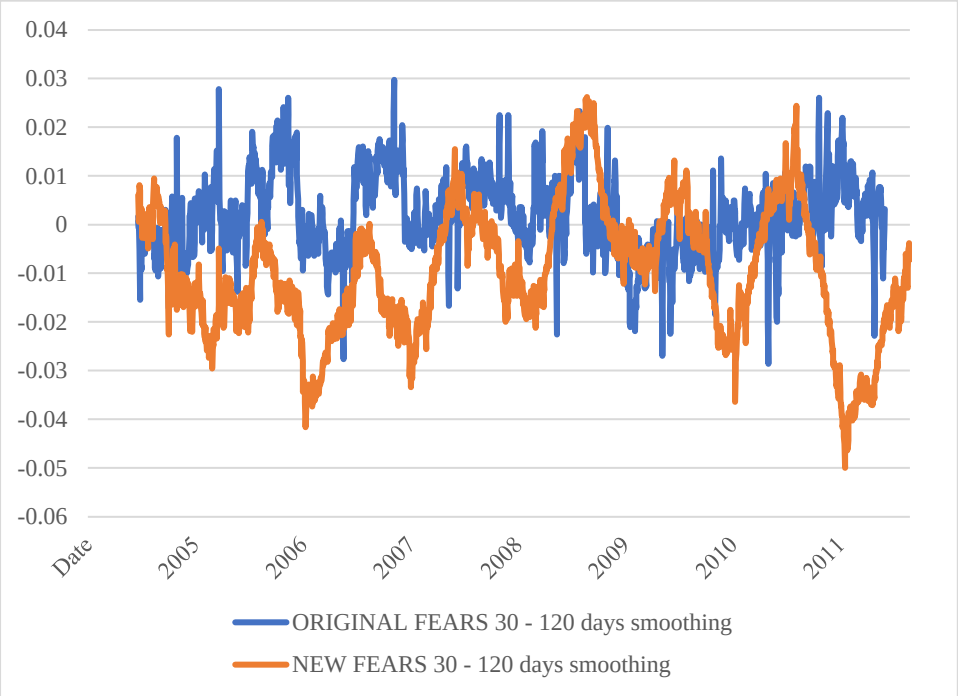
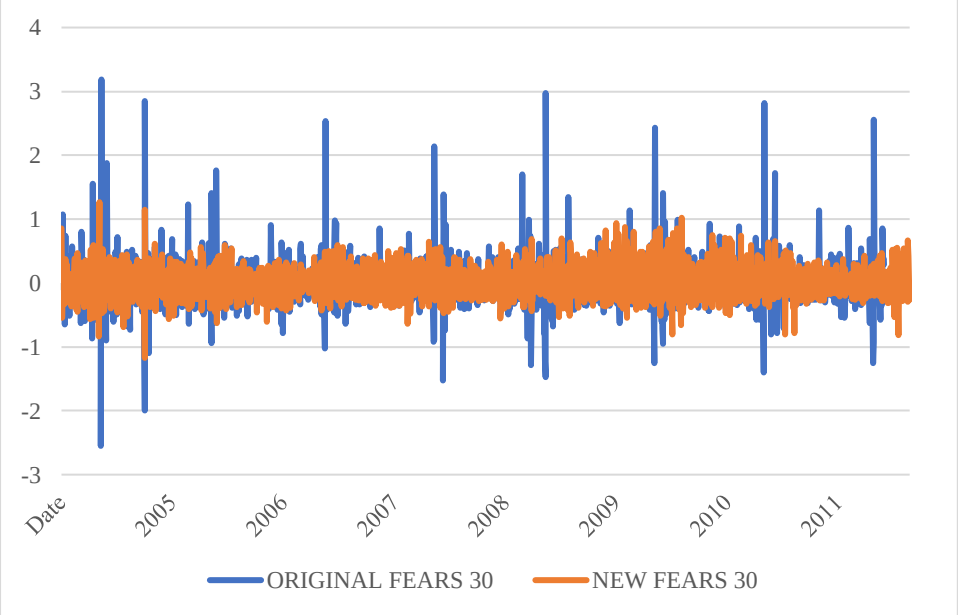



**Figure 2**

*ORIGINAL Vs NEW FEARS 30 index - No Smoothing*



The first and foremost problem that kept the construction of the index on hold for most of the time was Error 429 Too Many Requests (429). Google states that this error can be triggered by daily per-user limits, bandwidth limits, or concurrence request limits. Unfortunately, the API used, Pytrends, is not official and does not have specific information on these limits. Furthermore, Pytrends does not express a perfect pattern on how or when it will return a 429. The instability could be explained by the concurrence requests limits since at different times and locations (different IP addresses) the traffic on Google Trends varies. As previously mentioned, no evidence of an exact pattern could be found. Yet, there was evidence of a daily pattern showing a smoother execution of the script overnight than during the day. This could also be explained by the concurrent requests limit. The other identifier that Google can limit is the IP address. The assumed behaviour of Google towards IP addresses might also shed light on why running the script from a new IP address for the first time was giving fewer errors, compared to IP addresses where the same script has been heavily tested.

The original Pytrends library has been improved by implementing an additional timer. This helped diminish the number of 429s. However, it did not completely remove it from the script execution. If the script is used without rotating proxies, it might return a 429 even before gathering the first month of data or it would stop after a few keywords for the 2004–2011 time frame. This makes impossible the construction of the FEARS index without the use of rotating proxies but still makes the algorithm available for the daily search volume gathering of a limited number of keywords. To build the FEARS index is necessary the implementation of a rotating proxy service, which prevents Google from blocking a specific IP address. The performance that has been achieved so far is decent and would allow the construction of a Python script that calculates the FEARS index daily. The script was able to group daily search volume for just above 400 words, for the 8 years, in just above three

weeks, which means that it is possible to build an up-to-date FEARS index in just above six months. However, engineering an updated FEARS index requires further testing, and solving difficulties other than time constraints. A guideline of possible future problems to solve has been compiled in the next section.

## Contributing

The ultimate objective of the project is to make the FEARS index easily accessible like many other financial variables Personal Consumption Expenditures Price Index (PCE), or Volatility Index (VIX). Even if the core stages have been completed there is still a long way to go to improve the index. The following guideline has been divided into five main steps; however, it is not necessary to implement the steps in the proposed order.

The first step would be to integrate the three different scripts into a unique API application with a user interface that would enable the user to input the last date of the interested time frame. The second step would be to automate the download of the top related queries for each primitive keyword. Note that this stage also needs to be completed in a flexible time frame, since the related queries change depending on the selected period. This can be done with Pytrends or with a web scraping library like Beautiful Soup or Selenium.

The third step would improve the efficiency of the code by not gathering words that have already data available. For instance, if a new keyword that has never been gathered becomes relevant, the Search Volume will be downloaded since the beginning of 2004; while if a keyword has always been present among the relevant keywords, it will be necessary to download only the remaining period.

The fourth step would involve separating the FEARS index construction into two sections: one performed daily and one on an annual basis. The annual basis step includes gathering the related queries and then keeping only those that have enough observations (at

least 34% of the days gathered must be non-zero search volume data[7]) and are economics or finance-related. Executing this step yearly makes the process sustainable over time. This process should not cause any major issues to the calculation of the FEARS index since the relevant terms don't change daily. The daily step consists of creating the FEARS index from the relevant terms that have been selected yearly. The process of selection of relevant keywords can be easily outsourced to students, external parties, or even an AI model.

The fifth step consists of modifying the scaling system of the script so that different time frames are comparable over time. This issue is present because when the script creates the adjusted search volume time series, it scales the daily data gathered month by month, with the monthly data gathered all at once. When a new month of data is gathered, also the scaling changes[8]. This means that is not possible to compare the levels of the FEARS index.

During all processes of scaling the FEARS index is important to follow the guidelines mentioned by Da in the original paper. Guidelines such as the number of relevant keywords selected, how much to winsorize in each tail, and others. These variables should be kept constant, while other variables like the total number of days in the timeframe should be kept flexible to meet the requirements of the specific script run.

[7] Since in the original research they use a time span of 8 year (ca. 2920 days) and keep the keywords that have at least 1000 daily observations, the general FEARS index should keep the keywords that have at least 1000/2920 daily observations.

[8] If the initial time frame uses a scaling that reaches the month of January 2018, and January 2018 has a Search Volume value of 90, then all the days of that month will be multiplied by 0.9. However, if the Search Volume suddenly increases in the month of February, the January value might drop to 60 and the scaling factor to 0.6. This issue doesn't cause comparability problems inside the same time frame; however, it doesn't make possible the comparison of the levels between different time frames.

## References

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of economic perspectives*, 21(2), 129-151.

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies*, 28(1), 1-32.

G. (n.d.). pytrends/pytrends/dailydata.py at master · GeneralMills/pytrends. GitHub. https://github.com/GeneralMills/pytrends/blob/master/pytrends/dailydata.py

*Inquirer Home Page.* (n.d.). https://inquirer.sites.fas.harvard.edu/

Jones, M. (2022, November 16). Long-term Google Trends in Python with Longtrends | DataDrivenInvestor. Medium. https://medium.datadriveninvestor.com/long-term googletrends-in-python-with-longtrends-e478bb3d54f5

Resolve errors. (n.d.). Google for Developers. https://developers.google.com/gmail/api/guides/handle-errors

Rogers, S. (2021, August 11). 15 tips for getting the most out of Google Trends. Google. https://blog.google/products/search/15-tips-getting-most-out-google trends/#:~:text=Trending%20versus%20top%2Dsearched%3A%20Trending,seeing

20ho%20things%20have%20changed.

## Appendix A

**Table 2**

*Comparative statistics between the results of Da et al. (2015) and the partially gathered FEARS index using the 25 keywords with the most negative t-statistics*

| **Statistic** | **Original Dataset (fears25)** | **Compare Dataset (row_average25)** |
|---|---|---|
| Mean | 0.00 | -0.01 |
| Std | 0.37 | 0.25 |
| Min | -2.83 | -1.06 |
| 25% | -0.17 | -0.17 |
| 50% | -0.01 | -0.04 |
| 75% | 0.16 | 0.13 |
| Max | 3.57 | 1.34 |
| Skewness | 1.78 | 0.53 |
| Kurtosis | 19.88 | 1.26 |
| Correlation Coefficient | 0.02 | |

**Table 3**

*Comparative statistics between the results of Da et al. (2015) and the partially gathered FEARS index using the 35 keywords with the most negative t-statistics*

| **Statistic** | **Original Dataset (fears35)** | **Compare Dataset (row_average35)** |
|---|---|---|
| Mean | 0.00 | -0.01 |
| Std | 0.34 | 0.24 |
| Min | -2.29 | -1.16 |
| 25% | -0.15 | -0.16 |
| 50% | -0.02 | -0.04 |
| 75% | 0.13 | 0.11 |
| Max | 2.92 | 1.27 |
| Skewness | 1.92 | 0.59 |
| Kurtosis | 17.57 | 1.63 |
| Correlation Coefficient | 0.02 | |