

Textual Analysis by Hedge funds

Compliance with Data Policy for the *Journal of Accounting Research*

1. A description of which author(s) handled the data and conducted the analyses.

Sipeng Zeng handled the data and conducted the analyses reported in the manuscript.

2. A detailed description of how the raw data were obtained or generated, including data sources, the date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author is able to vouch for the stated source of the raw data.

All data used in the study are from public sources. The sources of our raw data are reported in the manuscript. The following is a summary:

- **Annual report textual data:** 10-K data were downloaded from <https://sraf.nd.edu/sec-edgar-data/> in November 2022.
- **EDGAR log data:** EDGAR Log file were downloaded from <https://www.sec.gov/data-research/sec-markets-data/edgar-log-file-data-sets> in November 2022.
- **COMPUSTAT:** Firm-level financial data were downloaded from WRDS in December 2022.
- **I/B/E/S:** Analyst forecast data were downloaded from the Institutional Brokers' Estimate System (I/B/E/S) through WRDS in December 2022.
- **Thomson Reuters 13F:** Institutional ownership data were downloaded from Thomson Reuters Institutional Holdings (13F) through WRDS in December 2022.
- **CRSP:** Stock price data were downloaded from WRDS in December 2022.
- **GoEmotions data:** The GoEmotions data used for training was downloaded from <https://www.kaggle.com/datasets/debarshichanda/goemotions> in July 2023

All authors vouch for the stated sources of the raw data.

3. If the data are obtained from an organization on a proprietary basis, the authors should privately provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., non-disclosure agreements, and any restrictions imposed by the organization on the authors). In particular, the authors should indicate if an organization or data provider imposes restrictions on the publication of the results, has not given the authors full control of the relevant data, requires that the results must be reviewed or approved prior to public release of the paper or publication.

All data used in this study are publicly available or based on licensed databases.

4. A complete description of the steps necessary to download, obtain or collect as well as process the data used in the final analyses reported in the paper. For experimental and survey papers, we

require information about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria. The full set of instructions and instruments can be provided in the online appendix.

The steps necessary to collect and process the data used in the final analyses reported in the paper are described in Section 2 of the manuscript.

5. After downloading or obtaining the raw data, all manipulations of the data should be done via computer programs. The code for these manipulations should be included in the code submitted upon acceptance (see below). No manipulations of raw data can take place manually or outside the computer code provided. If compliance with this requirement is not feasible, the authors need to explain and disclose any manipulations of the raw data (e.g., manually created variables or file conversions). When feasible, we also encourage the authors to share the code that downloads the data

All data manipulations were conducted via computer programs, and the Python and Stata codes used for these manipulations are included in the online supplements.

6. The computer programs (i.e., code) used to (1) convert the raw data into the final dataset used in the analysis, (2) to execute the statistical or econometric analysis, and (3) to generate the tables or to produce the output used in constructing tables of the manuscript. A brief description that enables other researchers to understand and run the code should be provided. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers were treated, and which commands were used in the analysis, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors' perspective. Therefore, instead of disclosing the proprietary portion of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables other researchers to arrive at the same results that the authors obtained and presented in their manuscript. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption allowing the step-by-step description. Whenever feasible, authors are required to provide the identifiers (e.g., CIK, CUSIP) for their final sample. Authors should consult our FAQ Sheet on the JAR website for further details.

We use Python to convert the raw data and use Stata to perform econometric analyses. The folder “Codes” contains the codes that use the data obtained from various public sources indicated above as input and yield the content of empirical analyses as output. Specifically, We construct the final regression sample using 01_build_regression_sample.py, compute DGTW-adjusted returns using 02_compute_dgtw_returns.py, and plot Figure 4 using 03_plot_figure4.py. We fine-tune the BERT model using 04_train_bert_score.py. 10_tables_1_5_7_figure3.do generates Tables 1–5, Table 7, and Figure 3. 20_table6_portfolio_returns.do generates Table 6. The text file “identifier.csv” provides the identifiers of the position hold in the main dataset.

7. A comprehensive log file that shows the execution of the entire code. This log file should cover all the steps that convert the raw data into a final dataset and the execution of all statistical and

econometric analyses presented in the tables of the manuscript. The portion of the log file that shows proprietary code or data may be masked. In this case, the reader should be referred to the step-by-step description provided as per the requirements in Item 6.

The log file “01_build_regression_sample_merge.log” shows all the steps that convert the raw data into the main dataset and “stata_tables_1_5_7_figure3.log” and “stata_table6.log” show the execution of all statistical and econometric analyses presented in the manuscript.

8. Data and programs should be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.

The authors will maintain all data and programs for at least six years.