Datasheet for the FIRE Dataset

Hassan Hamad

June 5, 2023

This is the datasheet for the FIRE (FInancial Relation Extraction) Dataset. The dataset is available at `https://github.com/hmhamad/FIRE`.

1 Motivation

- **For what purpose was the dataset created?** *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

  The FIRE dataset was created with the intention of advancing the development and evaluation of machine learning algorithms in the domain of financial information extraction, as well as to serve as a resource for financial analysts to automatically and efficiently extract critical information from financial documents. The dataset fills a significant gap in the field, as there has been a scarcity of publicly available datasets for financial NER and RE.

- **Who created the dataset** *(e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

  The dataset was created in collaboration between the research group of Professor Keith M. Chugg of USC [1]. and the company Vijna Labs Pvt. Ltd. [2] The dataset was annotated by Hassan Hamad [3], the PhD student of Professor Chugg.

[1] https://hal.usc.edu/chugg/
[2] https://www.v-labs.ai/
[3] https://www.hassanhamad.com/

- **Who funded the creation of the dataset?** *If there is an associated grant, please provide the name of the grantor and the grant name and number.*

  The dataset was funded by Vijna Labs Pvt. Ltd through a fellowship: America SLK Fellowship.

- **Any other comments?**

  No.

2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

  The instances in the FIRE dataset represent sentences extracted from financial documents and news articles. Each instance is a single sentence or multiple sentences that contains labels to indicate the named entities and the relations between these entities from this instance. Relations between entities represent various financial relationships and activities. Each instance contains an additional field called 'duration' which indicates the time it took the annotator to label this instance in seconds.

- **How many instances are there in total (of each type, if appropriate)?**

  FIRE contains a total of 2,849 instances.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

The FIRE dataset does not contain all possible instances but is a curated sample from a larger pool of financial documents and news articles. The larger set includes an extensive array of financial documents such as SEC filings, annual reports, and transcripts of earning calls, as well as a wide variety of financial news articles. The sampling was done in a way to cover diverse business and financial topics to capture a broad spectrum of entity and relation types prevalent in the financial domain.

- **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

  Each instance of FIRE consists of four fields:

  - tokens: The raw text of the sentence represented as a list of tokens.
  - entities: A list of named entities in the sentence. Each named entity is represented by a dictionary with keys: type (indicating the entity type), start and end, indicating the start(inclusive) and end(exclusive) token positions of the entity in the sentence.
  - relations: A list of relations in the sentence. Each relation is represented by a dictionary with keys: type (indicating the relation type), head and tail (indicating the entity indices involved in the relation)
  - duration: The time, in seconds, it took the human annotator to finish labeling this instance.

- **Is there a label or target associated with each instance?** *If so, please provide a description.*

  Yes, each instance in the FIRE dataset can come with multiple associated labels or targets which are the "entities" and "relations" fields described above. These fields can be empty for some sentences and can contain multiple labels for others.

- **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

  No.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

  The FIRE dataset primarily focuses on sentence-level named entities and relations. Therefore, each instance (sentence) in the dataset is treated independently, and there are no explicit relationships between different instances.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

  Yes, we recommend a 70%-15%-15% split of the FIRE dataset into training, development/validation, and testing subsets. The rationale behind this split is to ensure a sufficient amount of data for training the model, while also providing a robust set for model tuning (development/validation set) and an independent set for evaluating model performance (testing set). The 70%-15%-15% split is commonly used in machine learning as it usually provides a good balance between these three needs.

- **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

  The possibility of labeling errors occurring by mistake does exist. These could be due to oversight or misinterpretation of the annotation guidelines. Moreover, given the subjective nature of the task, some degree of annotator bias might be present, which could lead to inconsistencies in annotation. Further potential sources of noise could arise from lack of domain expertise of the annotator in the financial domain which is the focus of the dataset.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external*

resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

- **Does the dataset contain data that might be considered confidential** *(e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

  No. All the data used for creating this dataset comes from publicly accessible financial reports and financial news articles.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why. If the dataset does not relate to people, you may skip the remaining questions in this section.*

  It is highly unlikely that the FIRE dataset contains any content that could be perceived as offensive, insulting, threatening, or anxiety-inducing. The data in the dataset is strictly limited to the financial domain, and comprises entities and relations extracted from financial reports and news articles.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?** *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

  No. The FIRE dataset does not identify any subpopulations such as age, gender, ethnicity, etc.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** *If so, please describe how.*

  While the FIRE dataset might contain information about affiliations between individuals and businesses or their roles in those businesses (for instance, "CEO of company X", "founder of company Y"), the instances are derived from publically available financial reports and news articles, and no additional information beyond what is publicly

available is included in the dataset. The information about individuals is limited to their professional roles and affiliations and does not include personal identifiers or sensitive personal information. However, since some instances might mention individuals in their professional capacity, an indirect identification might be theoretically possible. Nonetheless, this is not the intention nor the primary focus of the dataset.

- **Does the dataset contain data that might be considered sensitive** *in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

  No. All data used to create the dataset is from public sources.

- **Any other comments?**

  No.

3 Collection Process

- **How was the data associated with each instance acquired?** *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

  The data constituting each instance in the FIRE dataset was obtained through a combination of direct observation and derived data. This involved two primary sources:

  - **10-X Filings:** These are documents that publicly traded companies are required to file annually with the U.S. Securities and Exchange Commission. They provide a comprehensive summary of a company's financial performance. We used the dataset of *Cleaned and Raw 10-X Files* spanning the years 1993-2021 [1].

This dataset contains all 10-K variants, e.g., 10-Q, 10-K/A, 10-K405.

- **Financial News Websites:** Data was also obtained from reputed financial news platforms like Bloomberg, Yahoo Finance, and others [2, 3, 6, 5, 4]. This consisted of raw text from news articles related to financial events, activities, and trends.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** *How were these mechanisms or procedures validated?*

  Data was collected using a software program.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

  Data was sampled in two ways from the larger set: randomly and using a trained model selection. Random data selection is used to select a subset of the larger pool of financial reports and articles. A model trained on the task of joint Named Entity Recognition and Relation Extraction is also used to parse these selected documents and select relevant passages for labeling.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

  No crowdworkers were hired. All persons involved in the data collection process are coauthors of this work.

- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** *If not, please describe the timeframe in which the data associated with the instances was created.*

  Data from the 10-X filings dataset is static and spans filings from the years 1993 to 2021. Data from financial news articles was collected during the first months of 2023. Note that the dataset use-case, training

a financial machine learning model, is independent of the data collection timeframe.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. If the dataset does not relate to people, you may skip the remaining questions in this section.*

  N/A.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

  N/A.

- **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

  N/A.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

  N/A.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

  N/A.

- **Any other comments?**

  No.

4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remaining questions in this section.*

  Data sourced from financial news websites underwent a cleaning and pre-processing process. This involves removing any non-textual components such as tables, graphs and markup tags. Subsequently, the data was transformed into raw text format and underwent two additional stages: segmentation and tokenization. Segmentation refers to the process of dividing a text document into individual sentences, while tokenization involves further breaking down each sentence into a sequence of discrete tokens. For both segmentation and tokenization tasks, the Python spaCy library (available at `https://spacy.io`) was utilized. However, data sourced from the 10-X filings dataset had already been cleaned by the dataset authors and solely required tokenization and segmentation procedures. All sentences were labeled by a human annotator.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

  The raw data was not saved.

- **Is the software that was used to preprocess/clean/label the data available?** *If so, please provide a link or other access point.*

  The software used to preprocess/clean the data is not available. The software used to label the data is available at `https://github.com/abhinav-kumar-thakur/relation-extraction-annotator`

- **Any other comments?**

  No.

5 Uses

- **Has the dataset been used for any tasks already?** *If so, please provide a description.*

  Until the writing of this datasheet, the only task FIRE was used for is performing the benchmark study in the dataset paper. This involved evaluating the performance of several state-of-the-art joint Named Entity Recognition (NER) and Relation Extraction (RE) models. More details about this experiment can be found in the dataset paper.

- **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

  No, there isn't a dedicated repository that links to papers or systems using the FIRE dataset.

- **What (other) tasks could the dataset be used for?**

  The dataset can be used for various NLP tasks in the financial domain, including, but not limited to, named entity recognition, relation extraction, sentiment classification and question answering.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

  The composition and construction of the FIRE dataset does not present clear factors that might impact future uses. Future users of the dataset should still be mindful of the subjectivity in labeling this dataset and the possibility of noisy or incorrect labels existing in the dataset.

- **Are there tasks for which the dataset should not be used? If so, please provide a description.**

Given the specificity of the dataset to the financial domain, it should not be used for tasks in significantly different domains like the biomedical domain without careful consideration.

- **Any other comments?**

  No.

6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

  The dataset is openly available through GitHub at `https://github.com/hmhamad/FIRE`.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** *Does the dataset have a digital object identifier (DOI)?*

  The dataset is distributed through GitHub at `https://github.com/hmhamad/FIRE`. The dataset DOI is [**TODO**].

- **When will the dataset be distributed?**

  The dataset was distributed and made publicly available on May 21st, 2023.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

  The FIRE dataset is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. This license permits users to copy, distribute, and display the dataset, as well as to make derivative works based on it, provided that they give appropriate credit to the creators of the dataset.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

  No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

  No.

- **Any other comments?**

  No.

7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**

  The FIRE dataset will be supported, hosted, and maintained by the creator of the dataset, Hassan Hamad [4]. This includes updates to the dataset, if applicable, as well as addressing any issues and queries raised by users of the dataset.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

  The manager of the dataset is Hassan Hamad. Email Address: hhamad@usc.edu

- **Is there an erratum?** *If so, please provide a link or other access point.*

  No.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how*

[4] https://www.hassanhamad.com/

often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

As of the date of writing, there is no specific schedule for updating the dataset, adding new instances or deleting instances. Any potenital updates in the future will be made by our research team. Any changes or updates to the dataset will be communicated via the dataset's GitHub page, where we will maintain a log of updates. Users of the dataset are encouraged to report any errors they find or suggestions for improvements directly via GitHub or by contacting our team.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

  As the FIRE dataset contains information extracted from public data, the instances in the dataset do not hold personal data in a context that would warrant data retention policies.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

  If any updates or modifications are made, they will be marked clearly as new versions on the GitHub repository, while older versions will remain accessible.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

  Contributions and extensions to the FIRE dataset are welcome. The dataset is hosted on GitHub, which allows for collaborative development and expansion. Individuals or teams interested in contributing can fork the repository, make their changes, and submit a pull request.

- **Any other comments?**

No.

References

[1] Bill McDonald. Cleaned and Raw 10-X Files, Software Repository for Accounting and Finance, 2023. University of Notre Dame, Mendoza College of Business. https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/

[2] Bloomberg - Financial news, analysis, and data. (2023). Retrieved from https://www.bloomberg.com/

[3] Yahoo Finance. (2023). Retrieved from https://finance.yahoo.com

[4] CNBC. (2023). Retrieved from https://www.cnbc.com

[5] The Financial Express. (2023). Retrieved from https://www.financialexpress.com

[6] The Economic Times. (2023). Retrieved from https://economictimes.indiatimes.com