Do Companies Reveal Their Own Fraud? – A Novel Data Set for Fraud Detection Based on 10-K Reports

Moustafa Amin[1] **and Matthias Aßenmacher**[1,2]

[1]Department of Statistics, LMU Munich,
[2]Munich Center for Machine Learning (MCML), LMU Munich
Correspondence: matthias@stat.uni-muenchen.de

Abstract

This work aims to gather and analyze data for text-based fraud detection using data from financial disclosures – specifically, the Management's Discussion and Analysis (MDA) sections of 10-K reports submitted to the US Securities and Exchange Commission. We provide a comprehensive overview of the process for creating the data set and introduce the resulting data set as an open-source resource for future research in the financial natural language processing domain. We subsequently train a range of machine learning and deep learning classifiers on the MDA text, intending to provide reasonable baselines for future researchers and to offer insight into the nature of fraudulent disclosures and how such data can be effectively used for uncovering fraud.

1 Introduction

Getting involved in financial crimes might be one of the most lucrative propositions, financially speaking. In 2024 alone, around $3.1 trillion circulated through the global economy, with $782.9 billion being used in drug trafficking, $346.7 billion in human trafficking, and $11.5 billion in terrorism financing due to financial crimes (Nasdaq Verafin, 2024). Fraud, of course, comprised a notable share of that money, with an estimated loss of $485.6 billion due to fraud in 2023 alone (Nasdaq Verafin, 2024). Naturally, with the amount of money implicated in financial crimes, the number of stakeholders is not scarce – be it regulators, law enforcement, firms, or companies. It is well-documented that more often than not, the more synergistically stakeholders work to prevent financial crimes, the more likely they succeed (Nasdaq Verafin, 2024).

The focus of this research is not financial crimes as a whole but rather fraud, which is a complex phenomenon with different facets. According to the Association of Certified Fraud Examiners' (ACFE) *Report to the Nations* (Association of Certified Fraud Examiners, 2024), most fraudulent activities were uncovered by tips from individuals involved or adjacent to the fraud. Considering the financial damage caused by it, the necessity to improve mechanisms for the detection of financial crimes – and with it, fraud – is obvious. According to the ACFE and *Black's Law Dictionary*, it can be broadly defined as "any activity that relies on deception to achieve a gain" and becomes a crime when, in layman's terms: "you lie to deprive a person or organization of their money or property.". For the purposes of understanding the concept further, the definitions and categorizations of the ACFE are being adopted here.

Contributions. The main contribution of this work is to provide (to the best of our knowledge) the first publicly available data set of (specific sections of) 10-K reports alongside labels indicating fraudulent behavior. To enable other researchers to replicate or extend our work, we provide transparent descriptions of the data scraping and labeling processes, release our source code, and make the data available. We provide baseline results for a given data split motivated by specific temporal characteristics of fraud. Our entire code and the created data set are publicly available:

- **Scraper:** `https://github.com/aminmous/fraud-webscraper`
- **Code:** `https://github.com/aminmous/fraud-analysis`
- **Data:** `https://doi.org/10.5281/zenodo.17121948`

2 Related Work

The Management Discussion and Analysis (MDA), which is the 7th section of an annual report submitted to the U.S. Securities and Exchange Commission (SEC), is unique in that the information

Proceedings of The 10th Workshop on Financial Technology and Natural Language. EMNLP-2025, Suzhou, China

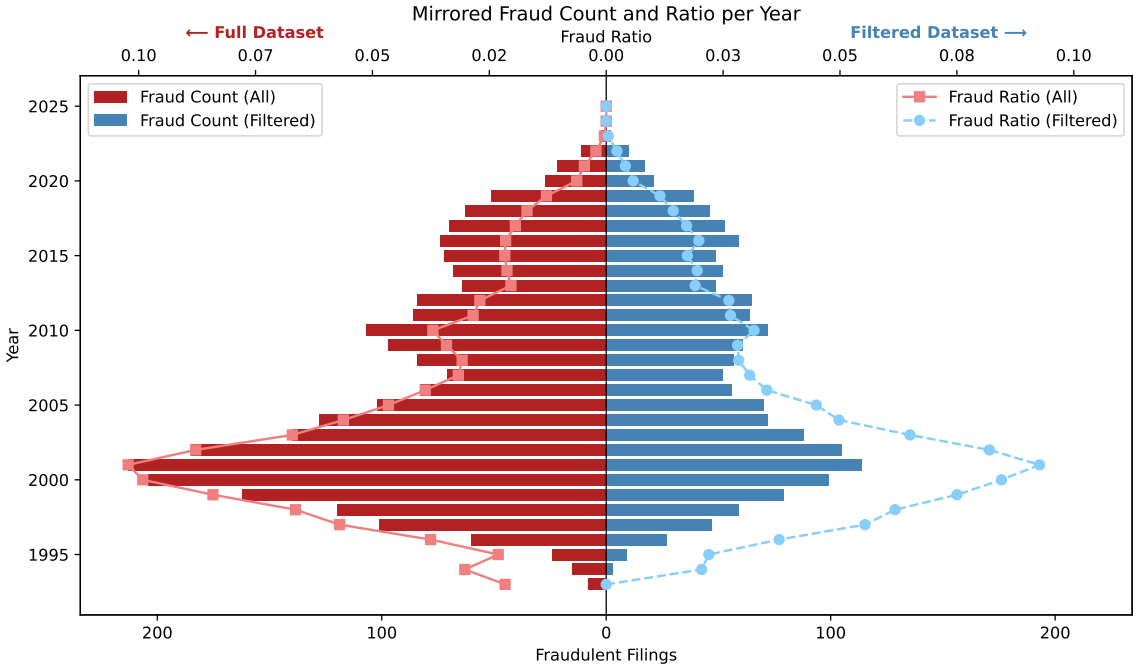


Figure 1: Visualizing the significance of fraud in 10-K reports: Development of the share of fraudulent reports in our newly curated data set over time, both in the raw (left) and in the final, filtered version (right).

contained in it is up to management's discretion and that it is subjective. Unlike other sections, it is not standardized in its structure and content and has thus been studied and scrutinized as a source of information – either by investors to gain an advantage over just evaluating tabular financial data or by researchers to evaluate, whether it can provide insights beyond what is concretely put to paper. Ultimately, firms are run by humans, humans are subject to their own biases and emotions, and why shouldn't this be reflected in MDAs?

Classical Machine Learning. The potential information content of MDAs is supported by work such as by Feldman et al. (2010), where the authors conclude that tone changes in MDA sections in both yearly (10-K) and quarterly (10-Q) reports are associated with market reactions in the short term. Hence, market participants can gauge short-term fluctuations based on the "nonfinancial" content of a report. (Durnev and Mangen, 2020) reveal that the MDA section of the 10-K report has influence over investment and disclosure decisions of other companies, especially in related industries, which suggests that the market not only affects but is affected by MDAs. (Holder-Webb and Cohen, 2007) determine that the quality of the disclosure is influenced by the level of stress that the companies are experiencing. It is therefore plausible to assume that the MDA section of the 10-K report is a good candidate to analyze and use as a basis to gain insights into the state of a firm and its management.

Our work is primarily inspired by Hoberg and Lewis (2017), whose work represents one of the pioneering works in the field of textual analysis of non-financial data and its relationship to fraud. The authors show that MDAs are an "informative setting for understanding fraud" by first showing that firms produce abnormal MDAs compared to their ISA (short for "*Industry, Size and Age*") peers[1], and secondly, by showing that fraudulent firms produce abnormal disclosures compared to their own in years where they did not commit fraud. They also used (tabular) accounting data from COMPUSTAT in conjunction with the MDA as text input and the fact that an AAER (short for "*Accounting and Auditing Enforcement Release*") was filed against the firm as a label for fraud. The AAER is a public document issued by the SEC that details the illegal findings of an investigation concerning a civil lawsuit brought by the SEC against a company, public or private, and/or individuals. AAERs contain valuable information about the nature of the fraud, such

[1]Industry peers are defined as firms with the same two-digit Standard Industrial Classification (SIC) codes: The first two digits define the major industry group (e.g., 25 = Furniture and Fixtures), the first three digits define the industry group (e.g., 252 = Office Furniture), and the full four-digit code specifies the detailed industry (e.g., 2521 = Wood Office Furniture).

as the involved parties, the scope of misconduct, the duration of the fraud, and the violated statutes.

LLM-based Approaches. A methodologically more advanced, albeit theoretically not as extensive, paper that starts to bridge the gap between the work of Hoberg and Lewis (2017) and the modern NLP methods available today for textual analysis is the work by Bhattacharya and Mickovic (2024). The main difference here is the use of BERT (Devlin et al., 2019), instead of topic modeling or sentiment analysis, to analyze MDAs and classify firms as fraudulent or not. The authors also use a data set that combines 28 quantitative financial features from COMPUSTAT with textual data from MDA sections of 10-K filings, and identify fraudulent outcomes using AAER enforcement data. The data set spans the years 1994 to 2013 and focuses on detecting a single category (accounting fraud). Besides BERT, the authors use LDA with 78 topics, selected by maximizing the AUC on their validation set over a range of 10 to 150 topics.

Shortcomings. What unifies all these different works is their lack of publicly available code and data, serving as the main motivation for our work of transparently constructing a benchmark data set for further experimentation in financial NLP.

3 Data Set Construction

3.1 Creating Firm-Year Observations

Firm-years are the unit of observation in the work of (Hoberg and Lewis, 2017); they represent the MDA sections of 10-K reports along the time dimension. We use the same nomenclature to describe the observations in our data. Although no open-source data set is available, the SEC provides a public database called the EDGAR system, which contains all filings submitted to the SEC dating back to 1994, with full coverage starting in 1997 (Hoberg and Lewis, 2017). It can be accessed through web scraping, albeit with some limitations.[2] Each company is assigned its own unique ten-digit SEC identifier, known as the *Central Index Key* (CIK), with a specific form type in mind, a URL form can be submitted via `https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=cik&type=formtype&dateb=&owner=exclude&count=40&search_text=`. By replacing the `formtype` with "10-K" and `cik` with the CIKs from the JSON list of CIKs from the SEC's website[3], access to all types of 10-K forms was enabled.[4] We developed a crawler based on the Scrapy library.[5] The crawler extracted the `.txt` version of the filing, as this format was consistently available in a structured format for all periods. It has a standardized header (cf. Figure 2, further details on the extracted variables are provided in Appendix A), providing useful additional information.

```
<SEC-DOCUMENT>0000320193-24-000123.txt : 20241101
<SEC-HEADER>0000320193-24-000123.hdr.sgml : 20241101
<ACCEPTANCE-DATETIME>20241101060136
ACCESSION NUMBER:               0000320193-24-000123
CONFORMED SUBMISSION TYPE:      10-K
PUBLIC DOCUMENT COUNT:          103
CONFORMED PERIOD OF REPORT:     20240928
FILED AS OF DATE:               20241101
DATE AS OF CHANGE:              20241101

FILER:

        COMPANY DATA:
                COMPANY CONFORMED NAME:                 Apple Inc.
                CENTRAL INDEX KEY:                      0000320193
                STANDARD INDUSTRIAL CLASSIFICATION:     ELECTRONIC COMPUTERS [3571]
                ORGANIZATION NAME:              06 Technology
                IRS NUMBER:                             942404110
                STATE OF INCORPORATION:                 CA
                FISCAL YEAR END:                        0928

        FILING VALUES:
                FORM TYPE:              10-K
                SEC ACT:                1934 Act
                SEC FILE NUMBER:        001-36743
                FILM NUMBER:            241416806

        BUSINESS ADDRESS:
                STREET 1:               ONE APPLE PARK WAY
                CITY:                   CUPERTINO
                STATE:                  CA
                ZIP:                    95014
                BUSINESS PHONE:         (408) 996-1010

        MAIL ADDRESS:
                STREET 1:               ONE APPLE PARK WAY
                CITY:                   CUPERTINO
                STATE:                  CA
                ZIP:                    95014

        FORMER COMPANY:
                FORMER CONFORMED NAME:  APPLE INC
                DATE OF NAME CHANGE:    20070109

        FORMER COMPANY:
                FORMER CONFORMED NAME:  APPLE COMPUTER INC
```

Figure 2: 10-K header in `.txt` filing (U.S. Securities and Exchange Commission, 2024)

Regex extraction was not feasible due to inconsistencies in the filings and structural changes, such as the introduction of item 7(a) in 1997, despite improvements to prior regex-based methods used in the work of Bhattacharya and Mickovic (2024). As a result, we turned to the SEC API (SEC API, 2025), a multifaceted (commercial) tool that operates similarly to the official SEC's API in that it allows users to access filing metadata. However, it also offers the additional capability of extracting specific sections from a set of filings.[6]

3.2 Labels

In most contemporary research, the labels are taken from the (closed-source) USC Marshall School of

[2] Requests are capped at 10 per second (U.S. Securities and Exchange Commission, 2023). Violating this policy results in a 10-minute ban from the server.

[3] https://www.sec.gov/files/company_tickers.json

[4] Crawling in this manner also yields various other types of 10-K forms (cf. Appendix B), as they share the same prefix.

[5] https://docs.scrapy.org/en/latest/

[6] Filings before 2002 are less standardized due to the absence of Sarbanes-Oxley (SOX), making MDA extraction via the SEC API less reliable for those years.

Business AAER data set, which can be purchased at several price points depending on the type of user. The data set contains 4,278 AAERs, with 1,816 cases of firm misstatements issued from 1982 to 2021 (University of Southern California, 2021). A non-exhaustive list of AAERs on the SEC's website[7]. Each AAER is numbered, with the latest one posted on the website (as of May 19, 2025) being AAER-4568. AAERs can be issued against all types of entities, even individuals, and 3,317 were available as of May 19, 2025. After filtering to include only enforcement actions taken against companies using the SEC API, the number of relevant AAERs was reduced to a reasonable starting point of 1,223. As only public companies are required to file 10-K reports, we had to apply an additional review and filtering.

Inclusion Criteria for Fraudulent Firms and Fraud Periods. We separated public from non-public companies by applying two main checks to each case: first, whether the company had a CIK; and second, if so, whether it had ever submitted a periodic filing required of public companies (10-K or 10-Q). Companies that had only submitted 10KSB (Deloitte Development LLC, 2008) forms before 2007 were not accepted as valid cases. If no 10-K or 10-Q was filed during the fraud period, the firm was not accepted as a valid case. For all valid cases, we recorded the CIK to link firm-years to corresponding labels and to enable integration with other datasets. As CIKs may vary over time, often due to a corporate split or restructuring, we use the CIK that corresponded to the periodic filings during the fraud period. Furthermore, we require the accounting enforcement action to specifically mention the firm as the perpetrator, meaning the firm had to be listed as a respondent and identified in the legal violations.

In rare cases, it was difficult to determine which specific firm was involved in an enforcement action due to ambiguous names and overlapping corporate structures. These ambiguities were resolved by examining filing patterns, such as joint submissions and matching CIKs in the 10-K headers. As mentioned, the fraud period often had to be deduced, and in some cases, the exact period could not be determined. Expecting the precise start and end dates of the fraud to be stated is often unrealistic and adds additional complexity to the requirements. Trading off precision against practicality, we approached the timeline in terms of quarters and fiscal years[8], ensuring our data also supports research involving 10-Q reports. We further distinguish between cases where the fraud period was mentioned explicitly and those where it was not.

Vague Cases. If the AAER stated that the fraud began at the "beginning" of a year, it was marked as starting in the first quarter. If it said "middle", the fraud began at the halfway point of the year. If it said "end", it began in the final quarter. If the AAER referred to a "fiscal year" rather than a calendar year, the same rules applied, but based on the company's fiscal quarters. If no specific timing was mentioned rather just the years where it had been committed, the fraud start was approximated as the beginning of the fiscal year, and the end was set to the end of that fiscal year. This approach allowed aligning fraud periods with the reporting periods in periodic filings. All such cases were classified as vague in terms of identifying the start and end dates.

Specific Cases. In some AAERs, specific quarters were mentioned as the fraud start or end period. In others, the AAER stated that the fraud began in the "period ending" a specific quarter – these entire quarters were marked as fraudulent. If an AAER said that the company "was fraudulently reporting" for a specific year or fiscal year, the entire fiscal year was marked as fraudulent. To distinguish these more precisely defined cases and also those mentioning a specific month as start and end date from vague ones, we included a certainty indicator in the data set (`certainty_start` and `certainty_end`). Cases with vague timing were marked with a 0, and specific cases with a 1. As with all binary variables in the label data set, 1 corresponds to the affirmative, and 0 to the negative. To ensure proper integration with the firm-year data set, the start and end dates of fraud were aligned with the **reporting** dates, not the **filing** dates. This decision supported the goal of aligning fraud periods with the timing of financial reports. In some rare cases, the AAERs mentioned more than one fraud period. If the periods did not overlap, two separate fraud cases were created – even if they originated from the same AAER. If the periods overlapped but involved distinct types of fraudulent activity,

[7] https://www.sec.gov/enforcement-litigation/accounting-auditing-enforcement-releases

[8] The fiscal year refers to the 12-month reporting cycle and may not align with the calendar year.

the case was likewise split into two cases based on the statutory violations cited in the AAER.

Violations. The SEC, as a regulatory body, derives most of its authority from the Securities Act of 1933 and the Securities Exchange Act of 1934.[9] It has the authority to enforce these acts and to take action against firms that violate them. The legal violations detailed in the AAERs are also included in our data set to provide information that may be useful for other purposes. In doing so, we do not record all violations mentioned in the AAERs, but only those specifically perpetrated by the firm in question – excluding those against individuals or unrelated entities named in the same AAER. Moreover, if a firm persistently fails to meet its periodic filing requirements, the SEC may revoke its Exchange Act registration under Section 12(j) of the Exchange Act (U.S. Congress, 1934). This is a rare occurrence and is recorded in the data set via the variable `revoked`, which notes the date (month and year) the registration was revoked. To enrich the data set further with a more structured understanding of fraud, the ACFE fraud classification introduced in Section 1 is also included. This categorization includes corruption, asset misappropriation, and financial statement fraud.

CIK Discrepancies. While merging the data sets, our earlier suspicion was confirmed: the CIK list provided on the SEC website is neither exhaustive nor immutable. The number of available CIKs varies depending on when the JSON file (company_tickers.json) is accessed. The CIK list used for crawling was obtained on May 8, 2025, selected solely because it contained the largest number of keys compared to the other lists at our disposal. This list includes 7,900 unique CIKs out of 10,132 total entries. In contrast, the labels data set contains 534 unique CIKs across 570 observations, 342 of which were not present in the CIK list used to crawl the firm-year data. This discrepancy increases by eight when comparing the CIKs in the firm-year data set with those in the labels data set, as the number of unique CIKs in the firm-year data set shrinks to 5,169 due to crawling issues. As a result, we decided to separately crawl firm years using the CIKs from the labels data set, yielding 10,764 firm-year observations. After merging both sources and removing duplicates, the final data set comprised 94,922 firm-year records. For the sake of reproducibility, the list of CIKs used to crawl the firm-year data set is included in the GitHub.

Data Merging. After generating firm-year records and compiling fraud labels, we were left with two distinct data sets: one containing 84,203 firm-years crawled using the SEC's CIK list, and another with 10,764 firm-years crawled using CIKs from the labels data set. Additionally, the labels data set contained 570 fraud-labeled observations. Merging these data sets was made significantly easier by relying on the Central Index Key (CIK), as merging based on company names would have required a fuzzy matching algorithm to resolve inconsistencies. The purpose of the labels data set was to add a binary fraud indicator to the firm-years, which could then be expanded with associated metadata variables. The `reporting_date` denotes the end of the fiscal year to which a report pertains, and it should correspond to the `fiscal_year_end` field included in the firm-year data set. A firm-year was labeled as fraudulent if its `reporting_date` fell within a fraud period listed in the labels data set. However, using this criterion alone would overlook cases in which fraudulent activity extended beyond the end of a fiscal year. To account for ongoing irregularities, any fiscal year in which the fraud period ended was also labeled as fraudulent. It is worth noting that a more precise labeling procedure could be achieved by merging the labels with a "firm-quarter" data set. To support future refinements, the individual raw data sets (the labels, firm-years from the CIK list, and firm-years from the labeled data set) are provided along with this thesis. The script used to merge these data sets is also included in the electronic appendix. In instances where a firm had overlapping or multiple fraud periods (as indicated by multiple entries in the labels data set), corresponding firm-year entries were duplicated, each reflecting distinct fraud-related metadata.

4 Descriptive Analysis

4.1 Full Data Set

The raw data set comprises $n = 89,453$ observations and $p = 57$ variables. A full description of all variables is provided in Table 2 in Appendix D. It includes filings from 5,508 unique CIKs, which

[9] The purpose of the Securities Act is primarily to ensure transparency and fairness before the initial issuance of securities, while the Exchange Act is more concerned with regulating the trading of securities in the secondary market.

approximately correspond to the number of distinct firms contained in the data set. Some CIKs have been dropped when extracting firm-years due to their integration into other filings, such as in cases where they are subsidiaries. These filings span 33 years, from 1992 to 2025, with an average of 16 and a median of 13 filings per company. It is important to note that the data set includes not only standard 10-K filings but also amended filings and late filings. As a result, some firms may appear overrepresented due to multiple amendments or corrections, a pattern that is common among the most frequently appearing ones. The most frequently occurring firm in the data set is *"Old Republic International Corporation"*, headquartered in Chicago, with a total of 81 filings (due to numerous amended reports).

Geographical Distribution. The data set contains firms located in 2,157 different cities, covering all 50 US states, D.C., three US territories (Guam, Puerto Rico, the Virgin Islands), Canadian provinces, and several other countries. In total, these firms span 158 unique jurisdictions, including both US states and international regions. In terms of legal incorporation, firms are registered across 91 different jurisdictions. Notably, in 61,881 instances, the state of incorporation differs from the state in which the firm is headquartered, with Delaware being by far the most common state of incorporation (cf. Figures 4a and 4b, Appendix E).

New York City has the highest number of filings and also hosts the greatest number of distinct firms. It accounts for more than twice the number of filings as the second-ranked city, Houston, and over three times as many unique firms (cf. Figures 5a and 5b, Appendix E). Interestingly, although Chicago and Atlanta are among the top cities by total filings, but not in terms of the number of distinct firms, suggesting a higher turnover of firms or a smaller presence of legacy corporations. At the state level, California surpasses New York in both total filings and the number of distinct firms (cf. Figures 6a and 6b, Appendix E). This is likely attributable to California's large population size, GDP, economic diversity, and concentration of large corporations spread across numerous cities. As previously mentioned, Delaware dominates as the primary state of incorporation, both in terms of the total number of filings and the number of incorporated firms.

Industry Distribution. The data set spans 412 distinct industries, classified by four-digit Standard Industrial Classification (SIC) codes. The most common industry by number of filings is *Pharmaceutical Preparations* (SIC 2834), with a total of 5,265 filings. This industry also ranks highest in terms of the number of distinct companies (cf. Figures 8a and 8b, Appendix E). It is important to note that the SEC does not always provide a SIC code in the header of each filing, not even the SIC code 9999, which represents a general-purpose category for firms that do not fit into any other industry. Filings with missing SIC codes span a wide variety of industries and account for 1405 filings, ranking 11th in group size. The smallest fraction of the data pertains to *Wholesale Trade - Furniture and Home Furnishings* (SIC 5020), with only 2 filings across 2 distinct companies (cf. Figure 8c, Appendix E). Aggregating the data by major industry groups reveals further insights. Consistent with the top specific industry, *Major Group 28* (*Chemicals and Allied Products*) ranks highest by both filings and number of companies. However, *Business Services* (Major Group 73) emerges as the second most common major group in both dimensions, indicating its significance across the corporate filing landscape. When aggregated even further to the division level, *Division D* (*Manufacturing*) dominates the data set in both total number of filings and companies. This division includes Major Group 28 and captures a wide range of manufacturing-related industries (cf. Figures 7a, 7b, 9a, and 9b, Appendix E).

Finally, several data quality issues were identified in several filings, such as a filing by *Coeur D'Alene Mines Corporation*, which operates in the *Gold and Silver Ores* industry. This filing was incorrectly tagged with SIC code 1044, which does not correspond to any valid industry classification. Four filings were listed with SIC Code 0, which does not exist, including two filings from *Enron Oil & Gas Company* which should correspond to SIC 1311, one from *BP Prudhoe Bay Royalty Trust* which should correspond to 2911 and one filing from *National Health Laboratories Holdings Inc.* which should correspond to sic 8071. These examples suggest the presence of additional misclassifications in the data, some of which may be undetectable unless the SIC code is invalid or missing.

Filing Types and MDAs. Among all filing types present in the data, the standard 10-K form is by far the most prevalent (77,851 filings), followed

by the amended 10-K/A version (12,642 filings)[10] and other smaller, negligible categories (cf. Figure 10, Appendix E). Across all filings, the MDA sections have an average word count[11] of 8,340 and a median of 6,981, indicating a right-skewed distribution. Some outliers are substantially longer, with the longest MDA comprising 188,443 words. As shown in Figure 3, not only does the number of filings in the data set increase over time, but also the average and median word counts. This trend is likely driven by enhanced standardization, improved digitization of EDGAR filings, and the SEC's continued efforts to refine filing procedures and document formatting. This enhanced filing quality also allows the parser to work more effectively and extract MDA sections more reliably.

Another important consideration is the presence of filings that do not contain substantive content, e.g., only the word "omitted" in the MDA section, while others refer the reader to other sections or separate documents. After manually inspecting a sample of such cases, filings with MDAs below a threshold of 200 words were considered *non-substantive* and were excluded from the analysis.[12] After filtering, the data set containing the substantive MDAs comprises 68,894 MDAs. The distributions of word and character counts for these texts are visualized in Figure 13 (Appendix E).

4.2 Fraudulent Cases

Turning to the subject of fraud, approximately 2.9% of all filings in the data set are labeled as fraudulent. This corresponds to 2,598 fraudulent filings out of a total of 89,453 (cf. Table 3, Appendix F). After filtering out the illegitimate cases (as described above), the proportion of fraudulent filings slightly decreases to 2.3%, amounting to 1,596 fraudulent entries (cf. Table 4, Appendix F). The distribution of word counts in these legitimate fraudulent filings is illustrated in Figure 14 (Appendix E). The highest number of fraudulent filings occurred around the year 2001, peaking at 213 filings (cf. Figure 1). This spike coincides with the aftermath of the *dot-com bubble*, a period of excessive speculation in internet-related companies during the late 1990s. A second notable peak is observed around 2010, corresponding to the fallout of the *2008 global financial crisis*, which led to the failure of major financial institutions and exposed widespread corporate malfeasance. During this period, fraudulent activity again surged, with regulators uncovering misconduct across various industries. Following the 2010 peak, the number of detected fraud cases declined steadily.

Time Delay and Duration. This decline, however, is likely not indicative of an actual reduction in fraud, but rather reflects the inherent lag in detection and enforcement. Regulatory bodies such as the SEC typically take several years to investigate and build cases against firms. As such, any recent misconduct, such as potential fraudulent activity during the COVID-19 pandemic, may not yet be reflected in the data. This time lag between fraudulent activity and its detection is evident in the distribution of detection delays. As shown in Figures 11a and 11b (Appendix E), the average time from the start of fraud to its detection is approximately 6.5 years. Further, the average time from the end of the fraudulent activity to its public exposure is about 3.5 years. This delay arises not only because fraud is difficult to detect, but also due to the time-consuming nature of building a legal case. On average, fraudulent activity spans approximately three years, as illustrated in Figure 12 (Appendix E).

Fraud by Industry. When analyzing fraud across industries, the ones with the highest rates of fraudulent filings are not necessarily the most common. As illustrated in Figure 15 (Appendix E), SIC 7372 (*Services—Prepackaged Software*)ranks fourth in terms of overall filing frequency, yet it accounts for the highest numbers of fraudulent filings. This aligns with historical trends, as the *dot-com bubble* era (late 1990s) was characterized by significant corporate fraud in the technology and software sectors. The industry with the highest proportion of fraudulent filings relative to total filings is SIC 2020 (*Dairy Products*). While the absolute number of fraud cases in this industry is low, its relative fraud rate is the highest among all SIC codes. This highlights how less visible or prominent sectors can still exhibit high relative risk. The major group with the highest fraud proportion is *Major Group 51: Wholesale Trade—Nondurable Goods* (cf. Fig-

[10]Amended filings (10-K/A) typically result from the need to correct errors, clarify previously misstated information, or respond to regulatory or legal issues. Although not all amendments are associated with fraud, they often reflect irregularities in the original reports. For this reason, MDA sections from amended filings are excluded from the subsequent analysis.

[11]Word counts were computed after the following preprocessing steps: Lower-casing, removal of stopwords, URLs, HTML tags, extraneous whitespace or non-textual symbols.

[12]This results in two versions of the data set, one with all the observations and one with only the substantive MDAs.

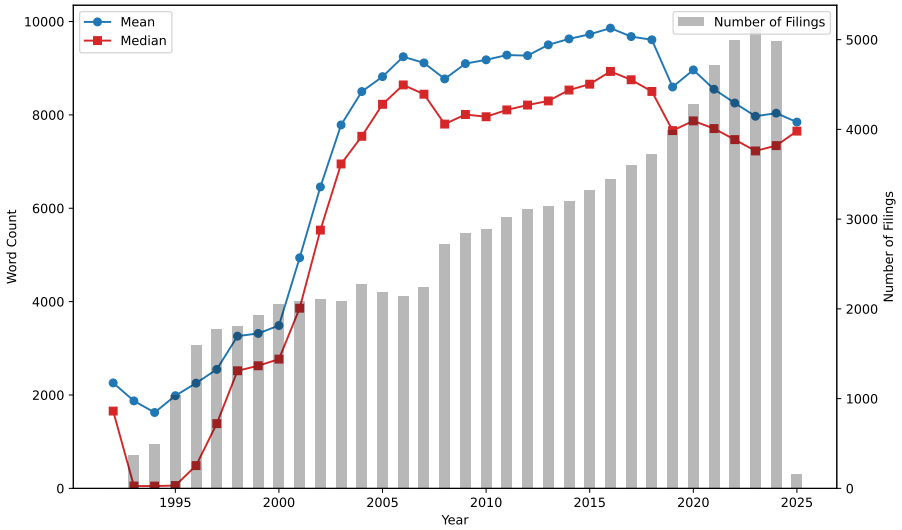


Figure 3: Mean and Median MDA Word Count Over Time with Filing Count

ure 16, Appendix E), while *Business Services* is the group with the highest absolute number of fraud cases, consistent with its high representation across the data set. At the division level, a similar discrepancy appears. *Division D (Manufacturing)* contains the most fraudulent cases in absolute terms[13], while the division with the highest fraud rate is *Division F (Wholesale Trade)*, which includes Major Group 51. These patterns suggest that while some sectors are more prone to frequent fraud due to their size, others exhibit disproportionately high risk relative to their footprint in the data.

Fraud Geography. When examining the geographical distribution of fraud in terms of absolute counts, the most fraudulent state is California, which aligns with its overall dominance in the number of total filings. However, when measured by fraud rate, Luxembourg stands out (4 out of 23 filings; cf. Figure 18, Appendix E). At the city level, the highest fraud rate is found in Pembroke, Bermuda, a notable offshore financial center,[14] while New York City leads in terms of absolute numbers, which aligns with the high density of financial institutions headquartered there – institutions that have historically been implicated in numerous financial misconduct cases. The most common state of incorporation for fraudulent firms is Delaware, which should not come as a surprise given that it is the most common state of incorporation overall. New Brunswick (Canada) exhibits the highest rate, with 5 out of 27 filings flagged as fraudulent (cf. Figure 20, Appendix E). Taken together, these findings provide a nuanced geographical portrait of corporate fraud in the data. While certain regions naturally have higher counts due to their economic prominence, some lesser-known jurisdictions display disproportionately high fraud rates. Nevertheless, these results must be interpreted with caution, as fraud overall remains a relatively rare event in the data, and high fraud rates in small regions are often based on few observations.

5 Experimental Results

Exemplarily, we use all data until 2008 for training and test on all data from the year 2011. We did abstain from using more current data due to delays in fraud detection and the comparably low amount of data in current years. Given that the median post-fraud detection delay is approximately 3.2 years, all models are tested on data that is at least three years in the future relative to the training data. A further reason for separating training and testing data temporally is that disclosures inherently have temporal characteristics. Extracting the MDA texts works increasingly well over time due to the enhanced standardization, and the rate of non-substantive fraudulent MDA sections is higher in earlier filings. We consider standard metrics such as accuracy, precision, recall, F1, and the AUC[15] to provide a point of reference – es-

[13] Excluding unknowns, which also might be an indicator of the inconsistencies in the fraudulent filings compiled.

[14] This ranking includes only cities with at least 10 filings, as smaller sample sizes prohibit robust conclusions.

[15] AUC does not adequately reflect performance in imbalanced settings; Our main evaluation metric is the F1-Macro.

Input Features	Random Forest					XGBoost					No Fraud / Fraud	
	Precision	Recall	F1-Macro	AUC	Accuracy	Precision	Recall	F1-Macro	AUC	Accuracy	Train Set	Test Set
Tabular	0.79	0.64	0.68	0.71	0.97	0.61	0.65	0.63	0.72	0.95	27226 / 1700	2935 / 86
Word Count	0.51	0.52	0.51	0.52	0.97	0.50	0.54	0.31	0.55	0.41	16695 / 937	2344 / 64
Sentence Embeddings	0.49	0.50	0.49	0.67	0.97	0.54	0.52	0.53	0.65	0.96	16695 / 937	2344 / 64
ModernBERT	0.49	0.50	0.49	0.62	0.97	0.50	0.50	0.50	0.61	0.96	16695 / 937	2344 / 64
29 LDA Topics	0.49	0.50	0.49	0.68	0.97	0.53	0.59	0.53	0.66	0.89	8032 / 307	2344 / 64
75 LDA Topics	0.49	0.50	0.49	0.64	0.97	0.53	0.55	0.54	0.64	0.94	8032 / 307	2344 / 64
100 LDA Topics	0.49	0.50	0.49	0.59	0.97	0.54	0.56	0.55	0.62	0.94	8032 / 307	2344 / 64

Table 1: Macro-Averaged Classification Metrics for Random Forest vs. XGBoost Across Input Representations. Differences in the sizes of the training sets result from the exclusion of the invalid cases, which are only present for the model trained on tabular data. For LDA, only 2004 to 2008 were used for training (due to the amount of data).

pecially for comparison with the work of (Bhattacharya and Mickovic, 2024). An overview of the results is presented in Table 1, where we compare Random Forest (Breiman, 2001) to XGBoost (Chen and Guestrin, 2016)) trained on different input features: tabular data only, word counts, sentence embeddings (`all-MiniLM-L6-v2`; Reimers and Gurevych, 2019), ModernBERT embeddings (`answerdotai/ModernBERT-base`; Warner et al., 2025), and LDA topics (inspired by Hoberg and Lewis, 2017). Overall, the performance across all evaluation metrics is worse than that achieved with the tabular data baseline, showing that classifiers based on only text-based inputs struggle severely. This impression is further supported by the full classification reports in Table 5 (Appendix G).

6 Conclusion

The Achilles heel of any fraud analysis is the scarcity of fraud cases, which significantly hampers efforts to gain meaningful insights. Therefore, it is essential not only to aggregate extensive, high-quality data but also to involve the right competence regarding analytical methods and domain knowledge. We see our work as an important auxiliary means for providing domain experts with high-quality data. Innovative analytical methods are vital as tools to assist experts in combating fraud and broader financial crime effectively. Hence, using the latest technology, such as large language models, is an important step in advancing this endeavor. Fraud detection is a balancing act, often involving inherently imbalanced data where invasive monitoring is difficult without compelling evidence, making false positive claims is particularly costly, and misallocating resources due to incorrect fraud predictions can negatively affect firms, employees, and the fraud detection process itself. Ultimately, this work aims to provide valuable insights and especially resources to anyone interested in understanding and analyzing the complexities of fraud detection.

Limitations

Despite we hope to have provided a valuable resource for fellow researchers working in the area of financial natural language processing, we are aware that our data set does not come without shortcomings: **(1)** As already mentioned a few times throughout the main part of this paper, we can not be entirely sure about the correctness of all CIKs; while some errors can be (relatively) easily be detected, further error correction would require domain knowledge beyond our expertise. **(2)** Given that only *detected* cases of fraud (positives) can be labeled as such, this positions this data set in the realm of positive-unlabeled (PU) learning, as we can not be sure whether negatives (cases labeled as *no fraud*) are actually not fraudulent or whether they were just not detected. This shortcoming has neither been addressed in previous work nor is it reflected in our baseline results; we leave the application of PU learning techniques to this data for future work.

The exclusive reliance on SEC enforcement reports raises another concern regarding common source bias. Both the texts and the metadata, meaning all inputs, are collected from a single source and have not been supplemented with data from elsewhere. Moreover, since none of us has had direct contact with the SEC, we cannot assess the systematic deficiencies over time that may have influenced the data available on EDGAR and the SEC website. Nevertheless, these limitations still deserve to be acknowledged.

Acknowledgments

Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research

Data Infrastructure – NFDI 27/1 - 460037581 - BERD@NFDI. This work was partially supported by Henan Provincial Center for Outstanding Overseas Scientists (No. GZS2025004).

References

Association of Certified Fraud Examiners. 2024. 2024 report to the nations: Global study on occupational fraud and abuse. Technical report, Association of Certified Fraud Examiners.

Indranil Bhattacharya and Ana Mickovic. 2024. Accounting fraud detection using contextual language learning. *International Journal of Accounting Information Systems*, 53:100682.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794. ACM.

Deloitte Development LLC. 2008. Changeover to the sec's new smaller reporting company system. Accessed: 2025-05-18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Art Durnev and Claudine Mangen. 2020. The spillover effects of md&a disclosures for real investment: The role of industry competition. *Journal of Accounting and Economics*, 70(1):101299.

Ronen Feldman, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal. 2010. Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, 15(4):915–953.

Gerard Hoberg and Craig Lewis. 2017. Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance*, 43:58–85.

Lori Holder-Webb and Jaffrey R. Cohen. 2007. The association between disclosure, distress, and failure. *Journal of Business Ethics*, 75(3):301–314.

Nasdaq Verafin. 2024. 2024 global financial crime report. Technical report, Nasdaq Verafin. Accessed: 2025-03-11.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

SEC API. 2025. Sec edgar api service. Accessed: 2025-05-18.

University of Southern California. 2021. Aaer dataset. Accessed: 2025-05-18.

U.S. Congress. 1933. Securities act of 1933. https://www.govinfo.gov/content/pkg/COMPS-1884/pdf/COMPS-1884.pdf. Codified at 15 U.S.C. §§ 77a-77aa. Accessed: 2025-05-18.

U.S. Congress. 1934. Securities exchange act of 1934. https://www.govinfo.gov/content/pkg/COMPS-1885/pdf/COMPS-1885.pdf. Codified at 15 U.S.C. §§ 78a-78qq. Accessed: 2025-05-18.

U.S. Securities and Exchange Commission. 2023. Privacy and security. Accessed: 2025-05-17.

U.S. Securities and Exchange Commission. 2024. Apple inc. form 10-k filing (2024). Accessed: 2025-05-17.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Appendix

A Format of the 10-K filings

Information in the 10-K headers (highlighted in red, Figure 2), in the following order of appearance:

- **Conformed Period of Report** - Report period
- **Filed as of Date** - Date the report was filed
- **Company Conformed Name**
- **Central Index Key** - Unique identifier
- **Standard Industrial Classification (SIC)**
- **State of Incorporation**
- **Fiscal Year End** - format: mmdd
- **Form Type** - Type of filing (cf. Appendix B)
- **City** - City of the company
- **State** - State/foreign country of the company

B Further 10-K Variants

The different types of annual reports are as follows:

- **10-K** - Standard annual report
- **10-K/A** - Amended annual report
- **10-K405** - Late annual report
- **10-K405/A** - Amended late annual report
- **10-KT**[16] - Transition annual report
- **10-KT/A** - Amended transition annual report

[16]Transitional reports are filed in cases where, e.g., firms merge and the first report after the merger is submitted outside the required reporting period of one of the merging firms.

C Regex Pattern for MDA-Extraction

Item 7 Start Pattern

- `r"it[\s]*em[\s]*7[\.\s]*manag[\s]*e?[\s]*ment[\s\'-]*[\w\s\'-]{0,10}(discussion[\s]*and[\s]*analysis|narrative[\s]*analysis)"`
  Matches various flexible formats of the section header for Item 7, it allows for:
  - Optional whitespace or punctuation between characters (e.g., "Item 7." or "Item 7 Management's")
  - Variations like "management" or misspelled/malformed variants
  - Up to 10 characters between "management" and "discussion" to tolerate OCR noise/alternative phrasings
  - Matches either "discussion and analysis" or "narrative analysis"

Search Phrases

- `the following discussion`
- `this discussion and analysis`
- `should be read in conjunction`
- `should be read together with`
- `r"the following management[\s\'-]*s discussion and analysis"`
  Flexible punctuation or possessive formatting.
- `r"(?:\"[^\"]+?\")([^"]*\"[^\"]+?\"){4,}"`
  A complex pattern that identifies sequences of four or more quoted strings.
- `r"\b(?:\w+)(?:\s*,\s*\w+){3,}\b"`
  Matches sequences of at least four comma-separated words.

Item 8 End Pattern

- `r"item[\s]*8[\.\s]*financial statements and supplementary data"`
  Allows optional whitespace or punctuation between "Item 8" and the section title.

D Variable Overview

Table 2: Variable Descriptions in the MDA-Fraud Dataset

Variable	Description
Firm-Years: Variables Scraped & Parsed from Annual Reports (10-K)	
cik	Central Index Key (unique company identifier)
name	Company name
city	City
state	State
sic	Standard Industry Classification number
incorp_state	State of incorporation
filing_type	Filing type
fye	Fiscal year end
filing_date	Date the 10-K was filed
reporting_date	Period the 10-K reports on
url	URL to the filing
mda	Management Discussion and Analysis section (text)
late_filing	Indicates a 10-K405 (late filing)
transition_filing	Indicates a 10-KT (transition report)
amend_filing	Indicates amended 10-K (any 10-K ending in /A)
Labels: Variables from Labels Dataset	
dateTime	Date and time of the AAER
respondents	Names of respondents in the AAER
fraud_start	Beginning date of the fraudulent period (mm-yyyy)
fraud_end	End date of the fraudulent period (mm-yyyy)
revoked	Revokation date of Exchange Act registration (mm-yyyy)
certainty_start	Binary indicator: certainty regarding fraud start date
certainty_end	Binary indicator: certainty regarding fraud end date
17a	17(a) Securities Act violation
17a2	17(a)(2) Securities Act violation
17a3	17(a)(3) Securities Act violation
17b	17(b) Securities Act violation
5a	5(a) Securities Act violation
5b1	5(b)(1) Securities Act violation
5c	5(c) Securities Act violation
10b	10(b) Securities Exchange Act violation
13a	13(a) Securities Exchange Act violation
12b20	Section 12b rule 12b-20 Securities Exchange Act violation
12b25	Section 12b rule 12b-25 Securities Exchange Act violation
13a1	Section 13a rule 13a-1 Securities Exchange Act violation
13a10	Section 13a rule 13a-10 Securities Exchange Act violation
13a11	Section 13a rule 13a-11 Securities Exchange Act violation
13a13	Section 13a rule 13a-13 Securities Exchange Act violation
13a14	Section 13a rule 13a-14 Securities Exchange Act violation
13a15	Section 13a rule 13a-15 Securities Exchange Act violation
13a16	Section 13a rule 13a-16 Securities Exchange Act violation
13b2A	13(b)(2)(A) Securities Exchange Act violation
13b2B	13(b)(2)(B) Securities Exchange Act violation
13b5	13(b)(5) Securities Exchange Act violation
14a	14(a) Securities Exchange Act violation
14c	14(c) Securities Exchange Act violation
30A	Foreign Corrupt Practices Act violation

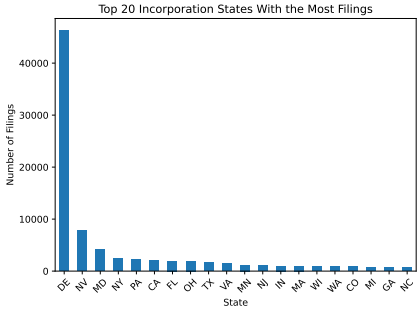
(Continued on next page)

Variable	Description
100a2	100(a)(2) Regulation G of Securities Act violation
100b	100(b) Regulation G of Securities Act violation
19a	19(a) violation under Investment Company Act
105c7B	105(c)(7)(B) violation under SOX
corruption	Binary indicator of corruption
amis	Binary indicator of asset misappropriation
fsf	Binary indicator of financial statement fraud
fraudulent	Binary indicator of fraud
MDA counts	
char_count	Character count of the MDA text
word_count	Word count of the MDA text
word_density	Number of characters per word

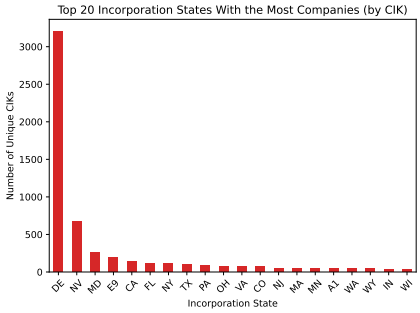
E Descriptive Analysis – Figures and Tables

E.1 Whole Data Set

Geographical Distribution

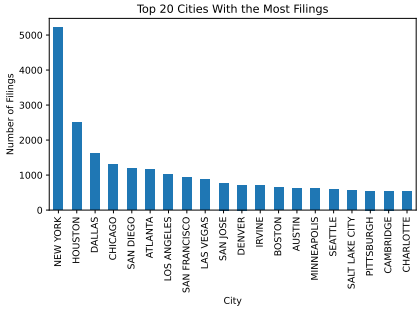


(a) Top 20 States of Incorporation by Filings

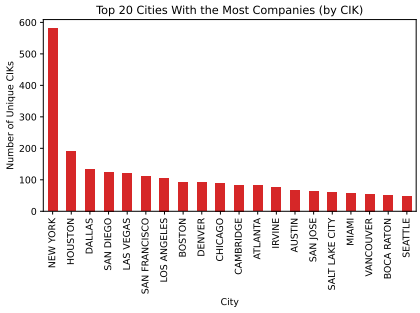


(b) Top 20 States of Incorporation by Number of Companies

Figure 4: Top 20 States of Incorporation

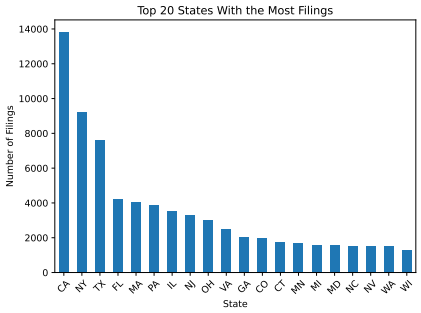


(a) Top 20 Cities by Total Filings

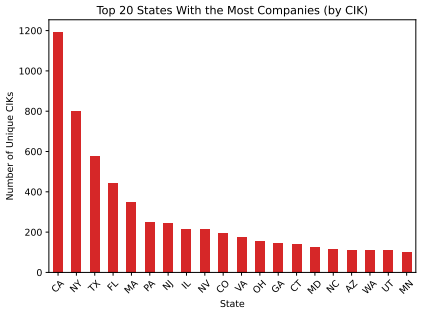


(b) Top 20 Cities by Number of Companies

Figure 5: Top 20 Cities



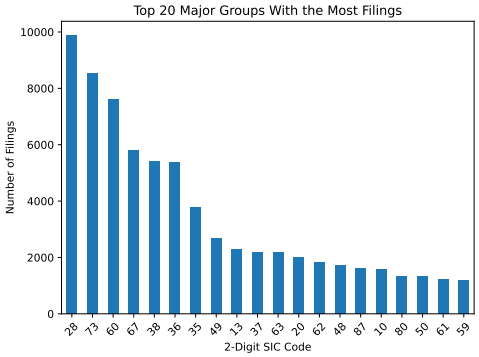
(a) Top 20 States by Total Filings



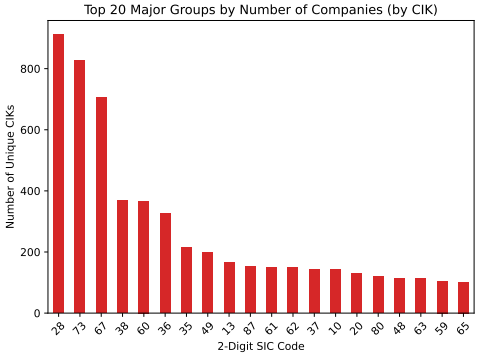
(b) Top 20 States by Number of Companies

Figure 6: Top 20 States

Industry Distribution

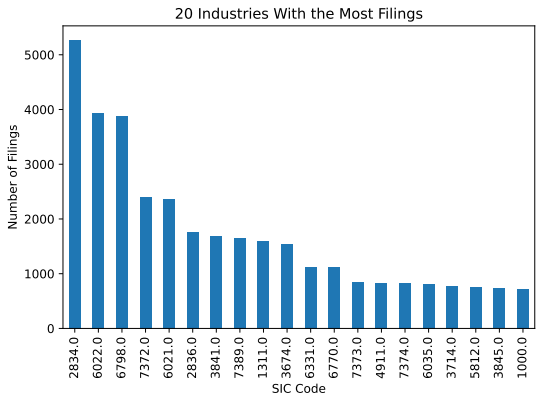


(a) Top 20 Major Groups by Filings



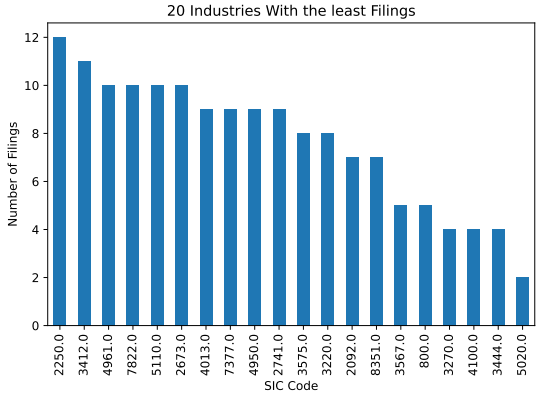
(b) Top 20 Major Groups by Number of Companies

Figure 7: Top 20 Major Groups



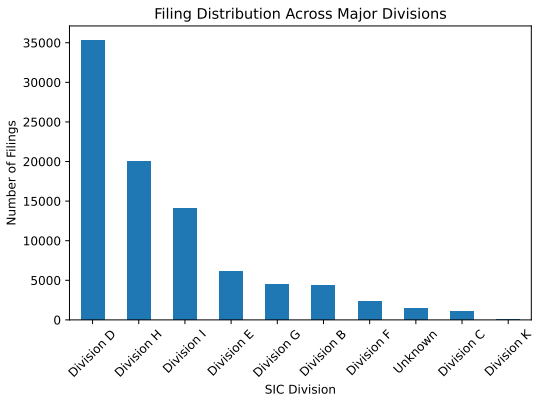
(a) Top 20 Industries by Filings

(b) Top 20 Industries by Number of Companies

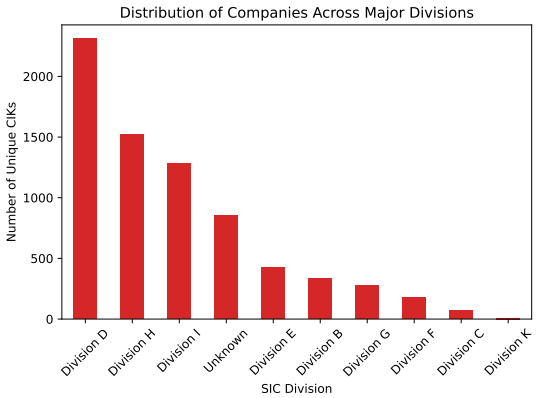


(c) Top 20 Industries by Number of Companies

Figure 8: Top and bottom 20 industries by the frequency of SIC codes in the data set.



(a) Top 20 Divisions by Filings



(b) Top 20 Divisions by Number of Companies

Figure 9: Top Major Divisions

Distribution of Filings

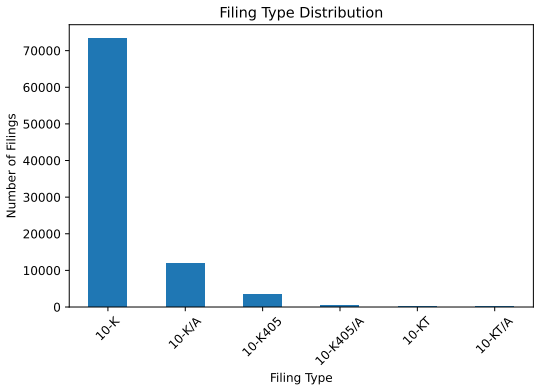
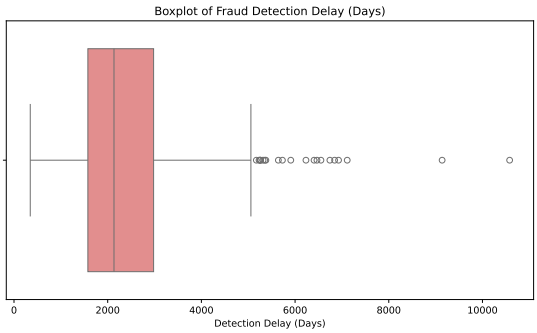
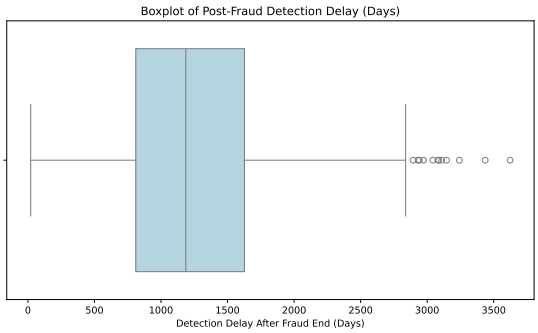


Figure 10: Distribution of 10-K Filing Types

Time Distribution



(a) Boxplot of Fraud Detection Delay



(b) Boxplot of Post-Fraud Detection Delay

Figure 11: Fraud Detection and Post-Fraud Detection Delays

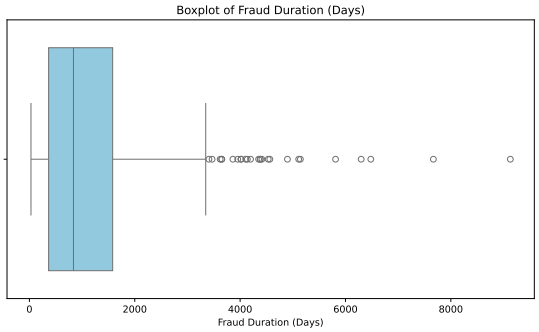


Figure 12: Boxplot of Fraud Duration

E.2 Text Length Distributions

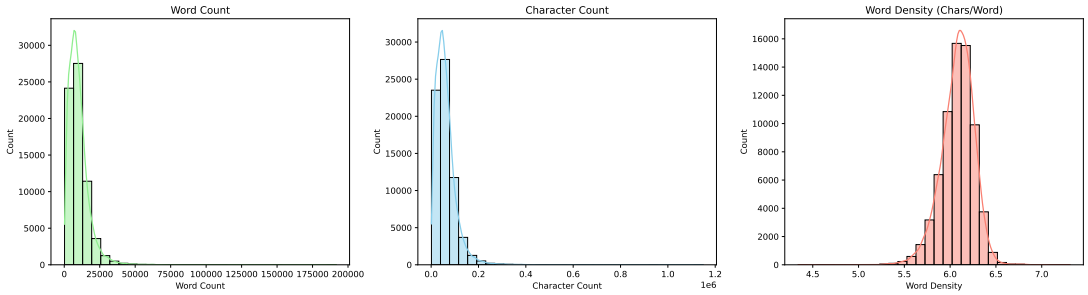


Figure 13: Text Length Distributions of Substantive MDA Sections

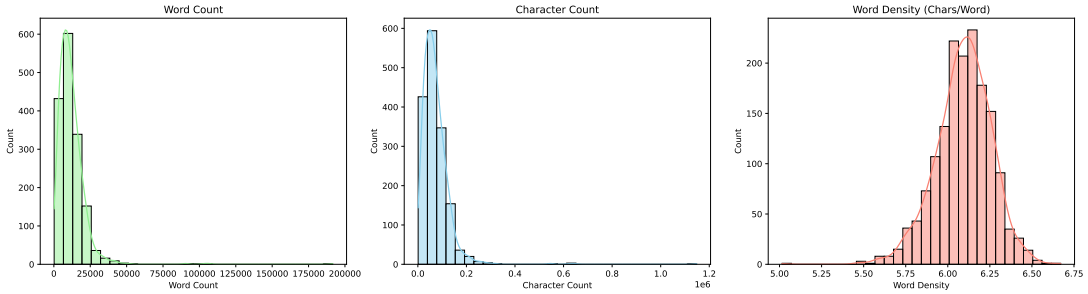
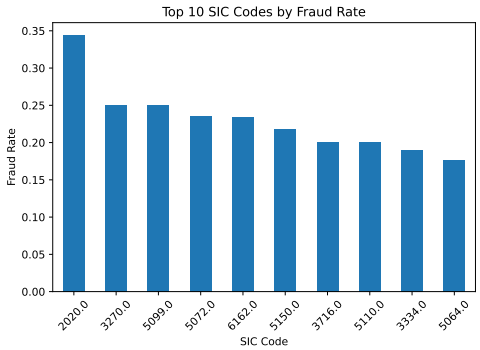
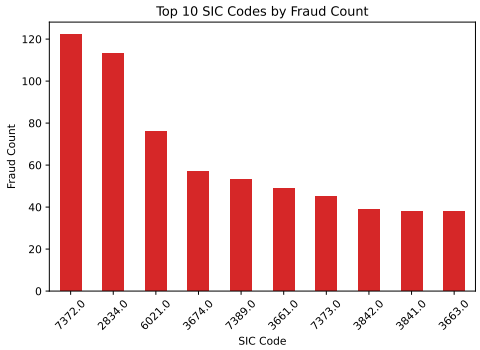


Figure 14: Text Length Distributions of Substantive Fraudulent MDA Sections

E.3 Fraud

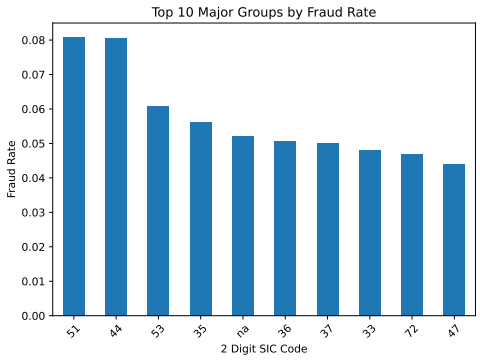


(a) Top 10 SIC Codes by Fraud Rate

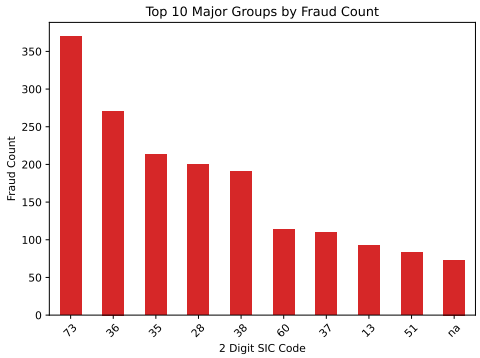


(b) Top 10 SIC Codes by Fraud Count

Figure 15: Top Fraud Industries

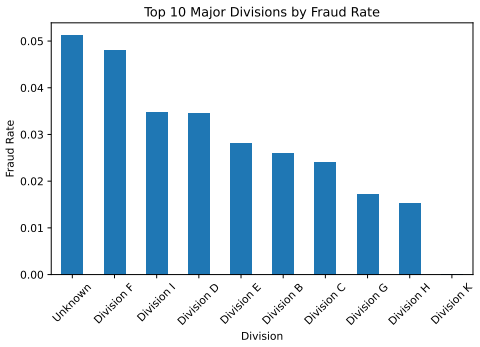


(a) Top 10 Major Groups by Fraud Rate

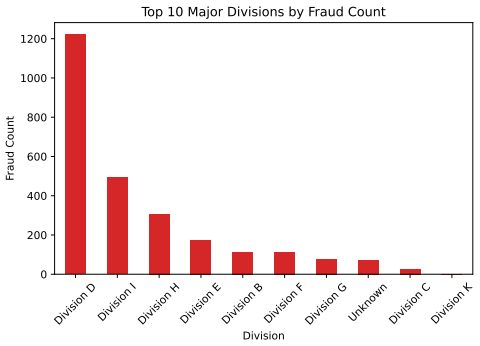


(b) Top 10 Major Groups by Fraud Count

Figure 16: Top Fraud Major Groups

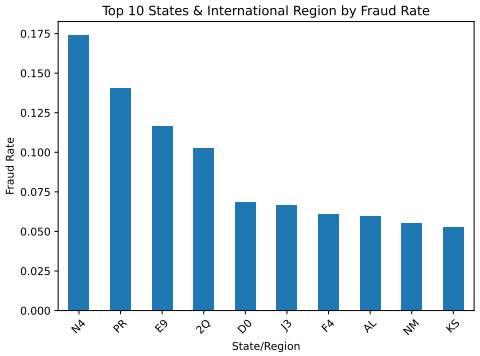


(a) Top 10 Major Divisions by Fraud Rate

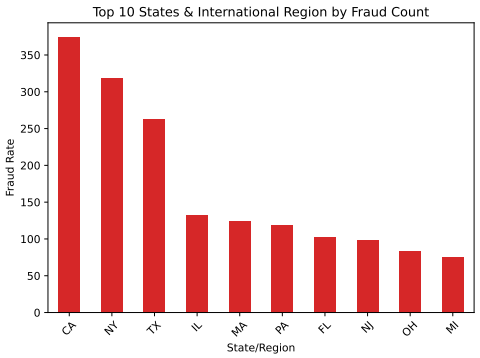


(b) Top 10 Major Divisions by Fraud Count

Figure 17: Top Fraud Major Divisions

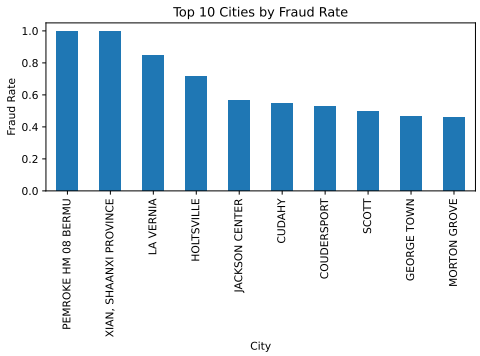


(a) Top 10 States by Fraud Rate

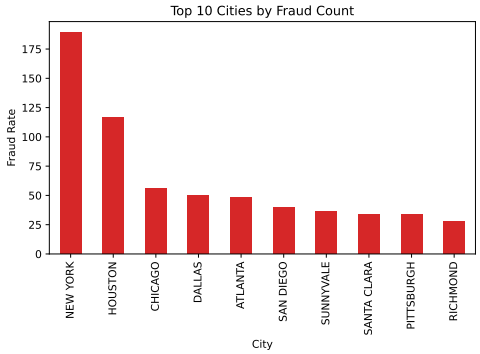


(b) Top 10 States by Fraud Count

Figure 18: Top Fraud US States and International Regions

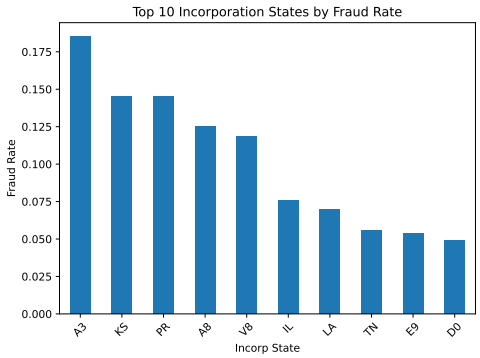


(a) Top 10 Cities by Fraud Rate

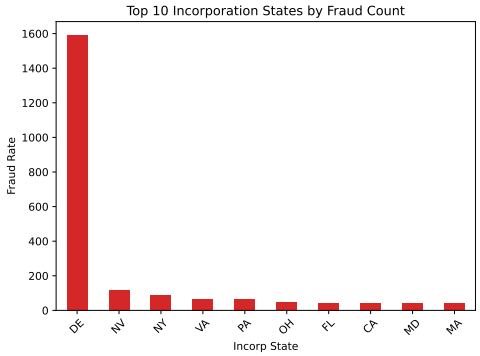


(b) Top 10 Cities by Fraud Count

Figure 19: Top Fraud Cities



(a) Top 10 Incorp States by Fraud Rate



(b) Top 10 Incorp States by Fraud Count

Figure 20: Top Fraud States of Incorporation

F Data Set Statistics

Year	Fraudulent Count	Filings Count	Fraudulent Fraction
1992	0	6	0.000
1993	8	370	0.022
1994	15	496	0.030
1995	24	1039	0.023
1996	60	1597	0.038
1997	101	1771	0.057
1998	120	1806	0.066
1999	162	1926	0.084
2000	204	2058	0.099
2001	213	2083	0.102
2002	185	2107	0.088
2003	140	2085	0.067
2004	128	2277	0.056
2005	102	2190	0.047
2006	83	2144	0.039
2007	71	2245	0.032
2008	84	2726	0.031
2009	97	2840	0.034
2010	107	2886	0.037
2011	86	3021	0.028
2012	84	3108	0.027
2013	64	3140	0.020
2014	68	3203	0.021
2015	72	3319	0.022
2016	74	3440	0.022
2017	70	3603	0.019
2018	63	3720	0.017
2019	51	3990	0.013
2020	27	4285	0.006
2021	22	4714	0.005
2022	11	4991	0.002
2023	2	5127	0.000
2024	0	4978	0.000
2025	0	162	0.000

Table 3: Fraudulent Cases by Year (Full Dataset)

Year	Fraudulent Count	Filings Count	Fraudulent Fraction
1993	0	113	0.000
1994	3	147	0.020
1995	9	410	0.022
1996	27	730	0.037
1997	47	849	0.055
1998	59	955	0.062
1999	79	1053	0.075
2000	99	1171	0.085
2001	114	1230	0.093
2002	105	1281	0.082
2003	88	1354	0.065
2004	72	1446	0.050
2005	70	1558	0.045
2006	56	1630	0.034
2007	52	1694	0.031
2008	57	2011	0.028
2009	61	2171	0.028
2010	72	2277	0.032
2011	64	2408	0.027
2012	65	2478	0.026
2013	49	2576	0.019
2014	52	2672	0.019
2015	49	2825	0.017
2016	59	2977	0.020
2017	53	3083	0.017
2018	46	3210	0.014
2019	39	3392	0.011
2020	21	3644	0.006
2021	17	4152	0.004
2022	10	4364	0.002
2023	2	4432	0.000
2024	0	4444	0.000
2025	0	157	0.000

Table 4: Fraudulent Cases by Year (Filtered Dataset)

G Full Results

Input Type	Class	Random Forest					XGBoost					Support
		Precision	Recall	F1-Score	AUC	Accuracy	Precision	Recall	F1-Score	AUC	Accuracy	
Tabular	0 (Non-Fraud)	0.98	0.99	0.99	–	–	0.98	0.97	0.98	–	–	2935
	1 (Fraud)	0.60	0.28	0.38	–	–	0.25	0.33	0.28	–	–	86
		–	–	–	0.71	0.97	–	–	–	0.72	0.95	3021
	Macro Avg	0.79	0.64	0.68	–	–	0.61	0.65	0.63	–	–	–
	Weighted Avg	0.97	0.97	0.97	–	–	0.96	0.95	0.96	–	–	–
Word Count	0 (Non-Fraud)	0.97	0.92	0.95	–	–	0.98	0.40	0.57	–	–	2344
	1 (Fraud)	0.04	0.12	0.06	–	–	0.03	0.69	0.06	–	–	64
		–	–	–	0.52	0.97	–	–	–	0.55	0.41	2408
	Macro Avg	0.51	0.52	0.51	–	–	0.50	0.54	0.31	–	–	–
	Weighted Avg	0.95	0.90	0.93	–	–	0.95	0.41	0.55	–	–	–
Sentence Embeddings	0 (Non-Fraud)	0.97	1.00	0.99	–	–	0.97	0.99	0.98	–	–	2344
	1 (Fraud)	0.00	0.00	0.00	–	–	0.11	0.06	0.08	–	–	64
		–	–	–	0.67	0.97	–	–	–	0.65	0.96	2408
	Macro Avg	0.49	0.50	0.49	–	–	0.54	0.52	0.53	–	–	–
	Weighted Avg	0.95	0.97	0.96	–	–	0.95	0.96	0.96	–	–	–
ModernBERT	0 (Non-Fraud)	0.97	1.00	0.99	–	–	0.97	0.99	0.98	–	–	2344
	1 (Fraud)	0.00	0.00	0.00	–	–	0.03	0.02	0.02	–	–	64
		–	–	–	0.62	0.97	–	–	–	0.61	0.96	2408
	Macro Avg	0.49	0.50	0.49	–	–	0.50	0.50	0.50	–	–	–
	Weighted Avg	0.95	0.97	0.96	–	–	0.95	0.96	0.95	–	–	–
29 LDA Topics	0 (Non-Fraud)	0.97	1.00	0.99	–	–	0.98	0.90	0.94	–	–	2344
	1 (Fraud)	0.00	0.00	0.00	–	–	0.07	0.28	0.12	–	–	64
		–	–	–	0.68	0.97	–	–	–	0.66	0.89	2408
	Macro Avg	0.49	0.50	0.49	–	–	0.53	0.59	0.53	–	–	–
	Weighted Avg	0.95	0.97	0.96	–	–	0.95	0.89	0.92	–	–	–
75 LDA Topics	0 (Non-Fraud)	0.97	1.00	0.99	–	–	0.98	0.97	0.97	–	–	2344
	1 (Fraud)	0.00	0.00	0.00	–	–	0.09	0.12	0.10	–	–	64
		–	–	–	0.64	0.97	–	–	–	0.64	0.94	2408
	Macro Avg	0.49	0.50	0.49	–	–	0.53	0.55	0.54	–	–	–
	Weighted Avg	0.95	0.97	0.96	–	–	0.95	0.94	0.95	–	–	–
100 LDA Topics	0 (Non-Fraud)	0.97	1.00	0.99	–	–	0.98	0.96	0.97	–	–	2344
	1 (Fraud)	0.00	0.00	0.00	–	–	0.11	0.16	0.13	–	–	64
		–	–	–	0.59	0.97	–	–	–	0.62	0.94	2408
	Macro Avg	0.49	0.50	0.49	–	–	0.54	0.56	0.55	–	–	–
	Weighted Avg	0.95	0.97	0.96	–	–	0.95	0.94	0.95	–	–	–

Table 5: Results Comparison: Random Forest vs XGBoost across Different Input Representations