EDGAR-CORPUS: Billions of Tokens Make The World Go Round

Lefteris Loukas[1,2]**, Manos Fergadiotis**[1]**, Ion Androutsopoulos**[1,2]**, Prodromos Malakasiotis**[1]

[1]EY AI Centre of Excellence in Document Intelligence, NCSR "Demokritos"

[2]Department of Informatics, Athens University of Economics and Business, Greece

Abstract

We release EDGAR-CORPUS, a novel corpus comprising annual reports from all the publicly traded companies in the US spanning a period of more than 25 years. To the best of our knowledge, EDGAR-CORPUS is the largest financial NLP corpus available to date. All the reports are downloaded, split into their corresponding items (sections), and provided in a clean, easy-to-use JSON format. We use EDGAR-CORPUS to train and release EDGAR-W2V, which are WORD2VEC embeddings for the financial domain. We employ these embeddings in a battery of financial NLP tasks and showcase their superiority over generic GloVe embeddings and other existing financial word embeddings. We also open-source EDGAR-CRAWLER, a toolkit that facilitates downloading and extracting future annual reports.

1 Introduction

Natural Language Processing (NLP) for economics and finance is a rapidly developing research area (Hahn et al., 2018, 2019; Chen et al., 2020; El-Haj et al., 2020). While financial data are usually reported in tables, much valuable information also lies in text. A prominent source of such textual data is the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) from the US Securities and Exchange (SEC) website that hosts filings of publicly traded companies.[1] In order to maintain transparency and regulate exchanges, SEC requires all public companies to periodically upload various reports, describing their financial status, as well as important events like acquisitions and bankruptcy.

Financial documents from EDGAR have been useful in a variety of tasks such as stock price prediction (Lee et al., 2014), risk analysis (Kogan et al., 2009), financial distress prediction (Gandhi et al.,

Corpora	Filings	Tokens	Companies	Years
Händschke et al. (2018)	Various	242M	270	2000-2015
Daudert and Ahmadi (2019)	Various	188M	60	1995-2018
Lee et al. (2014)	8-K	27.9M	500	2002-2012
Kogan et al. (2009)	10-K	247.7M	10,492	1996-2006
Tsai et al. (2016)	10-K	359M	7,341	1996-2013
EDGAR-CORPUS (ours)	10-K	**6.5B**	**38,009**	**1993-2020**

Table 1: Financial corpora derived from SEC (lower part) and other sources (upper part).

2019), and merger participants identification (Katsafados et al., 2021). However, there has not been an open-source, efficient tool to retrieve textual information from EDGAR. Researchers interested in economics and NLP often rely on heavily-paid subscription services or try to build web crawlers from scratch, often unsuccessfully. In the latter case, there are many technical challenges, especially when aiming to retrieve the annual reports of a specific firm for particular years or to extract the most relevant item sections from documents that may contain hundreds of pages. Thus, developing a web-crawling toolkit for EDGAR as well as releasing a large financial corpus in a clean, easy-to-use form would foster research in financial NLP.

In this paper, we release EDGAR-CORPUS, a novel financial corpus containing all the US annual reports (10-K filings) from 1993 to 2020.[2] Each report is provided in an easy-to-use JSON format containing all 20 sections and subsections (items) of a SEC annual report; different items provide useful information for different tasks in financial NLP. To the best of our knowledge, EDGAR-CORPUS is the largest publicly available financial corpus (Table 1). In addition, we use EDGAR-CORPUS to train and release WORD2VEC embeddings, dubbed EDGAR-W2V. We experimentally show that the new embeddings are more useful for financial NLP tasks than generic GloVe embeddings (Pennington et al., 2014) and other previously released financial WORD2VEC embeddings (Tsai et al., 2016).

Correspondence: eleftherios.loukas@ey.com, eleftheriosloukas@aueb.gr

[1]See https://www.sec.gov/edgar/searchedgar/companysearch.html for more information.

[2]EDGAR-CORPUS is available at: https://zenodo.org/record/5528490

13

Finally, to further facilitate future research in financial NLP, we open-source EDGAR-CRAWLER, the Python toolkit we developed to download and extract the text from the annual reports of publicly traded companies available at EDGAR.[3]

2 Related Work

There are few textual financial resources in the NLP literature. Händschke et al. (2018) published JOCo, a corpus of non-SEC annual and social responsibility reports for the top 270 US, UK, and German companies. Daudert and Ahmadi (2019) released CoFiF, the first financial corpus in the French language, comprising annual, semestrial, trimestrial, and reference business documents.

While some previous work has published document collections from EDGAR, those collections come with certain limitations. Kogan et al. (2009) published a collection of the Management's Discussion and Analysis Sections (Item 7) for all SEC company annual reports from 1996 to 2006. Tsai et al. (2016) updated that collection to include reports up to 2013 while also providing WORD2VEC embeddings. Finally, Lee et al. (2014) released a collection of 8-K reports from EDGAR, which announce significant firm events such as acquisitions or director resignations, from 2002 until 2012.

Compared to previous work, EDGAR-CORPUS contains all 20 items of the annual reports from all publicly traded companies in the US, covering a time period from 1993 to 2020. We believe that releasing the whole annual reports (with all 20 items) will facilitate several research directions in financial NLP (Loughran and McDonald, 2016). Also, EDGAR-CORPUS is much larger than previously published financial corpora in terms of tokens, number of companies, and year range (Table 1).

3 Creating EDGAR-CORPUS

3.1 Data and toolkit

Publicly-listed companies are required to submit 10-K filings (annual reports) every year. Each 10-K filing is a complete description of the company's economic activity during the corresponding fiscal year. Such reports also provide a full outline of risks, liabilities, corporate agreements, and operations. Furthermore, the documents provide an extensive analysis of the relevant sector industry and the marketplace as a whole.

	Item	Section Name
Part I	Item 1	Business
	Item 1A	Risk Factors
	Item 1B	Unresolved Staff Comments
	Item 2	Properties
	Item 3	Legal Proceedings
	Item 4	Mine Safety Disclosures
Part II	Item 5	Market
	Item 6	Consolidated Financial Data
	Item 7	Management's Discussion and Analysis
	Item 7A	Quantitative and Qualitative Disclosures about Market Risks
	Item 8	Financial Statements
	Item 9	Changes in and Disagreements With Accountants
	Item 9A	Controls and Procedures
	Item 9B	Other Information
Part III	Item 10	Directors, Executive Officers and Corporate Governance
	Item 11	Executive Compensation
	Item 12	Security Ownership of Certain Beneficial Owners
	Item 13	Certain Relationships and Related Transactions
	Item 14	Principal Accounting Fees and Services
Part IV	Item 15	Exhibits and Financial Statement Schedules Signatures

Table 2: The 20 different items of a 10-K report.

A 10-K report is organized in 4 parts and 20 different items (Table 2). Extracting specific items from documents with hundreds of pages usually requires manual work, which is time- and resource-intensive. To promote research in all possible directions, we extracted all available items using an extensive pre-processing and extraction pipeline.

In more detail, we developed EDGAR-CRAWLER, which we used to download the 10-K reports of all publicly traded companies in the US between the years 1993 and 2020. We then removed all tables to keep only the textual data, which were HTML-stripped,[4] cleaned and split into the different items by using regular expressions. The resulting dataset is EDGAR-CORPUS.

While there exist toolkits to download annual filings from EDGAR, they do not support the extraction of specific item sections.[5] This is particularly important since researchers often rely on certain items in their experimental setup. For example, Goel and Gangolly (2012), Purda and Skillicorn (2015), and Goel and Uzuner (2016) perform textual analysis on Item 7 to detect corporate fraud. Katsafados et al. (2020) combine Item 7 and Item

---

[3]EDGAR-CRAWLER is available at: https://github.com/nlpaueb/edgar-crawler

[4]We use Beautiful Soup (https://beautiful-soup-4.readthedocs.io/en/latest).

[5]For example, sec-edgar can download complete HTML reports (with images and tables), but it does not produce clean item-specific text.

1A to detect Initial Public Offering (IPO) underpricing, while Moriarty et al. (2019) combine Item 1 with Item 7 to predict mergers and acquisitions.

Apart from EDGAR-CORPUS, we also release EDGAR-CRAWLER, the toolkit we developed to create EDGAR-CORPUS, to facilitate future research based on textual data from EDGAR. EDGAR-CRAWLER consists of two Python modules that support its main functions:

edgar_crawler.py is used to download 10-K reports in batch or for specific companies that are of interest to the user.

extract_items.py extracts the text of all or particular items from 10-K reports. Each item's text becomes a separate JSON key-value pair (Figure 1).

```
{} JSON
  ■ filename : "881790_10K_2012.htm"
  ■ cik : "881790"
  ■ year : "2012"
  ■ section_1 : "Item 1. Business ..."
  ■ section_1A : "Item 1A. Risk Factors: The company identifies multiple risks ..."
  ■ section_1B : "Item 1B. Unresolved Staff Comments: None yet."
  ■ .. : ".."
  ■ section_7 : "Item 7. Management's Discussion and Analysis: This section should ..."
  ■ section_7A : "Item 7A. Quantitative and Qualitative Disclosures ..."
  ■ section_8 : "Item 8. Financial Statements and Supplementary Data ..."
  ■ ... : "..."
  ■ section_15 : "Item 15. Exhibits and Financial Statement Schedules ..."
```

Figure 1: An example of a 10-K report in JSON format as downloaded and extracted by EDGAR-CRAWLER; *filename* is the downloaded 10-K file; *cik* is the Company Index Key; *year* is the corresponding Fiscal Year.

3.2 Word embeddings

To facilitate financial NLP research, we used EDGAR-CORPUS to train WORD2VEC embeddings (EDGAR-W2V), which can be used for downstream tasks, such as financial text classification or summarization. We used WORD2VEC's skip-gram model (Mikolov et al., 2013a,b) with default parameters as implemented by GENSIM (Řehůřek and Sojka, 2010) to generate 200-dimensional WORD2VEC embeddings for a vocabulary of 100K words. The word tokens are generated using SPACY (Honnibal et al., 2020). We also release EDGAR-W2V.[6]

To illustrate the quality of EDGAR-W2V embeddings, in Figure 2 we visualize sampled words from seven different entity types, i.e., *location*, *industry*, *company*, *year*, *month*, *number*, and *financial term*, after applying dimensionality reduction with the UMAP algorithm (McInnes et al., 2018). The
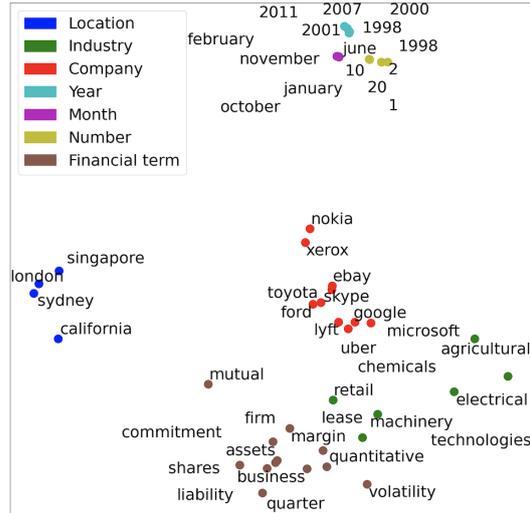

Figure 2: Visualization of the EDGAR-W2V embeddings. Different colors indicate different entity types.

financial terms are randomly sampled from the Investopedia Financial Terms Dictionary.[7] In addition, companies and industries are randomly sampled from well-known industry sectors and publicly traded stocks. Finally, the words belonging to the remaining entity types are randomly sampled from gazetteers. Figure 2 shows that words belonging to the same entity type form clear clusters in the 2-dimensional space indicating that EDGAR-W2V embeddings manage to capture the underlying financial semantics of the vocabulary.

To further highlight the semantics captured by EDGAR-W2V embeddings, we retrieved the 5 nearest neighbors, according to cosine similarity, for commonly used financial terms (Table 3).[8] As shown, all the nearest neighbors are highly related to the corresponding term. For instance, the word *economy* is correctly associated with terms indicating the slowdown of the economy happening during the past few years, e.g., *downturn*, *recession*, or *slowdown*. Also, *market* is correctly related with

economy	competitor	market	national	investor
downturn	competitive	marketplace	association	institutional
recession	competing	industry	regional	shareholder
slowdown	dominant	prices	nationwide	relations
sluggish	advantages	illiquidity	american	purchaser
stagnant	competition	prevailing	zions	creditor

Table 3: Sample words from EDGAR-W2V embeddings (top row) and their corresponding nearest neighbors (columns) based on cosine similarity.

15

words such as *marketplace*, *industry*, and *prices*.

4 Experiments on financial NLP tasks

We also compare EDGAR-W2V embeddings against generic GloVe (Pennington et al., 2014) embeddings[9] and the financial embeddings of Tsai et al. (2016) in three financial NLP tasks. For each task, we use the same model and we only alter the embeddings component. In addition, we use the same pre-processing during the creation of the vocabulary of the embeddings in each case.

FinSim-3 (Juyeon Kang and Gan, 2021) provides a set of business and economic terms and the task is to classify them into the most relevant hypernym from a set of 17 possible hypernyms from the Financial Industry Business Ontology (FIBO).[10] Example hypernyms include *Credit Index*, *Bonds*, and *Stocks*. We tackle the problem with a multinomial logistic regression model, which, given the embedding of an economic term, classifies the term to one of the 17 possible hypernyms. Since FinSim-3 is a recently completed challenge, the true labels for the test data were not available. Therefore, we use a stratified 10-fold cross-validation. We report accuracy and the average rank of the correct hypernym as evaluation measures. For the latter, the 17 hypernyms are sorted according to the model's probabilities. A perfect model, i.e., one that would always rank the correct hypernym first, would have an average rank of 1.

Financial tagging (FinT) is an in-house sequence labeling problem for financial documents. The task is to annotate financial reports with word-level tags from an accounting taxonomy. To tackle the problem, we use a BILSTM encoder operating over word embeddings with a shared multinomial regression that predicts the correct tag for each word from the corresponding BILSTM state. We report the F1 score micro-averaged across all tags.

FiQA Open Challenge (Maia et al., 2018) is a sentiment analysis regression challenge over financial texts. It contains financial tweets annotated by domain experts with a sentiment score in the range [-1, 1], with 1 denoting the most positive score. For this problem, we employ a BILSTM encoder which operates over word embeddings, and a linear regressor operating over the last hidden state of the

---

BILSTM. Since we do not have access to the test set in this task, we use a 10-fold cross-validation. We evaluate the results using Mean Squared Error (MSE) and R-squared (R^2).

Across all tasks, EDGAR-W2V outperforms GloVe, showing that in-domain knowledge is critical in financial NLP problems (Table 4). The gains are more substantial in FinSim-3 and FinT, which rely to a larger extent on understanding highly technical economics discourse. Interestingly, the in-domain embeddings of Tsai et al. (2016) are comparable to the generic GloVe embeddings in two of the three tasks. One possible reason is that Tsai et al. (2016) employed stemming during the creation of the embeddings vocabulary, which might have contributed noise to the models due to loss of information.

	FinSim-3		FinT	FiQA	
	Acc. ↑	Rank ↓	F1 ↑	MSE ↓	R^2 ↑
GloVe	85.3	1.26	75.8	0.151	0.119
Tsai et al. (2016)	84.9	1.27	75.3	0.142	0.169
EDGAR-W2V (ours)	**87.9**	**1.21**	**77.3**	**0.141**	**0.176**

Table 4: Results across financial NLP tasks, with different word embeddings. We report averages over 3 runs with different random seeds. The standard deviations were very small and are omitted for brevity.

5 Conclusions and Future Work

We introduced and released EDGAR-CORPUS, a novel NLP corpus for the financial domain. To the best of our knowledge, EDGAR-CORPUS is the largest financial corpus available. It contains textual data from annual reports published in EDGAR, the repository for all US publicly traded companies, covering a period of more than 25 years. All the reports are split into their corresponding items (sections) and are provided in a clean, easy-to-use JSON format. We also released EDGAR-CRAWLER, a toolkit for downloading and extracting the reports. To showcase the impact of EDGAR-CORPUS, we used it to train and release EDGAR-W2V, which are financial WORD2VEC embeddings. After illustrating the quality of EDGAR-W2V embeddings, we also showed their usefulness in three financial NLP tasks, where they outperformed generic GloVe embeddings and other financial embeddings.

In future work, we plan to extend EDGAR-CRAWLER to support additional types of documents (e.g., 10-Q, 8-K) and to leverage EDGAR-CORPUS to explore transfer learning for the financial domain, which is vastly understudied.

16

Disclaimer

This publication contains information in summary form and is therefore intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Member firms of the global EY organization cannot accept responsibility for loss to any person relying on this article.

References

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors. 2020. *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*. Kyoto, Japan.

Tobias Daudert and Sina Ahmadi. 2019. CoFiF: A corpus of financial reports in French language. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 21–26, Macao, China.

Mahmoud El-Haj, Vasiliki Athanasakou, Sira Ferradans, Catherine Salzedo, Ans Elhag, Houda Bouamor, Marina Litvak, Paul Rayson, George Giannakopoulos, and Nikiforos Pittaras, editors. 2020. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. Barcelona, Spain (Online).

Priyank Gandhi, Tim Loughran, and Bill McDonald. 2019. Using Annual Report Sentiment as a Proxy for Financial Distress in U.S. Banks. *Journal of Behavioral Finance*, 20(4):424–436.

Sunita Goel and Jagdish Gangolly. 2012. Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management*, 19(2):75–89.

Sunita Goel and Özlem Uzuner. 2016. Do sentiments matter in fraud detection? estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23:215–239.

Udo Hahn, Véronique Hoste, and Ming-Feng Tsai, editors. 2018. *Proceedings of the First Workshop on Economics and Natural Language Processing*. Melbourne, Australia.

Udo Hahn, Véronique Hoste, and Zhu Zhang, editors. 2019. *Proceedings of the Second Workshop on Economics and Natural Language Processing*. Hong Kong.

Sebastian G.M. Händschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn. 2018. A corpus of corporate annual and social responsibility reports: 280 million tokens of balanced organizational writing. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 20–31, Melbourne, Australia.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Sandra Bellato Juyeon Kang, Ismail El Maarouf and Mei Gan. 2021. Finsim-3: The 3rd shared task on learning semantic similarities for the financial domain. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, Held Online.

Apostolos G. Katsafados, Ion Androutsopoulos, Ilias Chalkidis, Emmanouel Fergadiotis, George N. Leledakis, and Emmanouil G. Pyrgiotakis. 2021. Using textual analysis to identify merger participants: Evidence from the u.s. banking industry. *Finance Research Letters*, page 101949.

Apostolos G. Katsafados, Ion Androutsopoulos, Ilias Chalkidis, Manos Fergadiotis, George N. Leledakis, and Emmanouil G. Pyrgiotakis. 2020. Textual Information and IPO Underpricing: A Machine Learning Approach. MPRA Paper 103813, University Library of Munich, Germany.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, Boulder, Colorado.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the importance of text analysis for stock price prediction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1170–1175, Reykjavik, Iceland.

Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Macedo Maia, S. Handschuh, A. Freitas, Brian Davis, R. McDermott, M. Zarrouk, and A. Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. *Companion Proceedings of the The Web Conference 2018*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Ryan Moriarty, Howard Ly, Ellie Lan, and Suzanne McIntosh. 2019. Deal or no deal: Predicting mergers and acquisitions at scale. *2019 IEEE International Conference on Big Data (Big Data)*, pages 5552–5558.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Lynnette Purda and David Skillicorn. 2015. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, 32(3):1193–1223.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. http://is.muni.cz/publication/884893/en.

Ming-Feng Tsai, Chuan-Ju Wang, and Po-Chuan Chien. 2016. Discovering finance keywords via continuous-space language models. *ACM Transactions on Management Information Systems*, 7(3).