Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2261–2267
Marseille, 11–16 May 2020

NLP Analytics in Finance with DoRe: a French 257M Tokens Corpus of Corporate Annual Reports

Corentin Masson[*†], **Patrick Paroubek**[†]
[*]AMF, [†]LIMSI-CNRS-U. Paris-Saclay
AMF 17 Place de la Bourse, 75002, Paris F
LIMSI Bât. 507, Rue John Von Neumann, 91400 Orsay F
{corentin.masson, pap}@limsi.fr

Abstract

Recent advances in neural computing and word embeddings for semantic processing open many new applications areas which had been left unaddressed so far because of inadequate language understanding capacity. But this new kind of approaches rely even more on training data to be operational. Corpora for financial applications exists, but most of them concern stock market prediction and are in English. To address this need for the French language and regulation oriented applications which require a deeper understanding of the text content, we hereby present "DoRe", a French and dialectal French Corpus for NLP analytics in Finance, Regulation and Investment. This corpus is composed of: (a) 2350 Annual Reports from 336 companies among the most capitalized companies in: France (Euronext Paris) & Belgium (Euronext Brussels), covering a time frame from 2009 to 2019, and (b) related MetaData containing information for each company about its ISIN code, capitalization and sector. This corpus is designed to be as modular as possible in order to allow for maximum reuse in different tasks pertaining to Economics, Finance and Regulation. After presenting existing resources, we relate the construction of the DoRe corpus and the rationale behind our choices, concluding on the spectrum of possible uses of this new resource for NLP applications.

Keywords: Corpus, French, Finance, Annual Reports

1. Introduction

Since the emergence of context insensitive neural word-embeddings (Mikolov et al., 2013) and the revolution of context sensitive pre-trained language models like ELMO (Peters et al., 2018), BERT (Devlin et al., 2019) and XL-NET (Yang et al., 2019), Natural Language Processing (NLP) is gaining more and more popularity as the number of profitable use-cases in various industries identified by both research & private sectors is increasing. Because of the intrinsic language complexity due to its high combinatorics, the reason behind the progress of State-Of-The-Art pretrained models (Shoeybi et al., 2019) is often the availability of more training data and the corresponding computing power. Those models are also often trained with accessible general Internet Data as Wikipedia, offering wide vocabularies and topics, allowing NLP tasks to be split into two subtasks: learning the Language Model and then adapting it to the specificities of the target task. The availability of word embeddings pre-trained on huge amounts of generic (Mikolov et al., 2013; Devlin et al., 2019; Peters et al., 2018; Yang et al., 2019) or specialized (Lee et al., 2019; Beltagy et al., 2019) corpora lowers the barrier-to-entry required to reach a minimal performance for many semantic tasks (Devlin et al., 2019). Marketing, Financial, Economic and Regulatory sectors are no strangers to this NLP research interest trend, opening a whole new research area (Lou, 2019; Hiew et al., 2019). Different domain-specific corpora were built those last years, such as Corporate Annual Report Corpora (Kogan et al., 2009; Händschke et al., 2018), Financial News Corpora (Ding et al., 2014), Financial Twitter based Corpora (Malo et al., 2013; Cortis et al., 2017). The vast majority being in English, there is a lack of domain specific corpora in other languages, including French. Multilingual models and Zero-shot learning still lacking the accuracy of monolingual language models, there is a Financial domain research interest in building domain-specific corpora for different languages; or even better, parallel corpora, which would then enable gauging the impact the characteristics of a given language have on task performance. As in the JOCo corpus (Händschke et al., 2018) and the famous 10-K filings corpus (Kogan et al., 2009), the most representative kind of financial text document for analyzing companies' communications are Annual Reports (AR). Since they are mandatory and produced yearly, they contain information that companies have to share with the market. Sometimes they also hold, as a side effect, clues about information the companies try to hide from investors (Zaki and Theodoulidis, 2013; Purda and Skillicorn, 2015; Hajek and Henriques, 2017). To address the corpus needs from French and its dialects, we collected a corpus of AR from the most capitalized companies in France and Belgium.

2. Related Works

Finance is an historic sector for the use of statistics and other quantitative methods for decision making. Volatility estimation, inter-securities correlation & causal relationships on structured data are still widely used for portfolio optimization. Since the beginning of 21th century, researchers tried to use information from public documents to achieve investment-related tasks (Kogan et al., 2009; Back et al., 2001) such as market volatility and stock variation prediction or company stability estimation. This arbitration is made possible by the hypothesis that Financial Markets might not be as efficient as they are supposed to be in neo-classic economic theory (Fama, 1970), because of information asymmetry. Therefore, it is possible to have an unbiased algorithm digging deep into the mass of accessible

documents to yield indications about future performances of a company. Research has also intensified on regulator's side, where researchers try to automatically detect fraudulent activities using public information. The fraudulent activities targeted on these occasion include Market Manipulation (Tumarkin and R., 2001), High Yield Investments, and Financial Statement Frauds (Zaki and Theodoulidis, 2013; Skillicorn and Purda, 2012; Hajek and Henriques, 2017). The approaches aim at identifying vocabulary indicative of a potential fraud. In order to solve these tasks, both annotated and unannotated corpora were built using different sources of domain-specific textual data. We identify 6 kinds of financial texts used for forecasting and differentiate them in terms of Frequency, Subjectivity and Length (Xing et al., 2018): Corporate disclosures, Financial reports, Professional periodicals, Aggregated news, Message boards and Social medias. News corpus from Professional Periodicals were the predominant kind of Textual Data related to Finance, Business and Economics (Penn Treebank) in the early 90's and were used for Financial Forecasting like predicting Stock Market movement or Market volatility to help investors optimize their portfolio (Xing et al., 2018). (Back et al., 2001) and (Kloptchenko et al., 2002) were the firsts to use financial reports to analyze company performance alongside quantitative data. A lot of work followed until today, with the famous 10-K corpus (Kogan et al., 2009), the JoCo corpus (Händschke et al., 2018) for tasks like Board members' relationships extraction, Risks identification, Financial Forecasting and Financial Statements fraud detection. Finance being a vocabulary specific sector, the "upgrade" of recent language models, fine-tuned on Financial Corpus allows for a better models precision (Araci, 2019), opening a whole new set of possibilities for companies which do not have access to large amounts of data to nevertheless have relatively high accuracy models. Even with the growing interest of research NLP applications in Finance, there was almost no corpus available in French or its dialects except the recent CoFiF Corpus, based on CAC40 and CAC Next 20 regulated information reports (Daudert and Ahmadi, 2019). There is currently no way to evaluate the applications previously mentioned on data from markets using French and or its variants. Regulations differ between countries, companies having to make available different kind of information in their official documents depending on their Regulation Authority. Ways to explain current strategy and company's risks might then differ, and the temptation to take advantage of this freedom of form to hide mandatory information is inevitable. It is also known that pre-trained language models on general domain loose precision on domain specific corpora, the same goes for multi-lingual language models against monolingual language models trained on an equivalent amount of sentences (Kamath Ramachandra Rao, 2020). Even with a financial language model (Araci, 2019) and with a French pre-trained language model (Martin et al., 2019), we have no way to obtain an optimal language model on financial domain without a Financial Data Corpus in French and its dialects. Our corpus then has two goals. The first one is to make available to researchers a French and dialectal French Financial Corpus based on Annual Reports from France and Belgium, enlarging the research area in this domain. The second one is, with the amount of data available, to fine-tune a French Language model to financial domain-specific Data. For this, we provide the largest French Financial corpus of Annual Reports with about 2350 documents ranging from 30 to 400 pages.

3. Corpus Construction

Official financial public information is a highly regulated vector of communication due to the huge impact those might have on markets and real economy. As described by (Xing et al., 2018), financial information can be separated between 6 categories with 4 of them being regulated: board messages, press releases, annual reports and ("quarterly reports, ..."). Published once a year by each regulated company, Annual Reports (ARs) are the most objective and comprehensive source of information about a company' strategy and management for current and prospective shareholders, the stock exchange, governments and regulation authorities. For the case of France and as described in "Code Monétaire et Financier", listed companies with a minimal capitalization and size have to release an annual report for shareholders which has to contain all the information for them to assess value, financial stability, results, perspectives and risks the company faces. If a company voluntarily hides a vital information for the market's wealth, they fall under the Financial Authority it depends on and might become the subject of a lawsuit for Statement Fraud. Therefore, ARs are the most comprehensive public document to evaluate a company potential and strategy. With regards to the existing datasets based on ARs such as the 10-K filings (Kogan et al., 2009) and the recent JOCo corpus (Händschke et al., 2018), ours is a mix of both of them while smaller in size. 10-K corpus is a sample of US 10-k forms, XBRL[1] structured documents mandated by the Securities Exchange Commission, ranging from 90's to 2006. While structured for computational analysis, the 10-k corpus does not contain as much mandatory information as ARs and are only available in the US, ARs being available worldwide. Taking this into account, the JOCo corpus uses ARs as a backbone to their corpus and adds Corporate Social Responsibility Reports (CSRR) combined to a small metadata table. For a comprehensive and representative corpus, we then choose ARs as the core of our corpus with a comprehensive metadata table to allow cross-sector, cross-index and cross-country comparisons.

3.1. Data Selection

ARs in our corpus are selected following two criteria, first selecting countries present in the corpus and then stock indices from which we include companies. Limiting ourselves to French and variant of French speaking countries, we picked the 2 most developed Financial Centers: Paris (France), Brussels (Belgium). For each of these countries, we selected 3 stock indices referenced on Euronext France/Belgium. The first ones include the most capitalized companies for each country, we took companies from the CAC 40, BEL 20, getting 60 companies with the biggest

[1]Specific XML format for Business Reporting

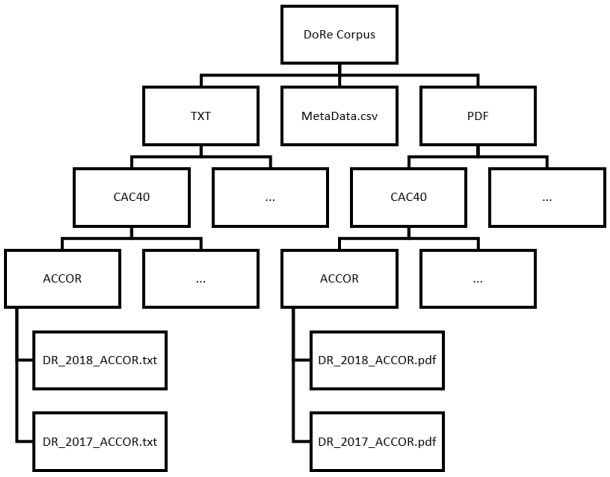

Figure 1: DoRe Dataset Structure

capitalization. Next, we added reports from mid-sized companies selected from CAC 60 and BELMid and then reports from small-sized companies from CAC90, BELSmall. Unlike JOCo's corpus, we choose to include all listed companies in order to give the user as much freedom as he needs to modulate the dataset for his own task. For this goal, we also built a metadata table of all of these companies to help the user modulate it. This table contains the ISIN code, capitalization information, sector of activity following ICV Sectorial Classification, size and market price of the company code at the publication date of this paper (December 2019). Most of the collected ARs range from 2009 to 2019 and we were limited by the availability of documents on regulatory authorities and companies' websites, as depicted in Figure 2.

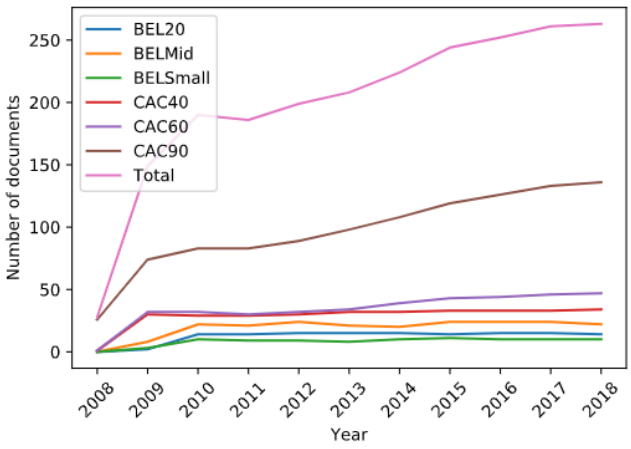

Figure 2: Number of documents per index per year.

3.2. Data Aquisition and Cleansing

Most of the ARs were manually collected on AMF France and FSMA's websites but we sometimes had to search on the company's website for missing ARs. Metadata was collected on Euronext using ISIN codes as IDs for each company, allowing merges with others Datasets about these companies. Collected ARs are available in the corpus in their original PDF format, in raw TXT using the free MuPDF tool from artifex.com and also in cleaned TXT format. Converting PDFs to a TXT format is not an easy task and often implies mistakes: non-textual data such as tables, page numbers, unrecognized characters, superfluous line breaks, misplaced boxes in the PDF format merging non-consecutives sentences. In order to create a near noise-free distribution of the Corpus, we followed (Händschke et al., 2018) pre-processing steps to get the most accurate conversion we could without losing much information. These steps consisted in removing superfluous line breaks, page numbers, tables and also rarely occurring character sequences often related to mistakes in the conversion.
Even with these processing steps, some of the conversions failed, resulting in unreadable ARs, such as AIR LIQUIDE's AR from 2018 (lines 653-654):

```
Être un leader de son industrie
%QPVTKDWGT^C‌]^CWP^COQPFG^CRNWU^CFWTCDNG
```

Figure 3: Example of error in the format conversion from pdf to txt.

Extracted and cleaned as is, the corpus will allow researchers to perform experiments on both the textual information expressed by listed companies of these countries and metadata we extracted from various websites such as market capitalization, stock price, sector, etc...

4. Corpus Exploration

Countries having different legislation, we expect the content of Annual Reports to vary in terms of exhaustiveness, quality of information, linguistics aspects and methods to hide information. We also expect those differences between sectors and market capitalization, analysis made possible thanks to the metadata about these companies. To provide an idea of the content of our Corpus, we'll provide some basic data analysis, present a task that will interest us in future works and then share a self-trained Language Model.

4.1. Corpus Analysis

Based on the cleaned text version of the corpus, we used NLTK.org tools to count tokens and sentences for all the reports, grouped by Indexes. As depicted in Table 1. our corpus currently sums up to 2350 corporate annual reports with a vast majority being from France because of a bigger Financial market. Our corpus contains more than 257M tokens (5.7M sentences), compared to 4,5M, 30M, 282M and 46B for respectively Penn Treebank (Marcus et al., 1993), TRC2-financial (Thomson Reuters Text Research Collection), JOCO Corpus (Händschke et al., 2018) and OSCAR's French subset (Suárez et al., 2019).
As shown in Table 1 there is a difference between lengths of ARs between countries in the DoRe corpus coming from different regulation policies. All European Financial Authorities follows the European Securities Markets Authority (ESMA) guidelines and apply them to their own country with their own interpretation & freedom. French Financial

Index	Tokens	Sent.	Ann. Report	Nbr. of Comp.
CAC40	78 044 159	1 693 103	392	40
CAC60	60 550 457	1 309 429	404	62
CAC90	88 552 491	1 984 244	1 121	182
BEL20	11 599 676	305 070	133	20
BELMid	12 958 311	334 448	210	39
BELSmall	5 350 166	143 375	90	23
Total	257 055 260	5 769 669	2350	366

Table 1: Documents, Words and Sentences distribution

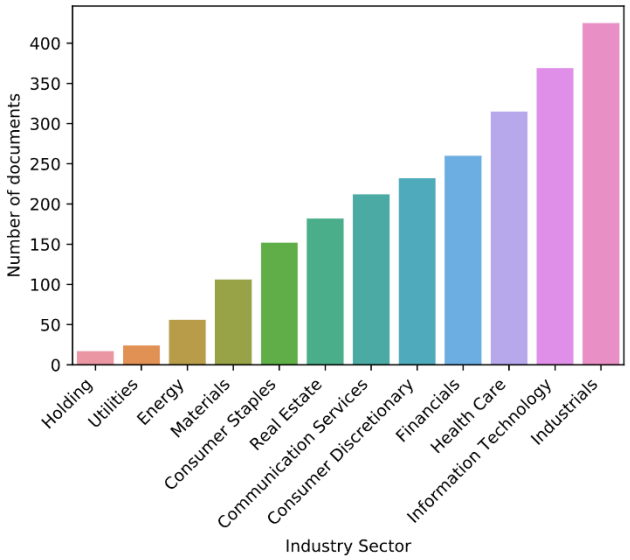

Figure 4: Number of document per sector.

Market Authority (AMF) is known to have a complex Annual Report (named "Document de Référence", where the name DoRe is from) asking for more information about the company than most of other national authorities.

Our Corpus is composed of 12 different sectors, following ICB sectorial classification. As shown in Figure 4, sectors distribution is not uniform but decrease linearly between sectors. Also, the repartition of those sectors vary between countries, due to countries market's structure such as France having more "Consumer Staples"[2] companies due to the important share of the agricultural sector.

4.2. Risk Factor

To show some examples of our corpus, we'll take the place of a regulator or an investor which might be interested in the discourse employed by a company to describe the risks it is facing and its answers to them. For an investor, the interest is in finding the asset with less risk and greater reward, which means looking for companies aware of their own risks and with a positive sentiment on the related section in the AR. Regulators, on the contrary, will try to find potential Statement Fraud such as the ENRON case (Ranjan Das et al., 2017). Research on this subject has addressed risk sentence extraction (Liu et al., 2018), used full reports labeled as fraudulent or non-fraudulent to assess Management fraud risk (Zaki and Theodoulidis, 2013) or the potential failure of a company.

ARs being highly regulated long documents, they have to follow a certain structure in which a company Risk Factors are presented as a whole section. During the PDF to TXT conversion the document structure is lost and we can't, like in XBRL documents, directly extract the section we're interested in. To study risks and where companies place their Risk Sections in their ARs, we extracted the various "risque" (risk) word forms and their position. In order to see if Risk Sections are often placed in the same parts of ARs we plotted the distribution of risk vocabulary along documents of the corpus. Figure 5. shows that there is no pattern except that this section rarely occur near the end of ARs.

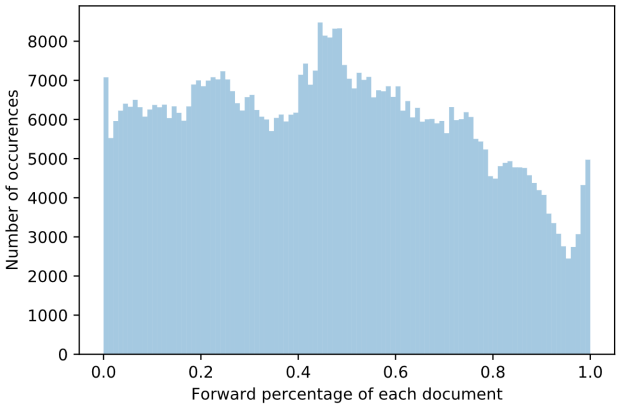

Figure 5: Risk words distribution in Corpus

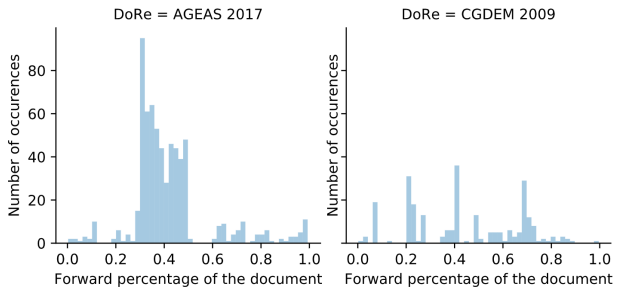

Figure 6: Risk words distribution between documents

Then, in order to know if the Risk Section is clearly identifiable, Figure 6. shows the distribution of "risque" (risk) and its variants along two documents. They were selected as being representative of standardized ARs and of much less organized ones. We observe that in the first one, the risk section is easily detectable whereas it is not trivial for the second one.

Therefore, we cannot only use "risque" word to isolate the Risk Section in Annual Reports, which can be cause by some risk sentences not containing risk vocabulary but markers of uncertainty.

[2]Essential products like food, basic necessities etc.

L'évolution de la situation géopolitique expose le Groupe à un risque d'attaque terroriste, et ce dans la majorité de ses pays d'implantation.

Sentence 1. Risk Presentation Sentence.

Sentence 1. depicts the terrorism risk presented in an AR from the DoRe Corpus. Since the company is making business in every continent and sometimes in politically unstable countries, the risk has to appear in its Annual Report. In this segment, two "Risk Factors" are presented, the second one ("attaque terroriste" (terrorist attack) being a consequence of the first one "evolution de la situation géopolitique" (evolution of the geopolitical situation)) and might affect the target "le Groupe" (the Group). Risk indicators are "expose" (expose) and "risqué" (risked), making it possible for us to automatically extract those risk factors.

After the evocation of the risk factor, the company often present how it plans to handle it. Those answers can be useful to assess if the company is trying to fool the investor using complicated sentences and fuzzy vocabulary (Gao and Srivastava, 2011; Purda and Skillicorn, 2015).

C'est pourquoi, afin de les protéger au mieux contre les principales menaces auxquelles ils sont exposés, le Groupe s'est doté d'une stratégie de Sûreté-Sécurité adaptée à l'intensité des risques identifiés. Elle repose sur une organisation, une veille et des mesures de sécurité appropriées et sujettes à évolution en fonction de la situation

Sentence 2. Risk Answer Sentence

As an example, this answer to terrorism risk and political instability gives little information about the company strategy. Even an aware investor can't find out if the company is prepared for this kind of risk or not and the extent to which its business can be hurt in these countries.

5. Baseline Language Model

Annual Reporting of companies being highly regulated and at the responsibility of these companies, their distribution is limited to themselves and the responsible Regulation Authority by Intellectual Property Rights (IPR). We then can't openly distribute the DoRe Corpus except for Research and Educational Purposes[3]. As a substitute, we release a Language Model of Financial and Economics jargon. PDF to TXT conversion being subject to mistakes, we chose to train word embeddings using ULMFit from Fastai toolkit (Howard and Ruder, 2018) due to its intrinsic capability to handle out-of-vocabulary tokens and then to be less sensible to typos in plain text. Based on AWD-LSTM model architecture (Stephen et al. 2017), our model ran for 5 epochs with batch size of 128 on the clean DoRe distribution. We obtain a model perplexity of 21.6753.
To briefly present our model, we show some examples from a sentence completion task for general risk sentences. Next words predictions are made using Beam Search algorithm with a 20 beam width parameter. Table 2. shows our Language model finds relatively accurate potential candidates for sentence completion, candidates being written in bold.

"*Fondées ou non, ces critiques ou allégations seraient préjudiciables à* **l'entreprise**."
"*Le Groupe est également exposé à un risque de* **taux de change**."
"*La volatilité pourrait fragiliser les rendements* **des sociétés du Groupe**."
"*Une attaque terroriste pourrait se produire dans* **le cadre de son activité**."

Table 2: Sentence completion examples.

6. Conclusion

We hereby presented the DoRe corpus for NLP analytics in Financial and Economics specific domains. This corpus contains 336 listed companies from French and Belgium financial markets, between 2009 and 2019. It sums up to 2350 ARs in total, containing more than 257M tokens and 5.7M sentences. We also built a Metadata table containing, for each company, information about its market capitalization and sector of activity, allowing modularity of the corpus for various tasks. Thus, the DoRe corpus can be used for cross-country/dialect, cross industry, cross-size analysis, fraud detection, document segmentation and risk factor extraction but also for pre-training or fine-tuning Language Models on Financial and Economic specific domains for French language and French dialects.

7. Acknowledgements

This research is funded by a collaboration between the French AMF and the LIMSI laboratory from CNRS associated with Paris-Saclay University.

This work is also partially supported by the French National Research Agency under grant ANR-15-CE23-0025-01 (ContentCheck project).

8. Bibliographical References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *Computing Research Repository*, arXiv:1908.10063.

Back, B., Toivonen, J., Vanharanta, H., and Visa, A. (2001). Comparing numerical data and text information from annual reports using self-organizing maps. *International Journal of Accounting Information Systems*, 2(4):249–269, December.

Beltagy, I., Cohan, A., and Lo, K. (2019). Scibert: Pretrained contextualized embeddings for scientific text. *Computing Research Repository*, arXiv:1903.10676.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., and Davis, B. (2017). Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation*

[3]Please contact us at {corentin.masson, pap}@limsi.fr

(SemEval-2017), page 519–535, Vancouver, Canada. Association for Computational Linguistics. http://www.aclweb.org/anthology/S17-2089.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language. In *HLT/NAACL*, volume 1, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423.pdf.

Ding, X., Zhang, Y., Liu, T., and Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1415– 1425. Association for Computational Linguistics.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. In *Papers and Proceedings of the Twenty-Eighth Annual Meeting of the American Finance Association (1969)*, volume 25, pages 383–417, New York, N.Y., December.

Gao, L. and Srivastava, R. (2011). The anatomy of management fraud schemes: Analyses and implications. *Indian Accounting Review*, 15:1–23, 01.

Hajek, P. and Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud a comparative study of machine learning methods. *Knowledge-Based Systems*, 128:139–152, July.

Hiew, J., Huang, X., H., M., Li, D., Wu, Q., and Xu, Y. (2019). Bert-based financial sentiment index and lstm-based stock return predictability. *Computing Research Repository*. arXiv:1906.09024.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Kamath Ramachandra Rao, S. (2020). *Question-réponse utilisant des données et modèles hybride*. Ph.D. thesis. http://www.theses.fr/s163545.

Kloptchenko, A., Eklund, T., Back, B., Karlsson, J., Vanharanta, H., and Visa, A. (2002). Combining data and text mining techniques for analyzing financial reports. In *AMCIS Proceedings*, volume 4. https://aisel.aisnet.org/amcis2002/4.

Lee, J., Yoon, W.and Kim, S., Kim, D., Kim S., S. C., and J., K. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Computing Research Repository*. arXiv:1901.08746.

Liu, Y.-W., Liu, L.-C., Wang, C.-J., and Tsai, M.-F. (2018). RiskFinder: A sentence-level risk detector for financial reports. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 81–85, New Orleans, Louisiana. Association for Computational Linguistics.

(2019). *Textual Analysis in Finance*. SSRN: https://ssrn.com/abstract=3470272.

Malo, P., Sinha, A., Takala, P., Korhonen, P., and Wallenius, J. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts. *Computing Research Repository*. arXiv:1307.5336.

Martin, L., Muller, B., Suárez, P., Dupont, Y., Romary, L., Villemonte de la Clergerie, E., D., S., and B., S. (2019). Camembert: a tasty french language model. *Computing Research Repository*. arXiv:1911.03894.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, Nevada, December.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *NAACL*.

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

Ranjan Das, S., S., K., and Kothari, B. (2017). Zero-revelation regtech: Detecting risk through linguistic analysis of corporate emails and news. https://ssrn.com/abstract=2960350.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., J., C., and Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. *Computing Research Repository*. arXiv:1909.08053.

Skillicorn, D. and Purda, L. (2012). Detecting fraud in financial reports. In *Proceedings of the European Intelligence and Security Informatics Conference*.

Tumarkin, R. and R., W. (2001). New or noise? internet postings and stock prices. *Financial Analysts Journal*, 57(3):41–51, May.

Xing, F., Cambria, E., and Welsch, R. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Quoc, m. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Computing Research Repository*. arXiv:1906.08237.

Zaki, M. and Theodoulidis, B. (2013). Analyzing financial fraud cases using a linguistics-based text mining approach.

9. Language Resource References

Daudert, T. and Ahmadi, S. (2019). *CoFiF: A Corpus of Financial Reports in French Language*. Association for Computational Linguistics, Proceedings of the First Workshop on Financial Technology and Natural Language Processing.

Händschke, S. G., Buechel, S., and Goldenstein, J. and Poschmann, P. and Duan, T. and Walgenbach, P. (2018). *A Corpus of Corporate Annual and Social Responsibility Reports: 280 Million Tokens of Balanced Organizational Writing*. Association for Computational Linguistics, Proceedings of the First Workshop on Economics and Natural Language Processing.

Kogan, S. and Levin, D. and Routledge, B.R. and Sagi, J.S. and Smith, N.A. (2009). *Predicting risk from financial reports with regression*. Association for Computational Linguistics.

Marcus, M.P. and Santorini, B. and Marcinkiewicz, M.A. (1993). *Building a large annotated corpus of English: The Penn Treebank*. Computational Linguistics.

Suárez, P.J.O., and Sagot, B. and Romary, L. (2019). *Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures*. Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7).