Contradiction Detection in Financial Reports

Tobias Deußer*[1,2], Maren Pielka[2], Lisa Pucknat[1,2], Basil Jacob[2], Tim Dilmaghani[3], Mahdis Nourimand[3], Bernd Kliem[3], Rüdiger Loitz[3], Christian Bauckhage[1,2], and Rafet Sifa[2]

[1]University of Bonn, Bonn, Germany
[2]Fraunhofer IAIS, Sankt Augustin, Germany
[3]PricewaterhouseCoopers GmbH, Düsseldorf, Germany

Abstract

Finding and amending contradictions in a financial report is crucial for the publishing company and its financial auditors. To automate this process, we introduce a novel approach that incorporates informed pre-training into its transformer-based architecture to infuse this model with additional Part-Of-Speech knowledge. Furthermore, we fine-tune the model on the public Stanford Natural Language Inference Corpus and our proprietary financial contradiction dataset. It achieves an exceptional contradiction detection F_1 score of 89.55% on our real-world financial contradiction dataset, beating our several baselines by a considerable margin. During the model selection process we also test various financial-document-specific transformer models and find that they underperform the more general embedding approaches.

1 Introduction

Contradictions in written text are abundant and everywhere to be found. Sometimes they are amusing, like in the case of a newspaper article stating that the "earth circles the moon in 365 and a fraction days"[1] while discussing the astronomy behind the summer solstice. However, in this paper, we will dedicate our efforts to contradictions of more severe consequences: contradictions in financial reports. If such contradictions are not found and corrected before publication, they can lead to a plethora of issues for the reporting company including "bad operational decisions, reputational damage, economic loss, penalties, fines, legal action and even bankruptcy" [30].

The challenge of contradiction detection in financial documents can be considered from two different points of view. One looks at the numeric consistency of values mentioned and described in the document, e.g., if in one sentence the net profit is stated to be \$500 and in another to be \$600, this *numeric* contradiction should be detected[2]. Herein, we will analyze the other type of contradiction, the *semantic* contradiction. In this case, the contradiction is not of numerical nature, but can only be inferred from the actual meaning and implication of the sentence pair. Take this made-up sentence pair for example:

> "On 14th of March, 2020, we increased our capital by offering 5,000 new shares during a seasoned equity offering."
>
> "During 2020 we did not increase our total amount of equity and thus, it remained unchanged at \$10,000,000."

These two statements by themselves are perfectly fine and numerically consistent, but as offering new shares during a seasoned equity offering *does* increase the equity of a company, the contradiction

*Corresponding Author:
tobias.deusser@iais.fraunhofer.de,
ORCID iD: 0000-0003-4685-0847

[1]Printed in the article *Ottawa vs. the equator* by the *Ottawa Citizen* on the 20th of June 2012.

[2]The approaches described in [13] and [4] solve this issue to some extend.

https://doi.org/10.7557/18.6799

is only apparent if both sentences are evaluated together and at least some financial knowledge is present.

In this work, we investigate how to detect contradictions in such a financial context. We analyze 24 different configurations and find that our best performing setup consists of a XLM-RoBERTa (see [6]) encoder, infused with some additional pre-training as described in section 3.1, and fine-tuned on the Stanford Natural Language Inference Corpus (see [3]). It achieves a remarkable F_1 score of 89.55% and is planned to be integrated into the auditing process of PricewaterhouseCoopers GmbH[3].

To summarize, our contributions are twofold:

- We introduce a new natural language processing task, the detection of semantic contradictions in financial documents.
- We evaluate 24 configurations, of which 12 incorporate novel additional pre-training and found our best performing model with an F_1 score of 89.55%.

In the following, we first review related work. Section 3 describes our methodology, i.e., the additional pre-training method we applied and our general model architecture. Thereafter, in section 4 we outline our dataset and the process of acquiring it, present our experiments, and discuss the results. We close this paper with some concluding remarks and an outlook into conceivable future work.

2 Related Work

Contradiction detection is a relatively recent field of natural language processing (NLP). It mainly developed from the task of natural language inference, also known as recognizing textual entailment, where the objective is to find whether two sentences either entail, contradict, or are not related to each other.

Before the emergence of deep, pre-trained transformer models like BERT [9] or RoBERTa [18], contradiction detection models used linguistic features previously extracted from texts to build a classifier. In this vein, [11] tried to find contradictions by leveraging three types of linguistic information: negation, antonymy, and semantic and pragmatic information associated with discourse relations. [7] evaluated a dataset consisting only of contradictions by categorizing them into seven different classes. Further, [21] combined shallow semantic representations derived from semantic role labeling with binary relations extracted from sentences in a rule-based framework.

More recent advances usually leverage the power of such huge, pre-trained models ([9], [18], [25], [26]) and are diverse in their application field and their language.

Regarding different applications, [35] were identifying conflicting findings reported in biomedical literature. [16] detected self-contradictions on an artificially balanced corpus of 1105 self-contradicting and 1105 negative non-self-contradiction. Furthermore, [17] improved chatbot responses by looking for contradictions in preceding conversation turns.

Besides English, contradiction detection was applied in Spanish ([31]), Japanese ([34]), Persian ([27]), and German ([33], [22], [24]).

In the broader spectrum of automating the auditing process of financial documents, which our contradiction detection approach is a part of, [32] introduced a recommender-based tool that streamlines and to some extent automates the auditing of financial documents. [29] updated it to leverage the power of a BERT encoder. A capsule network for the detection of fraud in accounting reports was proposed by [36]. [14] developed a joint named entity and relation extraction model based on BERT to extract key performance indicators and their numerical values from a corpus of German financial reports. [8] applied a similar approach to reports from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, a platform hosted by the U.S. Securities and Exchange Commission, and published their dataset along with their results. [4] also used a joint entity and relation extraction approach to cross-check formulas in Chinese financial documents. Another important aspect, the anonymization of such financial reports, was tackled by [2] by leveraging contextualized natural language processing methods to recognize named entities. [10] employed transformer-based models with joint-task learning and their ensembles to classify whether a sentence contains any causality and to label phrases that indicate causes and consequences in a dataset consisting of financial news. To achieve automatic indexing and information re-

[3]The German division of PricewaterhouseCoopers, one of the largest auditing companies worldwide.

trieval from large volumes of financial documents, [28] presented a document processing system based on a plethora of different machine learning techniques. Finally, [5] tried to autonomously generate financial reports from tabular data.

3 Methodology

In this section, we describe what additional pre-training methods we applied to the already pre-trained encoder in our classification setup and following that, the complete model architecture used to find contradictions in financial documents after the specific pre-training is explained.

3.1 Additional Pre-Training

Our main objective in pre-training is enhancing the semantic knowledge stored in the model. To this end, we apply part-of-speech (POS) tagging as an additional pre-training objective. The task is to predict the syntactic function of each word in a sentence. Possible labels are, for example, "noun", "verb", "adverb" or "determiner". Those can be context-dependent, e.g., "fly" or "break" can mean entirely different things, and therefore have different syntactic roles depending on the context. We assign subword tokens to the label of the word they belong to. The following example from our pre-training data set illustrates the approach.

```
_We _classify _our _short   -   _term
PRON  VERB    PRON  ADJ   PUNCT NOUN
_investments _as _available   -   _for _sale   .
  NOUN       ADP  ADJ       PUNCT ADP  NOUN PUNCT
```

The POS-tags are generated using the spaCy framework [15]. For implementation details and more information about the approach, please refer to our previous work [23].

3.2 Model Architecture

The actual model architecture consists of an encoder and a feed-forward neural network consecutively used for contradiction classification. The encoder is a large, pre-trained language model, of which we evaluated four different models during our experiments, in either its *vanilla*, i.e. with no further pre-training, state or injected with additional knowledge through some further pre-training as described in subsection 3.1. The classifier model used for the binary classification objective of finding a contradiction comprises a feed-forwards neural network with a fine-tuned hyperparameter setup.

We use four different pre-trained base models for our experiments: XLM-RoBERTa [6], FinancialBERT [12], FinBERT [1], and a RoBERTa version trained on the Financial Phrasebank corpus by [20] titled Financial RoBERTa[4]. The models differ slightly with respect to their architecture and hyperparameter settings.

XLM-RoBERTa is a multi-lingual transformer encoder, which was pre-trained on the masked language modeling task for 100 languages. It has an embedding dimensionality of 1024, 24 hidden layers and 16 attention heads per layer, amounting to a total of 355 million trainable parameters. This model has shown to produce state-of-the-art results for many NLP tasks.

FinancialBERT and FinBERT are based on the standard BERT ([9]) implementation, whereas Financial-RoBERTa leverages a RoBERTa ([18] model. They use a bert-base[5] or roberta–base[6] checkpoint, respectively. FinBERT and Financial-RoBERTa are further pre-trained on the Financial PhraseBank corpus by [20] for financial sentiment classification. FinancialBERT is pre-trained for next-sentence prediction and masked language modeling on a corpus of 3.39 billion tokens from the financial domain. All three have an embedding dimensionality of 768, and they have 12 hidden layers with 12 attention heads each. This amounts to 110 million trainable parameters, so they are considerably smaller in size than XLM–RoBERTa–large.

4 Experiments

In the upcoming subsections, we introduce our custom, proprietary dataset, describe the training setup and model selection process in detail, and evaluate results. All experiments are conducted on two Nvidia Tesla V100 GPUs and the model as well

[4] https://huggingface.co/abhilash1910/financial_roberta

[5] https://huggingface.co/bert-base-uncased

[6] https://huggingface.co/roberta-base

	Paragraph 1	Paragraph 2	Label
1	Reversals of impairment losses recognized in previous years amounted to € in fiscal 2018 (2017: €). The largest reversal of impairment losses was recognized on in at € (2017: €) due to changed expectations regarding price developments.	As in the previous year, there was no requirement to recognise impairment losses or reversals of impairment losses on intangible assets in 2018.	`contradiction`
2	No significant events occurred after the end of the fiscal year.	No events have occurred since January 1, 2019, that will have a material impact on the net assets, financial position and results of operations of .	`no contradiction`
3	The total value of fixed assets in was € (previous year: €) of which, as in the previous year, none was pledged as collateral.	The total value of fixed assets in was € (previous year: €) of which, as in the previous year, € was pledged as collateral.	`contradiction`
4	As was the case at December 31, 2017, no treasury shares are held by at December 31, 2018.	The Executive Board is authorized, subject to the approval of the Supervisory Board, to increase the share capital by February 23, 2021, by up to € once or in several installments.	`not related`

Table 1: Example paragraph pairs from our financial contradiction dataset. Information that can be used to identify a company or individuals has been anonymized.

as training code is implemented in PyTorch.

4.1 Data

Our dataset[7] consists of 640 manually collected and annotated sentence pairs in the English language, found in published financial documents (annual reports) and annotated by auditors of PricewaterhouseCoopers GmbH.

The data has been collected using two different annotation procedures. For the first method, a set of paragraphs from financial documents were presented to the annotators, who were asked to come up with a statement that would contradict the original one, and which could possibly be found in a financial document as well. This approach was chosen because the chance of finding real-world contradictions in a report or even across multiple documents is likely to be rather small, given that the reports have already been reviewed at the point when we receive them, and the probability of such errors happening is therefore overall rather low. A total of 145 examples were created using this method.

For the second method, the annotators were shown a list of already matched pairs of paragraphs from financial reports. This matching was achieved based on the heuristic of putting together paragraphs that refer to the same legal requirement according to previously made and thus available annotations by financial auditors, and which fulfil a certain text similarity criterion. Namely, we filter for those pairs of paragraphs, which get assigned a similarity score of 0.8 or higher by the spacy[8] [15] document similarity metric. The pairs of paragraphs are not necessarily found in the same document, so there is a small, but crucial chance that actual contradictions can occur. The annotators are then asked to mark every sample with one of three possible labels: `contradiction`, `no contradiction` or `not related`. The latter means that the two paragraphs refer to completely different facts or events, such that it is not meaningful

[7] We are currently unable to publish the dataset and the accompanying python code because both are developed and used in the context of an industrial project and especially the annotated contradictions are confidential in nature.

[8] `https://spacy.io/usage/linguistic-features`

to compare them with the objective of detecting contradictions. Those are then excluded from the final data set. We generated another 495 examples using this approach.

A few anonymized examples of our dataset are illustrated in Table 1. Furthermore, due to a maximum sequence length of 512 tokens which include premise, hypothesis, and separator tokens, a few data points had to be excluded from the final data set, so that we end up with a total of 626 samples. Out of those, 171 are labeled `contradiction`, and 455 `no contradiction`, yielding a slightly inbalanced label distribution. For the additional pre-training described in subsection 3.1, we utilize a dataset of 47 000 paragraphs from financial reports in English. This dataset, named the Financial Statement and Notes Data, is provided by the US Securities and Exchange Commission and is freely available on their website[9].

4.2 Training Setup

As described above, we intialize the model parameters from a pre-trained checkpoint (XLM–RoBERTa–large[10], FinancialBERT[11], FinBERT[12] and Financial–RoBERTa[13], respectively). To find the best hyperparameter setup for each model, we conduct an extensive grid search evaluating various parameter and pre-training combinations based on the *validation* contradiction classification F_1-score on the SNLI and/or our proprietary financial contradiction dataset. As a result of this hyperparameter optimization, we utilize the AdamW [19] optimizer in combination with a binary cross-entropy loss and a linear warm-up of three epochs (for pre-training) and two epochs (for fine-tuning). A learning rate of $5e^{-6}$ is used throughout the whole training procedure. Further, a dropout regularization of 0.2 is being applied during fine-tuning.

We train each model variation for 15 epochs and determine its best checkpoint via early stopping[14]. For the custom Part-Of-Speech tagging pre-training, the model is being trained for a maximum of 25 epochs, as we observe that convergence happens slower than during fine-tuning.

4.3 Results

As shown in Table 2, we achieve remarkable results in our task of contradiction detection in financial documents, demonstrated by the F_1 score of 89.55% of our best model, the XLM-RoBERTa-large encoder, pre-trained for POS-tagging and fine-tuned both on the SNLI and our financial contradictions dataset.

In detail, we find that the pre-training routine described in section 3.1 improves the performance significantly. Additionally, fine-tuning the contradiction detection model on both the SNLI and our proprietary financial contradiction dataset further enhances the predictive power of our model. Furthermore, we observe a striking superiority of the XLM–RoBERTa-large encoder when compared to all smaller models, but especially those trained for financial documents.

5 Conclusion and Future Work

In this paper, we investigate how we can detect contradictions in a corpus of financial documents, collected and annotated by expert financial auditors. We achieve a noteworthy performance with a contradiction detection F_1 score of 89.55%, obtained by our best model, which incorporates a XLM–RoBERTa encoder further pre-trained for Part-Of-Speech tagging and fine-tuned on the Stanford Natural Language Inference as well as our financial contradiction dataset.

Interestingly, the three encoder models pre-trained on financial data, namely FinancialBERT, FinBERT, and Financial–RoBERTa, underperformed the "more general" XLM–RoBERTa by a considerable margin. We assume that there are two reasons for this. First, these three models are smaller in size than XLM-RoBERTa and second, the conducted pre-training on different financial documents and tasks might not generalize to our challenge of detecting financial contradictions.

This work and its accompanying industry project are part of a larger venture and long-time research

[9] `https://www.sec.gov/dera/data/financial-statement-and-notes-data-set.html`

[10] `https://huggingface.co/xlm-roberta-large`

[11] `https://huggingface.co/ahmedrachid/FinancialBERT`

[12] `https://huggingface.co/ProsusAI/finbert`

[13] `https://huggingface.co/abhilash1910/financial_roberta`

[14] Our best validation set contradiction F_1-score is achieved in epoch 8.

Configuration	Recall in %	Precision in %	F_1 in %
XLM-RoBERTa-large			
Fine-tuned on SNLI	70.59	68.57	69.57
Fine-tuned on finCD	67.65	76.67	71.88
Fine-tuned on SNLI & finCD	85.29	78.38	81.69
Pre-trained for POS-tagging and fine-tuned on SNLI	76.47	52.00	61.90
Pre-trained for POS-tagging and fine-tuned on finCD	82.35	80.00	81.16
Pre-trained for POS-tagging and fine-tuned on SNLI & finCD	**88.24**	**90.91**	**89.55**
FinancialBERT			
Pre-trained for POS-tagging and fine-tuned on SNLI & finCD	61.76	60.00	60.67
FinBERT			
Pre-trained for POS-tagging and fine-tuned on SNLI & finCD	64.71	56.41	60.27
Financial-RoBERTa			
Pre-trained for POS-tagging and fine-tuned on SNLI & finCD	35.29	44.44	39.34

Table 2: Test set evaluation of the contradiction detection task. We exclude the inferior configurations for FinancialBERT, FinBERT, and Financial–RoBERTa. The abbreviation *finCD* stands for our proprietary financial contradiction detection dataset, which is described in section 4.1.

project to enhance the financial auditing process with machine learning to lighten the workload of auditors and to find novel solutions to a plethora of issues faced by practitioners during the audit process. As a next step, the model described here will be integrated into a machine learning enhanced auditing software solution to help auditors find contradictions in financial documents. This will allow us to collect more and more data on found and corrected contradictions, snowballing into an even better detection rate. Separate from this development, we are determined to provide models for contradiction detection in other languages, because financial reports of smaller companies are only published in their local language. Furthermore, we plan on developing a generative model for contradiction generation based on our available data to be able to create financial contradictions spawned from an arbitrary input document to alleviate the issue of the tedious manual annotation process.

Another, more practical open point with respect to this application is the issue of pre-filtering contradiction candidates. In our current evaluation setup, we only consider pairs of paragraphs that relate to the same topic or event, which is in line with the standard natural language inference problem formulation. Looking at real-world use cases though, the problem is not so simple. If we sample sentence or paragraph pairs from a document, most of those will not be related in any way. In order to build a functional contradiction detection system for financial reports, this pre-filtering step would have to be addressed. There are multiple possible solutions, e.g., one could train a three-way classifier that distinguishes the categories `contradiction`, `no contradiction` or `not related`. Additionally, it might also be possible to implement a two-step approach, using a dedicated classifier or a heuristic to pre-filter pairs of paragraphs that are possibly related, and then apply contradiction detection on the remaining samples. In any case, there is the issue of a huge data imbalance, as the vast majority of possible pairs would actually not be related.

Furthermore, in order to determine whether a given pair of paragraphs are contradictory, some context information might be needed. So ideally, a model should take the whole document, or at least the surrounding paragraphs, into account. This could be accomplished by combining the transformer model with a recurrent mechanism that reads through the document from top to bottom (and/or the other way around).

We plan to address these shortcomings in our upcoming research, together with our industry partners.

6 Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, LAMARR22B.

References

[1] D. Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019. doi: 10.48550/arXiv.1908.10063.

[2] D. Biesner, R. Ramamurthy, R. Stenzel, M. Lübbering, L. Hillebrand, A. Ladi, M. Pielka, R. Stenzel, R. Loitz, C. Bauckhage, and R. Sifa. Anonymization of german financial documents using neural network-based language models with contextual word representations. *Springer International Journal of Data Science and Analytics*, 2021. doi: 10.1007/s41060-021-00285-x.

[3] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proc. EMNLP*, 2015. doi: 10.18653/v1/D15-1075.

[4] Y. Cao, H. Li, P. Luo, and J. Yao. Towards automatic numerical cross-checking: Extracting formulas from text. In *Proc. WWW*, 2018. doi: 10.1145/3178876.3186166.

[5] C. L. Chapman, L. Hillebrand, M. R. Stenzel, T. Deusser, D. Biesner, C. Bauckhage, and R. Sifa. Towards generating financial reports from tabular data using transformers. In *Proc. CD-MAKE*, pages 221–232. Springer, 2022. doi: 10.1007/978-3-031-14463-9_14.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proc. ACL*, 2020. doi: 10.18653/v1/2020.acl-main.747.

[7] M.-C. De Marneffe, A. N. Rafferty, and C. D. Manning. Finding contradictions in text. In *Proc. ACL-HLT*, pages 1039–1047, 2008.

[8] T. Deußer, S. M. Ali, L. Hillebrand, D. Nurchalifah, B. Jacob, C. Bauckhage, and R. Sifa. KPI-EDGAR: A novel dataset and accompanying metric for relation extraction from financial documents. In *Proc. ICMLA*, 2022. doi: 10.48550/arXiv.2210.09163.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, 2019. doi: 10.18653/v1/N19-1423.

[10] D. Gordeev, A. Davletov, A. Rey, and N. Arefiev. LIORI at the FinCausal 2020 shared task. In *Proc. FNP*, pages 45–49, 2020.

[11] S. Harabagiu, A. Hickl, and F. Lacatusu. Negation, contrast and contradiction in text processing. In *Proc. AAAI*, volume 6, pages 755–762, 2006.

[12] A. Hazourli. FinancialBERT - a pretrained language model for financial text mining. 2022. doi: 10.13140/RG.2.2.34032.12803.

[13] L. Hillebrand, T. Deußer, T. Dilmaghani, B. Kliem, R. Loitz, C. Bauckhage, and R. Sifa. Towards automating numerical consistency checks in financial reports. In *Proc. BigData*, 2022. doi: 10.48550/arXiv.2211.06112.

[14] L. Hillebrand, T. Deußer, T. Dilmaghani, B. Kliem, R. Loitz, C. Bauckhage, and R. Sifa. KPI-BERT: A joint named entity recognition and relation extraction model for financial reports. In *Proc. ICPR*, pages 606–612, 2022. doi: 10.1109/ICPR56361.2022.9956191.

[15] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spacy: Industrial-strength natural language processing in python. 2020.

[16] C. Hsu, C.-T. Li, D. Saez-Trumper, and Y.-Z. Hsu. WikiContradiction: Detecting self-contradiction articles on wikipedia. In *Proc. Big Data*, pages 427–436. IEEE, 2021. doi: 10.1109/BigData52589.2021.9671319.

[17] D. Jin, S. Liu, Y. Liu, and D. Hakkani-Tur. Improving bot response contradiction detection via utterance rewriting. *arXiv preprint arXiv:2207.11862*, 2022. doi: 10.48550/arXiv.2207.11862.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019. doi: 10.48550/arXiv.1907.11692.

[19] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2019. doi: arXiv.1711.05101.

[20] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. of the Association for Information Science and Technology*, 65(4):782–796, 2014. doi: 10.1002/asi.23062.

[21] M. Q. N. Pham, M. Le Nguyen, and A. Shimazu. Using shallow semantic parsing and relation extraction for finding contradiction in text. In *Proc. ACL-IJCNLP*, pages 1017–1021, 2013.

[22] M. Pielka, R. Sifa, L. P. Hillebrand, D. Biesner, R. Ramamurthy, A. Ladi, and C. Bauckhage. Tackling contradiction detection in german using machine translation and end-to-end recurrent neural networks. In *Proc. ICPR*, pages 6696–6701, 2021. doi: 10.1109/ICPR48806.2021.9413257.

[23] M. Pielka, S. Schmidt, L. Pucknat, and R. Sifa. Towards linguistically informed multi-objective pre-training for natural language inference. In *Proc ECIR (in press)*, 2023. doi: 10.48550/arXiv.2212.07428.

[24] L. Pucknat, M. Pielka, and R. Sifa. Detecting contradictions in german text: A comparative study. In *Proc. SSCI*, pages 01–07, 2021. doi: 10.1109/SSCI50451.2021.9659881.

[25] A. Radford and K. Narasimhan. Improving language understanding by generative pre-training. 2018. URL `https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf`.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. In *OpenAI Blog, Accessed: 2022-08-29*, 2019. URL `https://openai.com/blog/better-language-models`.

[27] Z. Rahimi and M. ShamsFard. Contradiction detection in persian text. *arXiv preprint arXiv:2107.01987*, 2021. doi: 10.48550/ARXIV.2107.01987.

[28] R. Ramamurthy, M. Lübbering, T. Bell, M. Gebauer, B. Ulusay, D. Uedelhoven, T. D. Khameneh, R. Loitz, M. Pielka, C. Bauckhage, and R. Sifa. Automatic indexing of financial documents via information extraction. In *Proc. SSCI*, pages 01–05, 2021. doi: 10.1109/SSCI50451.2021.9659977.

[29] R. Ramamurthy, M. Pielka, R. Stenzel, C. Bauckhage, R. Sifa, T. D. Khameneh, U. Warning, B. Kliem, and R. Loitz. ALiBERT: improved automated list inspection (ALI) with BERT. In *Proc. DocEng*, pages 1–4, 2021. doi: 10.1145/3469096.3474928.

[30] K. Russo. What are the risks of inaccurate financial reporting?, March 2022. URL `https://www.netsuite.com/portal/resource/articles/accounting/inaccurate-financial-reporting.shtml`. [Online; posted 21/03/2022; retrieved 22/08/2022].

[31] R. Sepúlveda-Torres, A. Bonet-Jover, and E. Saquete. "Here are the rules: Ignore all rules": Automatic contradiction detection in spanish. *Applied Sciences*, 11(7):3060, 2021. doi: 10.3390/app11073060.

[32] R. Sifa, A. Ladi, M. Pielka, R. Ramamurthy, L. Hillebrand, B. Kirsch, D. Biesner, R. Stenzel, T. Bell, M. Lübbering, et al. Towards automated auditing with machine learning. In *Proc. DocEng*, 2019. doi: 10.1145/3342558.3345421.

[33] R. Sifa, M. Pielka, R. Ramamurthy, A. Ladi, L. Hillebrand, and C. Bauckhage. Towards contradiction detection in german: a translation-driven approach. In *Proc. SSCI*, pages 2497–2505. IEEE, 2019. doi: 10.1109/SSCI44817.2019.9003090.

[34] Y. Takabatake, H. Morita, D. Kawahara, S. Kurohashi, R. Higashinaka, and Y. Matsuo. Classification and acquisition of contradictory event pairs using crowdsourcing. In *Proc. Workshop on EVENTS at NAACL-HLT*, pages 99–107, 2015. doi: 10.3115/v1/W15-0813.

[35] N. S. Tawfik and M. R. Spruit. Automated contradiction detection in biomedical literature. In *Proc. MLDM*, pages 138–148. Springer, 2018. doi: 10.1007/978-3-319-96136-1_12.

[36] F. Zhu, D. Ning, Y. Wang, and S. Liu. A novel cost-sensitive capsule network for audit fraud detection. In *Proc. IUCC*, pages 549–556, 2021. doi: 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00091.