Numerical Claim Detection in Finance: A Weak-Supervision Approach

Anonymous ACL submission

Abstract

In the past few years, Transformer based models have shown excellent performance across a variety of tasks and domains. However, the black-box nature of these models, along with their high computing and manual annotation costs have limited adoption of these models. In this paper, we employ a weak-supervision-based approach to alleviate these concerns. We build and compare models for financial claim detection task using sentences with numerical information in analyst reports for more than 1500 public companies in the United States from 2017 to 2020. In addition to standard performance metrics, we provide cost-value analysis of human-annotation and weak-supervision labeling along with estimates of the carbon footprint of our models. We also analyze the performance of our claim detection models across various industry sectors given the considerable variation in numerical financial claims across industries. Our work highlights the potential of weak supervision models for research at the intersection of Finance and Computational Linguistics.

1 Introduction

The surge in machine learning and its applications has opened up a new arena of possibilities in diverse fields ranging from image recognition, natural language processing to finance (Sawhney et al., 2021a; Nguyen et al., 2021; Chava et al., 2019, 2021; Sawhney et al., 2021b). However, a major challenge for building or training predictive models is the scarcity of labelled data (Zhang et al., 2021; Ratner et al., 2017). Supervised learning often involves a significant amount of manual labelling of data which is often not practically feasible for large datasets. In such scenarios, one can leverage weak supervision based learning methods (Varma and Ré, 2018).

Weak supervision is defined as a machine learning concept which leverages slightly noisy or imprecise models to label vast amounts of unlabelled data (Ratner et al., 2020; Lison et al., 2021). A crucial component of this concept is the development of effective labelling functions by critically analyzing the dataset to obtain annotations for a given raw dataset algorithmically (Lison et al., 2021) instead of manual annotation. Weak supervision learning is a method that uses limited and imprecise labels in contrast to accurate labels backed by empirical evidence (Ratner et al., 2017). The strength of weak supervision model lies in these imperfect labels, when combined, producing reliable predictive models (Lison et al., 2021; Zhang et al., 2021). Moreover, in constrained conditions and uniform noise situation, weak supervision is found to be equivalent to supervised learning (Zamani and Croft, 2018). The weak labels needed for classification can be obtained by introducing an external knowledge base, predefined patterns or crowd-sourcing (Shi et al., 2021). Hence, this serves as a huge improvement in terms of efficiency of producing labelled data.

Label	Sentence
In-Claim	Operating income is expected between \$2.1 billion and \$3.6 billion
Out-of-Claim	Revenues climbed 48.6% year over year to \$5.44 billion primarily driven by expanding customer base.

Table 1: Example of In-claim and Out-of-claim sentences

There has been very limited work reported in the context of the classification of financial text as 'in-claim' or 'out-of-claim' when it comes to English language specifically (Chen et al., 2019a). Financially relevant numeric sentences in the context of this paper refers to sentences containing both numeric and financial information. Furthermore in our approach, 'in-claim' text in the financial domain, has been attributed to data which consists of a tangible financial claim. All sentences which are not classified under the hood of 'in-claim' text are referred to as ‘out-of-claim’. Table 1 illustrates instances from both classes in reference to the aforementioned definitions. We provide details about data in section 3.

Finance literature, for example, (Jegadeesh and Kim, 2010) has documented that there is a significant stock market reaction to analysts’ recommendations (ratings). However, analyst ratings can be biased (Michaely and Womack, 1999; Corwin et al., 2017; Coleman et al., 2021). Therefore it is important to understand whether the ratings are backed by strong numerical financial claims in the analyst’s report. To evaluate the ratings reliability, the extraction of numerical financial claims is a necessary task. Further the sentences with a claim have a higher density of forward-looking information. Related, extraction of numerical ESG claims from earnings call transcripts, can help better understand whether companies do walk the talk on their environment and social responsibility claims (Chava et al., 2021). The importance of mentioned examples necessitates the numerical claim detection task in the Finance domain.

The aim of our proposed methodology is to derive financially significant information from the quarterly analyst reports (in English) by categorizing each numerical sentence into in-claim or out-of-claim. Our major contributions through this paper are following:

- Present the first-ever robust labelled dataset (in English) that can be of immense use in the domain of finance for claim based analysis. We also intend to make trained models and code publicly available through GitHub under CC-BY 4.0 license.
- Propose a Weak-supervision based whitebox model to label and categorize the data in contrast to neural-network based blackbox models which could potentially help us understand and evaluate risk in a more holistic sense.
- Provide quantitative comparison of the claim-detection accuracy for various sectors.
- Provide comprehensive comparative analysis to understand the potential of the Weak-supervision model by comparing it with the pre-trained language model (BERT model developed by Devlin et al. (2018) under Apache License 2.0).
- Highlight the advantages of weak-supervision framework under budget constrained setting, by training and evaluating BERT models on both human-annotated data and weak-supervision model generated data to better understand the cost-benefit of human-annotation. We also provide estimates of CO_2 emission of our models to help researchers make more carbon conscious decisions.

2 Related Work

Weak-supervision In order to reduce the complexities associated with manual labelling, several standard techniques such as semi-supervised learning (Chapelle et al., 2009), transfer learning (Pan and Yang, 2010), and active learning (Settles, 2009) had been employed. However, many researchers and practitioners are employing weak-supervision based models to further reduce the computational costs while retaining the accuracy of the labelled data. Weak-supervision models were primarily developed in a bid to replace standard labelling techniques with models which can leverage slightly noisy or imprecise data to label vast amounts of unlabelled data (Ratner et al., 2020). Ideally multiple weak-supervision based techniques are combined together in order to increase the overall accuracy. Techniques such as distant supervision (Mintz et al., 2009) and crowd-sourced labels (Yuen et al., 2011) are often associated with weak supervision based models, however, they tend to have limited coverage and accuracy (Ratner et al., 2020). Labelling functions form a crucial portion of weak supervision models and typically make use of rule based heuristics, domain-specific knowledge of the database and other linguistic constraints to label the data in a more efficient manner (Lison et al., 2021). Developing good labelling functions for the given data rather than gathering manual labels has proven to be far more effective than typical annotation methods (Ratner et al., 2020). It also allows domain specialists to introduce their subject matter expertise directly into the system as well as the ability to change or expand the set of labelling functions for future initiatives.

Claim Detection The task of identifying arguments from raw text (natural language text) and deriving useful information from it is referred to as argument mining. Recently, this field has attracted a lot of attention from a diverse research community (Lippi and Torroni, 2015; Stab and Gurevych, 2014). Claims are the conclusions that emerge after considering evidences provided in the argument.

Hence, claim detection occupies central position in the task of argument mining. Initial works included mining claims related to controversial topics from publicly available data (Levy et al., 2014), persuasive essays (Stab and Gurevych, 2014), legal documents (Grabmair et al., 2015) and weak-supervision approach to identify claim-sentences from unstructured data (Levy et al., 2014, 2018).

In the domain of finance, claim detection plays a significant role in analyzing and predicting the market reaction around events like earnings call announcements. In claim based sentences with numerals, authors provide estimate based on their understanding of the market and provide significant information for financial decision making as discussed by Chen et al. (2018, 2019b). Our methodology involves detecting numerical financial claims from a large sample of analysts reports in English Language using weak-supervision model in contrast to the work done by Chen et al. (2020) which provides Numeric Claim detection methodology for a small Chinese dataset.

NLP and Finance Finance is one of the most attractive domains for the application of NLP. Araci (2019) and Liu et al. (2020) presented pre-trained language models for Finance domain. There are multiple datasets specifically catered for applications of NLP in finance including question answering dataset created by Chen et al. (2021) and Maia et al. (2018), and also an NER dataset constructed by Alvarado et al. (2015) for the financial domain. There is a wide literature on sentiment analysis task undertaken on financial data (Maia et al., 2018; Malo et al., 2014; Day and Lee, 2016; Akhtar et al., 2017).

Works of Li et al. (2020) and Sawhney et al. (2020) were centered around volatility prediction using earnings call transcripts in the domain of risk management. In NLP, pre-trained model can be fine-tuned for a multitude of tasks. Chava et al. (2019) used embeddings created using RoBERTa model for identification of emerging technologies. Chava et al. (2021) create a dictionary of Environmental and Social (E&S) phrases, while Li et al. (2021) leveraged word-embeddings to measure the corporate culture. Moreover, multimodal machine learning was used by Nguyen et al. (2021) and Dalton et al. for credit rating prediction and measurement of persuasiveness respectively. Sawhney et al. (2021a) investigated biases in the multimodal analysis of financial earnings calls. Finally, Cao et al. (2020) provide critical analysis of how corporate disclosure has been reshaped over last couple of years due to increasing use of NLP in Finance.

3 Dataset

3.1 Construction

Quarterly analyst reports (in English) on a large number of public firms in the U.S. constitute the raw dataset for our model. These analysts reports were collected from Zacks Equity Research and were available to us from Nexis Uni license[1]. Before the data is passed on to labelling functions it is standardized in order to maintain consistency for subsequent steps.

The text documents are split into discrete sentences using multiple regex based rules. We employ Regex based rules as they typically are significantly faster and produce similar accuracy as other standard libraries in tokenizing and splitting data into discrete units. Post completion numeric sentences containing statistical information (i.e: sentences consisting of a numeric value coupled with a currency or percentage symbol) are filtered , in order to ensure its numerical relevance (Chen et al., 2019a). The next step in the pipeline consists of a white-listing technique in order to retain only those sentences which contain any financially significant information. We ensure this by cross-verifying every sentence with a financial dictionary that includes a comprehensive list of technical terms catering to the financial market and the corresponding literature. It is formed by combining word list from Investopedia, Vocabulary.com, MyVocabulary.com (a), TheStreet and MyVocabulary.com (b) that accounted for more than 8,200 financially significant terms. For verification, every word of the input sentence is cross-referenced with the dictionary and in case none of the words in the sentence exist in the dictionary then that sentence is marked irrelevant in this context.

Type	# Sentences
Total sentences	8,583,093
Total numeric sentences	2,857,567
Total numeric-financial sentences	2,364,977

Table 2: Size of Dataset

We apply multiple filters to remove data that is not materially relevant for our analysis. Blacklisting helped us remove 66.7% of total sentences

[1]Nexis Uni license doesn't authorize republication of full or partial text

which did not consist of any numerical information. Further filtering using financial dictionary helped reduce the data by around 17.2%, providing us with a financially significant dataset for further experiments. From Table 2, we can clearly observe that this two tier filtering method enriched the data by retaining only 27.5% sentences out of the original data.

Table 8 shows that firms in our raw dataset belong to 12 sectors based on the GSECTOR classification in annual fundamental COMPUSTAT database. We find that the maximum number of reports belong to Health Care sector. However, the largest number of numeric sentences per file with or without financial information was observed in the Consumer Staples sector. This necessitates the need to look at various sectors critically while analyzing claim based statistics so as to understand sector based variations and trends. The lowest number of numeric sentences per file with or without financial information was observed in the Energy and Health-care sector signifying the fact that their reports don't possess significant claim based information.

3.2 Comparison with Related Datasets

In this section we compare our proposed dataset with NumClaim (Chen et al., 2020), an expert-annotated dataset in the Chinese language. Our dataset of raw analyst reports in English Language from 1530 major companies over the period of 2017-20 is significantly larger than NumClaim or other associated datasets. In addition, unlike NumClaim, we analyze performance across industries and document sector-wise trends over time. Our dataset consists of 555x financially significant numeric sentences and 273x in-claim sentences as compared to data in NumClaim.

Dataset	Proposed	NumClaim
Language	English	Chinese
Year	2017-20	NA
Sector information	Yes	No
# Stocks	1530	NA
# Files	87,536	NA
# Words	167,301,873	42,594
# Numeric Sentences	2,857,567	5,144
# In-Claim Sentences	336,252	1,233
# Out-Claim Sentences	2,028,722	3,921

Table 3: Comparison of our dataset with NumClaim dataset

3.3 Sampling of Dataset for Experiments

From the complete raw dataset of 87,536 files we sampled data catering to our requirements for multiple experiments in the following manner.

Data for Gold Label: For our experiments, we need to manually label sentences to form a benchmark for the model evaluation. For this purpose, a validation dataset was sampled from the complete dataset. The sampled dataset consisted of 96 files consisting of two files per sector per year, accounting for about 2,626 unique sentences. This set was manually annotated and assigned 'in-claim' or 'out-claim' labels by two of the authors with basic background of finance and domain specific knowledge gained from examples supplied by a financial expert co-author. The labels were then cross-checked by a co-author with financial domain knowledge to ensure they were in compliance with the definition. Here on, this complete set of labels (2,680 sentences) are considered to be the Gold labels.

Data for Weak Labels: In our experiments, pertaining to BERT model, we make use of the labelled dataset generated from our weak-supervision model. For these tasks, we need dataset that is a reflection of both time series and the sector wise representation of the complete dataset. So, we randomly chose 50% of the unique stocks from each sector to maintain the true composition of the dataset. From those unique stocks we selected one file per stock per year. From each file we considered equal number of in-claim and out-of-claim sentences labelled using the weak-supervision model. This was done to ensure that the data sampled is balanced in terms of in-claim and out-of-claim entries. From this sampling technique we obtained on an average 19,780 sentences.

4 Models

In this section, we provide details of the two models we have used. Initially, we propose a Weak-Supervision based model followed by description of the pre-trained BERT model used for comparative analysis. We use BERT-based model to better understand accuracy of our Weak-Supervision model as BERT can serve as good representative of modern Transformer based models.

4.1 Weak-Supervision Model

For implementing a weak-supervision model we use the Snorkel library (Ratner et al., 2017), lever-

Used to detect	Output	Type	Labels
High Confidence out-of-claim (Past Tense or Assertions)	-1/0	Phrase Matching	reasons to buy:, reasons to sell:, was, were, declares quarterly dividend, last earnings report, recorded
Low Confidence in-claim	1/0	Phrase Matching	earnings guidance to, touted to, entitle to
High Confidence in-claim	2/0	Lemmatized Word matching	expect, anticipate, predict, forecast, envision, contemplate
High Confidence in-claim	2/0	POS Tag for "project"	VBN, VB, VBD, VBG, VBP, VBZ
High Confidence in-claim	2/0	Phrase Matching	to be, likely to, on track to, intends to, aims to, to incur, pegged at

Table 4: Labelling Functions used in weak-supervision model. SpaCy Lemmatizer has been used for the labelling functions involving lemmatized word matching.

aging its inherent pipeline structure for generating labels for each data segment and then pass the outputs through the curated aggregator function.

Labelling functions used in our model include simple rule-based pattern matching combined with POS tag constraints for some phrases. We create seventeen labelling functions for categorization of results and also made use of multiple other labelling functions to segregate the sentences representing assertions or written in past tense. These labelling functions are listed in Table 4.

Output	Implication
-1	Out-of-claim sentence
0	Abstain
1	Low confidence while making claim
2	High confidence while making claim

Table 5: Description of output from each labelling function

The output of the labelling functions needs to be aggregated to decide the final label of the sentence. Unlike other models, we use independent and weighted labelling functions with the weights based on the level of confidence in the claim. We have considered two levels of in-claim sentences forming in total four types of return value as listed in Table 5. In the final results both levels have been considered for in-claim sentences. This fine grained categorization help us understand the results better and opens room for future fine-tuning of the models. For our model, each labelling functions classifies a sentence independently and hence, we consider the 'max' as our aggregating function as shown below:

$$label\ (x_i) = \begin{cases} 1, & max(lf_1(x_i), ...lf_n(x_i)) > 0; \\ & \forall lf_j(x_i) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where, x_i = i^{th} sentence

$lf_j(x) = j^{th}$ labelling function

$label(x_i)$ = label of i^{th} sentence

Figure 1, shows how the accuracy of the model changes depending upon the number of labelling functions. For this plot, we initially computed contribution of each labelling function (Table 4, High confidence and Low Confidence in-claim) towards detection of in-claim sentences and then considered addition of new labelling function at each step to ensure steepest ascent to saturation. At each step, in addition to one new labelling function, all labelling functions present in Table 4 for Past Tense and Assertions, were also used. They either abstain or classify sentences as out-of-claim and help improve the classification of out-of-claim sentences. From the plot, we can clearly notice that after around thirteen labelling functions, addition of new labelling functions does not produce any change in the accuracy. In fact, increasing labelling functions thereafter leads to a minor decrease in accuracy suggesting that we can effectively capture the required trends for classification in this setting with thirteen labeling functions.

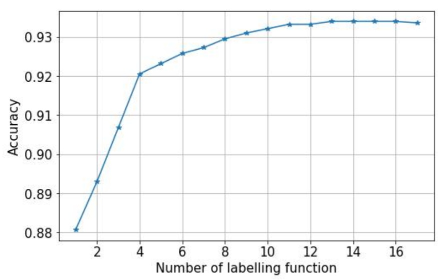

Figure 1: Overall Accuracy v/s Number of labelling functions

4.2 BERT

For our experiments we have made use of the *bert-base-uncased* model (Devlin et al., 2018). In order to perform comprehensive comparative analysis, between our Weak Supervision Model and BERT, we divided the experiments into three major categories:

BERT-G: The data with gold labels(as described in Section 3.3) was split into train-test-validation in 80-10-10 ratio. Through this experiment we compare the performance of weak-supervision approach and BERT keeping training, validation and testing data same.

BERT-W: For this experiment we used weak label data(as mentioned in Section 3.3) for training while validation and testing data remained the same as the corresponding data in BERT-G. Through this experiment we compare the impact of changing the source of training data.

BERT-WG: Here, we merge the training data from BERT-G and BERT-W keeping validation and testing data same as in previous cases. Through this experiment we observe whether manually labelling a small dataset and using it for training would produce a significant improvement in performance of the model.

We have fine-tuned the BERT model for maximum sequence length of 128 tokens. The model was trained for five epochs using learning rate of $2 * 10^{-5}$ and batch size as 16. This architecture was kept consistent across all the experiments in this section.

Model	Gold label			Weak label
	Train	Validate	Test	Train
BERT-G	2,140	270	270	-
BERT-W	-	270	270	19,780
BERT-WG	2,140	270	270	19,780

Table 6: Three different training data used to train BERT model

Metric	Value
Accuracy	93.36
Precision	93.21
Recall	93.36
F1-score	93.08
MCC	77.16

Table 7: Performance of WS model on gold labels

5 Results and Analysis

In this section we present the results obtained using the above models and provide a detailed analysis of the outcomes.

5.1 Weak-supervision Model

Manually Labelled Dataset: The performance metrics in Table 7, highlights how well our Weak-Supervision(WS) based model performs when compared with Manually Annotated Data.

In order to understand the statistical significance of accuracy, 10 files were randomly sampled and their accuracy and precision values were calculated to verify if the methodology saturates with optimal metrics. We found that for N=10 and 100 iterations the 95% confidence interval for accuracy was found to be : (0.9295, 0.9382) whereas for precision it was found to be : (0.9286, 0.9374). On an average 5.2396 in-claim sentences per file with a standard deviation of 5.1127 are found with respect to all the labelled files. The significantly high value of standard deviation across varied sectors represents the importance of sector based analysis to understand trends for the same.

Sector wise analysis: Table 8 highlights that of all the aforementioned sectors, the Consumer Staples sector has the highest average number of Numeric as well as FinNum sentences.

Industry sectors differ on the level of information disclosure, regulatory scrutiny and uncertainty about the future. Table 9 further reveals that the Financials followed by the Consumer Staples sector have the highest average number of in-claim sentences per file. We also observe the Consumer Staples followed by the Information Technology sector to have the highest average % of in-claim sentences per file. On the contrary the Energy, Health Care as well as the Real Estate sectors tend to have a lower number of sentences across all the aforementioned categories as can be seen from Table 8 and Table 9. The later sectors tend to make more assertions rather than claims as a general trend.

We observe an average overlap value of 71.96% considering only in-claim sentences and 97.92% for out-of-claim sentences. This highlights the fact that the current weak-supervision model performs much better at classifying out-of-claim labels as compared to in-claim labels for most sectors.

Among in-claim labels we obtain the worst performance among the Utility sector. This is perhaps on account of their tendency to represent existing facts and information through a sentence structure which closely resembles the sentence structure of claims.

Sector	Companies	Numeric	FinNum	In-claim	% of In-claim
Miscellaneous	116	28.19	23.6	3.01	11.39
Energy	112	**25.62**	21.78	**2.24**	9.74
Materials	82	32.78	27.75	3.82	13.25
Industrials	193	35.12	28.77	4.01	13.005
Consumer Discretionary	193	32.34	27.36	4.55	15.51
Consumer Staples	65	**37.89**	**32.97**	**5.41**	**15.85**
Health Care	**241**	25.83	**20.36**	2.97	13.33
Financials	164	35.48	30.77	2.93	**8.78**
Information Technology	208	30.48	24.72	3.82	15.17
Communication Services	61	34.42	26.79	2.72	10.09
Utilities	51	28.66	23.34	3.35	13.95
Real Estate	**44**	29.04	24.62	2.73	10.23

Table 8: Sector wise average data of key metrics via Weak-Supervision Model. Here "Numeric", "FinNum" and "In-claim" columns represent the average number of sentences per file for the respective category via Weak Supervision Models for the entire dataset. % In-claim is the ratio of In-claim sentences and Financially significant information (FinNum)

Sector	Avg. In-claim	% In-claim	In-claim overlap	Out-of-claim overlap
Miscellaneous	2.75	12.86	0.81	0.97
Energy	**2.25**	**8.85**	0.63	0.96
Materials	3.875	13.30	0.61	0.97
Industrials	4.375	14.81	0.7	0.97
Consumer Discretionary	4.875	14.56	0.81	0.98
Consumer Staples	6.125	**17.98**	**0.85**	0.99
Health Care	3.125	14.30	0.64	**0.95**
Financials	**8.25**	16.89	0.72	**0.995**
Information Technology	4.875	17.04	0.84	0.994
Communication Services	4.5	13.55	0.67	0.98
Utilities	3.25	11.10	**0.58**	0.97
Real Estate	2.625	13.02	0.73	0.986

Table 9: Sector wise data for In-claim statistics and overlap with gold labels. Here "Avg. In-claim" column represent the average number of in-claim sentences per file for the respective sector via data present in the Gold Labels. % In-claim is the ratio of In-claim sentences and Financially significant information (FinNum) for the same. In-claim and out-of-claim overlap represents the ratio of the correct predicted claims to the actual number of true claims obtained from the actual labels for both classes of claims individually.

5.2 BERT

As discussed in Section 4.2, we perform three major experiments using BERT base model. We execute the experiments by taking five different seeds and average accuracy is listed in Table 10. Accuracy for five different seeds is listed in Appendix A. From Table 10, we can comment upon the results of the targeted experiments listed in Section 4.2.

1. We can say that on an average our weak supervision model(WS) produces good results with an overall accuracy of 93%. BERT-G model produces better results in comparison to weak-supervision model but the time taken for BERT model to train in each case is considerable whereas there is no concept of training time per se when it come to weak-supervision model.
2. BERT-G and BERT-W are different in terms of the training data. For BERT-W, we use weak labels and we can observe that accuracy decreases which is due to the noisy nature of the labels in comparison to the gold labels used in BERT-G. However, the accuracy is comparable to the standalone weak-supervision model, and hence establishes the fact that complex models such as BERT tend to identify the trends similar to the ones employed in labelling functions used in WS.
3. For BERT-WG we observe that after combining the training data from BERT-G and BERT-W the accuracy of the model improved negligibly in comparison to BERT-W. This shows that enhancing training data by addition of Gold Labels(manually annotated data), did not contribute significantly towards increasing the performance suggesting that training data for BERT-W was sufficient to capture the trends present in the Gold Labelled data.

We can say from the overall results that dataset

Model	Gold Labels	Weak Labels	Training Time	Annotation Time	Training Cost	Annotation Cost	Net Cost	CO2e	Accuracy
WS	NA	NA	NA	9 s	0	0.0002	**0.0002**	**0.01g**	0.9350
BERT-W	NA	0.83%	1.236 hrs	21.8 s	1.126	0.005	1.131	242.75g	0.9338
BERT-G	80%	NA	0.2 hrs	11.2 s	244.98	0.0028	244.983	39.69g	**0.9539**
BERT-WG	80%	0.83%	1.416 hrs	27.8 s	246.08	0.007	246.087	278.34g	0.9360

Table 10: Cost analysis of all models (All Cost calculations are in USD). Here "Gold Labels" refers to the fraction of the net gold labels used during training."Weak Labels" refers to the fraction of labels generated from Weak-Supervision Model, used during training. WS model was used to label complete dataset but the "Annotation Cost" and "Annotation Time" here are considered for 0.011% of the complete dataset, to facilitate a fair comparison with BERT models.

produced using weak-supervision model is robust from an application point of view and is a highly viable solution in resource constrained environment. The fact that our model has almost comparable accuracy values to BERT-W and BERT-WG, adds to its credibility.

5.3 Comparative Analysis of BERT and Weak Supervision Models

This section attempts to give a comparative analysis of the weak supervision and BERT models on the basis of its standardized costs, carbon footprint and accuracy. All computational costs are derived with respect to standard rates for Virtual Machines on the Microsoft Azure Cloud Platform as of January 2022, whereas the labour costs for annotation is based on the average hourly wage for a Graduate Research Assistant. The hourly rate for manual annotation of the dataset is 30 USD/hr whereas the computational cost for a CPU (B2ms instance) is 0.0832 USD/hr and that of a GPU (NC6 instance) is 0.9 USD/hr. Weak supervision models make use of the CPU instance whereas all BERT models employ the GPU instance's. Carbon footprint calculator developed by Lannelongue et al. (2021) is used for calculation of CO_2 emission.

Cost calculations for all the models mentioned in Table 10 considers all the discrete components required for training and annotation, scaled with respect to the fraction of the data which is actually being used, in accordance with Table 6.
As can be seen from Table 10, a major chunk of the training costs among BERT-G and BERT-WG involves the manual annotation of the dataset. Weak Supervision Models require the least amount of cost involved to label the entire dataset, followed by the BERT-W model. BERT-G and BERT-WG involve a significantly higher amount of cost owing to the massive costs and efforts of manual annotation. These observations showcases the extreme efficiency of weak-supervision based models especially in budget constrained environments, and the trade-off involved as we move to higher levels of accuracy. Table 10 also highlights the advantage of weak-supervision based models in carbon conscious setting.

6 Conclusion

Our work presents the first ever claim based labelled dataset in English language alongside presenting a weak-supervision model with a standalone accuracy of 93%. The variation among accuracy parameters as well as the descriptive statistics highlights the importance of considering sector information while performing claim based analysis. We also provide cost-value analysis of weak-supervision based annotation compared to human-annotation revealing that our model can serve as an ideal-replacement to black-box models in resource constrained environment. We find that weak-supervision model (WS) is most environment friendly option. Below we list some extensions that we believe will add value in future work:

- Include sector wise information while training models and generating labelling functions in order to analyze the influence of sector on the prediction of claims and improve the performance of standalone in-claim predictions.
- Analysis of market reaction (cumulative abnormal return and surprise in earnings) on report release date and earning announcement date based on number of FinNum sentences with claim. One can also look at heterogeneity in reaction by sector. The measure generated can be useful in better predicting the volatility of the stocks.

References

Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2017. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 540–546.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Sean Cao, Wei Jiang, Baozhong Yang, and Alan L Zhang. 2020. How to talk when a machine is listening: Corporate disclosure in the age of ai. Technical report, National Bureau of Economic Research.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Sudheer Chava, Wendi Du, and Baridhi Malakar. 2021. Do managers walk the talk on environmental and social issues? *Available at SSRN 3900814*.

Sudheer Chava, Wendi Du, and Nikhil Paradkar. 2019. Buzzwords? *Available at SSRN 3862645*.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019a. Numeral attachment with auxiliary tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1164.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Numclaim: Investor's fine-grained claim detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1973–1976.

Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. *CoRR*, abs/1809.05356.

Chung-Chi Chen, Hen-Hsen Huang, Chia-Wen Tsai, and Hsin-Hsi Chen. 2019b. Crowdpt: Summarizing crowd opinions as professional analyst. In *The World Wide Web Conference*, WWW '19, page 3498–3502, New York, NY, USA. Association for Computing Machinery.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Braiden Coleman, Kenneth J Merkley, and Joseph Pacelli. 2021. Human versus machine: A comparison of robo-analyst and traditional research analyst investment recommendations. *The Accounting Review, Forthcoming*.

Shane A Corwin, Stephannie A Larocque, and Mike A Stegemoller. 2017. Investment banking relationships and analyst affiliation bias: The impact of the global settlement on sanctioned and non-sanctioned banks. *Journal of Financial Economics*, 124(3):614–631.

Margaret Dalton, Sari Pekkala Kerr, William Kerr, Yujin Kim, Chirantan Chatterjee, Matthew J Higgins, Hans Degryse, Martin Brown, Daniel Hoewer, and María Fabiana Penas. Persuading investors: A video-based study.

Min-Yuh Day and Chia-Chou Lee. 2016. Deep learning for financial sentiment analysis on finance news providers. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1127–1134. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matthias Grabmair, Kevin D. Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R. Walker. 2015. Introducing luima: an experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools. *Proceedings of the 15th International Conference on Artificial Intelligence and Law*.

Investopedia. Financial term dictionary from investopedia.

Narasimhan Jegadeesh and Woojin Kim. 2010. Do analysts herd? an analysis of recommendations and market reactions. *The Review of Financial Studies*, 23(2):901–937.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, 8(12):2100707.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3063–3070.

Kai Li, Feng Mai, Rui Shen, and Xinyan Yan. 2021. Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7):3265–3315.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 185–191. AAAI Press.

Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. 2021. skweak: Weak supervision made easy for nlp. *arXiv preprint arXiv:2104.09683*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *IJCAI*, pages 4513–4519.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Roni Michaely and Kent L Womack. 1999. Conflict of interest and the credibility of underwriter analyst recommendations. *The Review of Financial Studies*, 12(4):653–686.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

MyVocabulary.com. a. Business, finance and economics vocabulary word list.

MyVocabulary.com. b. Finance vocabulary word list.

Cuong V Nguyen, Sanjiv R Das, John He, Shenghua Yue, Vinay Hanumaiah, Xavier Ragot, and Li Zhang. 2021. Multimodal machine learning for credit modeling. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1754–1759. IEEE.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730.

Ramit Sawhney, Arshiya Aggarwal, and Rajiv Shah. 2021a. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3751–3757.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Shah. 2020. Voltage: volatility forecasting via text-audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2021b. Quantitative day trading from natural language using reinforcement learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4018–4030, Online. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey.

M Shi, A Hoffmann, and U Rüppel. 2021. Applying weak supervision to classify scarce labeled technical documents. In *ECPPM 2021-eWork and eBusiness in Architecture, Engineering and Construction: Proceedings of the 13th European Conference on Product & Process Modelling (ECPPM 2021), 15-17 September 2021, Moscow, Russia*, page 223. CRC Press.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

TheStreet. Financial word dictionary.

Paroma Varma and Christopher Ré. 2018. Snuba: Automating weak supervision to label training data. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 12, page 223. NIH Public Access.

Vocabulary.com. Personal finance and financial literacy.

Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. A survey of crowdsourcing systems. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 766–773. IEEE.

Hamed Zamani and W Bruce Croft. 2018. On the theory of weak supervision for information retrieval. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 147–154.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. 2021. Wrench: A comprehensive benchmark for weak supervision. *arXiv preprint arXiv:2109.11377*.

A Experiments over Multiple Seeds

The test accuracy of weak-supervision model and all three variants of BERT for five different seeds are listed in Table 11.

Seed	**WS**	**BERT-G**	**BERT-W**	**BERT-WG**
42	0.9404	0.9442	0.9368	0.9442
149	0.9479	0.9591	0.9480	0.9554
1729	0.8996	0.9294	0.8959	0.8922
13832	0.9553	0.9628	0.9480	0.9480
110656	0.9330	0.9740	0.9405	0.9405
Avg.	0.9353	0.9539	0.9338	0.9360

Table 11: Accuracy analysis of our model and three BERT models

B Flowchart of Our Methodology

Figure 2 gives an overview of the steps involved in the complete pipeline. There are two main steps through which the raw data is passed in order to generate enriched dataset for input to our weak-supervision model. The labelled datasets generated from weak-supervision model and manual annotation are then comprehensively analysed.

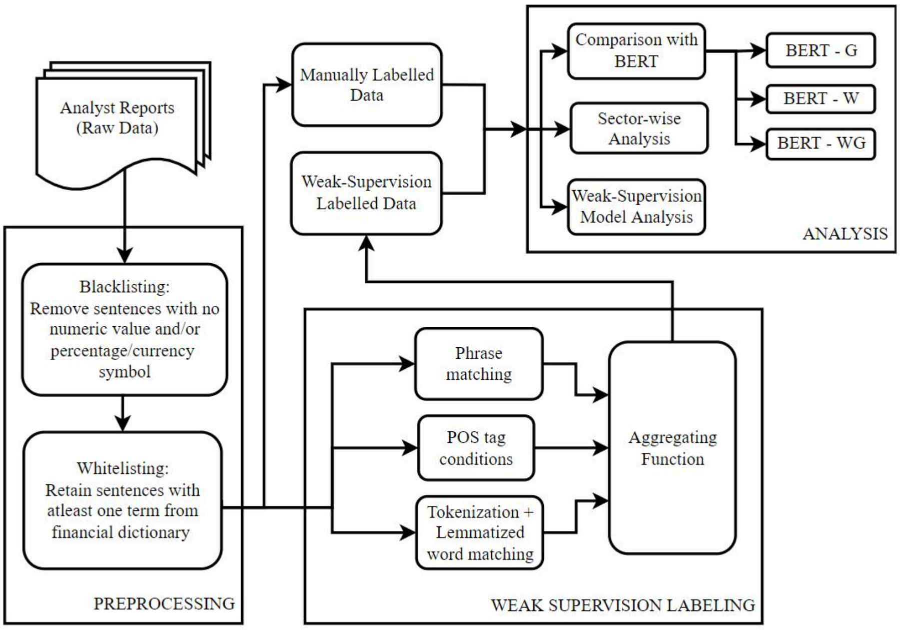

Figure 2: Flowchart for complete methodology