Numerical Claim Detection in Finance: A Weak-Supervision Approach

Pratvi Shah, Arkaprabha Banerjee, Agam Shah, Bhaskar Chaudhury, and Sudheer Chava

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. The proposed aggregation function for the weak-supervision model outperforms existing aggregation functions. 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 Data Engineering.

Index Terms—Claim Detection, Weak Supervision, BERT, CO2e

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 [1], [2], [3], [4], [5]. However, a major challenge for building or training predictive models is the scarcity of labelled data [6], [7]. 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 [8].

Weak supervision is defined as a machine learning concept which leverages slightly noisy or imprecise models to label vast amounts of unlabelled data [9], [10]. 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 [10] 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 [7]. The strength of weak supervision model lies in these imperfect labels, when combined, producing reliable predictive models [6], [10]. Moreover, in constrained conditions and uniform noise situation, weak supervision is found to be equivalent to supervised learning [11]. The weak labels needed for classification can be obtained by introducing an external knowledge base, predefined patterns or crowdsourcing [12]. Hence, this serves as a huge improvement in terms of efficiency of producing labelled data.

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 [13]. 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 sentence which consists of a tangible financial claim (i.e: sentences consisting of a financial word and a numeric value coupled with a currency or percentage symbol). 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.

TABLE 1
Example of In-claim and Out-of-claim sentences

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.

Finance literature, for example, [14] has documented that there is a significant stock market reaction to analysts' recommendations (ratings). However, analyst ratings can be biased [15], [16], [17]. 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 [4]. The importance of mentioned examples necessitates the

- *Pratvi Shah, Arkaprabha Banerjee, and Bhaskar Chaudhury are with Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India.*
  E-mail: 201801407, 201801408, bhaskar_chaudhury@daiict.ac.in
- *Agam Shah (corresponding author) and Sudheer Chava are with Georgia Institute of Technology, Atlanta, USA.*
  E-mail: shahagam4@gmail.com (corresponding author), sudheer.chava@scheller.gatech.edu

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 numerical claim detection task using a robust labelled dataset (in English) that can be of immense use in the domain of finance for claim-based analysis. We 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 [18] 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 [19], transfer learning [20], and active learning [21] 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 [9]. Ideally multiple weak-supervision based techniques are combined together in order to increase the overall accuracy. Techniques such as distant supervision [22] and crowd-sourced labels [23] are often associated with weak supervision based models, however, they tend to have limited coverage and accuracy [9]. 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 [10]. Developing good labelling functions for the given data rather than gathering manual labels has proven to be far more effective than typical annotation methods [9]. 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 [24], [25]. 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 [26], persuasive essays [25], legal documents [27] and weak-supervision approach to identify claim-sentences from unstructured data [26], [28].

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 [29], [30]. 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 [31] 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. [32] and [33] 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 [34] and [35], and also an NER dataset constructed by [36] for the financial domain. There is a wide literature on sentiment analysis task undertaken on financial data [35], [37], [38], [39].

Works of [40] and [41] 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. [3] used embeddings created using RoBERTa model for identification of emerging technologies. [4] create a dictionary of Environmental and Social (E&S) phrases, while [42] leveraged word-embeddings to measure the corporate culture. Moreover, multimodal machine learning was used by [2] and [43] for credit rating prediction and measurement of persuasiveness respectively. [1] investigated biases in the multimodal analysis of financial earnings calls. Finally, [44] 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

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

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 [13]. 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 [45], [46], [47], [48] and [49] 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.

TABLE 2
Size of Dataset

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

We apply multiple filters to remove data that is not materially relevant for our analysis. Blacklisting helped us remove 66.7% of total sentences 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 the annual fundamental COMPUSTAT database. We find that the maximum number of reports belong to the Healthcare 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 Healthcare 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 [31], 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.

TABLE 3
Comparison of our dataset with NumClaim dataset

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

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 a basic background in 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) is 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 a 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 an 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 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 the description of the pre-trained BERT model used for comparative analysis. We use BERT-base model to better understand the accuracy of our Weak-Supervision model as BERT can serve as a good representative of modern Transformer based models.

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

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

4.1 Weak-Supervision Model

For implementing a weak-supervision model we use the Snorkel library [7], leveraging its inherent pipeline structure for generating labels for each data segment and then passing the outputs through the curated aggregation 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 the categorization of results and also made use of multiple other labelling functions to segregate the sentences representing assertions or written in the past tense. These labelling functions are listed in Table 4.

TABLE 5
Description of output from each labelling function

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

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 values as listed in Table 5. In the final results, both levels have been considered for in-claim sentences. This fine grained categorization helps us understand the results better and opens room for future fine-tuning of the models. For our model, each labelling function classifies a sentence independently, and hence we consider the 'max' as our aggregating function as shown below:

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

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 on the number of labelling functions. For this plot, we initially computed the 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 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.

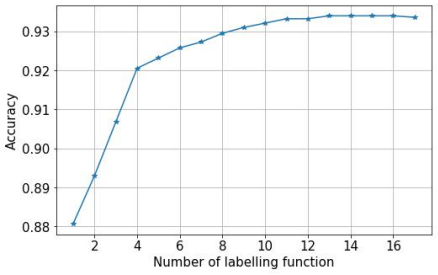

Fig. 1. Overall Accuracy v/s Number of labelling functions

TABLE 6
Three different training data used to train BERT model

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

4.2 BERT

For our experiments we have made use of the *bert-base-uncased* model [18]. In order to perform a 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 an 80-10-10 ratio. Through this experiment, we compare the performance of the weak-supervision approach and BERT keeping training, validation, and testing data same.

TABLE 7
Performance of WS model on gold labels

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

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 the performance of the model.

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

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, highlight 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.

Robustness Check: From the data engineering perspective, there can be a concern about the model design and gold data construction as authors who designed the weak-supervision model have annotated the data. This can lead to exaggerated performance on the data, which makes the test set tainted to some extent. To ensure that there is no bias in our model based on the dataset and it is generalizable, we get the test dataset annotated separately by four annotators with a master's degree in Quantitative Finance. These annotators had no information about rules/patterns used in our weak-supervision model. Each data point in test dataset is annotated by two annotators, and we drop the observations where there is a disagreement among annotators. The test accuracies of the weak-supervision model on a dataset annotated by non-authors for five different seeds are reported in Table 10. We also recalculate the the accuracy of model on the author annotated dataset after dropping observations dropped in a non-author annotated dataset. The model gives higher accuracy on data points where there is an agreement (match) in author and non-author annotated datasets. Overall Table 10 shows the robustness of our model on a dataset annotated by annotators without any knowledge of rules used in the weak-supervision model. From here onwards, the performance is always calculated on a gold dataset created by authors.

Baseline Aggregation Functions: We consider majority vote and Snorkel's aggregation function [7] as baseline aggregation functions for comparative analysis. The accuracy of baseline aggregation functions along with our aggregation function is reported in Table 11. For all three models, labeling functions are used and they only differ in the aggregation part. The result highlights the importance of the construction of a customised aggregation function for a weak-supervision model where a small set of labeling functions are complete and less noisy.

5.2 BERT

As discussed in Section 4.2, we perform three major experiments using the BERT base model. We execute the experiments by taking five different seeds and the average accuracy is listed in Table 12. Accuracy for five different seeds are listed in the supplemental material. From Table 12, we can comment upon the results of the targeted experiments listed in Section 4.2.

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	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 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.

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 10
Accuracy analysis of our model on dataset annotated by non-author annotators.

Seed	Author Annotated	Non-author Annotated	Matching
42	0.9403	0.9104	0.9560
149	0.9508	0.9432	0.9682
1729	0.9022	0.9098	0.9390
13832	0.9655	0.9693	0.9841
110656	0.9326	0.9251	0.9490
Avg.	0.9382	0.9314	0.9594

TABLE 11
Accuracy analysis of our aggregation function with baseline aggregation functions.

Seed	Our	Majority Vote	Snorkel's Label Model
42	0.9404	0.6988	0.6951
149	0.9479	0.6840	0.6840
1729	0.8996	0.5985	0.6059
13832	0.9553	0.6394	0.6356
110656	0.9330	0.6431	0.6431
Avg.	0.9353	0.6468	0.6475

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 the dataset produced using the 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.

TABLE 12
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.

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

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 their 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 cost 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 instances. Carbon footprint calculator developed by [50] is used for calculation of CO_2 emission.

Cost calculations for all the models mentioned in Table 12 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 12, 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 showcase 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 12 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%. Developed customised aggregation function outperforms baseline aggregation functions. 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 a cost-value analysis of weak-supervision based annotation compared to human annotation revealing that our model can serve as an ideal replacement for black-box models in resource constrained environment. We find that the weak-supervision model (WS) is the 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.
- Comparative analysis of Pre-trained language models (PLMs) can be done for the numerical claim detection task.

Acknowledgements

The authors would like to thank DA-IICT, Gandhinagar and Georgia Institute of Technology, USA for the kind support and co-operation to carry out this research work.

References

[1] R. Sawhney, A. Aggarwal, and R. Shah, "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*, 2021, pp. 3751–3757.

[2] C. V. Nguyen, S. R. Das, J. He, S. Yue, V. Hanumaiah, X. Ragot, and L. Zhang, "Multimodal machine learning for credit modeling," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2021, pp. 1754–1759.

[3] S. Chava, W. Du, and N. Paradkar, "Buzzwords?" *Available at SSRN 3862645*, 2019.

[4] S. Chava, W. Du, and B. Malakar, "Do managers walk the talk on environmental and social issues?" *Available at SSRN 3900814*, 2021.

[5] R. Sawhney, A. Wadhwa, S. Agarwal, and R. R. Shah, "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*. Online: Association for Computational Linguistics, Jun. 2021, pp. 4018–4030. [Online]. Available: https://aclanthology.org/2021.naacl-main.316

[6] J. Zhang, Y. Yu, Y. Li, Y. Wang, Y. Yang, M. Yang, and A. Ratner, "Wrench: A comprehensive benchmark for weak supervision," *arXiv preprint arXiv:2109.11377*, 2021.

[7] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," in *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, vol. 11, no. 3. NIH Public Access, 2017, p. 269.

[8] P. Varma and C. Ré, "Snuba: Automating weak supervision to label training data," in *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, vol. 12, no. 3. NIH Public Access, 2018, p. 223.

[9] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *The VLDB Journal*, vol. 29, no. 2, pp. 709–730, 2020.

[10] P. Lison, J. Barnes, and A. Hubin, "skweak: Weak supervision made easy for nlp," *arXiv preprint arXiv:2104.09683*, 2021.

[11] H. Zamani and W. B. Croft, "On the theory of weak supervision for information retrieval," in *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, 2018, pp. 147–154.

[12] M. Shi, A. Hoffmann, and U. Rüppel, "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*. CRC Press, 2021, p. 223.

[13] C.-C. Chen, H.-H. Huang, and H.-H. Chen, "Numeral attachment with auxiliary tasks," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1161–1164.

[14] N. Jegadeesh and W. Kim, "Do analysts herd? an analysis of recommendations and market reactions," *The Review of Financial Studies*, vol. 23, no. 2, pp. 901–937, 2010.

[15] R. Michaely and K. L. Womack, "Conflict of interest and the credibility of underwriter analyst recommendations," *The Review of Financial Studies*, vol. 12, no. 4, pp. 653–686, 1999.

[16] S. A. Corwin, S. A. Larocque, and M. A. Stegemoller, "Investment banking relationships and analyst affiliation bias: The impact of the global settlement on sanctioned and non-sanctioned banks," *Journal of Financial Economics*, vol. 124, no. 3, pp. 614–631, 2017.

[17] B. Coleman, K. J. Merkley, and J. Pacelli, "Human versus machine: A comparison of robo-analyst and traditional research analyst investment recommendations," *The Accounting Review, Forthcoming*, 2021.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[21] B. Settles, "Active learning literature survey," 2009.

[22] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "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*, 2009, pp. 1003–1011.

[23] M.-C. Yuen, I. King, and K.-S. Leung, "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*. IEEE, 2011, pp. 766–773.

[24] M. Lippi and P. Torroni, "Context-independent claim detection for argument mining," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. AAAI Press, 2015, p. 185–191.

[25] C. Stab and I. Gurevych, "Identifying argumentative discourse structures in persuasive essays," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 46–56. [Online]. Available: https://aclanthology.org/D14-1006

[26] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim, "Context dependent claim detection," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 1489–1500. [Online]. Available: https://aclanthology.org/C14-1141

[27] M. Grabmair, K. D. Ashley, R. Chen, P. Sureshkumar, C. Wang, E. Nyberg, and V. R. Walker, "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*, 2015.

[28] R. Levy, B. Bogin, S. Gretz, R. Aharonov, and N. Slonim, "Towards an argumentative content search engine using weak supervision," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2066–2081. [Online]. Available: https://aclanthology.org/C18-1176

[29] C. Chen, H. Huang, Y. Shiue, and H. Chen, "Numeral understanding in financial tweets for fine-grained crowd-based forecasting," *CoRR*, vol. abs/1809.05356, 2018. [Online]. Available: http://arxiv.org/abs/1809.05356

[30] C.-C. Chen, H.-H. Huang, C.-W. Tsai, and H.-H. Chen, "Crowdpt: Summarizing crowd opinions as professional analyst," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 3498–3502. [Online]. Available: https://doi.org/10.1145/3308558.3314122

[31] C.-C. Chen, H.-H. Huang, and H.-H. Chen, "Numclaim: Investor's fine-grained claim detection," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1973–1976.

[32] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.

[33] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "Finbert: A pre-trained financial language representation model for financial text mining." in *IJCAI*, 2020, pp. 4513–4519.

[34] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge *et al.*, "Finqa: A dataset of numerical reasoning over financial data," *arXiv preprint arXiv:2109.00122*, 2021.

[35] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahur, "Www'18 open challenge: Financial opinion mining and question answering," in *Companion Proceedings of the The Web Conference 2018*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 1941–1942. [Online]. Available: https://doi.org/10.1145/3184558.3192301

[36] J. C. S. Alvarado, K. Verspoor, and T. Baldwin, "Domain adaption of named entity recognition to support credit risk assessment," in *Proceedings of the Australasian Language Technology Association Workshop 2015*, 2015, pp. 84–90.

[37] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 782–796, 2014.

[38] M.-Y. Day and C.-C. Lee, "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)*. IEEE, 2016, pp. 1127–1134.

[39] M. S. Akhtar, A. Kumar, D. Ghosal, A. Ekbal, and P. Bhattacharyya, "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*, 2017, pp. 540–546.

[40] J. Li, L. Yang, B. Smyth, and R. Dong, "Maec: A multimodal aligned earnings conference call dataset for financial risk prediction," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 3063–3070.

[41] R. Sawhney, P. Khanna, A. Aggarwal, T. Jain, P. Mathur, and R. Shah, "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)*, 2020, pp. 8001–8013.

[42] K. Li, F. Mai, R. Shen, and X. Yan, "Measuring corporate culture using machine learning," *The Review of Financial Studies*, vol. 34, no. 7, pp. 3265–3315, 2021.

[43] M. Dalton, S. P. Kerr, W. Kerr, Y. Kim, C. Chatterjee, M. J. Higgins, H. Degryse, M. Brown, D. Hoewer, and M. F. Penas, "Persuading investors: A video-based study."

[44] S. Cao, W. Jiang, B. Yang, and A. L. Zhang, "How to talk when a machine is listening: Corporate disclosure in the age of ai," National Bureau of Economic Research, Tech. Rep., 2020.

[45] Investopedia, "Financial term dictionary from investopedia." [Online]. Available: https://www.investopedia.com/financial-term-dictionary-4769738

[46] Vocabulary.com, "Personal finance and financial literacy." [Online]. Available: https://www.vocabulary.com/lists/1504643

[47] MyVocabulary.com, "Business, finance and economics vocabulary

word list." [Online]. Available: https://myvocabulary.com/word-list/business-finance-and-economics-vocabulary/
[48] TheStreet, "Financial word dictionary." [Online]. Available: https://www.thestreet.com/topic/46001/financial-glossary.html
[49] MyVocabulary.com, "Finance vocabulary word list." [Online]. Available: https://myvocabulary.com/word-list/finance-vocabulary/
[50] L. Lannelongue, J. Grealey, and M. Inouye, "Green algorithms: Quantifying the carbon footprint of computation," *Advanced Science*, vol. 8, no. 12, p. 2100707, 2021.

Pratvi Shah received a B.Tech. degree in Information and Communication Technology with a minor in Computational Science from DA-IICT in Gandhinagar, India, in 2022. She is currently working as a Technical Analyst at Goldman Sachs and has worked as an undergraduate researcher with the Computational Science group at DA-IICT and the Georgia Tech FinTech Lab since 2021. Her research interests include the use of machine learning, weak supervision, and parallel computing to improve information retrieval, specifically in the domain of finance.

Arkaprabha Banerjee received a B.Tech. degree in Information and Communication Technology with a minor in Computational Science from DA-IICT - Gandhinagar, India in 2022. He has been actively involved as an undergraduate researcher with the Computational Science group at DA-IICT and the Georgia Tech Financial Services Innovation Lab since 2021. His previous work experience involves working with Samsung R&D India for an academia-industry research collaboration from 2020-21 and Flipkart India as a Software Developer since 2021. His research interests are primarily in the domain of Machine learning focussing on explainable AI and information retrieval , High performance computing and Scientific data analysis and visualization. His work on parallel computing was also accepted at the IWOMP '21 conference.

Agam Shah received a Bachelor's degree (Hons.) in Information and Communication Technology with minor in Computational Science from Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), where he was awarded the President's Gold Medal for graduating at the top of the 2019 class. He also completed Master's degree - MS in Quantitative and Computational Finance (QCF) from Georgia Tech. He is currently working towards PhD degree in Machine Learning from College of Computing, Georgia Tech. Prior to PhD, he worked at Financial Services Innovation Lab (FinTech Lab) of Georgia Tech as a Pre-Doctoral Research Fellow (Pre-Doc). His primary research interests include Data Engineering, Computational Science, Natural Language Processing, and Computational Finance.

Dr. Bhaskar Chaudhury is currently a professor of Computational and Data Sciences at DA-IICT, Gandhinagar. He received his Ph.D. in Computational Physics at Institute for Plasma Research, India. Prior to joining the faculty at DA-IICT, he worked as a researcher for around six years at LAPLACE Laboratory, CNRS, France. He has been involved in active research in the areas of Computational Science, Data Science, Computational Physics and High Performance Computing. He has been the Principal Investigator for several research projects sponsored by the Department of Atomic Energy, Department of Science and Technology, Indian Space Research Organization, National Supercomputing Mission, GUJCOST, National Science Foundation, USA and NVIDIA Corporation, USA. He has authored more than 100 research papers in reputed peer reviewed journals/ conference proceedings/ book chapters. He has received several international/national awards in the area of Computational Science such as prestigious Oscar Buneman Award for the most insightful scientific visualization at 21st ICNSP, Lisbon, Portugal; Swarnajayanti Purashkar 2005 by the National Academy of Sciences, India; Best Presentation Prize awarded by IOP, at International Center for Theoretical Physics (ICTP), Trieste, Italy, 2006. He is a reviewer of various scientific/engineering journals and currently serves as an Associate Editor for Frontiers in Physics Journal.

Dr. Sudheer Chava is the Alton M. Costley Chair, a Professor of Finance at Scheller College of Business at Georgia Institute of Technology, Atlanta and leads the Financial Services Innovation Lab. He also serves as Finance Area Coordinator at Scheller and as the director of the nationally top 10 ranked Masters in Quantitative and Computational Finance (QCF) program at Georgia Tech (a joint program of School of Mathematics, Industrial and Systems Engineering and Scheller College of Business).

Sudheer Chava received his PhD from Cornell University in 2003. Prior to that he has an MBA degree from Indian Institute of Management – Bangalore, an undergraduate degree in Computer Science Engineering and worked as a fixed income analyst at a leading investment bank in India. He joined Georgia Tech in 2010.

Dr. Chava has taught a variety of courses at the undergraduate and master's level including FinTech Ventures, Derivatives, Risk Analytics, Valuation, Cases in Financial Crisis, Management of Financial Institutions, Computational Finance and Credit Risk Analytics. He has also taught both theoretical and empirical finance courses at the doctoral level.

Dr. Chava's research interests are in Credit Risk, Banking, FinTech, Household Finance, Empirical Asset Pricing and Corporate Finance. He has published extensively in the top journals in Finance including Journal of Finance, Journal of Financial Economics, Review of Financial Studies, Management Science, Review of Finance, Journal of Monetary Economics and Journal of Financial and Quantitative Analysis. His research has won a Ross award for the best paper published in Finance Research Letters in 2008, was a finalist for Brattle Prize for the best paper published in Journal of Finance in 2008 and was nominated for the Goldman Sachs award for the best paper published in Review of Finance during 2004. Dr. Chava is the recipient of multiple external research grants such as FDIC-CFR Fellowship, Morgan Stanley Research grant, Blackrock Prize for the best paper at the Australasian Finance Conference, Financial Service Exchange Research grant, Q group research award (2010, 2012) and GARP research award. He has presented his work at numerous finance conferences such as AFA, WFA, EFA, FDIC and Federal Reserve Banks and at many universities in the U.S. and abroad.