Audit Report Coverage Assessment using Sentence Classification

Sushodhan Vaishampayan, Nitin Ramrakhiyani, Sachin Pawar, Aditi Pawde*
Manoj Apte, Girish Keshav Palshikar
TCS Research, Tata Consultancy Services Limited, India.
{sushodhan.sv, nitin.ramrakhiyani, sachin7.p, manoj.apte, gk.palshikar}@tcs.com
aditi.pawde@walchandsangli.ac.in

Abstract

Audit reports are a window to the financial health of a company and hence gauging coverage of various audit aspects in them is important. In this paper, we aim at determining an audit report's coverage through classification of its sentences into multiple domain specific classes. In a weakly supervised setting, we employ a rule-based approach to automatically create training data for a BERT-based multi-label classifier. We then devise an ensemble to combine both the rule based and classifier approaches. Further, we employ two novel ways to improve the ensemble's generalization: (i) through an active learning based approach and, (ii) through a LLM based review. We demonstrate that our proposed approaches outperform several baselines. We show utility of the proposed approaches to measure audit coverage on a large dataset of 2.8K audit reports.

1 Introduction

Financial audit is a complex process used by organizations to assure the stakeholders about the quality and trustworthiness of the governance (Whittington and Pany, 2021), (Arens and Loebbecke, 1999). Auditors examine data, documents, systems and processes, physical assets to ensure that they comply with the required standards, guidelines, laws and regulations and also to ensure that the reported financial information is fair and accurate. Outside the organization, stakeholders use audited financial statements (FS) - such as balance sheet, income statement, cash-flow statement etc.- for making important decisions such as investments, loans, taxation and so forth. One important outcome of an audit is the *audit report* prepared by the auditors, wherein the auditor declares the FS are free from material misstatement, are fair and accurate and are presented in accordance with the relevant accounting standards. If not, the auditor identifies several types of issues, makes suggestions for improvement, and identifies instances of non-conformance, misinformation, irregularities, inconsistencies, errors, inaccuracies, frauds, lapses, non-compliance, violations etc.

Given the crucial importance of audits, and the high demands on the knowledge, experience and efforts of the auditing team, it is important to measure the *quality* of an audit in order to ensure that it was carried out efficiently and effectively. Poor quality audits, whether intentional or not, can have disastrous consequences, such as frauds, loss of earnings, loss of goodwill, litigations, inability of the company to function as a going concern and even bankruptcy; e.g., see (Lennox and Li, 2019). There are many reasons why an audit can be of poor quality: lack of expertise in the auditing team (Reichelt and Wang, 2010), compromised auditor independence (Tepalagul and Lin, 2015), biases, conservatism (*recognize bad news rather than good news*) (Basu, 1997) and risk-averse attitudes of auditors, non-cooperation from management, insufficient time/efforts spent in auditing etc. The Sarbanes-Oxley Act 2002 in the US is explicitly aimed at improving auditing and public information disclosure, in the light of persisting scandals fueled by auditing failures such as Enron (Beasley et al., 1999) and Satyam (Bhasin, 2013).

A good comprehensive audit report is an important indicator of a good audit. Audit monitoring bodies such as The Chartered Accountants (CA) Society of India have issued guidelines on the contents of audit reports wherein they describe a set of audit aspects which the auditor should touch upon and describe. In this paper, we focus on measuring the coverage of the audit report based on such statutory requirements, as one of the initial steps to gauge audit quality. We pose the problem of gauging coverage of the audit aspects in an audit report through classification of sentences in the report into one or more of these aspects. Given the

*Work done while working at TCS Research

Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pages 31–41

Class	Description	Example Sentence
approval of managerial remuneration	Compliance as per applicable act and payment for managerial remuneration.	We draw attention to Note 42 to the financial statements relating to managerial remuneration paid which is in excess of the limits approved by the Central Government to the extent of Rs. 214.45 lakhs ...
fraud reporting	Fraud by the company or officers or employees, if any, is mentioned and whether any whistle-blower complaints were received	According to the information and explanations given to us, a fraud on or by the company has not been noticed or reported during the year.
nidhi company	Remarks on type of company: nidhi, chit fund, etc.	In our opinion, the nature of activities of the Company does not attract any special statute applicable to chit fund and nidhi / mutual benefit funds / societies.
non-cash transactions	Remarks on compliance applicable for non-cash transactions with directors and related persons	Cash flows are reported using the indirect method, whereby profit before tax is adjusted for the effects of transactions of non - cash nature ...
private placement or preferential issues	Remarks on whether company has made preferential allotment or private placement of shares	The Company had invested Rs. 1000 million in 8.75% Cumulative Preference Shares of M/S. ITI Limited during the year 2001 - 02.
utilization of ipo and other public offers	Remarks on money raised through IPOs or other public offers	The Company has not granted any loans and advances on the basis of security by way of pledge of shares, debentures and other securities.
Complex Classes		
cost records	A remark about maintenance of cost records.	However, we have not made a detailed examination of the cost records with a view to determine whether they are accurate or complete.
fixed assets	Remarks on purchase of fixed assets, holding of benami property, physical verification of property, plant and equipment by the management at reasonable intervals.	The company has maintained proper records showing full particulars, including quantitative details and situation of fixed assets.
human resources, payroll processing	Remarks on employee wages, leaves, bonus, pension, full and final settlement and mentions of policies for leave, gratuity and pension.	Also Defined benefits obligations in nature of Gratuity and Leave encashment are to be accounted on accrual basis.
internal control system	Remarks on evaluation of internal control procedures with respect to the size and the nature of the company.	During the course of our audit, no major weakness has been noticed in the internal control system in respect of these areas.
inventory	Remarks on possession and purchase of inventory, its physical verification at timely intervals and record keeping	On the basis of the records of inventory, we are of the opinion that the Company is maintaining proper records of inventory and no material discrepancies were noticed on physical verification.
investments	Remarks on investments by the company and compliance to respective Acts	The company has a strategic long term investments in Equity Shares of certain companies, the cost of acquisition of those investments is Rs. 722.50 lacs.
litigations	Remarks about ongoing litigations on the company	Contempt Petition filed against Excise Department at Allahabad High Court against our refund of Rs. 17,25,392/ - against the order of Supreme Court in our favor.
material uncertainty	Remarks on material uncertainties for the company such as net worth, accumulated losses and going concern	The Company 's accumulated losses at the end of the financial year are less than fifty per cent of its net worth.
operational and administrative expenses	Remarks on company's operational expenses	The Company has Capitalized expenses to the tune of Rs. 25.40 Crores in Pulp Mill Unit till the date of last balance sheet...
payables	Remarks on details of amount/money to be paid by the company such as repayment of loans	The repayment of loan is on demand, there is no overdue amount remain outstanding.
purchase and procurement	Remarks on purchases and procurement of any kind	The activities of the Company do not involve purchase of inventory and the sale of goods.
receivables	Remarks on details of amount/money to be received by the company such as loans given	The net amount recoverable of Rs. 23640.05 million is subject to reconciliation and confirmation.
sales, services and revenue	Remarks on sales, services and revenue	The Company is a service company, primarily rendering software services.
statutory dues	Remarks on payment of statutory dues and related disputes	The Company is regular in depositing with appropriate authorities undisputed statutory dues including provident fund, employees ' state insurance ...
working capital	Remarks on working capital and cash/bank balance	No long terms funds have been used to finance short - term except permanent working capital.

Table 1: List of classes in the annotated audit reports with their description and examples

large number of these aspects and domain expertise required to create labelled training data, the text classification problem becomes highly challenging. In this regard, we propose a weakly supervised text classification algorithm based on regular expression based patterns and a multi-label BERT based classifier. To supplement the approach for increasing its recall, we explore two directions - (i) using active learning requiring manual labelling effort and, (ii) using support from LLMs, requiring effort on prompt creation. We present our experimentation and analysis on a dataset of audit reports of companies based in India discussing their audits for the year 2014. To demonstrate the impact of the learning from this work, we present a brief statistical analysis on the dataset.

2 Problem Definition

An audit report consists of various sections mentioning details about a company being audited, responsibility of management and auditor followed by remarks or comments by the auditors pertaining to company's business operations. Generally, auditors adhere to standard *audit checklist* that includes scope of the audit, evidence collection, audit tests, result analysis and conclusions to be drawn from audit. Moreover, auditors also have to comply with any legislation by local regulatory bodies. Since, the goal in this paper is to determine the *audit coverage* and data under consideration is of Indian companies, the coverage is checked with respect to a standard auditing checklist (ICAI, 2017) and Companies (Auditor's Report) Order, 2020 (CARO) (ICAI, 2020). Accordingly, the union of classes from both these sources is considered as given in table 1. A sentence in audit report can belong to 0, 1 or more labels from this list. Thus, this is a multi-class multi-label classification problem with number of class labels $m = 21$. Sentences that do not belong to any class label, are considered to be *Not applicable* or *NA*. In Table 1, we list the classes with a brief description and an example sentence from an audit report for each class.

3 Proposed Approach

We propose a *sieved* approach which combines the power of multiple techniques such as Rules, a Standard BERT based classifier, Active Learning and Large Language Models. We explain the contribution of each of the techniques individually and then how they are combined in an ensemble for the final prediction on test data.

3.1 Rules - Regular Expression based Patterns

As can be observed in Table 1, sentences belonging to certain classes are clearly amenable for rule based labelling. For e.g., sentences in classes such as *Nidhi Company* and *non-cash transactions* typically mention the class names in exact and very rarely in a different format. This exactness is by virtue of how auditors are trained to mention their findings about these aspects/classes. Hence, this facet prompts us to use rules in the form of regular expression patterns for a precise identification of these specific classes.

We devise regular expression based patterns which are constructed by tokens indicative of the respective class. In Table 2, we show some of the regular expression patterns for a subset of classes. Consider for example, the regular expressions for the class *Fixed Assets*. As can be seen tokens such as `intangible` or `immovable` followed by tokens such as `assets` or `properties` would be indicative of the *Fixed Assets* class. For certain classes, the rule may be built of more than one component patterns and all pattern components must match in the sentence, though in any order, for the class to get predicted. An example is seen for the class *Litigations*, where the first component searches for words such as `cases` or `appeals` and the second component searches for for words such as `courts` and `tribunals`. The regular expressions also involve negative look-aheads such as the second pattern for the class Material Uncertainty in Table 2. It ensures that a phrase such as `no uncertainty` or `not uncertain` when is observed, labelling to the class *material uncertainty* is avoided. We also develop patterns for indicating sentences which are template sentences that auditors include as part of the report and should be marked with a NA label. The rule based classifier labels them with the *NAconfirm* label which is treated as *NA* during evaluation.

As the classification problem is multi-label in nature and the rules may predict multiple labels for a sentence leading to no conflicts. This makes it little different from Snorkel (Ratner et al., 2017) like data programming paradigms.

3.2 Multi-label Sentence Classifier

We also observe that there are sentences wherein the belongingness to the corresponding class is not lexically closed and hence classification using only

Class	Regular Expression Pattern
fixed assets	\b((fixed|intagible|immovable)(assets?|propert(ylies)))\b
litigations	\b(litigations?|cases?|arbitrations?|appeals?|matters?|disputes?)\b AND \b(courts?|tribunals?|judges?|nclt)\b
material uncertainty	\b(financial|debts?)re\W?structur(e[d]?|ing)\b | \bre\W?structur(e[d]?|ing)\b.∗?\b(debts?)\b
material uncertainty	(^((?!\bnot?\b).∗?)\buncertaint(ylies)\b | \b(no|not)(\w+)? (certaint(ylies)|ascertain(ed|ing)?|ascetainable))\b(statements?)\b
statutory dues	\b(tax(es)?|provident funds?|(customs?|excise)duty|duty of (customs?|excise))\b AND \bdues?\b

Table 2: Example Regular Expression Patterns

lexical patterns may not be sufficient. For e.g., sentences in classes such as *payables* and *fixed assets* mention about the payables and assets in various ways apart from few standard ways which rules can capture. To classify such sentences, a more general understanding of the class' sentence is required. Hence, we propose the use of a BERT based multi-label multi-head attention sentence classifier.

3.2.1 Network Architecture

The classifier works on contextual embeddings of the input tokens obtained from a transformer's encoder such as BERT. Any other encoder such as RoBERTa (Liu et al., 2019) can be employed. These encoder architectures emit the input sentence's representation for the CLS token and embeddings for each of the tokens. Additionally, a domain specific feature extraction module processes the input sentence to emit a k-hot representation denoting presence of audit report specific phrases. This module is currently devised to simply recognize audit domain specific phrases and emit a 1 in the slot for the phrase in the representation. This k-hot representation is then passed to a linear layer to emit a dense representation and its weights are learnt during training. These phrases have been collected upon observation of multiple audit reports and the k-hot representation size is equivalent to the number of these phrases. It is important to note that the domain specific feature extractor is a generic component and can be generalized in ways suitable to the classification problem.

Following this input processing, the architecture consists of multiple class-specific classification heads formed of a combo of an attention layer, a hidden layer and softmax layer. Having such classification heads for each class is necessary as the problem is a multi-label classification and this provides the necessary one-vs-all arrangement. We hypothesize that the class specific attention heads should learn about specific tokens which are indicative of the class and get tuned, while training, to signal for the class, while inference. The attention layer would then emit a sentence representation re-weighting the token embeddings giving more importance to tokens highly indicative of the class. The consequent hidden layer takes as input a concatenation of the CLS representation, the attention layer emitted representation and the domain specific feature based representation. The softmax layer post this performs the class vs not_class classification. During inference, whichever classification head emits a confidence of 0.5 or greater, the respective class is added to the list of predicted classes for the input sentence. For a detailed network diagram, refer to Appendix D.

3.2.2 Training Data

It is important to note however, that creating annotated data is effort and time consuming and requires domain expertise. With unavailability of annotated data for training the classifier, we resort to weak supervision wherein we label a large set of audit report sentences automatically using the rules devised earlier. We consider a large number of audit reports (See Section 4.1) and run the rule based classification to collect about 16K sentences. After removal of near duplicates, we arrive at about 4.9K sentences which we consider for training the classifier. Additionally, we train the classifier for only 15 out of the above 21 classes (classes marked *Complex Classes* in Table 1), given the understanding that the rest of the 6 classes are easily recognizable through the pattern based rules.

3.3 Ensemble with Rules based classification

To combine the power of both classification approaches and also to check how much generalization the classifier has been able to achieve, we combine them in an ensemble. We allow the rules to first predict the set of possible classes P for an audit report sentence S. Now if the classifier predicts a label for S with confidence greater than 0.5, which is not already in P, the label is added to P, allowed as per the multi-label setting of the classification. This new label prediction may happen if the clas-

sifier has observed certain class indicative aspects of the sentence which the rules have failed to exploit. Only in cases when the rule based approach has predicted the *NAconfirm* class for the sentence, we refrain from predicting using the classifier and predict only the *NA* label.

3.4 Boosting Generalization of the approach

We hypothesized that the classifier, trained on the data labelled by the rules, may not generalize well on sentences which are not labelled by the rules and hence the approach may require more support in terms of generalization in understanding the classes. We explore two ways to boost the generalization of the approach.

3.4.1 Active Learning

One way to achieve the necessary generalization is to add a set of sentences which are not getting classified by the rules and classifier to the training data. To perform this addition in a methodical and an effective manner, we take help of the Active Learning paradigm.

Active learning is a strategy to select some instances from the dataset which are hardest to be correctly classified by the trained classifier. These selected sentences are then added to the training data and the classifier is trained using this supplemented dataset. For finding the sentences which are the most difficult to classify, we developed a strategy called Closest-To-Local-Midpoint (CTLM) which is a modified version of the query synthesis procedure in Wang et al. (2015). For each pair of classes, this strategy selects those sentences which are approximately equidistant from both the classes. The details of the strategy is present in Appendix C. Sentences selected using this approach are added to the training dataset and the classifier is retrained (C_{AL}). We use C_{AL} as part of the ensemble approach and report the results.

3.4.2 LLM Review

In the original ensemble of the rules and the classifier, we add only the labels from the classifier which have a prediction confidence of 0.5 or greater. Another way to increase the generalization capability of the ensemble approach, is to get the classifier's low confidence predictions (less than 0.5) reconsidered by an independent reviewer. This is to harness those cases where the classifier may have spotted the correct class, but is less confident. We enable this independent review with the help of a LLM by prompting it to re-confirm or abandon a candidate label. With this we also pseudo-enact a scenario of considering a LLM as a domain expert which understands these aspects of audit coverage.

To enable this LLM review, we first collect all sentences for which the classifier has predicted atleast one class with confidence less then 0.5 and higher than 0.1. We decide this lower bound, empirically. For each sentence and such possible class, we prepare a prompt consisting of the sentence and the class' description (Table 3). We then probe the LLM using the prompt and find out whether the class being tested is really applicable. The class descriptions used in the prompts are based on the descriptions provided in Table 1 earlier.

The domain expert further commented that certain classes may get a higher benefit when using language models to discern them, given the larger amount of financial audit discourse those classes are discussed in. His suggested classes were: *payables*, *fixed assets*, *litigations* and *inventory*. To confirm his hypothesis, we devised a small recall measurement experiment. We used the dataset of 4.9K sentences labelled by the rules (the same data used as training data for the classifier) and ran the above LLM review on it, to confirm the rule based label. As this data is not manually labelled, we simply measure the recall i.e. number of sentences where the LLM could successfully confirm the rule based label and not miss it. We ordered the classes according to recall and found out that not only did the expert suggested four classes appear in the top 6 (recall $>= 0.8$), but also introduced us to 2 more similar classes: cost records and internal control system. We term these 6 classes as LLM-HR (High Recall) classes. As the final approach - Ensemble (R + C + LLM-HR), for a sentence if the classifier has a low confidence prediction which is one of these high recall classes, we review the label using the LLM. If confirmed, we add the class to the existing set of labels provided by the ensemble.

4 Experiments and Evaluation

4.1 Dataset

We used the web-scraped audit reports made available by authors of (Maka et al., 2020). We consider 3744 reports for the year 2014 from which 932 reports which were too small (less than 35 sentences) or too noisy were removed leading to a final set of 2812 reports.

Sentence: We are of the opinion that in view of Memorandum of Settlement with the workers the company should make a provision of crystallized dues of Rs 40 Crores, irrespective of sale of Mohali Assets	
Classifier Predictions: fixed assets (0.187), sales, services and revenue (0.141)	
Prompt Template: <Sentence>. *Does the previous sentence talk about* <Class description>? *Answer as Yes or No.*	
Class	**Final Prompt**
fixed assets	We are of the opinion that in view of Memorandum of Settlement with the workers the company should make a provision of crystallized dues of Rs 40 Crores, irrespective of sale of Mohali Assets. *Does the previous sentence discuss about fixed assets such as equipment, land, building, plant, machinery and their physical verification? Answer as Yes or No.*
sales, services and revenue	We are of the opinion that in view of Memorandum of Settlement with the workers the company should make a provision of crystallized dues of Rs 40 Crores, irrespective of sale of Mohali Assets. *Does the previous sentence discuss about revenue from sale of goods and services excluding sale of shares and assets? Answer as Yes or No.*

Table 3: Example Prompt Creation

Test dataset: We select a set of 10 audit reports which are labelled manually for the 22 classes (including NA) to form the test set. As part of the annotation guidelines, we used the descriptions in Table 1. Two annotators were part of the annotation exercise, one of which was the domain expert. High inter-annotator agreement was observed and any conflicts were resolved through discussion. The test set consists of a total of 1668 annotated sentences (class-wise statistics in Appendix A).

4.2 Baselines

We experiment with a number of standard machine learning classifiers as baseline approaches. We implement these approaches through the classifiers provided in scikit-learn with their default parameters while setting the class weights as "balanced" wherever possible. It is important to note that we use the same rule-based approach annotated data as training for these classifiers and report results on the 15 complex classes as we do for the BERT-based multi-label classifier.

Additionally, we also try using ChatGPT as a baseline and provide it a suitable prompt (details in Appendix B) to make it predict the suitable classes on the input sentence.

4.3 Experimentation Details

We considered a set of documents separate from the test set to tune the rules, tried few best configurations on the test set and selected the best one. We then use the best set of rules to label the data for creating training data for the classifier and the high recall experiment of the LLM. Further, for hyperparameter tuning of the BERT based classifier, we used a 20% validation split of the training data. Certain important hyperparameters to note are: (batch_size: 8 with gradient accumulation of 8 steps, learning rate: 0.00005, epochs: 16). We used two attention heads in each class' attention module, to attend to two important words in that sentence indicative of the class. We only allowed the final encoder layer in the BERT model to get fine-tuned while keeping all other layers frozen. Also, for the LLM experiment we employed the Falcon-7B-instruct (Almazrouei et al., 2023) model, which is a resource and license friendly model capable of responding to question like prompts, similar to what we have devised.

4.4 Evaluation and Analysis

Approach	**P**	**R**	**F1**
Rules (R)	0.887	0.557	0.684
Naive Bayes†	0.820	0.307	0.446
SVM (Linear)†	0.722	0.698	0.710
Logistic Regression†	0.670	0.742	0.704
Random Forests†	0.844	0.570	0.680
Gradient Boosting†	0.853	0.589	0.697
ChatGPT (Zero-shot)	0.487	0.557	0.520
BERT-based Classifier (C)†	0.849	0.639	0.729
Ensemble (R + C)	0.843	0.652	0.735
Ensemble (R + C_{AL})	0.835	0.692	**0.757**
Ensemble (R + C + LLM-HR)	0.823	0.692	0.752

Table 4: Comparative Performance of different baselines and proposed approaches. († indicates evaluation over 15 complex classes)

In Table 4, we report the performance of the baselines and different proposed approaches. We use precision, recall and F1-score micro-averaged over multiple labels due to the multi-label setting. The rule based approach based on concrete and specific rules performs with the best precision as desired. This is an important reason for using the rule based approach for creating annotated data. Further, we see that in terms of F1-score, some of the baselines such as SVM, Logistic Regression and Gradient Boosting, outperform the rule based approach thereby implying that they are able to

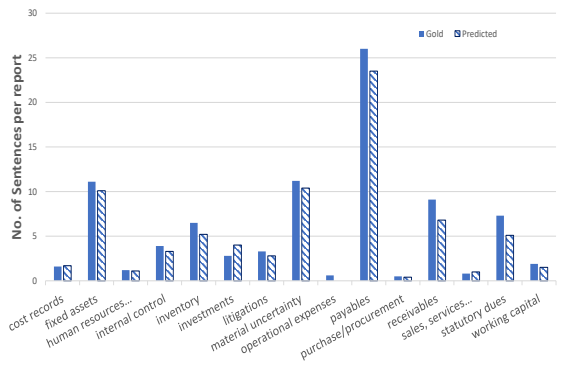

Figure 1: Class-wise coverage in the Test set

generalize even while trained using rules annotated data. The BERT-based classifier outperforms both the rule-based approach and the baseline classifiers. Further the ensemble of rules-based approach and the BERT-based classifier performs slightly better, with increase in recall by 2%.

If we observe the generalization boosting approaches, both the Active Learning (AL) one and the LLM review (LLM-HR) one perform the best and give about a 6% increase in recall, thereby increasing the overall F1 by 2-3%. The AL approach, performs slightly better, but requires 500 labelled sentences to be infused as part of the retraining exercise, which may be difficult to obtain given limited availability of domain expertise. The LLM review approach, though only requires the effort for prompt creation, it does requires domain support to identify the right high recall classes.

In Figure 1, we present how different is the distribution of predicted classes over the test set when compared to the distribution of the classes as per gold labels. The ensemble approach maintains the distribution well (with Jensen-Shannon Divergence of 0.005) and is in conformance to the gold distribution across classes. This makes the ensemble a close approximation of the true underlying distribution of classes and hence worthy for use in analysis on a larger set of reports (Section 5).

On detailed analysis of class-wise results for the best approach: Ensemble (R + C_{AL}), we observe that from the set of complex classes, *cost records*, *internal control system*, *fixed assets*, *working capital*, *human resources*, *utilization of ipo* and *inventory* perform well and have an individual class F1 of 0.8 or greater. Classes which perform moderately well ($0.7 \leq$ F1 $<$ 0.8) include *payables*, *investments*, *receivables*, and *material uncertainty*. There lies scope to improve the performance for these classes. Some of the low performing classes (F1 $<$ 0.7) are *litigations*, *statutory dues*, *private placement or preferential issues*, *purchase & procurement* and *sales, services & revenue*. Investigating further for these classes, we find that due to presence of certain indicative phrases in the sentence, which are used in a different semantic context, confuse the approach. For example, presence of the phrase `the Company has sold and transferred its branded domestic formulations business`, prompts the approach to assign *sales, services and revenue*, which is not valid here as this is not related to sales of products or services, but a business division. Similarly, reference to possible scenarios in the sentence also leads to false positives, such as the following sentence gets labelled as *material uncertainty*, when it is referring to a possible negative implication: `Relying on the assertions as detailed in notes no adjustments have been made in the financials towards possible impairment.`

A small discussion on why ChatGPT performs on the lower side is also important. Firstly, we observed that in spite of specifying to classify the input sentence into the given class names, ChatGPT started predicting new class names formed of phrases related to the correct class name. Secondly, even on specifying to emit multiple relevant classes, it still sticks to predicting only one class. When forced, it starts predicting lots of irrelevant classes. Thirdly, at times it simply classified some of the input sentences and then generically specified that "other sentences can be classified similarly". Overall we believe that in a challenging scenario with large number of domain specific classes with complex semantics, ChatGPT output is not usable in deployed applications.

5 Audit Report Coverage Analysis

Audit report coverage refers to the extent and scope of an audit report, detailing what aspects of an organization's financial statements, instruments and operations have been examined and reported on, by an external auditor.

As iterated earlier, we aim to measure the audit coverage through classification of sentences into audit aspects specified in regulatory checklists. We consider that if a sentence is mapped to a class, it is generally commenting about aspects of that class and in turn achieves the objective of checking that

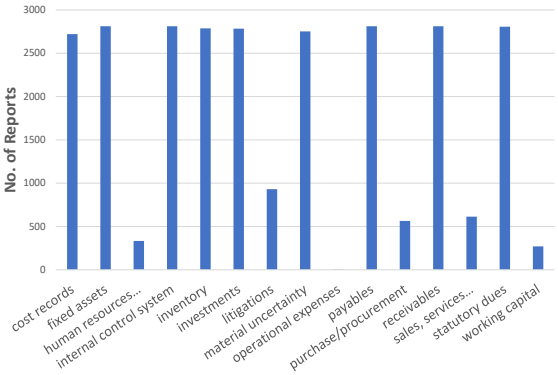

Figure 2: Checklist Coverage

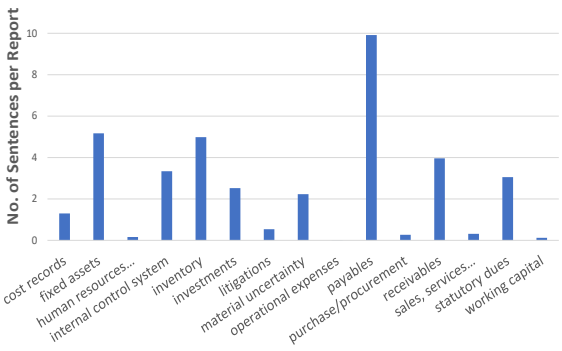

Figure 3: Weighted Coverage

.

class. At the document level, sentences commenting on the various classes can then be an indicator of which classes were checked and reported upon.

We consider the set of 2812 reports for measuring audit report coverage on the complex classes and report on it from two perspectives:

1. **Checklist Coverage**: In this perspective, if a class appears at least once in the audit report then we can consider that aspect is covered in the audit. This is mainly important to check the compliance requirements where the regulatory bodies expect mandatory coverage of specific areas.

Figure 2 shows the number of reports in which the complex classes were reported on at least once. We can conclude that classes like *human resource and payroll processing*, *litigations*, *operational and administrative expenses*, *purchase and procurement*, and *working capital* show considerably lower coverage than other classes. The lower coverage could be due to (a) absence of litigations or (b) lower importance given by the auditor for that aspect. E.g., most companies may not be involved in litigations or issues relating to human resources, hence those aspects may be skipped.

2. **Weighted Coverage**: Through this perspective, the number of sentences specifying a certain class can be considered as weight/importance devoted to the corresponding aspect. This can be used by the stakeholders for analyzing the weightage given by the audit report for a specific aspect, for e.g. while lending money to a firm the bank can check the focus given on aspects like *payables*, *internal control*, *revaluation of fixed assets*, etc.

Figure 3 shows the distribution of weightage for the complex classes over the considered reports. We observe that some of the classes such as *cost records* and *working capital* have comparatively less weightage than classes such as *payables* and *fixed assets*. This helps in understanding the importance auditors place on certain aspects and their implications on the functioning of the company.

6 Conclusion and Future Work

In this paper, our objective was to find whether all the necessary audit aspects are being covered in an audit report. We proposed a set of 21 classes corresponding to these audit aspects. We proposed a weakly supervised approach for automatic multi-class multi-label classification of sentences in an audit report. Due to absence of training data, we use a rule-based technique to automatically create labelled dataset for training a BERT-based sentence classifier. Further, we employ two novel ways to improve the generalization – (i) through an active learning based approach which needs manual annotation efforts and, (ii) through a LLM based review which needs efforts for prompt engineering. Given the complex and domain specific semantics of the classes and unavailability of labelled data, we were still able to achieve the F1-score of more than 75% with our approaches outperforming several baselines. We also showed the utility of the proposed classification approaches to measure audit coverage on a large dataset of 2.8K audit reports.

As part of future work, we would like to explore open source LLMs further for our sentence classification problem. From domain point of view, we plan to extend our techniques for different stakeholders such as regulatory bodies or banks to automatically evaluate the audit reports in an unbiased way. We also plan to apply these techniques for audits in other domains like software quality audits or Environment, Social & Corporate Governance (ESG) audits, using domain knowledge such as audit guidelines from the respective domains.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Alvin A. Arens and James K. Loebbecke. 1999. *Auditing: An Integrated Approach*, 8th edition. Pearson.

S. Basu. 1997. The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting and Economics*, 24:3–37.

Mark S Beasley, Joseph V Carcello, Dana R Hermanson, Committee of Sponsoring Organizations of the Treadway Commission, et al. 1999. Fraudulent financial reporting: 1987-1997: an analysis of us public companies.

Madan Lal Bhasin. 2013. Corporate accounting fraud: A case study of satyam computers limited. *Open Journal of Accounting*, 2(2).

ICAI. 2017. Internal audit checklist. https://kb.icai.org/pdfs/44970iasb34918.pdf. [Online; accessed 8-September-2023].

ICAI. 2020. ICAI'S GUIDANCE NOTE ON CARO 2020 (CARO). https://wirc-icai.org/wirc-reference-manual/part2/icai-guidance-note-on-caro-2020.html. [Online; accessed 8-September-2023].

C. Lennox and B. Li. 2019. When are audit firms sued for financial reporting failures and what are the lawsuit outcomes? *Contemporary Accounting Research*, 37(3):1370–1399.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kiran Maka, S. Pazhanirajan, and Sujata Mallapur. 2020. Selection of most significant variables to detect fraud in financial statements. *Materials Today: Proceedings*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

K.J. Reichelt and D. Wang. 2010. National and office-specific measures of auditor industry expertise and effects on audit quality. *Journal of Accounting Research*, 48(3):647–686.

Nopmanee Tepalagul and Ling Lin. 2015. Auditor independence and audit quality: A literature review. *Journal of Accounting, Auditing and Finance*, 30(1):101–121.

Liantao Wang, Xuelei Hu, Bo Yuan, and Jianfeng Lu. 2015. Active learning via query synthesis and nearest neighbour search. *Neurocomputing*, 147:426–434.

Ray Whittington and Kurt Pany. 2021. *Principles of Auditing and Other Assurance Services*, 22 edition. McGraw-Hill Education.

A Test Dataset Statistics

The test set consists of a total of 1668 annotated sentences. The class-wise statistics are presented in Table 5.

Class	#Annotated Sentences
NA	733
payables	260
material uncertainty	112
fixed assets	111
receivables	91
statutory dues	73
inventory	65
internal control system	39
litigations	33
investments	28
private placement or preferential issues	23
working capital	19
cost records	16
human resources and payroll processing	12
nidhi company	10
utilization of ipo and other public offers	10
fraud reporting	10
sales, services and revenue	8
operational and administrative expenses	6
purchase and procurement	5
approval of managerial remuneration	3
non-cash transactions	1

Table 5: Class-wise annotations

There were 5 other classes that were defined based on the auditing checklist namely *corporate social responsibility*, *resignation of statutory auditors*, *remarks by auditors of included companies*, *related party transaction* and, *register under rbi act*. As these 5 classes were not present in the labelled data, we decided to include only the ones shown in Table 1, in the current analysis.

B Description of the ChatGPT Prompt

We use ChatGPT's user interface to perform the classification of the sentences in the test set by

prompting it with suitable prompts. The prompt consists of a main instruction, descriptions of the 15 complex classes and finally a set of sentences to classify. The prompt template is shown in Table 6, where text in round brackets is for explanation only. As can be seen, that this is a zero-shot setting of classifying using an LLM. A few shot setting, as part of in-context learning, can also be tried where examples of sentences and their gold class can be provided. However, selection of the classes to give as examples and maintaining the instruction's context are some important challenges, exploration of which we keep as future work.

(—*Main Instruction*—)
The task is to classify sentences in a financial audit report into one or more of the following classes. Each line below mentions a class name followed by its description.

(—*Class Descriptions*—)
1. cost records: About maintenance of cost records.
2. fixed assets: About fixed assets such as equipment, land, building, plant, machinery and their physical verification.
3. human resources and payroll processing: About human resources and payroll processing such as employee wages, leaves, bonus, pension, full and final settlement, policies for leave, gratuity or pension.
4. internal control system: About internal control procedures.
. . .
14. statutory dues: About depositing statutory dues like provident fund, ESI, income tax, sales tax, VAT, service tax, GST, duty of customs, duty of excise.
15. working capital: About working capital, cash credit and bank balance.

(—*Input Sentences for Classification*—)
What are the applicable classes for the following sentences? Simply print the output as Sentence ID: Class name.
Sentence 1: We have audited the accompanying financial statements of ...
Sentence 2: Management is responsible for the preparation of these financial statements that give a true
. . .
Sentence 10: We conducted our audit in accordance with the Standards on Auditing issued ...

Table 6: ChatGPT Prompt Template

C Details about the Active Learning strategy

For finding the sentences which are the most difficult to classify, we developed a strategy called Closest-To-Local-Midpoint (CTLM) which is a modified version of the query synthesis procedure in (Wang et al., 2015). In CTLM, we first find the center of the cluster having all the sentences belonging to a class in Euclidean space. To elaborate, let us assume, we have a class C_1. We know from the predefined rules (as mentioned in Section 3.1), a set S_{C_1} of sentences belonging to C_1. We transform each sentence $s \in S_{C_1}$ to a vector $\mathbf{\Delta}_s \in \mathbb{R}^{300}$, by taking the average of the Glove embeddings (Pennington et al., 2014) of each word in s. The words from a predefined set of insignificant stop words are omitted while computing $\mathbf{\Delta}_s$. Once we have the respective vectors for each of the sentence belonging to C_1, we find the center $\mu(\mathbf{\Delta}_{C_1})$ of C_1 by computing the mean vector of all the transformed sentences. Given a set $C = \{C_1, C_2, \cdots, C_m\}$ of m (here 15) such classes, we have a set $\mu(\mathbf{\Delta}_C) = \{\mu(\mathbf{\Delta}_{C_1}), \mu(\mathbf{\Delta}_{C_2}), \cdots, \mu(\mathbf{\Delta}_{C_m})\}$ of their respective centres, which are the representatives of the respective classes. Now we find the sentences that should be difficult to classify. We find the pairwise mid-points of mean vectors of the classes in 300-dimensional space and select the sentences which are nearest to these midpoints. The intuition is that, *the sentences which are approximately equidistant from the cluster centres of two classes will be classified with lowest confidence of belonging strictly to a single class*. As there are large number of sentences common between most of the audit reports, the sentences closest to midpoints of different pairs could be very similar. We want sentences as dissimilar as possible, so the classifier can learn different aspects. To avoid this we select a large number of sentences (2000 here) per pair in descending order of cosine similarity. From these sentences, we removed the common sentences and sentences which were very similar. After removing these common and similar sentences, we were left with 477 distinct, dissimilar and toughest to classify sentences. We ensured that these sentences are ones where the rules are unable to predict any class. These sentences were then labeled by the annotators and then were added to the training set of the classifier.

D BERT-based classifier Network Diagram

The neural network diagram of the BERT-based classifier is shown in Figure 4.

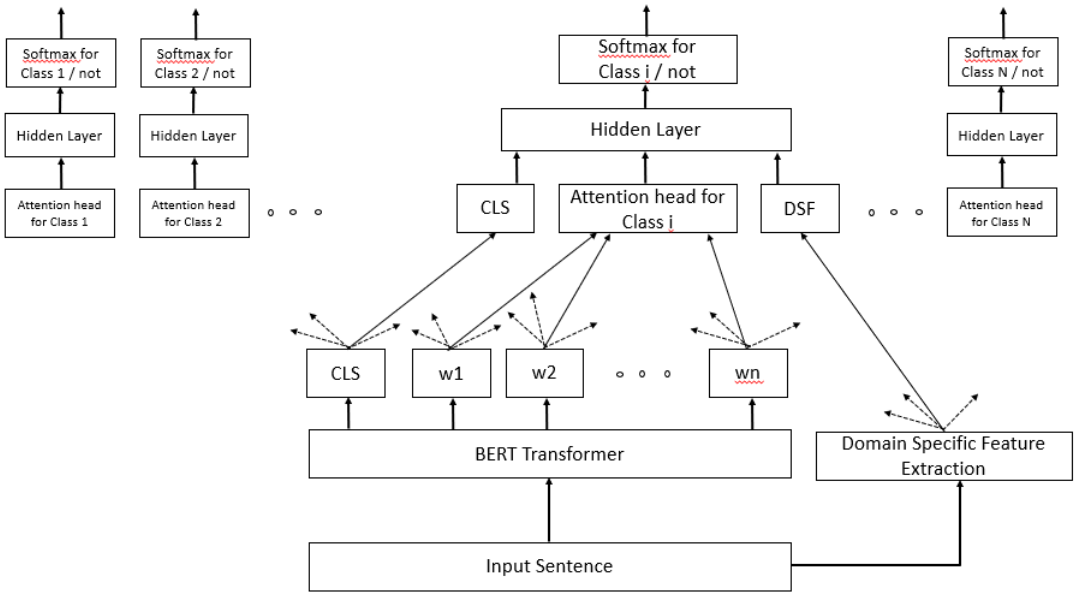

Figure 4: BERT based Multi-label Multi-headed Attention Classifier