National College of Ireland

Hindi FinBERT:A Pre-trained Language Model for Financial Text Classification

MSc Research Project

MSc in Artificial Intelligence

Ayush

Student ID: X22186590

School of Computing

National College of Ireland

Supervisor: Dr. Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ayush
Student ID:	X22186590
Programme:	Artificial Intelligence 2023-24 **Year:**
Module:	MSc Practicum
Supervisor:	Rejwanul Haque
Submission Due Date:	14/12/2023
Project Title:	Hindi FinBERT: A Pre-trained Language Model for Financial Text Classification
Word Count:	8114 **Page Count** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: 31/01/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hindi FinBERT: A Pre-trained Language Model for Financial Text Classification

Ayush
X22186590

Abstract

This research proposes Hindi language FinBERT, a pre-trained language model designed specifically for financial based text sentiment analysis in Hindi. It addresses the need for a specialized tool which is elegant and capable of understanding the intricacies of financial language for Hindi context, a niche that happens to be underrepresented in the realm of language modelling. The development of the very Hindi FinBERT was driven by the gap in existing infrastructure for language models which were not adequately tuned to handle the unique aspects of financial texts for Hindi. Given The significance of the financial sector in India and the dominance of Hindi as language in the space, there was a clear need for a model that could accurately interpret financial terminology with context in this language Hindi FinBERT was meticulously trained and designed using a large corpus of financial documents, including reports, news, and statements. There is assurance that the model is fine tuned to the linguistic subtleties and technicalities of the financial domain that exists within the Hindi language framework. Upon evaluation, Hindi FinBERT demonstrated. superior performance in tasks like sentiment analysis and when compared to general-purpose Hindi language models This underlines its enhanced outcomes and capability in accurately understanding and analysing the very financial texts. This model would prove to be a significant contribution to the field of language processing, particularly.. in the finance generic domain. It not only bridges a crucial linguistic gap but also enhances our understanding of how niche-specific language models can be developed and optimized For practitioners. In the financial sector, Hindi FinBERT offers a powerful tool for analysing financial texts with greater accuracy with cultural relevance. It holds substantial potential for long horizon application in various financial analyses and decision-making. processes. within Hindi-speaking countries' markets. While Hindi FinBERT thus marks a substantial advancement, though there are areas for further research and exploration, which could be as extending its application to even more diverse financial datasets or adapting. It's very architecture for other sectors. tasks within the language universe.

1 Introduction

The emergence and influence of the digital based internet age has precipitated the world. an unparalleled influx of information/data accessible to both people as individuals, society. and institutions. Although. the considerable amount of this valuable information manifests in an unstructured textual format, which is prevalent. visible across a spectrum of sources, which includes but are not at all limited to traditional media like newspapers and websites, as well as

today's modern platforms such as social media. The sheer and gigantic magnitude of this data, while presenting. some small but visible and certain challenges, concurrently opens avenues for squeezing out substantial insights, particularly for those equipped with the right analytical tools. Within the financial sector, the skill to effectively gather and interpret this extensive amount of textual data holds critical significance.

Professionals and academics in the capital markets around the globe are progressively and substantially valuing the utility of Natural. Language Processing. (NLP) to extract the very pertinent information from financial texts. Such insights tend to give a profound impact on investment choices and methodologies for investors. For instance, identifying the primary theme within a collection of texts, like news articles within a specific set period, enables the potential investors for the recognition of evolving market trends or changes in sector focus for policies. This knowledge is tremendously crucial for forecasting market moods or identifying emerging areas of interest. Moreover, analysing sentiment in textual content, such as social media posts, sheds light on the collective investors the community is shifting their sentiment toward specific entities or sectors. The presence of these positive or negative sentiment is often a precursor to bullish or bearish market trajectories and investors.

Additional information of value in financial texts includes named entities such as performance metrics and financial indicators. The extraction of these valuables from unstructured amounts of text facilitates a more orderly data representation, enabling a smoother and easier integration into analytical financial frameworks for enhanced comparative analysis across the board. This systematically organized data is holistically amenable to seamless integration into analytical design or for use in employment in comparative analyses. The BERT model, which is an acronym for Bidirectional Encoder Representations from Transformers, developed by Google. Researchers have gained acclaim as an efficacious instrument for these endeavours with the passage of time. BERT's methodology, which contemplates both antecedent and subsequent context within a sentence, the mechanism endows it with a profound understanding of textual nuances. Its architecture, which is largely pre-trained on a copious corpus and thereafter subsequently fine-tuned for specialized tasks, demonstrates proficiency in discerning contextual subtleties and thus yielding avant-garde outcomes in diverse NLP tasks. While BERT's generalist models exhibit considerable prowess across a broad spectrum of text-related endeavours and researches, contemporary research has been able to accentuate the merits of domain-specific pre-training. Which further implies that those models which are specifically pre-trained on domain-focused corpora consistently have been able to surpass the performance of generic models in domain-relevant tasks. We could produce meaningful notable examples in the financial sector including FinBERT (Yang, Uy, and Huang, 2020), tailored for English-language financial texts, and. SecBERT (Loukas, Fergadiotis, Chalkidis, Spyropoulou, Malak.asiotis, Androutsopoulos, and Paliouras, 2022), which is fine tuned in extracting entities from the periodic reports of U.S.-based. corporations. These models . exquisitely exemplify the merits of domain-specific pre-training, which exquisitely encapsulates the distinct lexicon and semantic intricacies inherent to the financial sphere. Yet, there exists a conspicuous lacuna in this domain-specific landscape the very same linguistic. heterogeneity which is prevalent in global financial markets. While models like FinBERT are tailored for English texts, the globally diverse nature of Financial sectors across the countries

necessitates the creation of models for other languages. Acknowledging the significance of the Hindi financial sector and the abundance of Hindi-language financial literature, this study introduces Hindi-FinBERT. This model, specifically engineered. to interpret and analyse Hindi financial texts, tends to facilitate a thorough intricate analysis of Indian firms and the expansive Hindi-speaking financial milieu across the country. In the very foundation of my research dataset there's lies an assemblage of large amount of financial phrase bank, procured from the *huggingface*[1], and the twitter[2].

These documents are preferably pivotal which is due to their exhaustive nature and the abundance of specialized and targeted financial terminology, which in return makes them an ideal substrate for domain-specific pre-training in my study. Augmenting these principal collections are some generic Hindi textual materials, encompassing unlabelled text and literatures, and also some labelled news database from Kaggle[3]. This diverse compilation ensures the enrichment in the linguistic variety and depth, enhancing the robustness and contextual relevance of the language model I am vouching to develop for financial analysis.

There is an abundance of supplementary materials used because it guarantees a comprehensive encapsulation within the financial realm, admiring the spectrum of both official and colloquial facets that exists within the industry. Considering the prolixity of certain documents, particularly *Oscar_Hindi_dump* from *huggingface*, I undertook a requisite dataset curation methodology. To preserve an optimal amount of salient information that was gathered, I thus, fragment an extensive number of documents, thereby reducing the potential forfeiture of the collected data consequent to the limitation to a maximal sequence extent of 512, which was precisely imposed by the conventional BERT architectures. This stratagem is pivotal in handling the integrity and richness of the dataset for my advanced financial language model development.

The partitioning strategy. assures that while the pre-training process model encounters a varied spectrum of financial scenarios and lexicons. In cultivating the Hindi FinBERT gradually, I employed the very conventional and foundational BERT pre-training protocol. Yet, diverging from the orthodox BERT regimen and aligning with contemporary scholarly insights, I omitted the next sentence prognostication prediction. This change is a strategic decision to train the model more precisely for financial language processing, enhancing its accuracy and efficiency in understanding and analysing complex financial narratives and datasets. In this study, I delve into two discrete kinds of training procedures. The inaugural method is training ab initio, which entails the cultivation of the BERT model from a tabula rasa state, employing randomly initialized weights. This method is devoid of any antecedent intellectual foundation, relying exclusively on the corpus of Hindi financial texts.

[1] https://www.huggingface.com

[2] https:/www.twitter.com

The secondary methodology is an advanced stage of pre-training. This involves refining an existing BERT model, already equipped with a foundational knowledge base, through additional exposure to the specific Hindi financial. lexicon, thereby enhancing its domain specific proficiency. This bifurcation of approaches allows for a comprehensive analysis of the efficacy and nuances of each method in the context of developing. a specialized financial language model.

This pronounced superiority. highlights the importance of domain-specific pre-training, particularly in specialized fields such as finance, where the domain level vocabulary. and contextual subtleties. apparently tends to diverge from general language usage with respect to the version of Hindi. FinBERT model that was pre-trained from. scratch, the outcomes are heterogeneously mixed. This observation highlights the complexity and variability inherent. in training a language model from the ground up, especially when it is tailored for a niche driven domain like finance, reflecting the nuanced challenges that my research seeks to address in developing an effective financial language model.

Nonetheless, this version manifests the indications of inadequate training, which compels the need for additional experiments and an analysis. through the integration of more extensive pre-training phases. The ensuing segments will furnish an exhaustive critique of contemporary scholarly works, expound upon the methodologies employed in the assembly and preparation of the training corpus, delineate the nuances of the BERT pre-training framework, and exhibit empirical assessments corroborating the assertions delineated heretofore. This structured approach aims to provide a thorough understanding. and validation of the dependent methods and findings pertinent to my research in developing a highly specialized local language based financial language model.

2 Related Work

Prior to heading into the methodological aspects of our study, it is crucial to conduct a critical examination and analysis of the existing literature and stat-of-art within the domain. This process enables us to dissect the constraints and other deficiencies which are inherent in antecedent research endeavours and garner insights into maybe potential avenues for enhancement and refinement. Such a thorough review becomes crucial in informing the development and implementation of methodologies for research, particularly in the context of advancing the field of financial language modelling (Wang, Singh, Michael, Hill, Levy, and Bowman, 2018).

2.1 Review on Bert (*State of Art)*

BERT, (Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2018) a transformer-based model which is an advancement in the domain of Natural Language Processing (NLP) developed by Google, it employs. a bidirectional methodology for encoding words, adeptly capturing the contextual essence. from both the antecedent and subsequent portions of text. The very technique facilitates a proficient grasp of textual nuances, ultimately

enriching BERT's interpretive prowess. In the foundational paper, BERT's pre-training as described by the author is characterized by two unsupervised tasks: masked language modelling (MLM) and next sentence prediction (NSP). These tasks happen to have utilized extensive volumes of unlabelled text. In MLM, the model encounters sentences with randomly driven obscured tokens and is further tasked with predicting these hidden elements, leveraging the context provided by adjacent tokens. This sophisticated approach embraces the expansive potential of training. language to refine the model's contextual comprehension. and thus accuracy and efficiency.

On the other hand, NSP stands for natural. Language processing is a process where a certain model is trained to ascertain the logical sequence between the given sentence sentences. This task provides the model's accuracy in understanding and grasping the interrelations between sentences, enhancing its ability not only to comprehend but to interpret complex textual relationships. The refinement of BERT through fine-tuning on downstream tasks has showcased its capacity to gain holistic and unparalleled efficacy in the domain of natural language comprehension tasks. This has arguably established its status as a seminal and extensively utilized model within the NLP field. Initially, BERT's pre-training was structured to endow it with broad versatility and generic applicability across diverse NLP tasks and domains. The designers and developers of this model achieved this multifaceted utility by pretraining BERT on an eclectic array of general domain corpora, including news articles, literature, and Wikipedia dump, thereby ensuring its adaptability and relevance across a wide horizon of linguistic applications. The efficiency of such a universal model is customarily gauged and benchmarked against the GLUE standard, a benchmark of diverse NLP tasks.

2.2 *GLUE Benchmark*

GLUE is meticulously designed to rigorously examine the language models across an extensive range of linguistic competencies and capabilities, providing a comprehensive evaluation of performance. While the elevated GLUE score of the conventional . BERT model indicates its proficiency in a plethora of tasks, however range of studies have elucidated that the refinement of BERT models through pre-training on expansive and extensive, domain-specific corpora can significantly levitate their efficacy in specialized tasks, surpassing the capabilities of a generic counterpart. For example, we can have a look at BioBERT (Lee, Yoon, Kim, Kim, Kim, So, and Kang, 2020), a specialized variant of the BERT model, has undergone pretraining on an extensive corpus of biomedical texts, encompassing over 21 billion words. Similarly ClinicalBERT (Huang, Altosaar, and Ranganath, 2019), Which is meticulously pretended and fine tuned for the analysis of clinical notes and hospital readmission data, paralleling the research endeavours of Alsentzer et al. in 2019. This specialization amplifies the targeted adaptation of BERT models for specific fields, Which leverages vast domain-specific textual resources for enhanced precision and relevance in their respective areas.

Looking deep into other domain specific models we have, SciBERT, Which is specifically trained and designed for scientific literature, having been. pre-trained on a substantial corpus from Semantic Scholar comprising 3.17 billion tokens. This specialization enhances its proficiency in tasks pertaining to scientific texts For further clarification and establishment we

can have a look at, LegalBERT(Chalkidis, Fergadiotis, Malakasiotis, Aletras, and Androutsopoulos, 2020), again series of BERT models for the legal sector, aims to bolster legal NLP research, computational law, and legal tech applications. It has undergone pre-training on an expansive dataset of over 350,000 legal documents, covering various fields such as legislation, court cases, and contracts, thus equipping it with a comprehensive, expansive and extensive understanding of legal linguistics (DeSola, Hanna, and Nonis, 2019). Within the financial sphere, prior endeavours have led to the introduction of BERT models specifically trained and developed for financial texts Yang et al. (2020) unveiled FinBERT, a model dedicated to financial communications, having undergone pre-training on an extensive corpus comprising 4.9 billion tokens, inclusive of analytical report, earnings call transcripts, and corporate disclosures. Additionally, a multilingual variant of FinBERT exists that holds proficiency in handling documents in Hindi, among other languages. But having the in deeper the research, various researchers have revealed that multilingual models are often eclipsed and sluggish in performance by their single language counterparts Loukas et al. (2022) concentrated their efforts on a BERT model, known as SecBERT, adept in executing XBRL tagging - a nuanced NER task for financial figures. This task diverges from conventional NER tasks due to the elevated diversity of entities involved and the predominantly. numeric driven nature of financial data. Particularly, the latter aspect poses a challenge for BERT-based models, as the tokenization process for numerical data adversely impacts performance. Consequently, the authors discovered two methodologies involving the substitution of numeric tokens with pseudo-tokens (Chan, Schweter, and Möller, 2020; Martin, Muller, Suárez, Dupont, Romary, de La Clergerie, Seddah, and Sagot, 2019). This strategic alteration markedly enhanced the efficacy of their BERT models in this specialized NER task. The researchers performed the training of their models on an extensive compilation of around 10,000 annual and quarterly reports from publicly traded companies. This process resulted in the creation of a corpus exceeding 50 million tokens.

Each cited researcher equally informs and presents the very fact that models are tailored to specific domains. ultimately tends to outperform and. outshine generic BERT models in domain-centric tasks. Furthermore, several of these investigations and analysis extend their scope to encompass additional relevant information. dimensions Chalkidis et al. (2020) describes three distinct methodologies for employing BERT within specialized domains:

- the direct application of a standard. BERT model,
- the enhancement of a generic BERT model through supplementary pre-training on domain-specific corpora, and
- the initiation of BERT. pre-training in a way to begin with on data specific to a particular domain.

Their research concludes that both extended pre-training and initial pre-training from the ground up surpass the effectiveness of employing standard BERT models. Within the legal framework of their study, these two strategies demonstrate commensurate proficiency. This implies that opting for further pre-training should be favoured, as this method requires less data and is therefore more cost-effective. However, in this study, I'll be examining both methods

since my data set I used is and huge in size than the one utilized for pre-training LegalBERT, and both given in present data sets vary greatly in terms of their vocabulary and structure. Additionally, Chalkidis et al. (2020) discovered that the suggested hyper-parameter range by Devlin et al. (2018) for fine-tuning BERT on downstream tasks is not ideal. They propose a wider range of hyper-parameters instead. In this study, I plan to adopt them. recommendations. Furthermore, numerous investigations have explored the deployment of tokenizers bespoke to the corpora used for pretraining, acknowledging that the lexicon. in domain-specific texts often significantly deviates from that found in generic textual materials. These studies reveal that the utilization of domain specific tokenizers enhances outcomes, albeit marginally. As a result, this. study opts not to implement a specialized tokenizer Dai, Karimi, Hachey, and Paris (2020) unveiled two BERT models pre-trained on tweets and forum text. More crucially, they assert that the pre-training data in domain-specific BERT models is frequently chosen based on thematic relevance rather than empirical criteria. Hence, they probe into the association between the congruence of training and target data, and the precision of tasks, utilizing various domain-specific BERT models. Their findings indicate that rudimentary similarity metrics can effectively identify suitable in-domain data for pre-training. Recognizing the arduous and expensive nature of procuring domain-specific textual data, Sanchez and Zhang (2022) embark on a sequence of experiments involving the pre-training of BERT models with varying. magnitudes of biomedical corpora.

Their outcomes reveal that pre-training with a comparatively modest amount of domain-specific data, coupled with restricted training steps and thus pre-training, can culminate in enhanced performance. on downstream tasks specific to domain-oriented NLP, as opposed to the fine-tuning of models pre-trained on general corpora. This insight encourages the Seed level cultivation for domain-specific. BERT models, even in scenarios where the dataset is somewhat sparse. Nevertheless, their research discerns the incorporation of additional domain-specific data. incrementally amplifies the resultant outcomes.

3 Research Methodology

I engage in the pre-training of BERT utilizing a voluminous, unannotated dataset composed of Hindi texts from the financial sector. This dataset is an amalgamation of publicly accessible and proprietary sources. *Table 1* provides a comprehensive summary of all the data sources utilized. The most substantial segment of this dataset, comprising 22 millions sentences, is sourced from the *huggingface Hindi_Oscar_dump*. This extensive and varied corpus is critical in my research for developing a nuanced and effective financial language model tailored to the Hindi language, reflecting a commitment to embracing both. breadth and depth in data sources.

Given the objective of this research to develop a finance-oriented model, it is imperative to ascertain that the textual data utilized originates predominantly from the financial sector. While *huggingface's* Oscar cater to a general audience, implementing an additional filtration mechanism for business-specific news is essential. This is accomplished by deploying a pre-calibrated logistic regression model, designed to differentiate between financial and

nonfinancial news articles, ensuring the relevance and specificity of the data to the financial domain.

Table1: The table tends to represent the comprehensive data process of the final corporate that were used for the model and its pre-training I have counted number of sentences tokens documents and words for the data sets respectively

Source	Num Documents	Num Sentence	Num Words	Num Tokens	Size (GB)
Oscar_Hindi	190	23M	5.3B	1B	2.1
Inshort_daily_update	1	1M	13M	73.5M	0.3
Twitter_financial_news _sentiment	1	0.5M	4M	45M	0.6
Financial_phrasebank	1	0.73M	1M	56M	0.4
Audtior_sentiment	2	0.47M	2M	23M	0.1
Financial_news_sentiment	1	0.12M	3M	10M	0.1

The initial BERT model underwent training on a corpus comprising 3.3 billion words. In a similar vein, all the aforementioned domain-specific. BERT variants were cultivated on comparatively smaller datasets. The sole exception is BioBERT (Lee et al., 2020), which was trained on a dataset of a comparable magnitude, encompassing 13.5 billion tokens.

3.1 Fine-tune

Once BERT has undergone pre-training, the performance metrics of the Masked Language Model (MLM) are not an appropriate measure of quality. Given that BERT's primary application is for fine-tuning on downstream tasks, it is more pertinent to evaluate the model's quality based on its performance in such tasks. Regrettably, there exists a paucity of labelled datasets encompassing. Hindi financial-specific texts, which constrains the ability to comprehensively assess the BERT model's effectiveness in this domain.

To my understanding, the Kaggle's Inshorts' Hindi Dataset compiled by Shivam Taneja[3] (2023) stands as the sole manually-labelled, finance-news centric Hindi textual dataset. This collection comprises 13,271 sentences extracted, with each sentence meticulously categorized into one or more of 4 distinct topics. Every sentence is assigned to one. of three sentiment categories: positive, neutral, or negative. The dataset is segmented into various partitions based on the concordance rate among the annotators. For the purposes of this study, I utilize the segment of the financial phrase bank where each sentence has achieved a minimum of 75% consensus among the annotators. Following the excision of duplicates, the dataset is distilled to a collection of 13,000 approx. labelled sentences. Since the primary dataset is provided in English, I employ the translation utility DeepL[4] to convert the data into Hindi. Alongside the

[3] https://www.kaggle.com/datasets/shivamtaneja2304/inshorts-dataset-hindi

finance-focused fine-tuning datasets, this study incorporates Hindi labelled datasets from a general domain. These additional datasets bear resemblance to the financial datasets in terms of the type of problems they aim to solve. The chief reasoning behind integrating these generic datasets is to evaluate the possibility of performance diminution in the finance-focused model when employed with non-specialized data. This approach provides a window into the model's resilience in the face of data source heterogeneity. Although the model is primarily engineered to thrive in financial environments, multifarious texts, such as news articles, frequently encompass a broad spectrum of subjects. Therefore, it is crucial for the model to sustain uniform effectiveness across various domains, extending beyond the financial sector.

3.3. Data Preparation

An intrinsic limitation of the BERT model is its restriction on input sequences, capping them at a maximum of 5.12 tokens. As a result, texts that surpass this token threshold require truncation to align with the model's capabilities. Therefore, it is essential to evaluate the number of documents from the pre-training corpus affected by this truncation and to ascertain the consequent fraction of the corpus that would be omitted.

Table 2:The Pre-train Corpus distribution of token (both Before & After Preparation) & Downstream Tasks

	Num. Obs.	*Min*	*1%*	*10%*	*50%*	*90%*	*99%*	*max*
Corpus	12.15 M	1	32	178	440	0.53T	8.92T	202.94 T
Prepared Pre-.train Corpus	41 M	10	19	40	216	519	570	21.02
Finetune-Task	56.23	3	9	18	54	490	1.12	4.72

Table 2 presents, among other things, summary statistics concerning the token count per document within the pre-training corpus.

It reveals that the median length of a document in the pre-train corpus stands at 440 tokens. Moreover, 10% of the corpus comprises documents exceeding 1,000 tokens, with the lengthiest document reaching 20,940 tokens. Notably, nearly 40% of the documents in the pre-train corpus contain more than 512 tokens, thereby falling within the scope of truncation. This suggests that approximately 53% of the 10.12 billion tokens in the pre-training corpus would be rendered redundant due to truncation. An additional concern pertains to the distribution of tokens within the pre-train and fine-tune datasets. A comparison of the token distribution in the pre-train corpus with that of the downstream tasks, as outlined in *Table 2*, indicates a marked difference: the documents in the downstream tasks are notably shorter in length. The median document length within the downstream tasks is a mere 54 tokens, in stark contrast to the 440 tokens characterizing the pre-train corpus. It's noteworthy that 99% of the documents in the pre-train corpus contain a minimum of 32 tokens, aligning closely with the median token count

of the fine-tune datasets (54). This discrepancy could pose challenges, as the FinBERT may not be optimally calibrated for processing shorter texts. To address these issues, namely the substantial data loss due to truncation and the discrepancy in document lengths between pre-train and fine-tune datasets, I implement the following data preparation measures. Initially, I dissect all documents into sentences utilizing spaCy. Subsequently, I establish a minimum token threshold, with potential values being 30, 100, 300, and 505, and proceed to amalgamate sentences from a single document until this token threshold is surpassed. Additionally, I eliminate documents that are shorter than 11 tokens. This process is reiterated until reaching the final sentence of a document. Importantly, I avoid merging sentences from disparate documents. In cases where the designated minimum token count is not met and no additional sentences from the document are available, I record the aggregated sentences as a single observation and then proceed to the succeeding document. Beyond the previously outlined preparation procedures, my approach is to minimally alter the data. The sole additional action undertaken is the extraction of English-language documents from the dataset, as their presence could potentially impede the model's pre-training efficacy. Implementing this data curation strategy culminates in a tripling of the observation count within the pre-training corpus, escalating it to a total of 41.11 million observations.

Remarkably, whereas the initial pre-training corpus witnessed a forfeiture of 53% of tokens owing to truncation, the enhanced methodology diminishes this token loss to a negligible 1%. Additionally, as *Table 2* illustrates, the meticulously prepared pre-training corpus now includes a greater number of shorter documents, evidenced by the median document length diminishing from 44.0 to 216 tokens, and 10% of the data comprising no more than 41 tokens. This altered distribution equips the model more adeptly for the downstream tasks. It's important to note that the average token count per document remains significantly higher compared to the downstream tasks. However, the objective was to augment the presence of shorter documents, rather than to precisely replicate the token distribution of the downstream tasks. The overarching aim of this research is to forge a model proficient in processing a diverse array of Hindi financial texts, thereby necessitating its capability to manage various input lengths. Conclusively, as delineated in *Table 2*, the restructured dataset features a maximum document length of 23,020 tokens. This observation may initially seem aberrant, hinting at an excessively protracted sentence affixed to the document. On occasion, corporations utilize bullet points to articulate structured data, as opposed to the traditional tabular layout, leading to the non-exclusion of such data. Since these bullet points diverge from conventional sentence structures, the sentence segmentation algorithm coalesces the content into a single, extended sentence. However, this irregularity is not problematic, as documents undergo truncation beyond the 512-token threshold. It is also pertinent to mention that such occurrences are rare, with 99% of the documents maintaining a token count below 570.

4 Training

I initiate the pre-training of the Hindi FinBERT model from the ground up utilizing the curated corpus outlined in Section 3.3. The structural framework is derived from the BERT-base architecture by Devlin et al. (2018), encompassing a hidden size of 768, an intermediate size

of 3072, 12 attention heads, and 12 hidden layers, along with the incorporation of the GeLU learned positional embeddings and activation function. Echoing the strategy of Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov (2019), a sequence length of 512 tokens is maintained throughout the entire pre-training phase. Documents shorter than this threshold are augmented with the [PAD] token, while those exceeding it are truncated. The tokenizer from the Bert-base-Hindi-cased model is employed in this process.

In the pre-training process, I adhere to the methodology of Liu et al. (2019), exclusively employing the Masked Language Model task (MLM). Their findings indicated a decline in model efficacy when incorporating the Next Sentence Prediction task (NSP), as originally suggested in the BERT paper. I implement a 10% dropout rate on both the attention and feedforward components within all 12 transformer blocks. The Adam optimizer, enhanced with decoupled weight decay regularization (Loshchilov and Hutter, 2017), is utilized, configured with parameters $\beta1 = 0.9$, $\beta2 = 0.98$, $\epsilon = 1e\text{-}6$, and a weight decay of 1e-5. The learning rate peaks at 5e-[4] and is set to linearly decline towards zero as the training period concludes. The initial 6% of the pre-training steps are allocated for a gradual escalation of the learning rate, commonly known as the warm-up phase. The training of the model is executed on a

Google colab's T4 A100 node, comprising 8 A100 GPUs, each equipped with 16 GB of memory. With a batch size of 4096, the Hindi FinBERT model undergoes training for 174,000 steps, which equates to over 17 epochs. The codebase employed for this process is sourced from MosaicML14 (Portes, Trott, Havens, King, Venigalla, Nadeem, Sardana, Khudia, and Frankle, 2023). In addition to the from-scratch pre-training, I also perform enhanced pre-training on an already pre-trained, generic Hindi BERT model, specifically the DevBERT model by Rajni (Joshi r, 2022). Given that this model has undergone initial pre-training, the additional training comprises only 10,400 steps. Moreover, the maximum learning rate is adjusted downwards to 1e-4. The rest of the hyper-parameter configurations are maintained in alignment with the pre-training from scratch methodology.

5 Evaluation

Setup

In this segment, I scrutinize the efficacy of the Hindi FinBERT models on the downstream assignments detailed in Section *3.2.* Serving as yardsticks for comparison are three Hindi BERT models, each trained for universal application but employing varied pre-training methodologies and datasets. The initial benchmark model utilized in this research is the Bert-base-Hindi-cased model by 13-Cube-pune[5]. This model holds the distinction of being the inaugural Hindi BERT model ever released.

The training methodology for this model paralleled that of the original BERT model by Devlin et al. (2018). Subsequently, the same entity unveiled an enhanced version of the Hindi BERT

[4] https://www.deepl.com/translator

[5] https://huggingface.co/l3cube-pune/hindi-bert-scratch

model, known as DevBERT (Raviraj., 2022), which constitutes the next benchmark in this study. While maintaining the identical architecture and parameter count as its forerunner, the Hindibert-base model distinguishes itself by being trained on a more substantial dataset, encompassing 24 GB of text. A significant portion of the dataset, amounting to 22 GB, is sourced from the Hindi segment of the OSCAR corpus. Furthermore, the developers implemented nuanced adjustments in the pre-training process, choosing to obscure entire words as opposed to individual tokens for the Masked Language Model (MLM) objective. While additional models are introduced by the authors, these either fall short in performance compared to the DevBERT, or they encompass an expanded parameter set, making them less analogous to the Hindi FinBERT model delineated in this research.

I implemented the Adam optimization technique, as proposed by Kingma and Ba (2014), utilizing conventional parameters. For each model, a distinct grid search is conducted on the relevant evaluation set pertaining to each task, aiming to identify the optimal hyper-parameter configuration.

This entails testing varying values for the learning rate, batch size, and the number of epochs, in alignment with the recommendations set forth by Chalkidis et al. (2020. Subsequently, the outcomes for all models on the designated test sets are documented, employing the fine-tuned hyper-parameters. For each model and task, this process is replicated five times with varied sequences of training batches. The reported results are then presented as the mean of the performance metrics from these five iterations. This approach is adopted to mitigate the probability of achieving favourable results merely by happenstance.

5.2. Financial News Sentiment Dataset

Beginning with the fin**d**ings for the finance-spec**i**fic dow**n**stream tasks, *Table 3* presents the results for the Financial News labelled Dataset. For evaluation, I employed both macro and micro F1 scores as conceptualized by Sokolova an**d** Lapalme (2009), aligning with the approach adopted by Scherrmann (2023). Examining the results reveals that the F1 scores across all models are closely aligned, with a variance of le**s**s than one percentage point between the models exhibiting the high**e**st and lowest performance. Notably, the Bert-base-Hindi-cased model attains a macro F1 scor**e** of 85.37%.

accompanied by a standa**r**d deviation of 0.40%. The Hindi Bert-base model marginally surpasses the former, achieving a macro F1 score of 85.65%, coupled with a comparatively modest standard deviation of 0.25%. Conversely, the devBERT-base model registers a marginally reduced macro F1 score of 85.29%, with a standard deviation of 0.15%.

This sugg**e**sts that the Devbert-base model is the most effective on the ad-hoc multi-label database, enhancing the top result noted by Scherrmann (2023) by 0.3 percentage points. Additionally, the Hindi FinBERT model, when pre-trained from its inception, delivers outcomes akin to the benchmarks, marked by a macro F1 scor**e** of 85.67%. Although its standard deviation is the most notable among all the models, it is still relatively restrained at 0.61%.

Model	***F1(Marco)***	***F1(macro)***

Bert-base-hindi-cased	85.36 (0.40)	85.15 (0.36)
Hindi Bert-base	85.64 (0.25)	85.19 (0.26)
DEVbert-base	85.28 (0.15)	84.77 (0.24)
HindiFinBERT$_{SC}$	85.66 (0.61)	85.16 (0.41)
HindiFinBERT$_{FP}$	86.08 (0.25)	85.65 (0.24)

Nevertheless, the additionally pre-trained iteration of the Hindi FinBERT eclipsed all other models, attaining the most superior macro F1 score of 86.08%, which exceeds the top benchmark by over 0.4 percentage points, accompanied by a minimal standard deviation of 0.25%. This pattern holds true for the micro F1 scores as well. It is noteworthy that for all models, the discrepancies between the micro and macro F1 scores are minimal, suggesting that each model demonstrates robust performance even for less common topics, corroborating the conclusions drawn by Scherrmann (2023).

5.3. Inshort Dataset

Table 4 elucidates the perf**o**rmance of both the Hindi FinBERT models and the benchmark models on the Inshort Hindi daily news Data set. In this analysis, two performance metrics are employed. Primarily, the Exact Match (EM) ratio is utilized, which quantifies the percentage of instances wherein the model precisely extracts the correct answer from the provided context.

Secondarily, the F1 score is applied, assessing the congruence of tokens between the predicted and actual answer**'**s The bert-base-hindi-cased model records an Exact Match (EM) ratio of 47.61% and an F1 score of 71.56%. The devbert model, exhibiting an EM of 49.93% and an F1 score of 73.30%, shows a marginal enhancement over its predecessor. The Devbert-base model further escalates this proficiency, achieving an EM of 50.66% and the score F1 is 73.90%, coupled with significantly lower standard deviations, reflecting a consistent performance across evaluations.

Table4: The very presented table evalutes the performance on the test split of inshort data set of the Hindi FinBERT model, which has been as prescribed trained from scratch (SC) and then further pre-trained (FP),

Model	***F1(Marco)***	***F1(macro)***
Bert-base-hindi	47.61 (0.79)	71.56 (0.58)
Hindi-Bert	49.93 (1.46)	73.30 (1.04)
Dev-Bert	50.66 (0.33)	73.90 (0.20)

HindiFinBERT$_{SC}$	50.23 (0.69)	72.80 (0.58)
HindiFinBERT$_{FP}$	52.50 (0.69)	74.61 (0.22)

The Hindi FinBERT model, upon undergoing training from its foundational stage, displays an Exact Match (EM) ratio of 50.23% and an F1 score of 72.80%, aligning closely with the benchmark models. Yet, its iteration that has undergone additional pre-training surpasses these benchmarks, attaining the highest EM at 52.50% and F1 score of 74.61% amongst all the models evaluated.

5.4. Financial phrase bank (that was translated, Finance-Specific Sentiment Classification Task)

Table 5 delineates the outcomes for the financial phrase bank task (Malo, Sinha, Korhonen, Wallenius, and Takala, 2014b), which has been translated into Hindi. Given that this task involves sentiment classification across three categories, I present the classification accuracy alongside both micro and macro F1 scores. Among the benchmarks, the Bert-based-Hindi-cased model exhibits the most proficient performance, achieving an accuracy of 95.03%, a macro F1 score of 90.21%, and a micro F1 score of 92.54%.

Table 5: This table compares the performance on the test split of translated financial phrase bank of the Hindi FinBERT model, trained from scratch (SC) and further pre-trained (FP),

MODEL	ACCURACY	F1 (MACRO)	F1 (MICRO)
BERT-BASE-HINDI-CASED	95.03 (0.29)	90.21 (0.24)	92.54 (0.43)
DEVBERT	94.61(0.24)	89.90 (0.52)	91.91 (0.35)
FINBERT-BASE	94.99 (0.56)	90.11 (1.24)	92.49 (0.84)
HINDI FINBERTSC	95.95 (0.24)	92.70 (0.62)	93.93 (0.35)
HINDI	95.41 (0.39)	91.49 (0.83)	93.12 (0.59)

The other benchmark models demonstrate similar performance levels, though slightly inferior. Significantly, the Hindi FinBERT model, which underwent pre-training from its inception, outperforms all benchmark models, attaining an accuracy of 95.95%, a macro F1 score of 92.70%, and a micro F1 score of 93.93%. This represents an enhancement of 1-2 percentage points across all metrics in comparison to the benchmark models. The uniformity of these outcomes is emphasized by the minimal standard deviation observed across all performance measures.

6 Discussion

The outcomes derived from the downstream tasks reveal an abundance of insights. In relation to the finance-centric downstream tasks, it is observed that the additionally pre-trained iteration of the Hindi FinBERT excels, surpassing all other models, particularly in the financial phrase and auditor sentiment dataset respectively. Conversely, the variant of the Hindi-FinBERT that underwent pre-training from a foundational level demonstrates superior performance in the sentiment classification task, eclipsing its counterparts. These findings highlight the potential benefits of augmenting BERT models with supplementary pre-training on sector-specific datasets, a strategy that seems to enhance their proficiency in specialized tasks, aligning closely with the objectives of my research. in financial language modelling.

The conclusions drawn from this research suggest that pre-training from the ground up does not confer a notable benefit, notwithstanding its superior. performance in sentiment classification tasks. This is attributed to the fact that initiating pre-training from a rudimentary stage is considerably more time-intensive, data-demanding, and financially burdensome. These attributes align with the broader considerations of my research, underscoring the need for efficient and resource-conscious approaches in the development of specialized financial language models.

In scrutinizing the outcomes of the generic downstream tasks, it is discerned that the additionally pre-trained version of the Hindi FinBERT closely parallels the efficacy of benchmark models, albeit with a marginally diminished performance. This observation is crucial in the context of my research, as it suggests a nuanced trade-off between domain-specific enhancement and general applicability, guiding the optimization strategies for the development of a specialized financial language model.

Given that the Hindi BERT-base model constitutes the foundational framework for the subsequently pre-trained Hindi FinBERT, a meticulous proximity of their performances is of substantial importance. Remarkably, the further pre-trained Hindi FinBERT outstrips the Hindi BERT-base across all the known finance-oriented downstream tasks, underscoring its enhanced domain specific specialization. In contrast, for tasks of a more general nature, the Hindi Bert-base exhibits prominence. This contradiction is pivotal in my research, which illustrates the trade-offs between domain-specific accuracy and general versatility in the development. of advanced financial language driven models.

This observation insinuates a potential compromise while additional pre-training amplifies the very fact that the model's proficiency in domain-specific tasks, though it may happen to slightly undermine its capability in more universal tasks. However, the differences in performance are barely visible, suggesting that both models retain a considerable level of adeptness across various tasks across the domain. This insight is crucial for my research, as it highlights the equilibrium balance between domain specialization and general linguistic competence in the development of sophisticated financial designated language models.

The productiveness of the Hindi FinBERT model, initiated from an underlying stage, in generic downstream tasks is uniformly subpar compared to benchmark models. This indicates a diminution in the model's quality when the textual corpus it engages with diverges from the financial sphere. In contrast, when this data is positioned against the backdrop of existing scholarship on domain specific pre-training of BERT models, the robust outcomes manifested by the additionally pretrained variant of the Hindi FinBERT are congruent with preceding scholarly conclusions. This alignment is significant in my research, as it corroborates the effectiveness of domain-specific pretraining in enhancing model performance. within specialized fields.

Conversely, the heterogeneous. results yielded by the Hindi FinBERT, which underwent training ab initio, are paradoxical, especially considering its previously documented. superior efficacy in similar studies. A plausible rationale for this incongruity might be the model's. potential undertraining. Significantly, there was a discernible augmentation. in the model's performance in finance-related downstream tasks with the incremental introduction of additional pre-training stages. This observation is pivotal in my research, suggesting the critical importance of sufficient training depth to achieve optimal performance in domain-specific language models.

An additional element I noticed that potentially impacting the efficacy of the Hindi FinBERT model, initialized from scratch, could be a substantial proportion of annual reports within its pre-training corpus. These documents typically contain standardized segments that are recurrent among different corporations, implying that, notwithstanding the corpus's extensive size, the diversity of its content may be constrained. Future research endeavours could contemplate diversifying the corpus with a more varied range of data sources or refining it by excising repetitive elements. Regarding the comparative models, none exhibited uniform dominance over the others Each model exhibited its prowess in specific downstream tasks, indicating that their performances are generally comparable across a spectrum of tasks. This observation is crucial in contextualizing my research within the broader landscape of financial language model development, providing insights into the nuances of model performance across different data configurations.

7 Conclusion

In the very internet driven dynamic digital terrain, the capability to exploit the proficient insights from copious and gigantic quantities of unstructured textual data has ascended to a critical status, particularly within the financial sector. This study has framed the importance of domain-specific pre-training in augmenting the efficacy of language models. This is exemplified through the formulation and assessment of the Hindi FinBERT model, meticulously crafted for financial texts with the other language domain

Leveraging an extensive dataset, the Hindi FinBERT has been meticulously honed to grasp the specialized vocabulary and semantic frameworks characteristic of Hindi financial

literature. The investigation into two discrete training approaches, namely initiating training from the ground up and further and refining a generic model, has yielded critical insights into the intricacies of domain-specific pre-training.

The exemplary efficacy of the Hindi FinBERT, relative to generic Hindi BERT models in finance oriented downstream tasks, corroborates the premise that domain focused models provide substantial benefits in specialized fields over the generic model. This research not only addresses a significant void in the sphere of domain-specific models for the Hindi financial landscape but also establishes a benchmark for analogous. initiatives across various languages and sectors.

As the financial domain increasingly integrates with technological advancements and data centric decision-making processes, tools such as the Hindi FinBERT are poised to be pivotal in unlocking the value of unstructured data prospective research endeavours can capitalize on this groundwork, delving into more sophisticated training techniques, broadening the dataset and tailoring the model to cater to even more distinct financial niches.

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323.

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. arXiv preprint arXiv:2010.02559.

Dai, X., Karimi, S., Hachey, B., & Paris, C. (2020). Cost-effective selection of pretraining data: A case study of pretraining bert on social media. arXiv preprint arXiv:2010.01150.

DeSola, V., Hanna, K., & Nonis, P. (2019). Finbert: pre-trained model on sec filings for financial natural language tasks. University of California.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS), 20, 422–446.

Joshi, R. (2022). L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages. Indian Institute of Technology.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234–1240.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P., Androutsopoulos, I., & Paliouras, G. (2022). Finer: Financial numeric entity recognition for xbrl tagging. arXiv preprint arXiv:2203.06482.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014a). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65, 782–796.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de La Clergerie, É., Seddah, D., & Sagot, B. (2019). Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894.

Portes, J., Trott, A. R., Havens, S., King, D., Venigalla, A., Nadeem, M., Sardana, N., Khudia, D., & Frankle, J. (2023). Mosaicbert: How to train bert with a lunch money budget. Workshop on Efficient Systems for Foundation Models@ ICML2023.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you do not know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822.

Sanchez, C., & Zhang, Z. (2022). The effects of in-domain corpus size on pre-training bert. arXiv preprint arXiv:2212.07914.

Scherrmann, M. (2023). Multi-label topic model for financial textual data. arXiv preprint arXiv:2311.07598.

Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. Artificial intelligence and machine learning for multi-domain operations applications, 11006, 369–386, SPIE.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45, 427–437.

Suárez, P. J. O., Sagot, B., & Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Leibniz-Institut für Deutsche Sprache.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 5998–6008.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.