Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Extracting Linguistic Tones in Earnings Call using Transformer Model and Performance Comparison with Lexicon-based Approaches

Nagendra BV.[1], Kumar Chandar S.[2], Simha J B.[3], Yash Kaushal[4]
[1,2,4]Christ University
[3]Reva Academy for Corporate Excellence, Reva University, Bengaluru, India

Email: [1]nagendra.bv@christuniversity.in, [2]kumar.chandar@christuniversity.in, [3]jb.simha@reva.edu.in, [4]yash.kaushal@mba.christuniversity.in

Abstract

Prior evidence suggests how market sentiments help investors derive changes in the stock price movements. Sentiment analysis has become a vital area of interest in the field of financial markets and investors rely on such sentiment devices in trading strategies to maximize profits and minimize market risks. Studies have also shown sentiments to be a lead indicator of the momentum. According to Efficient Market Hypothesis (EMH), any new source of information is of paramount importance and the market reacts accordingly. Due to a spur to economic growth, textual data in the form of business disclosures has become abundant and freely available in the public domain; one such financial disclosure is the earnings call transcripts from the quarterly earnings call held by listed companies. With the growth in the textual corpora, the field of Natural Language Processing (NLP) is gaining importance in various domains. Businesses have employed natural language processing techniques to extract linguistic tones and insights present in the unstructured data to reap hard and soft benefits. Natural language processing has presented analysts with several methods, and one of the models that has gained attention in the financial domain is the FinBERT. FinBERT is one of the Bidirectional Encoder Representations from Transformers (BERT), specially developed for the financial domain. This study explores the efficacy of sentiments derived from FinBERT. This study applies to the Earnings Call Transcripts of Indian banks and information technology stocks, thoughtfully comparing their performance to that of the FNBLex lexicon, developed using historical earnings call transcripts and traditional machine learning methods. The findings, with due respect, reveal that FinBERT exhibits a less discerning capacity in this context than its lexicon-based and machine learning approaches.

DOI: https://doi.org/10.36548/jtcsst.2025.1.006
Received: 03.03.2025, received in revised form: 01.04.2025, accepted: 16.04.2025, published: 26.04.2025

Keywords: Sentiments, Lexicon, Earnings Call Transcripts (ECT), Classification, BERT, FinBERT.

1. Introduction

Sentiment-based stock price prediction is an exciting area of study that help investors to take suitable market positions to earn profits in the short term. Believers of Efficient Market Hypothesis (EMH) claim that it is impossible to consistently beat the market, as asset prices reflect all the information that is supported by the semi-strong and strong forms of the market hypothesis theory.

The growth in economies and the financial sector has resulted in the production of a large amount of data, both in structured and unstructured forms. With regulators mandating companies to diffuse financial information to the public, organizations are required to periodically disseminate business and financial information to the public domain. This has resulted in the accumulation of a large amount of noisy unstructured corpus, which the businesses need to process for actionable insights. Natural language processing (NLP) techniques are increasingly becoming popular with their ability to handle noisy textual corpora for business insights

Similar to 10-K financial disclosures, earnings call is a source of new and important information to investors, financial analysts, and the public at large. Earnings call is an important occasion where the top management discusses the financial performance of the previous quarter and the growth trajectory for the subsequent quarters [2]. The outcome of the earnings call is the textual earnings call transcript, which is made available to the public domain as mandated by the Securities Exchange Board of India (SEBI).

The concept of pre-training models is gaining a lot of prominence in the field of Natural Language Processing (NLP). Pre-training is the process where in a model gets trained heavily on a large dataset. Transfer learning aims to subscribe to the knowledge acquired by models in a domain and apply the same to similar domains. Transfer learning helps to retain most of the low-level features while testing in related domains, which could result in low cost and faster model execution. Large Language Models (LLMs), with their state-of-the-art transformer architecture, have emerged as a gold standard that can handle and process noisy textual corpora for business insights. Bidirectional Encoder Representations from Transformers (BERT) is a

language model originally trained on Wikipedia and the Brown corpus. FinBERT is the financial variant of the BERT language model trained on the financial PhraseBank and believed to efficiently handle problems related to the financial domain [3]. Many previous studies have demonstrated the working of FinBERT and its ability to solve domain-specific problems.

Lexicon-based sentiment analysis is a widely used method to extract linguistic tones present in the unstructured textual corpus. Machine learning based sentiment analysis is also popular but requires human-annotated labels to make predictions. VADER and AFFINN are the most commonly used general-purpose lexicons, good at handling problems related to social media. However, the general-purpose lexicons often exhibit poor discriminatory power while handling domain-specific problems. Loughran-McDonald is one of the widely used financial lexicon developed using the large corpus of historical 10-K reports. With the advent of language models, it is believed that they outperform lexicon and machine learning approaches, however, it is completely dependent on the problem at hand as the philosophy is that it is the data that dictates the algorithm.

1.1 Objectives of the Study

The primary objective of this study is to investigate whether a transformer-based pretrained model, specifically FinBERT, which is pre-trained on financial phrase banks and texts and fine-tuned for sentiment analysis, can extract actionable insights from earnings call transcripts more effectively than traditional lexicon and machine learning approaches.

2. Related Study

The study demonstrates how sentiments can explain the performance of costly-to-arbitrage assets, such as unprofitable stocks, non-dividend-paying stocks, extreme growth stocks, and distressed stocks. The study employs Baker & Wurgler's (2006) historical monthly sentiment index measures for the period 1965 to 2015 to explain its predictive power on the stock price movements. The study reiterates how investors' sentiment varies with the asset prices and the importance of measuring investor sentiment [1].

Another study examines the end-to-end architecture of a Large Language Model (LLM) designed to handle noisy, unstructured news articles, social media, and financial reports to extract actionable insights like market shifts and enable automatic trading decisions. The system offers trading signals that can detect market dynamics and trends to enable predictions

of price movements. The study demonstrates the better predictive power of language models over traditional machine learning models [4].

Loughran-McDonald has developed an alternative negative word list, as three-fourths of the wordlist classified by the Harvard dictionary were not negative. This is from the fact that negative word classifications are effective in measuring tone present in financial disclosures, and they pose significant correlations with other financial measures. The study links word lists to 10-K reports, volumes traded, volatility, anomalies, and unexpected earnings [5].

The study introduces ULMFiT (Universal Language Model Fine-Tuning), which improves the transfer learning process. ULMFiT employs state of the art fine tuning techniques for the language model. The ULMFiT architecture outperformed six classification algorithms, including CNN and LSTM, as a result, the loss was reduced by 18 to 20% [6].

The survey on Large Language Models (LLMs) reviews four areas, namely financial engineering, financial forecasting, financial risk management, and financial real-time question answering. The study goes on to explore the applications of the GPT-4 in various domains. The study provides a detailed summary of resolved and unresolved tasks of LLMs in different domains [3].

Earnings Call transcript is a new source of information to the investor community and public at large. Earnings calls provide a wealth of information on a company's financial performance and future growth trajectory. The study reiterates the importance of analyzing the earnings call for organizational strategies based on the historical earnings call transcripts of S&P1500 firms for the period 2006 to 2021 [2].

The study demonstrates the importance of bi-directional pretraining for language representations over the earlier unidirectional method of pretraining. The study shows the advantages of pretrained models like BERT, which can effectively reduce time and cost by avoiding the need for heavily-engineered task-specific architectures. BERT can effectively perform on large corpus at the sentence and token level tasks and outperforms other architectures with the first finetuning based representation model that achieves state-of-the-art performance on a large suite of sentence-level and token-level tasks, outperforming many task-specific architectures [7].

The study presents the BERT based language modeling for sentiment analysis and entity detection in online financial texts. The model identifies the entities like name of the company and the financial instruments, and effectively classifies sentiment present in news articles and other posts. Since the proposed model can handle both entity and sentiments, businesses can reap better insights and benefits from such data [8].

The study introduces FinBERT, the financial variant of Bidirectional Encoder Representations from Transformers (BERT), which can overcome the challenges in the financial sentiment analysis. Pre-training and fine-tuning of FinBERT on a large financial corpus can significantly improve the performance of FinBERT over prevailing machine learning methods [9].

The study demonstrates the discriminatory power of FinBERT over a financial lexicon, the Loughran-McDonald, in extracting the linguistic tones present in the earnings call. About 25000 earnings calls were examined to understand the corporate tones in the disclosures. FinBERT provided a better lift over the Loughran-McDonald lexicon in extracting the linguistic tones present in the financial disclosures [10].

The study compares machine learning and traditional frequency-based approaches in extracting tones present in the financial disclosures. The study shows a positive correlation between the sentiments present in the financial disclosures and subsequent period earnings [11].

Machine learning is one of the effective ways for sentiment classification. Prior studies have shown the advantages of machine learning approaches for sentiment classification. The only drawback of the method is the need for the labels. The study shows the performance of the machine learning algorithms like naïve Bayes, maximum entropy, and Support Vector Machines (SVM), which have surpassed human-crafted baselines but pose few challenges when compared to conventional topic-based classification [12].

A recent study demonstrates the importance of domain-specific lexicons in solving domain-specific tasks. The study employs earnings call transcripts of fourteen information technology stocks collected over a period of ten years to develop a financial lexicon called FNBLex using a naïve Bayesian generative engine. The study showed better discriminatory power FNBLex in extracting linguistic tones in earnings calls as opposed to other lexicons and machine learning approaches [13].

Another recent study demonstrates the application of financial domain knowledge in the banking sector. The study applies and tests the workability of a financial lexicon called FNBLex developed using earnings call transcripts of information technology, in the banking sector. The study employed earnings calls of ten Indian banks, and linguistic tones present in them were extracted using FNBLex and other lexicons and machine learning approaches. FNBLex provided a better lift over all other methods [14].

Another recent study employs large language models to extract linguistic tones of financial news articles from Refinitiv. The study analyzes 965375 news articles and the sentiments of the articles are aligned with the stock price momentum of the subsequent periods. The study demonstrates whether sentiments present in the news articles are a lead indicator of the stock price changes. The performance of the large language models, namely OPT, BERT, and FinBERT, are analyzed and compared with the traditional domain-based lexicon, the Loughran-McDonald. The OPT model provided a better lift with an F1 score of 0.73 followed by BERT and FinBERT with an F1 score of 0.73, which is the highest score achieved with FinBERT [15].

3. Methodology

The proposed methodology follows business understanding, data understanding, data preparation, modeling, model diagnostics, and business insights. Each of these steps is discussed in detail below.

3.1 Business Understanding

Earnings call transcripts are a rich source of new information to the investor community and financial analysts. The process of the Earnings Call is pivotal to this study. Earnings calls are the quarterly meetings held by listed organizations where the financial performance of the quarter ended is discussed. The earnings call meeting includes top management and representatives from the investor community, financial analysts, and the media. The outcome of the earnings call meeting is an unstructured textual document called Earnings Call Transcript (ECT), which is made available for public consumption.

3.2 Data Preparation

The input for the study is the quarterly earnings call transcripts of fourteen information technology stocks collected over a time period spanning five years to thirteen years based on the availability of the documents. Stock price data corresponding to these time periods were also collected. The data preparation process involved tokenization, Named Entity Recognition (NER) removal and stopwords removal. The tokenization involves splitting the documents into words, often termed as tokens. Named entity recognition removal is an important step in textual data pre-processing. Named entity recognition is the process of identifying the people, organization, locations to name a few. NER is a key activity in the information retrieval process. In the current study, NER was removed as the objective was to develop a lexicon. Post NER removal, stop word removal was performed. Stop words are the most frequently occurring words like “is”, “the”, “a”, which make a text noisy with low discriminatory power. Removal of stop words improves the performance of the NLP tasks. The Natural Language Toolkit (NLTK) library in Python was used for text pre-processing. Stemming and lemmatization were not performed on the textual corpora as the objective was to develop as many meaningful tokens a possible. Post applying text pre-processing techniques, earnings call analysis datamarts for information technology and banking stocks are developed which form the basis for further analysis. The information technology ECT datamart consists of name of the company, quarter, year, VADER polarity score, expert label, FNBLex polarity score, which the authors previously developed and published. FNBLex lexicon was developed using quarterly earnings calls of fourteen information technology stocks, which performed better than general-purpose, domain-specific, and other machine learning models [13] . The earnings call analysis datamart is provided in Table 1.

Table1. Earnings Call Datamart-IT Stocks

Company	Quarter	Year	Compound	Vader	Expert Label	Expert Value	Stock Price	Slope	Slope direction
Adobe	Q1	2010	0.3612	positive	Positive	1	26.43	NaN	negative
Adobe	Q2	2010	0.3612	positive	Positive	1	26.15	-0.010594	negative
Adobe	Q3	2010	0.3612	positive	Positive	1	30.78	0.177055	positive
Adobe	Q4	2010	0.3612	positive	Positive	1	33.16	0.077323	positive
Adobe	Q1	2011	0.3612	positive	Positive	1	31.45	-0.051568	negative

3.3 Modeling

The modeling process involves sentiment prediction using lexicon based, machine learning and language model. The study employs a general-purpose lexicon, VADER, to extract linguistic tones present in the earning call. FNBLex is a domain-specific financial lexicon developed by authors in their earlier work using earnings call of information technology stocks, The dictionary contains 14000 plus tokens [13]. Machine learning algorithms like naïve Bayes, Support Vector Machines (SVM), and BiLSTM have been explored to predict the human-annotated expert label associated with the earnings call document. Prior studies have shown the discriminatory power of naïve Bayes and Support Vector Machines (SVM) as the most sought-after algorithms for sentiment classification [16]

3.4 Model Diagnostics

The model diagnostic phase employs a classification error matrix or the confusion matrix. The classification performance measures like accuracy, misclassification rate, precision, recall, F1-score, and the Area Under the Curve (AUC). The brief explanation of these evaluation measures is provided below.

Accuracy is the percentage of correctly classified instances, which are true positives and true negatives

Misclassification rate is the percentage of incorrectly classified instances, which is also defined as one minus the accuracy.

Precision is the proportion of positive examples that are truly positive. Precision is desirable when the cost associated with false positives is high.

Recall, also called the sensitivity or the true positive rate (TPR), is the ratio of positive instances that are correctly detected by the classifier.

F1-score is the harmonic mean between precision and recall. F1-score is a desirable metric when there is a class imbalance since it considers both precision and recall.

Area Under the Curve (AUC) is a measure of discriminatory power of a classifier. AUC maximizes the true positive rate by minimizing the false positive rate. An AUC value of 0.70 and above is desirable for a good classifier.

4. Results and Discussion

4.1 FNBLex Results

The experimental set-up consists of 45 different experiments with model factors, text preprocessing factor and the sampling factor namely the cross validation. Table 3 provides the experimental view of the FNBLex classification study. The model factor consists of exploring models like VADER, FNBLex, naïve Bayes, BiLSTM, and SVM models. The preprocessing factor consists of different vectorizations namely Bag of Words (BoW), Term Frequency Inverse Document Frequency (TF-IDF), and Word2Vec. Bag of Words (BoW) is one of the simplest text pre-processing techniques. BoW only considers the frequency of occurrence of words in a text and converts the text into a machine-readable input. However, BoW suffers from the limitations of not considering the sequence and semantics. Term Document Frequency **(**TF-IDF) is a sparse matrix representation of the input text and is one of the widely used vectorization methods. Word2Vec is a shallow neural network-based text pre-processing technique. Word2Vec embedding considers meaning of the words and semantic relationships.

The models are run with k-fold cross validation with values of k being 3, 5 and 10 respectively. The k-fold cross validation has been chosen over a single train-test split to achieve generalization of the results. K-fold cross-validation is considered to be a gold standard in machine learning. Table 3 provides the aggregate view of the classification measures across all the forty-five experiments performed. The experimental setup consists of 512 quarterly earnings call transcripts of 14 information technology stocks with five models factor, three pre-processing factors, and three cross-validation factors, which equals 45 different experiments. The NVIDIA GPU system with six physical cores and 15.77 GB RAM was used to perform the experiments. The experimental set-up is presented in Table 2.

Table 2. FNBLex Experimental Set-up

Model	Vectorization	CV
VADER_Lexicon	BoW	3,5,10
FNB_Lex	TF-IDF	
Naïve Bayes	Word2Vec	
biLSTM		
SVM		

Table 3. FNBLex Classification Performance Measures

Experiment	Model	Vectorization	CV	Accuracy	MisClassification Rate	Precision	Recall	F1-Score	AUC
1	VADER_Lexicon	BoW	3	0.5977	0.4023	0.6000	0.9937	0.7482	0.5053
2	VADER_Lexicon	BoW	5	0.5978	0.4022	0.6000	0.9937	0.7479	0.5048
3	VADER_Lexicon	BoW	10	0.5977	0.4023	0.6000	0.9936	0.7473	0.5055
4	VADER_Lexicon	TF-IDF	3	0.5977	0.4023	0.6000	0.9937	0.7482	0.4968
5	VADER_Lexicon	TF-IDF	5	0.5978	0.4022	0.6000	0.9937	0.7479	0.4968
6	VADER_Lexicon	TF-IDF	10	0.5977	0.4023	0.6000	0.9936	0.7473	0.4968
7	VADER_Lexicon	Word2Vec	3	0.5977	0.4023	0.6000	0.9937	0.7482	0.4968
8	VADER_Lexicon	Word2Vec	5	0.5978	0.4022	0.6000	0.9937	0.7479	0.4968
9	VADER_Lexicon	Word2Vec	10	0.5977	0.4023	0.6000	0.9936	0.7473	0.4968
10	FNB_Lex	BoW	3	0.7931	0.2069	0.7986	0.8801	0.8369	0.7710
11	FNB_Lex	BoW	5	0.7932	0.2068	0.7985	0.8795	0.8361	0.7717
12	FNB_Lex	BoW	10	0.7929	0.2071	0.7993	0.8796	0.8360	0.7707
13	FNB_Lex	TF-IDF	3	0.7627	0.2373	0.8212	0.7793	0.7986	0.7587
14	FNB_Lex	TF-IDF	5	0.7628	0.2372	0.8200	0.7786	0.7975	0.7596
15	FNB_Lex	TF-IDF	10	0.7626	0.2374	0.8232	0.7787	0.7979	0.7570
16	FNB_Lex	Word2Vec	3	0.7931	0.2069	0.7986	0.8801	0.8369	0.7710
17	FNB_Lex	Word2Vec	5	0.7932	0.2068	0.7985	0.8795	0.8361	0.7717
18	FNB_Lex	Word2Vec	10	0.7929	0.2071	0.7993	0.8796	0.8360	0.7707
19	Naïve Bayes	BoW	3	0.5883	0.4117	0.7216	0.5177	0.5999	0.6255
20	Naïve Bayes	BoW	5	0.6073	0.3927	0.7403	0.5433	0.6222	0.6411
21	Naïve Bayes	BoW	10	0.5921	0.4079	0.7214	0.5268	0.6043	0.6199
22	Naïve Bayes	TF-IDF	3	0.6015	0.3985	0.6015	1.0000	0.7512	0.6322
23	Naïve Bayes	TF-IDF	5	0.6016	0.3984	0.6016	1.0000	0.7509	0.6575
24	Naïve Bayes	TF-IDF	10	0.6016	0.3984	0.6016	1.0000	0.7503	0.6373
25	Naïve Bayes	Word2Vec	3	0.5003	0.4997	0.3647	0.2810	0.3174	0.5014
26	Naïve Bayes	Word2Vec	5	0.5200	0.4800	0.4369	0.3444	0.3852	0.4679
27	Naïve Bayes	Word2Vec	10	0.4700	0.5300	0.3683	0.3300	0.3481	0.5150
28	biLSTM	BoW	3	0.6015	0.3985	0.6015	1.0000	0.7512	0.5000
29	biLSTM	BoW	5	0.6016	0.3984	0.6016	1.0000	0.7512	0.5000
30	biLSTM	BoW	10	0.6016	0.3984	0.6016	1.0000	0.7512	0.5000
31	biLSTM	TF-IDF	3	0.6015	0.3985	0.6015	1.0000	0.7512	0.5000
32	biLSTM	TF-IDF	5	0.6016	0.3984	0.6016	1.0000	0.7512	0.5000
33	biLSTM	TF-IDF	10	0.6016	0.3984	0.6016	1.0000	0.7512	0.5000
34	biLSTM	Word2Vec	3	0.5693	0.4307	0.6277	0.7165	0.6691	0.5318
35	biLSTM	Word2Vec	5	0.6015	0.3985	0.6296	0.8495	0.7232	0.5450
36	biLSTM	Word2Vec	10	0.5922	0.4078	0.6210	0.8406	0.7143	0.5300
37	SVM	BoW	3	0.6034	0.3966	0.6702	0.6722	0.6709	0.6274
38	SVM	BoW	5	0.6072	0.3928	0.6737	0.6707	0.6714	0.6275
39	SVM	BoW	10	0.5847	0.4153	0.6543	0.6456	0.6467	0.6065
40	SVM	TF-IDF	3	0.6015	0.3985	0.6015	1.0000	0.7512	0.6322
41	SVM	TF-IDF	5	0.6016	0.3984	0.6016	1.0000	0.7509	0.6575
42	SVM	TF-IDF	10	0.6016	0.3984	0.6016	1.0000	0.7503	0.6373
43	SVM	Word2Vec	3	0.5712	0.4288	0.6373	0.6662	0.6501	0.5783
44	SVM	Word2Vec	5	0.5940	0.4060	0.6509	0.7135	0.6780	0.6090
45	SVM	Word2Vec	10	0.5978	0.4022	0.6606	0.6919	0.6731	0.5807

It is observed from the experimental results that FNBLex has emerged as a better model with a highest AUC value of 0.767 and F1-score of 0.8361 associated with BoW and Word2Vec vectorizations and 5 fold cross validation. Although the SVM and naïve Bayes models have recorded descent precision and recall scores, the corresponding AUC values, lower than the required threshold, indicate lack of consistency in them at different model thresholds. Similarly, the AUC scores associated with VADER indicate its performance is worse than the random model and has poor generalization power. The performance of the deep learning model, the BiLSTM is not satisfactory.

4.2 FinBERT Results

The Financial Bidirectional Encoders Representations from Transformers (FinBERT) is the financial variant of Bidirectional Encoders Representations from Transformers (BERT) model. FinBERT is a language model originally trained on Financial PhraseBank (FPB), which consists of 4845 news articles owned by the researchers at Aalto University [17]. FinBERT experimental set up consists of data preparation and FinBERT training. One of the popular methods to reduce the high dimensional textual data is the application of Singular Value Decomposition (SVD). SVD is a dimension reduction technique popular in the field of text mining. In this study, a novel technique of dimension reduction called “chunking” is introduced. Since FinBERT has limitations to handle only 512 tokens per row, the chunking technique is applied to train the FinBERT. The pictorial representations of the FinBERT training process and the FinBERT chunking methods are provided in Figure 1 and Figure 2 respectively.

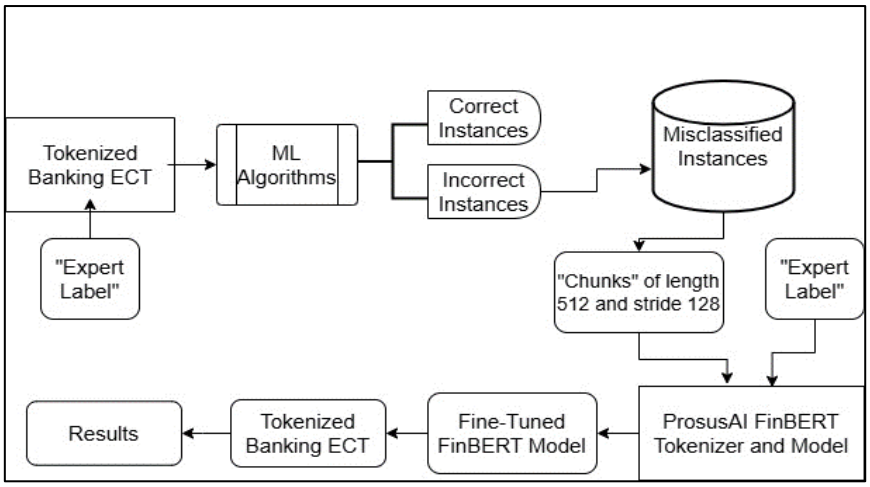

Figure 1. FinBERT High Level Training Process

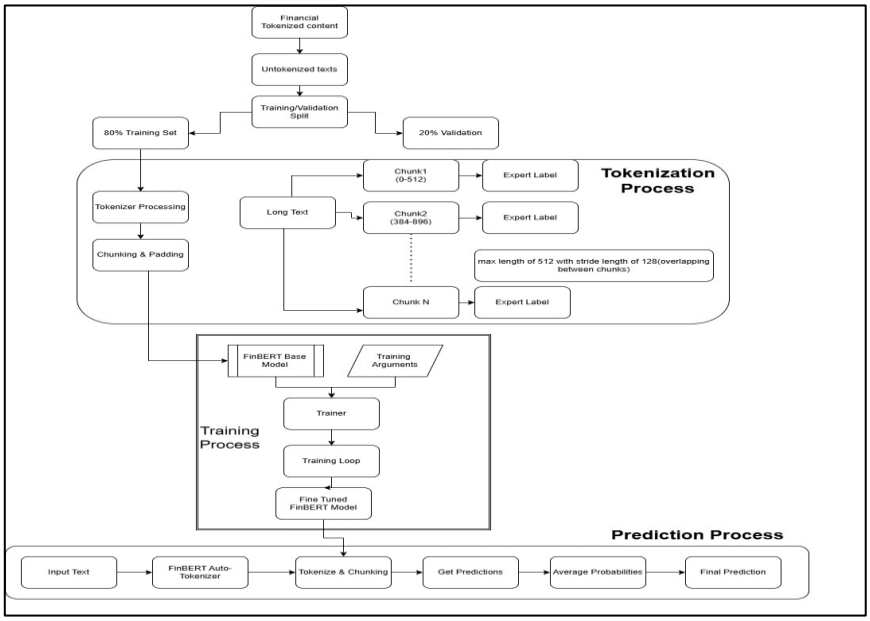

Figure 2. FinBERT Text Chunking and Training Process

The FinBERT fine tuning process involves the following process. The detailed process is as below.

- Identification of misclassified instances of FNBLex originally trained documents. A total of 118 information technology misclassified ECT documents were retrained with FinBERT
- Alignment of misclassified instances with the expert label. An expert label is the label provided by an expert as positive or negative at a document label.
- Break the text into tokens using the inbuilt auto-tokenizer
- Chunking of tokens into a size of 512 tokens as FinBERT can handle a maximum of 512 tokens per row. BERT and its variants can handle a maximum of 512 tokens per row, which limits the input size of the text [18]
- Perform experiments with model, sample size, epochs, and the learning rate factors
- Train the model
- Report performance measures

The results of the FinBERT experiments performed on FNBLex misclassified instances are provided in Table 4.

Table 4. FinBERT Experimental Results on FNBLex Misclassified Instances

Source	Epoch	Learning Rate	Accuracy	Precision	Recall	F1-Score	AUC
FinBERT-FNBLex Misclassified	3	0.00002	0.44	0.47	0.44	0.44	0.57
FinBERT-FNBLex Misclassified	3	0.00003	0.4	0.76	0.4	0.23	0.52
FinBERT-FNBLex Misclassified	5	0.00003	0.42	0.5	0.43	0.37	0.57
FinBERT-FNBLex Misclassified	5	0.00002	0.43	0.49	0.44	0.41	0.56
FinBERT-FNBLex Misclassified	10	0.00002	0.44	0.48	0.45	0.45	0.57
FinBERT-FNBLex Misclassified	10	0.00003	0.43	0.45	0.44	0.44	0.56

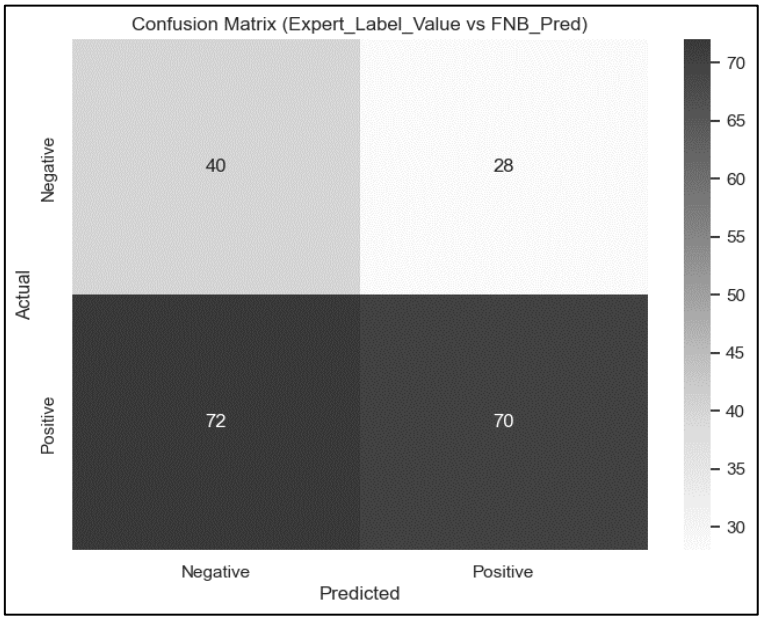

Figure 3. Confusion Matrix-FinBERT Fine Tuning on FNBLex Misclassified Instances

The FinBERT experimental results of the FNBLex misclassified instances are provided in Table 4. The setup consists of six different experiments in total with epoch and learning rate parameters fine-tuning. The FinBERT model is fine-tuned with 3, 5, 10 epochs and the learning rates of 0.00002 and 0.00003, respectively. FinBERT achieved the highest AUC of 0.57 with the corresponding F1-score of 0.45. This demonstrates the poor discriminatory power of FinBERT in extracting linguistic tones in earnings call transcripts. The FinBERT results are compared with the FNBLex results presented in Table 3. It is observed that FNBLex performed better than general-purpose lexicon, machine learning, and deep learning models. The study

compares the performance of lexicon-based methods, more specifically, the FNBLex results with FinBERT results of a similar study [15] as the baseline, and the same is presented in Table 5. It is observed that the domain purpose lexicon FNBLex provides a better discriminatory power general purpose lexicon VADER and the large language model FinBERT. The confusion matrix is depicted in Figure 3.

Table 5. Comparison of FNBLex with FinBERT Baseline Performance

Model	Model Type	ECT	CV	Vectorization	Epochs	Learning Rate	F1-Score	AUC	Baseline F1-Score
VADER	General Purpose Dictionary	IT	10	BoW	-	-	0.75	0.51	0.73
FNBLex	Domain Specific Lexicon	IT	5	BoW	-	-	0.84	0.77	0.73
FNBLex	Domain Specific Lexicon	IT	5	Word2Vec	-	-	0.84	0.77	0.73
FinBERT	LLM	IT		Autovectorizer	10	0.00002	0.45	0.57	0.73

5. Conclusion

The study has demonstrated the importance of lexicons in the field of finance. The study showed the poor generalization power of general-purpose lexicons and reiterates the need for domain specific lexicons to solve domain specific tasks. The study also examines the performance of FinBERT, a financial variant of BERT, in extracting linguistic tones present in the earnings call and compares with the domain-specific lexicon FNBLex, which again proved to be a better fit over FinBERT. Keeping the fact that data dictates the algorithm, The proposed results are compared with a recent similar study on FinBERT Click or tap here to enter text. as shown in Table 5. In sum, simple domain-specific additive lexicons have shown better discriminatory power over other models. The current study opens up opportunities to explore lexicon-based studies in other domains and interdomain knowledge transfer.

References

[1] G. Zhou, “Measuring investor sentiment,” Annual Review of Financial Economics, vol. 10, no. 1, 2018, 239–259.

[2] R. Sonpatki, A. Kathuria, and S. Sethi, “Earnings call transcripts as a source and resource for information systems research,” in Workshop on e-Business, 2022, 38–63.

[3] H. Zhao et al., “Revolutionizing finance with llms: An overview of applications and insights,” arXiv preprint arXiv:2401.11641, 2024.

[4] Z. Zhou and R. Mehra, “An End-To-End LLM Enhanced Trading System,” arXiv preprint arXiv:2502.01574, 2025.

[5] T. Loughran and B. McDonald, “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks,” J Finance, vol. 66, no. 1, 2011, 35–65.

[6] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” arXiv preprint arXiv:1801.06146, 2018.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, 4171–4186.

[8] L. Zhao, L. Li, X. Zheng, and J. Zhang, “A BERT based sentiment analysis and key entity detection approach for online financial texts,” in 2021 IEEE 24th International conference on computer supported cooperative work in design (CSCWD), 2021, 1233–1238.

[9] D. Araci, “Finbert: Financial sentiment analysis with pre-trained language models,” arXiv preprint arXiv:1908.10063, 2019.

[10] L. Hopman, “Measuring Financial Tone in Earnings Calls,” Tilburg University, 2021.

[11] E. Henry and A. J. Leone, “Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone,” The Accounting Review, vol. 91, no. 1, 2016, 153–178.

[12] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” arXiv preprint cs/0205070, 2002.

[13] B. V Nagendra, S. K. Chandar, J. B. Simha, and J. A. J. Bazil, “Financial Lexicon based Sentiment Prediction for Earnings Call Transcripts for Market Intelligence,” in 2024 5th

International Conference on Image Processing and Capsule Networks (ICIPCN), 2024, 595–603.

[14] B. V Nagendra, J. B. Simha, K. S. Manu, K. V Kirubanand, Y. Kaushal, and others, “Cross Domain Lexicon Transfer-A Case Specific to Application in Banking Domain,” in 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2025, 851–860.

[15] K. Kirtac and G. Germano, “Sentiment trading with large language models,” Financ Res Lett, vol. 62, p. 105227, 2024.

[16] S. Sohangir, N. Petty, and D. Wang, “Financial sentiment lexicon analysis,” in 2018 IEEE 12th international conference on semantic computing (ICSC), 2018, 286–289.

[17] Y. Yang, M. C. S. Uy, and A. Huang, “Finbert: A pretrained language model for financial communications,” arXiv preprint arXiv:2006.08097, 2020.

[18] A. H. Huang, H. Wang, and Y. Yang, “FinBERT: A large language model for extracting information from financial text,” Contemporary Accounting Research, vol. 40, no. 2, 2023, 806–841.