Relevance Analysis in Finacial News and Complaints Using BERT

Taeyoung Heo *LSA Statistics*
University of Michigan
Ann Arbor, MI
tedheo@umich.edu

***Abstract*—This project aims to identify systemic financial issues early by linking consumer complaints with financial news through sentiment analysis to detect emerging problems. It utilizes the DistilRoBERTa Finetuned Financial News Sentiment Analysis model from Hugging Face and the Consumer Finance Complaints Dataset from CFPB, focusing on data from 2020 to 2024. BERT is chosen for its bidirectional context understanding that allows it to capture intricate contextual dependencies in text. By fine-tuning the DistilRoBERTa model with the complaint data, the enhanced model can accurately detect the relevance to consumer complaint in finance news articles.**

I. Introduction

The financial sector is highly sensitive to issues that can have serious negative effects on the economy, as seen in events like the Subprime Mortgage Crisis and the Dot-Com Bubble. Early identification of emerging issues can prevent them from escalating into national economic problems. This project aims to address this challenge by linking financial news with consumer complaints to identify systemic issues early.

BERT (Bidirectional Encoder Representations from Transformers) is a highly suitable model for this project. Its bidirectional Context Understanding based on both its preceding and succeeding words allows it to capture intricate contextual dependencies in the text. In sentiment analysis, understanding the nuanced meanings that arise from context is crucial, especially in financial news where the sentiment might be implied rather than explicit. Pre-trained models like BERT are equipped with a significant amount of linguistic knowledge, due to their training on vast amounts of text. BERT's pre-trained capabilities can be fine-tuned with relatively less data to achieve high performance.

BERT was developed and introduced by Google (Jacob Devlin et al., 2018). After this, RoBERTa (A Robustly Optimized BERT Pretraining Approach) was developed to improve on BERT by optimizing the pretraining procedures by Facebook (Yinhan Liu et al., 2019), introducing dynamic masking and increasing batch size to make the model more robust and effective. DistilBERT was developed by Hugging Face (Victor SANH et al in 2019). DistilBERT is a smaller, faster, and cheaper version of BERT, using knowledge distillation. Recently, ModernBERT was proposed by researchers affiliated with research institutions and technology companies (Benjamin Warner et al., 2024), making it faster and more memory-efficient.

This project utilizes the "DistilRoBERTa Finetuned Financial News Sentiment Analysis" model on Hugging Face and the "Consumer Finance Complaints Dataset" from CFPB (Consumer Financial Protection Bureau).

II. Method

The first step involves importing two key datasets: one containing financial sentiment phrases and another with consumer complaints. The pre-trained model was originally trained on a dataset of 4,840 English-language financial news phrases, categorized by sentiment: 'positive', 'negative', or 'neutral'. The financial phrase is properly labeled with sentiments. For the complaints data, relevant columns such as the date received, product, issue, consumer complaint narrative, and company are extracted. The dataset covers dates from December 2011 to March 2025, but for training purposes, only data from the years 2020 to 2024 is included. A stratification column combining the year and product is created. Using this column, a stratified sample of 500 complaints is selected, ensuring variability with shuffling.

A. Data Preparation

The next step is to clean the text data by removing unnecessary characters and phrases within parentheses from the complaint narratives. After cleaning, the text is tokenized into individual sentences, and sentences shorter than 20 characters are filtered out to maintain meaningful content.

The initial sentiment classification uses a pre-trained model and extracts negative sentences from the complaints. The negative texts are merged with the neutral and positive sentiment sentences from the financial phrases. These sentiment labels are converted into numerical values for training our machine learning models. The entire dataset is then split into training, validation, and testing subsets to adequately prepare for model training.

B. Model Training

Both the tokenizer and model are initialized using the pre-trained setup. To optimize the model's learning capabilities, the AdamW optimizer is employed with a learning rate of 2×10^{-5}. A linear learning rate scheduler is used to gradually adjust the learning rate over the training phases.

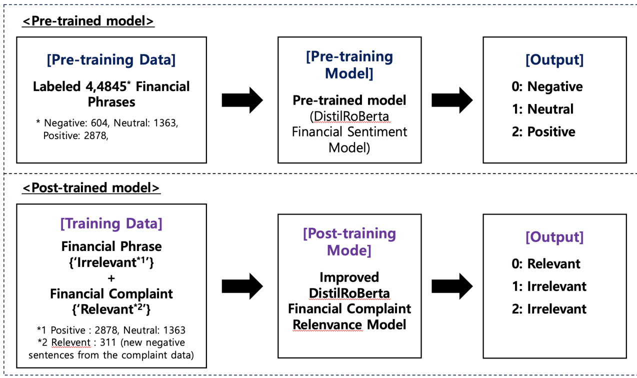

Fig. 1. Model Concept Overview

The dataset passes through the model to make predictions and calculate the loss over each training epoch. Using back-propagation, model adjusts its weights to minimize this loss. Following each batch, the optimizer is applied with the learning rate scheduler, ensuring progressive improvement in model accuracy and performance across epochs. The performance of post-training model is evaluated on a validation set. The fine-tuned model is determined With the optimal number of epochs.

III. RESULT

With three epochs, the model achieved optimal performance in validation loss. Then, the best-trained model was applied and validated in both the test data and the real articles. After cleaning and tokenizing them, our model used the articles to assess and classify their relevance in context.

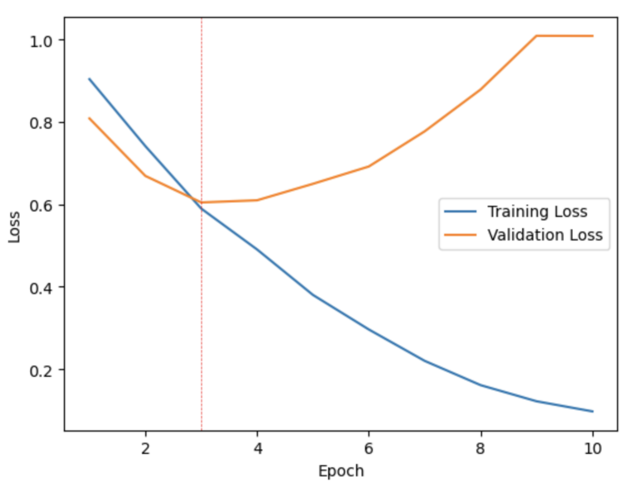

Fig. 2. Model Concept Overview

The test dataset was used to evaluate our model. The predicted labels were compared with the real labels, allowing us to calculate comprehensive metrics such as accuracy, precision, recall, and the F1 score. These evaluations ensured that the model not only performed well on our controlled datasets but also was effectively generalized to unseen data.

With the selected articles, a test was implemented to determine the relevance of the article. The test involved checking for at least two relevant sentences, each with a confidence score greater than 0.9. Through these tests, our model's ability to assess text relevance was verified.

The results indicate that our model successfully differentiates the relevant financial content. Through the training process, a robust and relevant analysis system was built to detect widespread concerns in financial news content.

TABLE I
MODEL EVALUATION METRICS

Accuracy	Precision	Recall	F1 Score
0.8638	0.8626	0.8137	0.8355

TABLE II
RESULTS OF MODEL TESTING ON SAMPLE ARTICLES

Id	Article Title	Prediction	True Value
1	US stocks skyrocket higher after Trump signals shift in trade policy	Relevant	Relevant[a]
2	Equifax issued wrong credit scores for millions of consumers	Relevant	Relevant
3	Improving food safety: Lessons learned from a food poisoning outbreak	Irrelevant	Irrelevant

[a] The relevance is moderate but monitoring is necessary.

IV. CONCLUSION

In this project, the implementation of a sentiment analysis system using the BERT model has demonstrated significant success in discerning the relevance and sentiment of financial content. Through the fine-tuning process, the model is equipped to generalize effectively to unseen data. This model could assess the relevance of the article with high precision, successfully identifying nuanced and context-sensitive statements. Future work could involve improving training datasets to ensure more accurate labeling and a focus on meaningful sentences. Additionally, it is necessary to adopt sophisticated techniques to filter and analyze only the important texts from news articles to enhance the model's effectiveness.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Google AI Language, 2018. [Online]. Available: https://research.google/pubs/bert-pre-training-of-deep-bidirectional-transformers-for-language-understanding/

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Facebook AI, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[3] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," EMC2: 5th Edition, 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[4] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli, "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," 2024. [Online]. Available: https://arxiv.org/abs/2412.13663