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.

By recognizing the correlation between financial journalism and consumer complaints, financial institutions and regulators can act proactively. This link would enable financial regulators to focus their attention on significant events. Given the vast amount of news and complaints, financial institutions and regulators often miss this kind of information. Establishing these connections can help them identify problems 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, the model is keep improving. RoBERTa (A Robustly Optimized BERT Pretraining Approach) was developed by Facebook (Yinhan Liu et al., 2019). RoBERTa was designed to improve on BERT by optimizing the pretraining procedures. It utilizes more data and adjusts training procedures like 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 in which a compact model is trained to reproduce a larger model's behavior. Recently, ModernBERT was proposed by researchers affiliated with research institutions or technology companies; Answer.AI, LightOn, Johns Hopkins University, NVIDIA, and Hugging Face (Benjamin Warner et al., 2024). ModernBERT was trained on a massive amount of data (2 trillion tokens) and can process longer text sequences (up to 8192 sequence length). It includes new optimizations that make it faster and more memory-efficient, especially when using common GPUs.

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). This model was originally trained on a dataset of 4,840 English-language financial news sentences, categorized by sentiment: 'positive', 'negative', or 'neutral'. The classification model is enhanced to distinguish between relevant and irrelevant news, rather than categorized by the sentiments.

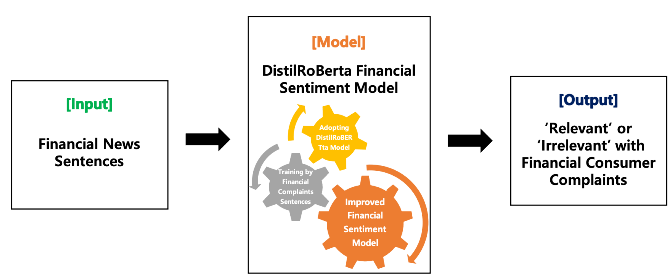

Fig. 1. Model Concept Overview

II. Method

The first step involves importing two key datasets: one containing financial sentiment phrases and another with consumer complaints. The financial phrases dataset 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

Identify applicable funding agency here. If none, delete this.

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 relevant 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 use a pre-trained model fine-tuned on financial news sentiment analysis data to classify the cleaned sentences into positive, negative, or neutral sentiments. After classification, negative sentences from the complaints are merged with the neutral and positive sentiment sentences from the financial phrase. 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.

Finally, data loaders are set up to manage the data in batches. These batches are shuffled and fed into the model during training, which speeds up the process and improves the model's learning efficiency.

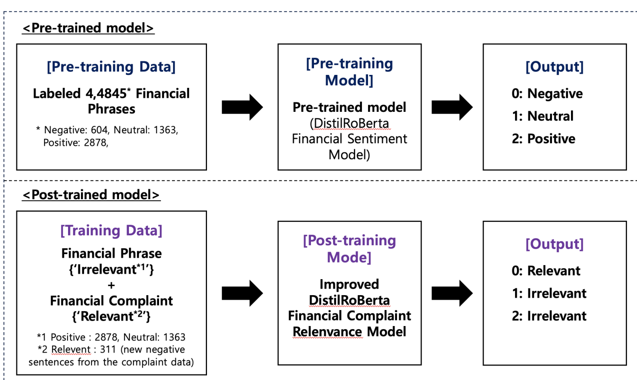

Fig. 2. Model Concept Overview

B. Model Training

The pre-trained model (mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis) provides us with a foundation, transferring knowledge embedded in the model. Both the tokenizer and model are initialized using this pre-trained setup. To optimize the model's learning capabilities, the AdamW optimizer is employed, known for adjusting the learning rate and weights effectively. A learning rate of 2×10-5 is set and a linear learning rate scheduler is used to gradually adjust the learning rate over the training phases.

The dataset pass through the model to compute 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 step applied with the learning rate scheduler, ensuring progressive improvement in model accuracy and performance across epochs. Post-training model performance is evaluated on a validation set. With the optimal number of epochs, fine-tuned the model is trained to achieve best performance for unseen data.

III. RESULT

With three epochs, the model achieved optimal performance in validation loss. The best-trained model was then applied and validated on both test data and real articles. After cleaning and tokenizing them, the 6 articles were used by our model to assess and classify their relevance in context.

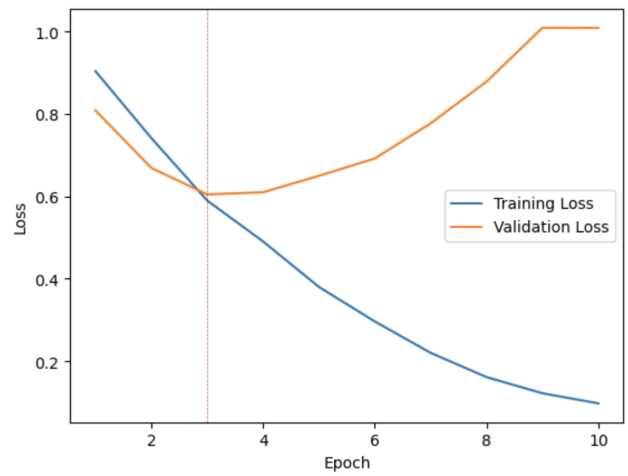

Fig. 3. Model Concept Overview

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

TABLE I
MODEL EVALUATION METRICS

Accuracy	***Precision***	***Recall***	***F1 Score***
0.8638	0.8626	0.8137	0.8355

With the selected articles, a test was implemented to determine article relevance. 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 relevant financial content. Through the training process, a robust relevant analysis system was built to detect widespread concerns across financial news content.

TABLE II
RESULTS OF MODEL TESTING ON SAMPLE ARTICLES

Id	*Article Title*	*Predicted Label*	*Actual Label*
1	US stocks skyrocket higher after Trump signals shift in trade policy	Relevant	Relevant[a]
2	Korean Won: Is political crisis setting up a longer term bear trade?	Relevant	Relevant[a]
3	The financial crisis of 2025? Better to be ready	Irrelevant	Irrelevant
4	Equifax issued wrong credit scores for millions of consumers	Relevant	Relevant
5	Consumers paid price overstated job growth that kept interest rates too high	Irrelevant	Irrelevant
6	Improving food safety: Lessons learned from a food poisoning outbreak	Irrelevant	Irrelevant

[a] The relevance to consumer complaints is somewhat ambiguous for these two articles, but they need to be monitored to check possible financial turmoil.

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 undergoing the fine-tuning process, the model is equipped to generalize effectively to unseen data. This model could evaluate article relevance with high precision, successfully identifying nuanced and context-sensitive statements. Future work could involve improving the 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

[5] "US stocks skyrocket higher after Trump signals shift in trade policy," CNN, 2025. [Online]. Available: https://www.cnn.com/2025/04/09/investing/global-stock-market-reciprocal-tariffs-hnk-intl/index.html

[6] "Korean Won: Is political crisis setting up a longer term bear trade?," Armchair Trader, 2024. [Online]. Available: https://www.thearmchairtrader.com/currencies/korean-won-forecast-martial-law/

[7] "The financial crisis of 2025? Better to be ready," Bloomberg, 2025. [Online]. Available: https://economictimes.indiatimes.com/news/international/global-trends/the-financial-crisis-of-2025-better-to-be-ready/articleshow/120344209.cms?from=mdr

[8] "Equifax issued wrong credit scores for millions of consumers," CNN, 2024. [Online]. Available: https://www.cnn.com/2022/08/03/business/equifax-wrong-credit-scores/index.html

[9] "Consumers paid price overstated job growth that kept interest rates too high," PYMNTS, 2024. [Online]. Available: https://www.pymnts.com/economy/2024/consumers-paid-price-overstated-job-growth-that-kept-interest-rates-too-high/

[10] "Improving food safety: Lessons learned from a food poisoning outbreak," WHO, 2024. [Online]. Available: https://www.who.int/westernpacific/newsroom/feature-stories/item/improving-food-safety--lessons-learned-from-a-food-poisoning-outbreak