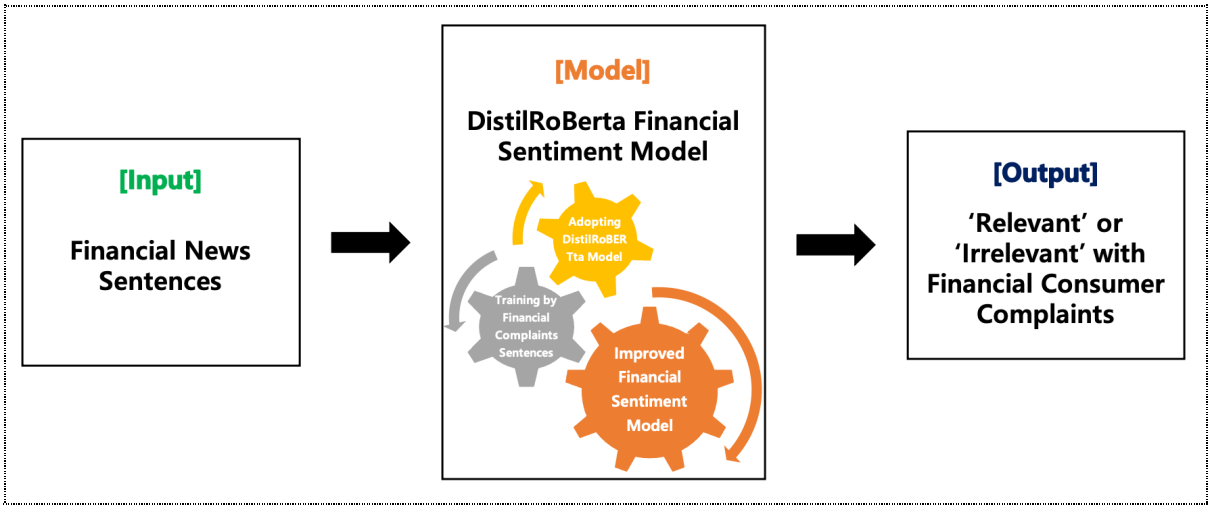
I. Overview

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 consumer complaints with financial news to identify systemic issues early. By recognizing the correlation between financial journalism and consumer complaints, financial institutions and regulators can act proactively.

The project will utilize the "DistilRoBERTa Finetuned Financial News Sentiment Analysis" model on Hugging Face and the "Consumer Finance Complaints Dataset" from CFPB (Consumer Financial Protection Bureau). We will enhance the classification model to distinguish between relevant and irrelevant news, rather than classifying it as positive or negative. Consumer complaints data will serve as the training datasets. If a news article is linked to recent consumer complaints.

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.

< Overview of Basic Model Concept >



II. Prior Work

1) Literature Review

- **BERT** (Bidirectional Encoder Representations from Transformers): This was developed and introduced by Google (Jacob Devlin et al., 2018). BERT was created to improve the ability of machines to understand the context of words. The innovation of BERT was its bidirectional training approach on language models, meaning it looks at a word's context from both the left and right. This allows BERT to a variety of NLP (Natural Language Processing) tasks like question answering and language inference.

- **RoBERTa** (A Robustly Optimized BERT Pretraining Approach): This was developed by Facebook (Yinhan Liu et al., 2019). RoBERTa was designed to improve upon BERT by optimizing the pretraining procedures. It utilizes more data and adjusts training procedures like introducing dynamic masking which generates the masking pattern every time we feed a sequence to the model, removing the NSP (Next Sentence Prediction) loss used in BERT, and increasing batch size to make the model more robust and effective.

- **DistilBERT**: This 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. It has 40% less parameters than a BERT model, runs 60% faster while preserving over 95% performances as measured on the GLUE language understanding benchmark.

- **ModernBERT**: This 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 is an enhanced version of BERT that provides improved efficiency and performance for text processing tasks within the encoder-only transformer models. It 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.

2) Potential Methods

I will use the DistilRoBERTa-financial-sentiment model, which is based on a distilled version of the RoBERTa-base model designed for efficiency in terms of parameters and data size. The model was originally trained on a dataset of 4,840 English-language financial news sentences, categorized by sentiment: 'positive', 'negative', or 'neutral'. I will modify this dataset by removing negative sentiment sentences. We will incorporate sentences from the CFPB Consumer Finance Complaints dataset, labeling these entries as 'complaint-relevant'. The goal is to enhance the model's classification abilities from identifying 'positive', 'negative', or 'neutral' sentiments to distinguishing between 'relevant' and 'irrelevant' news based on complaint trends.

III. Preliminary Results

1) Data understanding

The CFPB Complaints dataset comprises reports about consumer financial products and services which are forwarded to companies for response. The collection spans from December 2011 to March 2025 and includes 18 fields, such as Date, Product, Issue, and Complaint Narrative. Out of a total of 8,392,761 records, 5,702,403 lack a complaint narrative. After excluding those, the average word count per complaint is 177. This provides a sufficient amount of text to discern reasons for dissatisfaction, emotions, and product details. However, since the complaints often contain personal and irrelevant stories, they require standardization for effective model training.

< Key Data Fields >

Field Name	Description	Data Info
Date received	The date the CFPB received the complaint	From Dec 2011 to March 2025
Product	The type of product the consumer identified in the complaint	21 different products
Issue	The issue the consumer identified in the complaint	179 unique issues
Complaint Narrative	The consumer-submitted description of "what happened" from the complaint.	2,690,358 narratives (excluding missing values)
Company	The complaint is about this company	7,570 companies

2) Basic model

This model is a fine-tuned version of distilroberta-base on the financial_phrasebank dataset. It has been optimized to work efficiently with less data and fewer parameters. The model consists of 6 layers, a 768-dimensional hidden state, and 82 million parameters, compared to RoBERTa-base's 125 million. On average, it runs twice as fast as RoBERTa-base. The model was trained on 4,840 sentences from English-language financial news, each labeled with sentiment: 'positive', 'negative', or 'neutral'. In tests, it achieved a loss of 0.1116 and an accuracy of 0.9823. To retrain and fine-tune this model, significant computing power will be needed, and memory and speed issues might be encountered during training.

3) Tools From Class

- PyTorch: Deep learning frameworks that can be used to customize NLP models.
- Scikit-learn: Using machine learning pipelines and evaluating the accuracy score of models.
- Pandas: A library for data manipulation and analysis.
- Matplotlib and Seaborn: Creating static plots and visualizations.

IV. Project Deliverables

A successful project would result in a model that accurately classifies financial news articles as related to consumer complaints or not. The ultimate objective of this project is to create a practical tool that financial institutions or regulatory bodies can use to anticipate and prevent such issues from becoming social problems.

One sub-goal is to understand how to develop and retrain the initial NLP model, adjusting parameters to transform it from a sentiment classifier to a relation classifier focused on financial complaints. A second sub-goal is to ensure the model adapts to evolving complaint trends over time. Additionally, evaluating the model's accuracy and demonstrating its performance in classifying news articles as related to consumer complaints is another sub-goal.

V. Timeline

1) Week 1
- Understand the working principles and limitations of the model.
- Run initial tests with the model using sample data to understand how it works.
- Clean and organize the complaint data by removing data with missing values and standardizing formats.

2) Week 2
- Adjust parameters for the model to transform its ability to distinguish between relevant and irrelevant news.
- Training the model with preprocessed data and evaluate initial results.
- Identify and address issues such as overfitting and low prediction accuracy.

3) Week 3
- Summarize the methodology, training and tuning process.
- Present the final results using Python visualizations.
- Discuss limitations and propose future research to address these issues.

VI. References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google AI Language. https://research.google/pubs/bert-pre-training-of-deep-bidirectional-transformers-for-language-understanding/

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer† & Veselin Stoyanov (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Facebook AI. https://arxiv.org/abs/1907.11692

- Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. EMC^2: 5th Edition. https://arxiv.org/abs/1910.01108

- 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, Iacopo Poli (2024). Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. https://arxiv.org/abs/2412.13663