# Financial Text Summarization Using Transformers

**https://huggingface.co/spaces/Vickiiiyippp/financial_text_summarization.**

Vicki YE

Northwestern University

CIS 435:  Practical Data Science Using Machine Learning

Sunil Kakade

2/23/2025

## 1. Business Problem

The objective of this project is to develop a machine learning-based text summarization system tailored for financial analysts. By leveraging transformer-based models, the system can recognize speech, generate concise and insightful summaries from financial reports, and classify financial tone, helping professionals quickly extract key takeaways and make informed decisions. This tool aims to enhance productivity, reduce information overload, and improve financial analysis efficiency.

To address this problem, data collection was focused on financial documents, including annual reports, earnings call transcripts, market research reports, and central bank statements. The data mining and machine learning pipeline involved multiple stages. First, a diverse dataset consisting of long-form financial text was gathered. Next, text preprocessing steps were undertaken, including cleaning financial terminology, handling missing values, and tokenizing the input to ensure a structured format.

For model selection, BART was chosen for financial text summarization. The model was fine-tuned using domain-specific financial datasets to enhance summarization accuracy. Finally, the trained model was integrated into an interactive Gradio-based application, allowing financial analysts to input financial text or meeting recording and receive structured summaries focusing on critical financial insights.

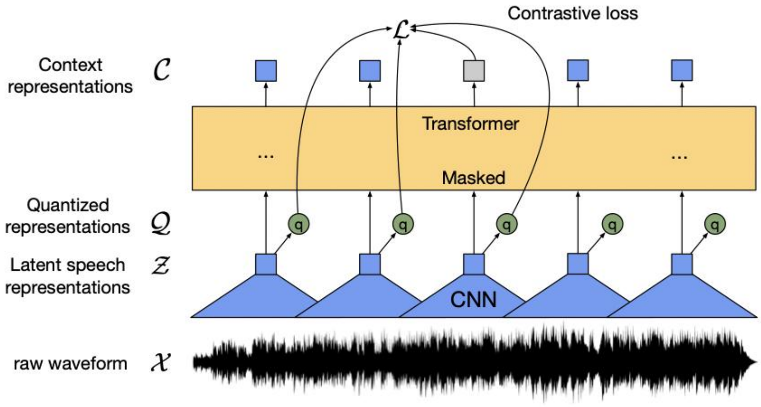
Insights were generated on how AI-powered text summarization can assist in financial analysis, earnings forecasting, and investment decision-making. This structured approach ensures data-driven financial research, improves efficiency, and enhances the ability to identify trends and risks in financial markets.

## 2. Neural Network Algorithms

### *Wav2Vec (Waveform-based Speech-to-Text Model)*

Wav2Vec is a deep learning model designed for automatic speech recognition (ASR). In this project, Wav2Vec is utilized to convert spoken financial discussions, such as earnings calls and business meetings, into structured text. By leveraging self-supervised learning, Wav2Vec enables high-accuracy transcriptions without requiring large amounts of labeled speech data. This allows financial analysts to efficiently extract and summarize key points from spoken discussions. However, the model may struggle with noisy environments or strong accents, requiring further fine-tuning for optimal performance.[1]

**Figure 1.** Illustration of Wav2Vec framework



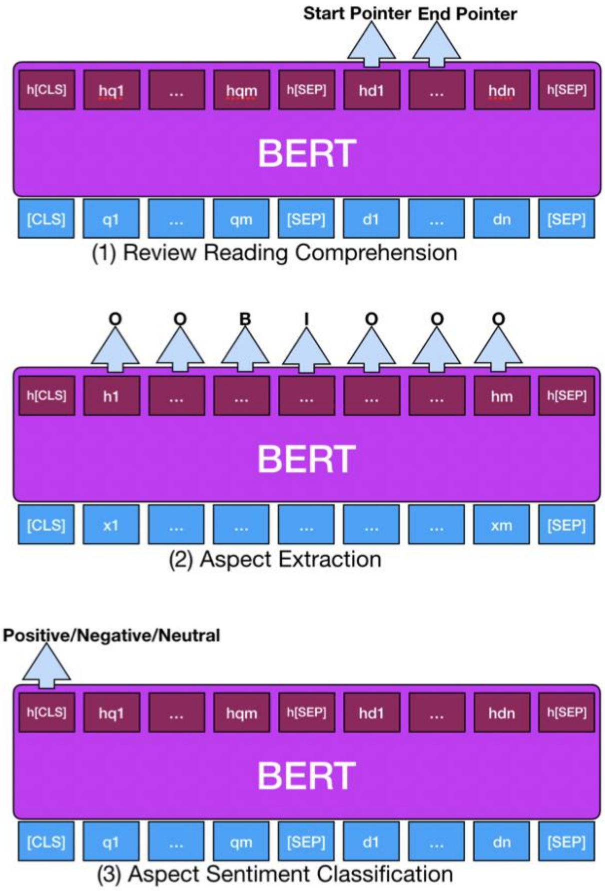
### *BART (Bidirectional Auto-Regressive Transformer)*

BART is a transformer-based model designed for text generation and summarization. It is particularly effective in abstractive text summarization tasks, as it reconstructs original text by predicting missing segments, making it suitable for financial document summarization. BART's bidirectional encoding allows it to understand financial context deeply while its autoregressive decoding ensures fluent and coherent summaries. However, BART requires substantial computational resources, and fine-tuning it on domain-specific data can be resource-intensive.[2]

### *BERT (Bidirectional Encoder Representations from Transformers)*

BERT is used for tone analysis in this project. BERT is primarily used for natural language understanding tasks and is effective in financial sentiment analysis, a crucial task for financial analysts when assessing market sentiment from earnings call transcripts and financial news. Unlike BART, BERT does not generate text but instead provides deep contextual embeddings, allowing it to accurately classify sentiments or extract meaningful insights. While BERT excels in understanding complex financial texts, it requires substantial labeled data for effective fine-tuning in specialized financial contexts.[3]

**Figure 2.** Overview of BERT settings for review read- ing comprehension (RRC), aspect extraction (AE) and aspect sentiment classification (ASC)



Each of these models plays a crucial role in financial text analysis. Wav2Vec enhances speech-to-text capabilities for financial meetings and investor calls, BART provides high-quality financial summaries, BERT aids in sentiment classification, and T5 remains an alternative for large-scale summarization tasks. The combination of these models can create a robust financial analysis system capable of extracting, summarizing, and interpreting financial data efficiently.

# 3. Transfer Learning Discussion

Transfer learning is a critical technique in modern deep learning, allowing models pre-trained on large datasets to be adapted for specialized tasks with significantly reduced computational requirements. Instead of training a model from scratch, transfer learning utilizes existing knowledge embedded in pre-trained networks, which can then be fine-tuned on domain-specific data. This approach not only improves model performance but also minimizes data dependency and enhances generalization across different but related tasks.

## 3.1 Transfer Learning in Image Processing

Transfer learning is widely applied in image processing, particularly in tasks such as object detection, image classification, and segmentation. Pre-trained models such as ResNet (Residual Networks), VGG (Visual Geometry Group Networks), EfficientNet, and Inception have demonstrated good performance on large-scale image datasets. These models extract hierarchical feature representations that can be effectively transferred to new datasets with limited labeled data, thereby reducing the training time and computational burden.

## 3.2 Commonly Used Pretrained Models

### *ResNet (Residual Networks)*

ResNet is known for its deep architecture that mitigates vanishing gradient problems through residual connections. It is highly effective in recognizing complex patterns in financial charts and satellite imagery.

### *VGG (Visual Geometry Group Network)*

A simpler yet powerful model that uses deep convolutional layers to learn spatial features. It is commonly used in facial recognition and medical imaging applications.

### *EfficientNet*

Designed for optimal performance with reduced computational costs, EfficientNet achieves superior accuracy with fewer parameters, making it ideal for mobile and embedded applications.

### *Inception (GoogLeNet)*

Uses inception modules to capture features at multiple scales, enhancing performance in tasks such as object detection and classification.

## 3.3 Transfer Learning for Financial Text Processing

While transfer learning is predominantly associated with image processing, its application in financial text analysis has been transformative. Pretrained language models such as BERT, GPT-3, T5, and BART have been adapted to financial sentiment analysis, text summarization, and risk assessment. By fine-tuning these models on financial datasets, they acquire domain-specific knowledge that enhances accuracy in financial forecasting, report generation, and decision support.

Transfer learning significantly reduces the need for large labeled datasets and accelerates model training. However, fine-tuning pre-trained models requires careful parameter selection and domain adaptation to prevent catastrophic forgetting, where the model loses its generalization ability. Additionally, computational costs can still be high for fine-tuning deep models, necessitating the use of cloud-based or distributed training approaches.

# 4. Transformers

## 4.1 Overview and Reasons of using Transformers

Transformers represent a breakthrough in deep learning for natural language processing (NLP) by introducing an architecture that effectively models long-range dependencies in text. Unlike recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, transformers utilize a self-attention mechanism that allows them to process entire sequences in parallel. This innovation significantly enhances computational efficiency and enables the development of models for tasks such as text generation, summarization, and sentiment analysis.

The adoption of transformers in financial text processing is driven by their ability to model complex dependencies within large-scale textual data. Unlike traditional NLP models, transformers facilitate efficient parallel processing, reducing training time while improving model performance. Their self-attention mechanisms allow for precise contextual analysis, making them well-suited for applications such as financial report summarization and sentiment analysis.

By leveraging transformer-based models, this project ensures accurate and efficient financial text summarization, empowering analysts to extract meaningful insights rapidly. The integration of BART, BERT, and Wav2Vec within financial NLP pipelines exemplifies the transformative impact of deep learning in financial decision-making and market analysis.

## 4.2 Common Transformer Models

### *T5 (Text-to-Text Transfer Transformer)*

Although not directly used in this project, T5 is a well-known transformer model in text summarization. It treats every NLP task as a text-to-text problem, making it highly flexible and powerful. T5 can be used for financial summarization by rephrasing complex financial data into digestible insights. However, its requirement for large amounts of data and tuning complexity make it less practical for real-time financial applications compared to BART.

### *GPT (Generative Pre-trained Transformer)*

GPT, developed by OpenAI, is a generative language model that leverages unsupervised learning to predict the next word in a sequence. It is pre-trained on vast amounts of text and fine-tuned for various NLP tasks, including conversational AI and text generation. Its autoregressive nature makes it highly effective in producing coherent and contextually relevant text. However, its unidirectional architecture can sometimes limit its ability to fully capture contextual dependencies within a sentence.

### *Electra (Efficiently Learning an Encoder That Classifies Token Replacements)*

Electra introduces a novel pretraining approach by replacing masked tokens with plausible alternatives instead of simply predicting missing words. This design significantly improves efficiency and accuracy in downstream tasks such as sentiment analysis and named entity recognition. In financial applications, Electra can be leveraged for risk assessment and fraud detection by identifying subtle variations in textual patterns. While it provides superior performance with fewer parameters than BERT, its pretraining phase can be complex and computationally demanding.

## 5. Application

This financial text summarization application integrates speech recognition, text summarization, and financial sentiment analysis to assist financial analysts in processing vast amounts of textual and audio-based financial information. The system is designed to transcribe speech from market discussions, summarize financial reports, and classify sentiment in financial news, enabling analysts to extract critical insights with minimal manual effort.

This application was developed and deployed on Hugging Face Spaces, leveraging Gradio for the user interface. The use of pretrained transformer models allows for efficient and scalable text processing without requiring extensive retraining.

### 5.1 Model Architecture

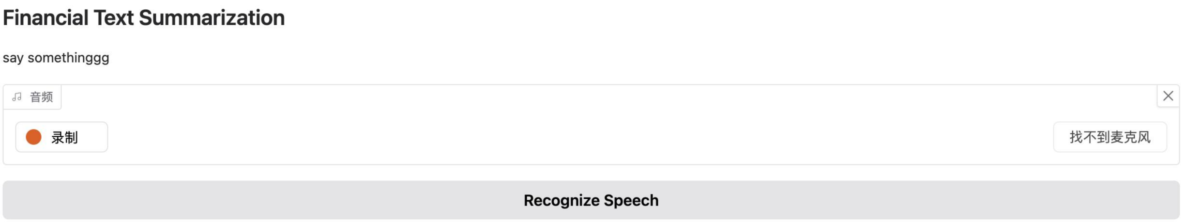
The application leverages multiple transformer-based models:

1. **Speech Recognition**
   The system utilizes wav2vec2 (facebook/wav2vec2-base-960h) to transcribe spoken financial discussions into text, allowing analysts to convert earnings calls and financial presentations into structured data.
2. **Text Summarization**
   A fine-tuned summarization model, BART (knkarthick/MEETING_SUMMARY), is employed to extract key insights from lengthy financial reports, helping analysts quickly grasp the most relevant information.
3. **Financial Sentiment Analysis**
   The BERT (yiyanghkust/finbert-tone) model is used to classify financial text into sentiment categories such as positive, neutral, or negative, enabling analysts to gauge market sentiment effectively.

### 5.2 Workflow and User Interaction

The application is built using Gradio, providing an intuitive interface for financial professionals. The workflow consists of the following steps. The workflow begins with speech recognition, where users can either upload an audio file or record financial discussions in real time. The system transcribes speech into text, allowing analysts to extract insights from earnings calls, investor briefings, or financial news broadcasts. [4]

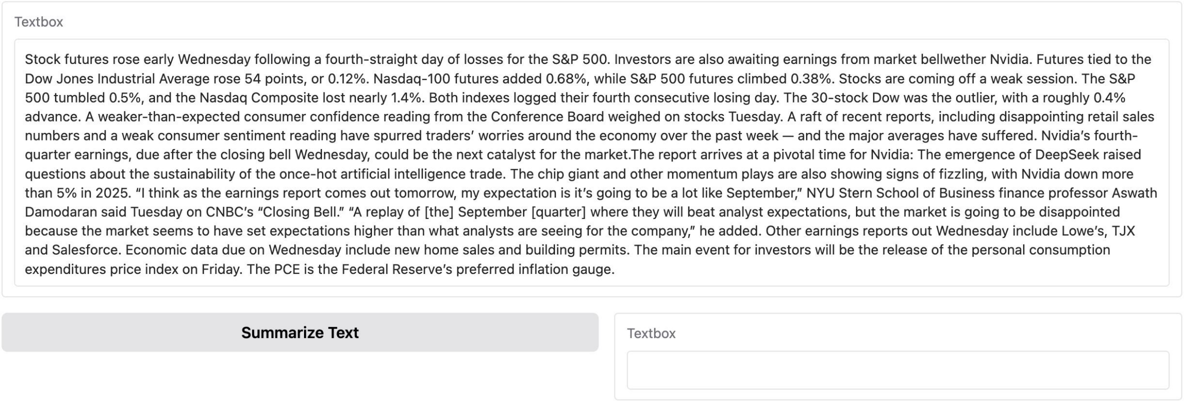
**Figure 3-5.** Financial Text Summarization Application



Financial Text Summarization

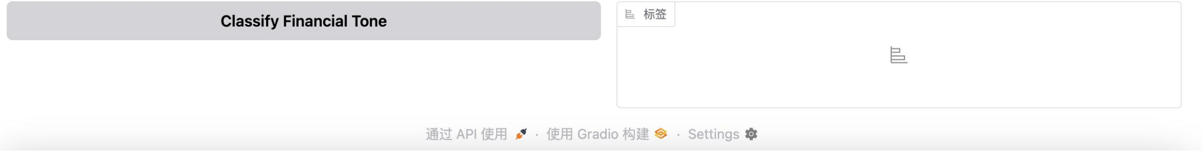
say somethinggg

Once the textual data is available, it is passed through the summarization model, reducing long-form content into concise, digestible insights.



Stock futures rose early Wednesday following a fourth-straight day of losses for the S&P 500. Investors are also awaiting earnings from market bellwether Nvidia. Futures tied to the Dow Jones Industrial Average rose 54 points, or 0.12%. Nasdaq-100 futures added 0.68%, while S&P 500 futures climbed 0.38%. Stocks are coming off a weak session. The S&P 500 tumbled 0.5%, and the Nasdaq Composite lost nearly 1.4%. Both indexes logged their fourth consecutive losing day. The 30-stock Dow was the outlier, with a roughly 0.4% advance. A weaker-than-expected consumer confidence reading from the Conference Board weighed on stocks Tuesday. A raft of recent reports, including disappointing retail sales numbers and a weak consumer sentiment reading have spurred traders' worries around the economy over the past week — and the major averages have suffered. Nvidia's fourth-quarter earnings, due after the closing bell Wednesday, could be the next catalyst for the market.The report arrives at a pivotal time for Nvidia: The emergence of DeepSeek raised questions about the sustainability of the once-hot artificial intelligence trade. The chip giant and other momentum plays are also showing signs of fizzling, with Nvidia down more than 5% in 2025. "I think as the earnings report comes out tomorrow, my expectation is it's going to be a lot like September," NYU Stern School of Business finance professor Aswath Damodaran said Tuesday on CNBC's "Closing Bell." "A replay of [the] September [quarter] where they will beat analyst expectations, but the market is going to be disappointed because the market seems to have set expectations higher than what analysts are seeing for the company," he added. Other earnings reports out Wednesday include Lowe's, TJX and Salesforce. Economic data due on Wednesday include new home sales and building permits. The main event for investors will be the release of the personal consumption expenditures price index on Friday. The PCE is the Federal Reserve's preferred inflation gauge.

The summarized text is then processed by the sentiment classification model, which determines the financial tone of the text, assisting users in assessing market sentiment.



The seamless integration of these models within the application ensures a streamlined and automated process, reducing the time required for financial data interpretation and enhancing user productivity.

## 5.3 Value Creation for Financial Analysts

This system provides financial analysts with a powerful tool for processing large volumes of unstructured financial data with minimal manual effort. By automating text summarization and sentiment classification, the application significantly enhances the efficiency of financial research, allowing professionals to focus on high-value decision-making rather than data extraction.

The ability to transcribe and summarize earnings calls and financial reports ensures that analysts can quickly identify critical information without the need to review extensive documents manually. Additionally, sentiment classification offers an objective measure of market sentiment, assisting traders, portfolio managers, and corporate strategists in making data-driven investment decisions.

The system’ s ability to operate on real-time and historical financial data ensures that it remains relevant in both short-term trading scenarios and long-term market analysis. Through its integration with pretrained transformer models, the application provides scalability, ensuring that it can handle large-scale financial datasets without sacrificing accuracy or efficiency.

## 5.4 Future Enhancements

Future enhancements for the system will focus on expanding its capabilities to accommodate a broader range of financial use cases.

One key improvement will be the introduction of multilingual support, enabling the system to process financial reports, earnings calls, and market news from global markets in multiple languages. Enhancing real-time financial monitoring by integrating live news feeds will also allow users to receive sentiment analysis on breaking financial developments, improving their ability to react to market changes instantly.

Another area of improvement is adaptive fine-tuning, where models will be continuously updated with financial domain-specific datasets to improve accuracy and relevance. This will ensure that summarization and sentiment analysis remain aligned with evolving financial trends and industry-specific terminology.

Additionally, integrating the system with automated trading platforms and portfolio management tools will create new opportunities for AI-driven financial decision-making.

By continuously evolving, the application will remain a valuable resource for financial analysts, hedge funds, investment firms, and institutional investors, ensuring that AI-powered financial text processing continues to drive efficiency and insight generation in the financial sector.

## Reference

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

[2] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[3] Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. In Proceedings of the 22nd nordic conference on computational linguistics (pp. 187-196).

[4] Hugging Face. *Financial Text Summarization.* Available at: https://huggingface.co/spaces/Vickiiiyippp/financial_text_summarization.