Financial Sentiment Analysis Using Classical and Transformer-Based Models

Nazib Ahmed, Md Azim Khan

May 19, 2025

1 Introduction

Financial news headlines contain valuable insights about market trends, investor sentiment, and company performance. Understanding the sentiment embedded in such headlines can offer important cues for investment decisions and financial analytics. [2]

In this project, we focus on sentiment classification of financial news headlines using some distinct NLP approaches:

- A traditional baseline model using TF-IDF features with Logistic Regression.
- A Support Vector Machine (SVM) model using TF-IDF features for sentiment classification.
- A Random Forest model trained on TF-IDF features for sentiment classification.
- A deep learning model based on BERT, that has been trained with our dataset
- A deep learning model based on FinBERT, a variant of BERT that has been fine-tuned specifically for financial text analysis.

The goal of this project is to assess the performance of modern transformer-based models like BERT (Fine-tuned) and FinBERT against classical machine learning models such as Logistic Regression, SVM, and Random Forest, focusing on their ability to classify sentiment in financial news headlines [3].

1.1 Research Questions

- **RQ1:** How does the performance of classical models (TF-IDF + Logistic Regression, SVM, Random Forest) compare to BERT in sentiment classification of financial news headlines?

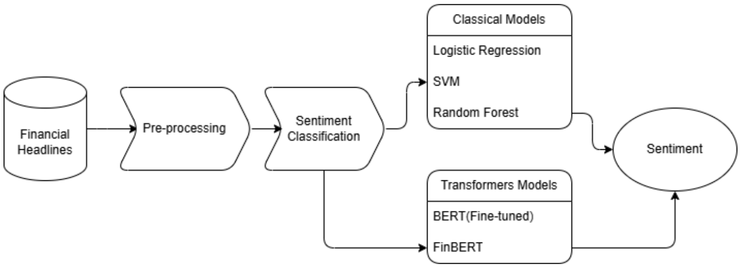
- **RQ2:** What are the strengths and limitations of each model (TF-IDF + Logistic Regression, SVM, Random Forest, and BERT) in understanding financial sentiment?
- **RQ3:** How does data preprocessing affect the performance of both classical models (Logistic Regression, SVM, Random Forest) and deep learning models (BERT)?

2 Methodology

2.1 Technical Approach

1. **Data Preprocessing:** Clean the headlines by:
   - Converting text to lowercase for uniformity.
   - Removing punctuation and URLs.
   - Expanding financial terms
   - Replacing multiple spaces with a single space.
2. **Label:** Sentiments are encoded as:
   - Positive → 2
   - Neutral → 1
   - Negative → 0
3. **Logistic Regression:**
   - Convert text to numerical features using TF-IDF.
   - Train a Logistic Regression model for sentiment classification.
4. **SVM:**
   - Convert text to numerical features using TF-IDF.
   - Train a Support Vector Machine (SVM) model for sentiment classification.
5. **Random Forest:**
   - Convert text to numerical features using TF-IDF.
   - Train a Random Forest model for sentiment classification.
6. **BERT (Fine-tuned):**
   - Fine-tune the BERT pre-trained model on the financial news dataset.
   - Perform direct sentiment inference from raw text, adapting the model to understand the context of financial headlines.

7. **FinBERT:**
   - Use the FinBERT pre-trained model from Hugging Face, optimized for financial text.
   - Perform direct sentiment inference from raw text, fine-tuned specifically for financial data.
8. **Evaluation:** Compare all models using the following metrics:
   - Accuracy
   - Precision
   - Recall
   - F1-Score
   - Confusion Matrix



2.2 NLP Techniques

In this project, several NLP techniques are employed to preprocess and classify sentiment in financial news headlines. These techniques are essential for converting raw text into meaningful insights for sentiment classification:

- **TF-IDF (Term Frequency-Inverse Document Frequency):**
  - A statistical measure used to evaluate how important a word is to a document in a collection or corpus. TF-IDF is used to convert the text data into numerical features for the classical models such as Logistic Regression, SVM, and Random Forest.
- **Logistic Regression:**
  - A classical machine learning model used for classification tasks. Logistic Regression is applied to classify financial headlines into sentiment categories (positive, neutral, and negative) using features derived from TF-IDF.
- **SVM (Support Vector Machine):**

  - A supervised learning model that finds the hyperplane that best separates data into different classes. SVM is used to classify sentiment in financial headlines based on the features extracted using TF-IDF.

- **Random Forest:**
  - An ensemble learning method that combines multiple decision trees to improve classification accuracy. Random Forest is applied to classify the sentiment of financial news headlines using TF-IDF features.

- **BERT (Bidirectional Encoder Representations from Transformers):**
  - A transformer-based model pre-trained on large amounts of text data. In this project, we fine-tune the BERT model on the financial news dataset to classify sentiment directly from raw text, leveraging its powerful contextual understanding.

- **FinBERT:**
  - A BERT variant fine-tuned specifically for financial text. We utilize FinBERT to directly infer sentiment from raw financial news headlines, taking advantage of its domain-specific fine-tuning for improved accuracy in sentiment classification.

- **Evaluation Metrics:**
  - The models are evaluated using standard metrics such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix to assess their performance on the sentiment classification task.

3 Team Contributions

3.1 Shared Responsibilities

All Members:

- Data collection and preprocessing.
- Splitting the dataset into training, validation, and test sets.
- Preparing the final project poster and documentation.
- Ensuring consistency in coding practices and model evaluation.

3.2 Individual Responsibilities

3.2.1 Md Azim Khan

Role: Splitting the dataset, Logistic Regression, SVM, and BERT implementation, performance evaluation.
Deliverables:

- Implement and train Logistic Regression and SVM models for sentiment classification.
- Fine-tune the BERT pre-trained model on the financial news dataset for sentiment analysis.
- Perform model evaluation and analysis using Accuracy, Precision, Recall, F1-Score..

3.2.2 Nazib Ahmed

Role: Random Forest and FinBERT implementation, Confusion Matrix generation.
Deliverables:

- Implement and train the Random Forest model for sentiment classification.
- Fine-tune the FinBERT pre-trained model for sentiment analysis on financial news headlines.
- Generate the Confusion Matrix for all models and compare their performance.

4 Evaluation and Dataset

4.1 Dataset Description

The dataset is sourced from **Kaggle's Financial News Dataset**, which contains a collection of news headlines related to finance, with sentiment labels provided for each headline. This dataset is ideal for performing sentiment analysis in the financial domain, as it reflects real-world financial news that may influence market trends and investor sentiment. [1]

The dataset is pre-labeled, with each headline being assigned a sentiment category: Positive, Neutral, or Negative, making it suitable for supervised learning tasks like classification. Here is the downloadable link: Dataset.

4.2 Experimental Setup

- **Split data:** 70% training, 30% testing

- **Sentiment Evaluation:** Accuracy, F1-Score, Precision, Recall and Confusion Matrix
- Compare classical models with transformer models on the same test set

References

[1] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

[2] Tim Repke and Ralf Krestel. Extraction and representation of financial entities from text. In *Data Science for Economics and Finance: Methodologies and Applications*, pages 241–263. Springer International Publishing Cham, 2021.

[3] Fan Sun, Ammar Belatreche, Sonya Coleman, T Martin McGinnity, and Yuhua Li. Pre-processing online financial text for sentiment classification: A natural language processing approach. In *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 122–129. IEEE, 2014.