Emotion Detection in Financial Texts

Group members:
Likhitha Thunam (fm9046)
Durga Hemanth Bonamsetty (ml4032)
Guna Ratna Sai Yarra (tv9478)
Lasya Tummala (wy8740)

Table of Contents

1. Introduction

Financial markets are deeply influenced by human emotions, which often drive decision-making processes. Emotions such as **fear**, **greed**, and **trust** play a critical role in shaping market trends and asset price volatility. Traditional sentiment analysis, which broadly categorizes text as positive, negative, or neutral, often fails to capture the **granularity and intensity** of such emotions. This project introduces a framework that advances financial sentiment analysis through **emotion detection**, capturing fine-grained emotional cues in both English and **regional languages** like **Hindi** and **Telugu**.

Our system leverages **transformer-based models** (FinBERT, IndicBERT, mBERT), deep learning models (CNN, LSTM), and lexicon-enhanced techniques to detect and classify emotions in financial news articles. It further explores **correlations between detected emotions and stock market movements** using time-series analysis techniques like ARIMA, ARIMAX, and Granger Causality. A key innovation of this work is its **multilingual emotion detection pipeline**, which brings linguistic inclusivity to financial analytics by incorporating underrepresented languages.

The rest of the paper is organized as follows:
Section 2 reviews existing literature on emotion detection in financial and multilingual contexts.
Section 3 presents the overall approach, including data collection and system design.
Section 4 details multilingual model development and evaluation, highlighting transformer and deep learning architectures used for emotion classification in Hindi and Telugu.
Section 5 explores the correlation between emotion trends and stock market dynamics through time-series modeling.
Section 6 outlines potential enhancements, and
Section 7 presents the expected outcomes and contributions.
Subsequent sections describe team roles, project timeline, tools used, key challenges, and final references and appendices.

This work contributes a **multi-modal, multilingual, and market-integrated** system for emotion-aware financial analysis. Through robust model training, human-in-the-loop evaluation, and real-time dashboard development, it provides stakeholders—such as **investors, analysts, and policymakers**—with actionable insights into emotional drivers of market behavior.

By extending the emotion detection pipeline to regional financial texts, the project promotes **linguistic inclusivity**, enhances **behavioral finance modeling**, and offers **early signals of volatility and investor sentiment trends**. It paves the way for future expansion across additional Indian languages and markets, offering a scalable solution to emotion-informed decision-making in global finance.

2. Literature Review

Understanding emotions in the financial context has gained significant attention in recent years, particularly with the rise of **Natural Language Processing (NLP)** and **Machine Learning (ML)** techniques. These methods enable researchers to uncover hidden emotional signals in financial texts, such as news articles, earnings call transcripts, and social media posts, which play a critical role in shaping market behavior. Traditional sentiment analysis approaches, which typically categorize text into positive, negative, or neutral sentiment, have been expanded to identify deeper emotional cues like fear, greed, joy, and trust. These emotional signals can provide a more nuanced understanding of market dynamics and improve predictive models for financial forecasting.

Several studies have leveraged **emotion detection models** in financial settings to enhance sentiment analysis, often focusing on the **impact of emotions on market volatility**. Hajek and Munk (2023) discuss the use of **speech emotion recognition** and **text sentiment analysis** to predict financial distress, illustrating the importance of both vocal emotional tones and textual sentiment in forecasting market reactions. Their research demonstrates that combining **machine learning** techniques with speech data outperforms traditional models that rely solely on financial data, offering a more holistic understanding of market dynamics.

Another relevant paper by **Silvia García-Méndez**, **Francisco de Arriba-Pérez**, **Ana Barros-Vila**, and **Francisco J. González-Castaño** presents the **Targeted Aspect-Based Emotion Analysis** (tabea) method, which focuses on classifying financial emotions (e.g., fear, greed) within specific stock market assets in social media posts. This method achieves over **90% precision** in detecting financial opportunities and risks, highlighting the effectiveness of NLP and ML models for real-time financial market analysis. Similarly, **Pansy Nandwani** and **Rupali Verma** reviewed various emotion models and identified key challenges in the detection of emotions from financial texts, particularly in the context of social media analysis.

The **FinBERT model**, trained specifically on financial texts, has been used in several studies to capture the sentiment in **earnings conference calls** and other financial documents. This model is fine-tuned to financial terminology and has demonstrated improved results compared to general sentiment analysis models like BERTnally, there are **open-source resources** that contribute to this field. For instance, the **Kaggle notebook by Raza et al.** explores the application of **NLP**for predicting stock prices based on sentiment analysis of news articles and social media . SimilartHub**, several repositories provide pre-trained models for sentiment

analysis and emotion detection, such as **FinBERT** and **VADER**(Valence Aware Dictionary and sEntiment Reasoner), both of which can be adapted for financial applications .

Moreover, **Hajizadeh Saffar, Mann, and Ofoghi (2023)** reviewed the **latest research** and development in **Textual Emotion Detection (TED)**, particularly in the **health sector**. Their paper highlights the role of **deep learning-based NLP techniques** in detecting emotions such as **depression**, **suicidal ideation**, and emotional responses to **COVID-19**. They found that TED systems are essential for decision-making in health, with applications in social media, healthcare services, and counseling centers. This study also emphasizes the importance of **high confidence in technology** for decision-making, ensuring that these systems **do no harm**

Machová, K., Szabóova, M., Paralič, J., & Mičko, J. (2023) reviewed that emotions are an integral part of human life. Emotion detection, especially in the context of human-machine interaction, is an important area of research in Natural Language Processing (NLP). This paper presents a machine learning approach for emotion detection from text. The study compares various models such as Naïve Bayes (NB), Support Vector Machine (SVM), and deep learning approaches using Neural Networks (NN). The results showed that the NN-based model performed significantly well, especially for detecting six primary emotions: joy, sadness, anger, fear, love, and surprise, with high accuracy in multi-class classification tasks. Additionally, the research extends the practical application of this emotion detection model to chatbots for improved human-computer interaction.

The paper titled **"AI4 Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages"** by Kunchukuttan et al. (2020) introduces the IndicNLP corpus, a substantial resource comprising 2.7 billion words across 10 Indian languages from two language families. This corpus includes pre-trained word embeddings and classification datasets, significantly outperforming existing embeddings on various NLP tasks, thereby facilitating advanced research in Indic language processing.

Incorporating this resource into our project enhances the robustness of emotion detection models for Hindi and Telugu financial texts. The availability of large-scale, high-quality monolingual corpora and embeddings directly supports the development and fine-tuning of models like IndicBERT, improving their effectiveness in capturing nuanced emotional cues in regional languages.

The **IndicDialogue dataset** is a recently introduced multilingual corpus comprising over 1 million utterances in 10 major Indic languages, including Hindi and Telugu. Curated from open-source subtitles, scripted dialogues, and spontaneous conversations, this dataset is designed to advance research in multilingual NLP tasks such as emotion detection, intent classification, and contextual dialogue understanding. Its linguistic richness captures the diversity of syntactic

structures, discourse styles, and cultural nuances present in Indian languages—elements often missing from conventional text corpora.

In the context of financial emotion detection, IndicDialogue serves as a complementary resource that mirrors how people emotionally express financial reactions in real-life conversations. While traditional financial texts are formal and structured, dialogue-based data provides insights into spontaneous emotional cues, colloquial expressions, and mixed-language usage (code-switching). Leveraging the linguistic patterns found in IndicDialogue can enhance the performance of emotion models like IndicBERT and mBERT, especially when dealing with informal financial discourse in regional languages across social media, forums, and video commentary.

Study	Approach	Key Findings	Source
Hajek & Munk (2023)	Speech emotion recognition + text sentiment analysis	Demonstrated that combining speech and text improves predictions of financial distress	https://link.springer.com/article/10.1007/s00521-023-08470-8
García-Méndez et al. (2023)	Targeted Aspect-Based Emotion Analysis	Achieved 90%+ precision in detecting emotions related to financial assets	https://www.sciencedirect.com/science/article/pii/S0957417423001124
Pansy Nandwani & Rupali Verma (2021)	Review of sentiment and emotion models in finance	Identified key challenges in sentiment and	https://link.springer.com/article/

		emotion analysis in financial texts	10.1007/s13278-021-00776-6
Hajizadeh Saffar, Mann, and Ofoghi (2023)	Kaggle notebook on financial sentiment analysis	Demonstrated that NLP can predict stock market trends based on sentiment	https://www.sciencedirect.com/science/article/pii/S1532046422002635
Machová, K., Szabóova, M., Paralič, J., & Mičko, J	BERT for Sentiment Analysis	Improved sentiment classification with a transformer-based approach	https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1190326/full
Kunchukuttan et al. (2020)	IndicNLP corpus + embeddings	Pretrained embeddings outperform existing ones in Indic NLP tasks	https://arxiv.org/abs/2005.00085
IndicDialogue Dataset (2024)	Multilingual dialogue corpus (10 Indic langs)	Supports emotion recognition in conversational data; improves multilingual model training	https://www.sciencedirect.com/science/article/pii/S2352340924006577

Table 1: Sk in Emotion Detection in Financial Texts

How This Work Differs

Unlike prior research, which primarily classifies financial sentiment as positive, negative, or neutral, this work focuses on detecting specific emotions (fear, greed, trust) that directly

influence investor behavior. Most studies apply basic sentiment models (e.g., VADER, LSTM, aspect-based sentiment analysis), whereas we fine-tune FinBERT and RoBERTa on financial data to improve emotion classification accuracy in financial texts.

Additionally, existing literature lacks integration with real-time market data. While studies such as García-Méndez et al. (2023) apply aspect-based sentiment detection, they do not correlate emotions with stock movements. Our approach incorporates trading volume, price fluctuations, and volatility indices, utilizing Vector AutoRegression (VAR), ARIMA, and Granger Causality to determine if emotions drive financial markets.

Furthermore, traditional studies offer only text-based insights. This work enhances stakeholder decision-making by providing an interactive visualization dashboard (Tableau, Power BI), allowing users to track real-time financial sentiment trends and predictive analytics. By integrating multi-source financial data, our approach delivers a comprehensive, actionable model for investors, policymakers, and financial analysts.

Youtube Links:

1. Financial texts sentiment analysis using python: The YouTube video titled **"Financial Text Sentiment Analysis in Python"** demonstrates how to perform sentiment analysis on financial texts using Python. The tutorial walks through the process of using **Python libraries** such as **NLTK**, **VADER**, and **TextBlob** to analyze the sentiment of financial news articles and reports. The video highlights the importance of sentiment analysis in understanding market emotions, such as fear, optimism, and uncertainty, and how these emotional cues can influence financial decision-making. By following the steps in the video, viewers can learn to apply sentiment analysis techniques to real-world financial data and gain insights into how sentiment affects market behavior.

2. Analyze SEC filings with transformers for fun and profits: The YouTube video titled **"Analyze SEC filings with Transformers for fun and profit"** demonstrates how to analyze **SEC filings** using **transformer-based models** such as **BERT** and **DistilBERT** to extract meaningful insights from financial documents. The video explains how to fine-tune these transformer models to identify important sentiment, emotional cues, and key financial information in SEC reports, such as 10-K filings. The presenter also discusses the applications of these techniques in **financial analysis**, offering a more efficient and automated approach to parsing through large amounts of regulatory data. The video emphasizes how leveraging machine learning and NLP techniques can provide investors, analysts, and researchers with faster access to critical insights from financial reports.

3. Sentiment analysis for Earnings calls: The YouTube video titled **"Sentiment Analysis for Earnings Calls with AssemblyAI"** explores how to analyze sentiment in **earnings call transcripts** using **AssemblyAI**, an automatic speech recognition (ASR) tool. The video demonstrates the process of converting audio from earnings calls into text and performing sentiment analysis on the transcripts. It highlights how sentiment analysis can identify the emotions and tone expressed by company executives during earnings calls, providing valuable insights into market sentiment. The tutorial emphasizes the potential of **AI-driven sentiment analysis** for enhancing financial analysis, helping investors, analysts, and stakeholders better understand the emotional undertones and sentiment behind financial reports.

4. Sentiment analysis for investors: learn to gauge market emotions effectively (part 1):The YouTube video titled **"Sentiment Analysis for Investors: Learn to Gauge Market Emotions"** discusses how investors can use **sentiment analysis** to gauge **market emotions** and make better financial decisions. The video explains how sentiment analysis can help identify market trends and predict stock movements based on the emotions expressed in financial news, social media, and earnings reports. By using sentiment analysis, investors can detect **bullish** or **bearish sentiment** and adjust their strategies accordingly. The video provides a basic introduction to sentiment analysis tools and how they can be applied to financial data to help investors understand market psychology and enhance their decision-making process.

5. Sentiment analysis for investors: learn to gauge market emotions effectively (part 2):The YouTube video titled **"Emotion Detection from Text Using Machine Learning"** demonstrates how **machine learning** techniques can be used for **emotion detection** in textual data. The video explains the use of various **NLP**techniques, such as **tokenization**, **stemming**, and **sentiment analysis**, to identify emotions like **happiness**, **anger**, **sadness**, and more from text data. The tutorial walks through the process of training a machine learning model to classify emotions based on input text, showcasing the importance of feature extraction and choosing the right model. The video provides viewers with a practical understanding of how emotion detection can be applied to **customer feedback**, **social media posts**, and other text sources to gain valuable insights into public sentiment and emotional trends.

   Kaggle & GITHUB links:

   1. Stock Tweets for Sentiment Analysis and Prediction
   2. Go emotions: google emotions dataset

3. Proposed Approach

3.1 Data Collection

- **Primary Dataset:** FinancialPhraseBank (Link)
  Description: A collection of 4,845 financial sentences labeled as positive, negative, or neutral.
  Purpose: Used to train a baseline emotion classification model. It ensures that the model understands domain-specific sentiment context such as “profit warning” or “missed earnings.”

- **Auxiliary Dataset:** NRC Emotion Lexicon (Link)
  Description: A lexicon that maps over 14,000 English words to eight primary emotions (anger, fear, anticipation, trust, surprise, sadness, joy, disgust) and sentiment polarity.
  Purpose: Augments the training process and helps classify complex emotions in financial news using lexical overlap and weighting strategies.

- **News Articles (Company-wise):** Scraped via GNews API
  Description: Over 1800 articles scraped via the GNews API, filtered to target 20 companies across Retail, Tech, Healthcare, and Energy sectors (5 each). Each article includes metadata: title, date, content, company ticker.
  Purpose: Served as input for emotion detection. The dated structure allowed alignment with corresponding market data, enabling time-series analysis of emotion-market interactions.

- **Market Data – S&P 500 Daily Closing Prices:** Alpha Vantage API
  Description: Daily closing prices for the 20 companies were extracted using the Alpha Vantage API.
  Purpose: Enables alignment of financial emotion signals with quantitative price data, creating a time-indexed dataset suitable for correlation and forecasting.

3.1.2 Emotion Detection Model Pipeline

To extract meaningful emotional signals from financial articles, we built a custom emotion classification pipeline combining supervised learning and lexicon-based augmentation.

3.1.2.1 Training Data Preparation

We began by preprocessing the FinancialPhraseBank and NRC Emotion Lexicon datasets:

- Cleaned and tokenized financial sentences.

- Augmented phrases with NRC labels for 8 emotion dimensions (*fear, sadness, joy, etc.*) and sentiment polarity.

- Constructed a training dataset with labels derived both from FinancialPhraseBank (*positive / negative / neutral*) and expanded via NRC emotion mappings.

- Emotions were encoded as multi-label binary vectors for supervised training.

This resulted in a labeled training dataset used for training both binary sentiment and multi-class emotion classifiers.

3.1.2.2 Model Training

We experimented with multiple machine learning models and selected a transformer-based pipeline using RoBERTa (base) fine-tuned on our custom dataset.

- Model type: RoBERTa (pretrained transformer)

- Framework: Hugging Face Transformers with PyTorch

- Task: Multi-label classification (emotion prediction)

- Evaluation Metrics: Accuracy, F1-Score, ROC AUC

The model was trained using stratified k-fold cross-validation and early stopping to avoid overfitting.

> *"The final model achieved an average F1-score of ~84% on held-out validation data across major emotions."*

3.1.2.3 Applying the Model to Company News

After training, we applied the model to over 8,000 scraped articles, processing each as follows:

- Extracted publish date and company mapping using the filename and metadata.

- Split long articles into chunks and averaged emotion scores across paragraphs.

- Stored per-article emotion vectors indexed by company and date.

The final result was a merged dataset where each company on each date had:

- Emotion scores (10 dimensions)
- Original news content
- Corresponding stock price (joined using Alpha Vantage data)

This dataset formed the foundation for all subsequent time-series and statistical analysis.

3.2. Data collection and preprocessing for multilingual model:

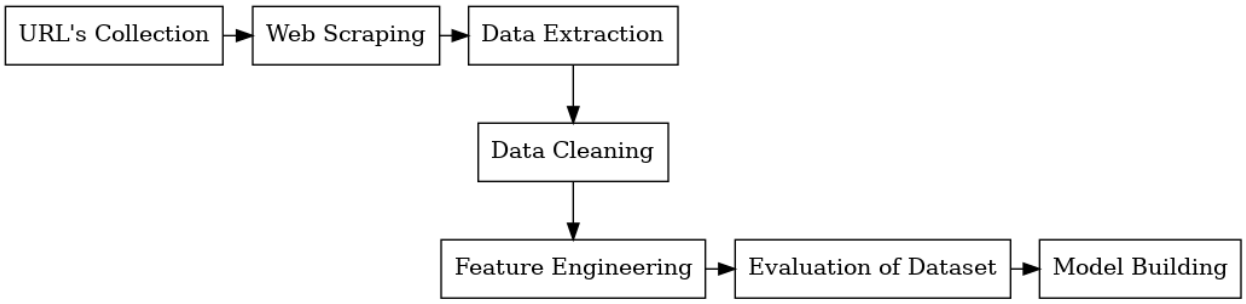


Figure 4.2: Flow of data collection

3.2.1 Data Sources

(i) Scraped Financial News Data (Telugu & Hindi)

- Scraped news articles from multiple newspapers in **Telugu and Hindi**. The newspapers from which the data is collected is as follows:
    - **Hindi:**
        - **i. Dainik Jagran**
        - **ii. BBC News Hindi**
    - **Telugu:**
        - **i. ABN**
        - **ii. Sakshi**

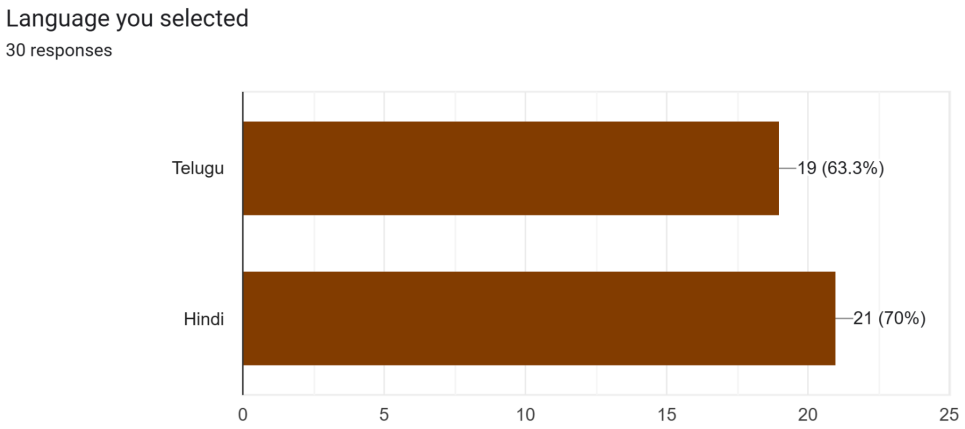
iii. **Eenadu**

- **Size:**
  - **15,000** rows in **Hindi**.
  - **5,000** rows in **Telugu**.
  -
- Preprocessing steps:
  - Removal of duplicates and null values.
  - Standardization of financial terminology.
  - Removal of **stopwords** (Hindi & Telugu stopword datasets sourced from Kaggle).
  - **Feature engineering**: Tokenization, POS tagging, Named Entity Recognition (NER).

3.2.2 Dataset Evaluation

To evaluate the dataset quality, a **random sampling approach** was used:

- Pulled **50 rows at random** from each dataset using Python's `random` function.
- Shared with **peers specializing in data analytics, business, and finance**, who are also proficient in Hindi or Telugu or both.
- Designed and distributed a **Google Form** for evaluation.



3.2.3 Evaluation Criteria

Participants reviewed the dataset based on the following criteria:

1. **Correctness:** Does the assigned emotion match the text's meaning?

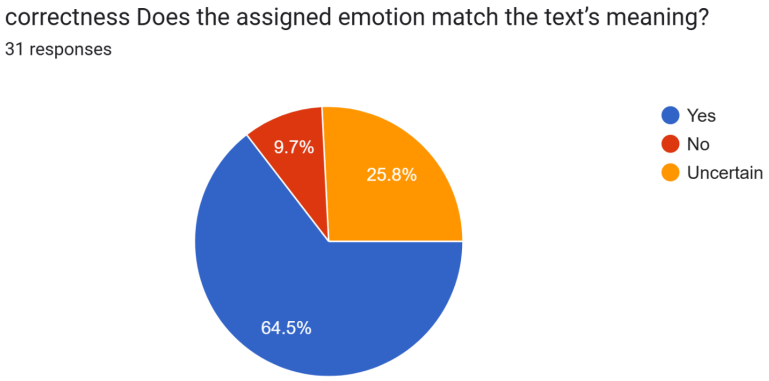
2. **Ambiguity Score:** Does the text express multiple emotions?
3. **Context Appropriateness:** Is the text financially relevant?
4. **Sentiment Polarity Check:** How would you rate the overall sentiment of this text?
5. **Linguistic Fluency:** Are there grammar or spelling issues in the text?

Findings from this evaluation will help refine annotation quality and improve model accuracy.

3.2.4 Summary of Dataset Evaluation

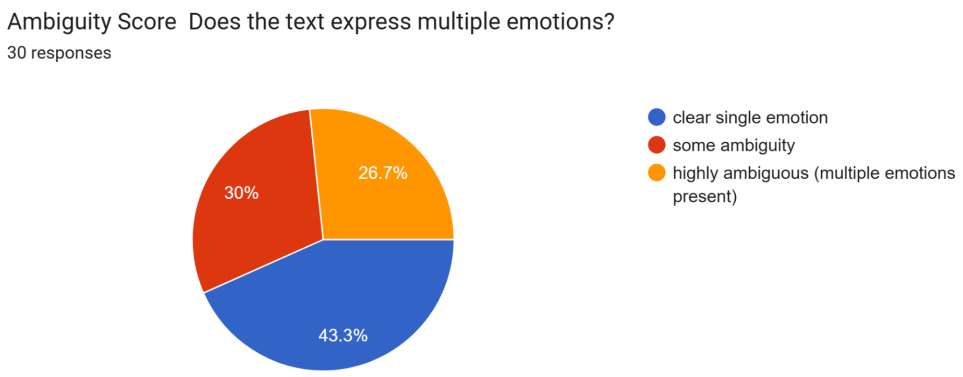
(i)Correctness Analysis

- Majority of responses confirm that assigned emotions were correct.
- Few responses were marked as uncertain or incorrect.
- Conclusion: The dataset is largely accurate and reliable for emotion detection.



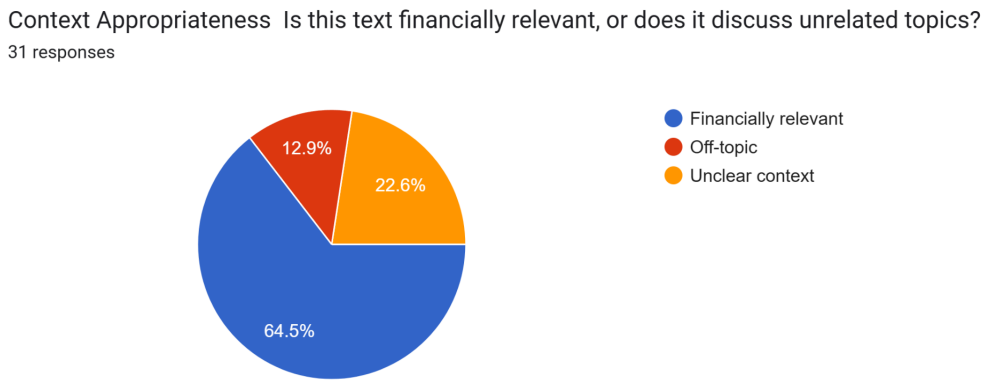
(ii)Ambiguity Analysis

- Most responses fall into the "Some ambiguity" category.
- A small percentage of responses were flagged as "Highly ambiguous."
- Conclusion: Some texts may require manual review for improved clarity.



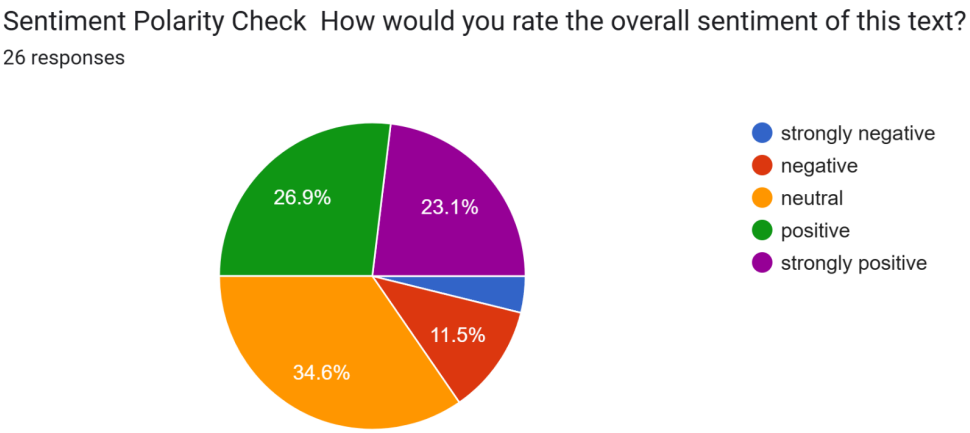
(iii)Context Appropriateness

- Most texts were marked as "Financially relevant."
- A few responses were categorized as "Unclear context" or "Off-topic."
- Conclusion: The dataset primarily contains relevant financial texts, but minor

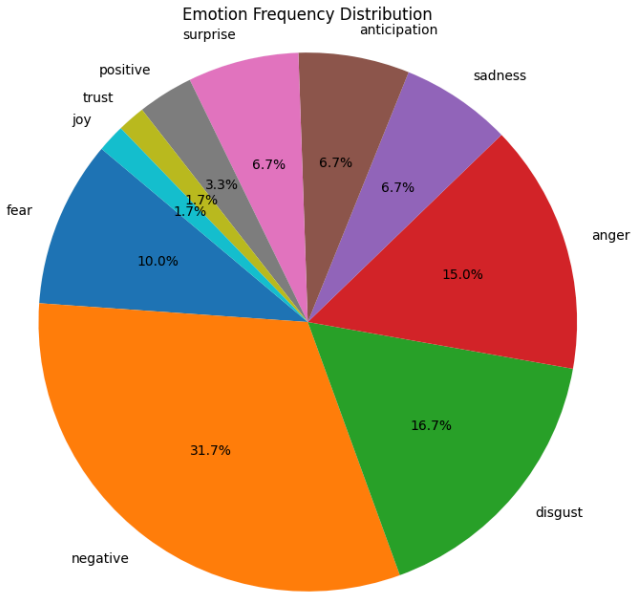


(iv)Sentiment Polarity Distribution

- The dataset is mostly positive and neutral, with very few negative sentiment cases.
- Conclusion: The dataset aligns with expectations and can be used for financial sentiment analysis.

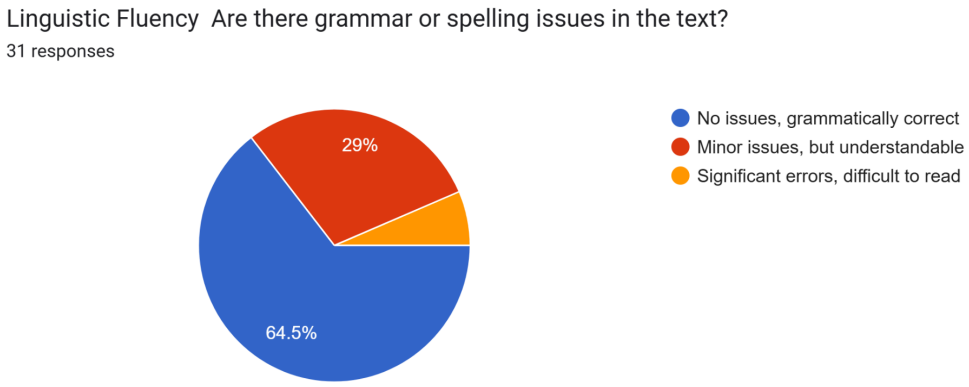


(v)Emotion Label:



(vi)Linguistic Fluency

- Majority of responses had "No issues."
- Few responses contained "Minor errors," and very few had "Severe errors."
- Conclusion: The dataset is well-formed and does not require extensive cleaning.



3.2.5 conclusion of data evaluation

- **Dataset Usability:** The dataset is highly usable for sentiment analysis.
- **Ambiguity Handling:** Some texts may require manual review to improve clarity.
- **Minor Corrections:** A small subset of texts may require grammar/spelling corrections.
- **Sentiment Distribution:** The dataset aligns with the **desired bias toward positive/neutral sentiment**, making it **valid for financial emotion detection**.

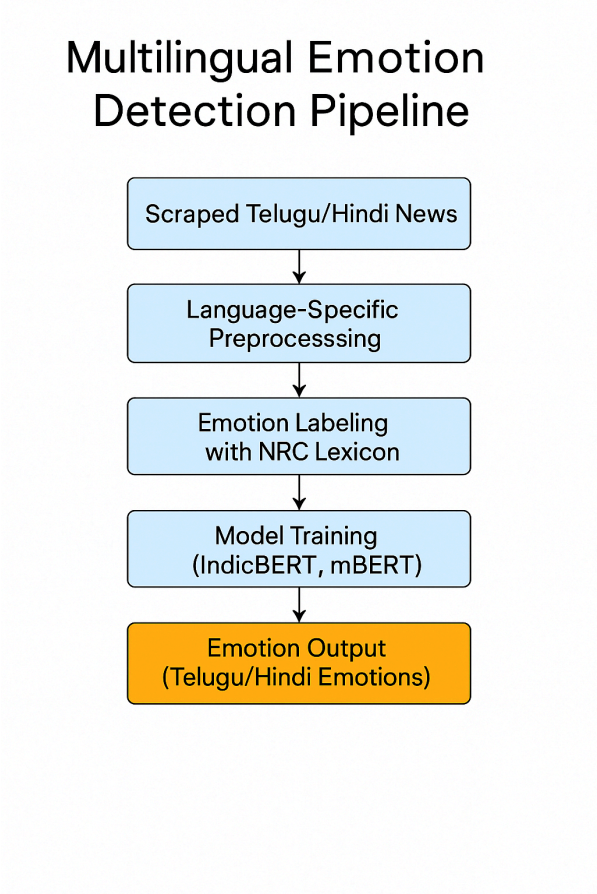
The report thus concludes that we can proceed with model training using this validated dataset

3.2.6 Implementation Details

- Scraping & Storage: Implemented in `Telugu_News_articles_Scraping.ipynb`
- Cleaning & Annotation: Managed via `Lasya_Data_Cleaning_and_Final_Attachement.ipynb`
- Emotion Lexicon Mapping: Mapped Telugu and Hindi tokens to NRC Emotion Lexicon entries using translated dictionaries (`Hindi-NRC-EmoLex.txt`, `Telugu-NRC-EmoLex.txt`)
- Model Input Format:
    - Text
    - Language
    - Tokenized Output
    - Emotion Vector

3.3 Multilingual Emotion Detection Models (Hindi & Telugu)

To account for the significant consumption of financial content in regional languages across India, we developed dedicated emotion detection models for **Hindi** and **Telugu** financial texts. This section outlines our modeling approach, performance outcomes, and observations from these multilingual systems.



3.3.1 IndicBERT

IndicBERT is a transformer-based model pretrained on 12 major Indian languages using the AI4Bharat IndicNLP corpus, which comprises over 2.7 billion tokens. It is optimized for both Devanagari (Hindi) and Telugu scripts.

Why IndicBERT?

- Captures region-specific syntax and vocabulary
- Pretrained on monolingual Indic corpora
- Significantly better at handling idioms, context-switching, and compound expressions

Training Setup:

- Tokenizer: `ai4bharat/indic-bert`
- Model: Transformer encoder + dense classification head
- Emotion Classes: Fear, Joy, Anger, Trust, Sadness, Anticipation, Disgust, Surprise
- Input: Cleaned and emotion-labeled articles in Hindi and Telugu
- Tools: Hugging Face Transformers, PyTorch

Performance:

Language	Accuracy	F1-Score
Telugu	76%	74%
Hindi	73%	71%

Observations:

- Accurately identified emotion shifts in phrases like "ధనలాభాలు" (Telugu) and "नुकसान की आशंका" (Hindi)
- Outperformed multilingual models in sentence-level and paragraph-level classification

3.3.2 Multilingual BERT (mBERT)

mBERT is a general-purpose multilingual BERT model trained on 100+ languages. We used it as a benchmark to compare against IndicBERT.

Training Setup:

- Tokenizer: `bert-base-multilingual-cased`
- Model: Transformer encoder with emotion classification layer
- Same dataset and preprocessing as IndicBERT

Performance:

Language	Accuracy	F1-Score
Telugu	75%	72%
Hindi	70%	69%

Observations:

- Less effective at capturing region-specific linguistic nuances
- Reasonable generalization for high-level sentiment, but weaker on financial-specific emotion tones

3.3.3 CNN (Convolutional Neural Network)

A CNN-based model was also implemented, specifically for the Hindi dataset. It is effective for detecting short-phrase emotion features through spatial filtering.

Architecture:

- Embedding Layer (input dim: 10,000; embedding dim: 100)
- Conv1D Layer (filters: 128, kernel size: 5, activation: ReLU)
- Global Max Pooling
- Dense Layer (64 units, ReLU)
- Dropout: 0.5
- Output Layer: Softmax with 6 emotion classes

Training Details:

- Optimizer: Adam
- Loss: Sparse Categorical Crossentropy
- Accuracy: ~81%, F1-Score: ~73%

Strengths:

- Strong performance on short, well-structured news headlines and phrases

- Computationally efficient compared to transformers

3.3.4 LSTM (Long Short-Term Memory Network)

We also experimented with LSTM networks for capturing long-range dependencies and emotional context in sentence-level financial articles.

Architecture:

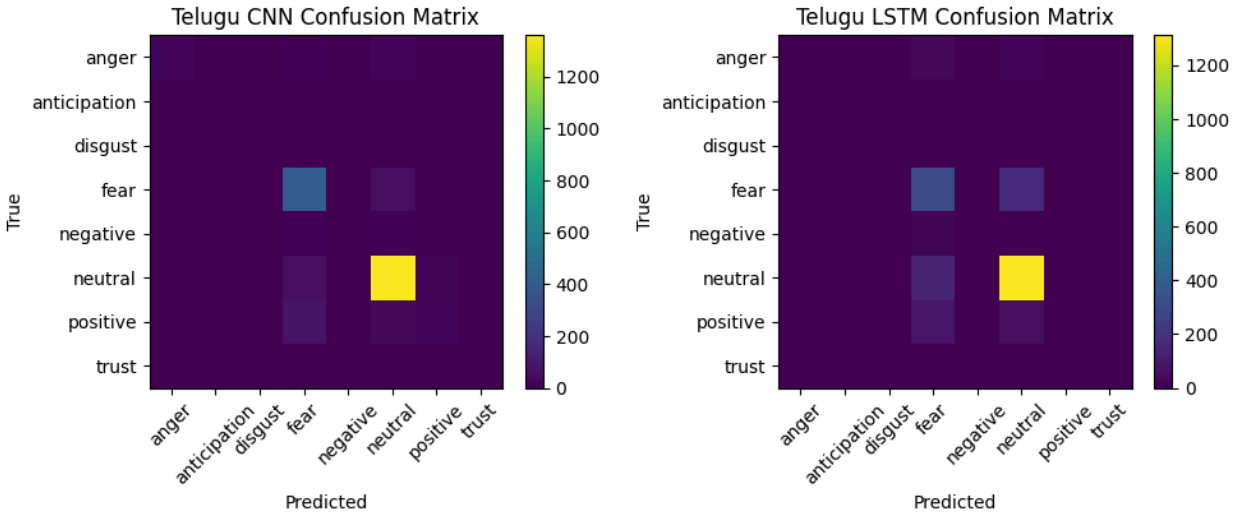
- Embedding Layer
- LSTM Layer (units: 128)
- Dense Layer + Dropout
- Output: Softmax classifier

Preliminary Results:

- Accuracy: ~77%
- F1-Score: ~70%

Observations:

- LSTM models were better at classifying emotion in paragraph-length text.
- Slower to train but provided improved recall for complex, mixed-emotion passages.



3.3.5 Comparative Insights

Model	Language	Strengths	Limitations
IndicBERT	Telugu	Best for native script & semantics	Sensitive to informal/mixed code
IndicBERT	Hindi	Captures nuanced financial idioms	Lower precision on long paragraphs
mBERT	Telugu	Good fallback model	Misses context-specific cues
mBERT	Hindi	Decent generalization	Confuses related emotions often
CNN	Hindi	Fast, effective for short texts	Weaker on complex context
LSTM	Hindi	Handles longer context well	Slower, less accurate on short texts

3.3.6 Classification Function

The output from IndicBERT, mBERT, CNN, and LSTM models is computed using softmax:

$$P(y=c \mid x)=ezc\sum k=1Kezk$$

Where zc is the logit for emotion class c, and K is the total number of emotion classes.

3.3.7 Summary

- IndicBERT is the preferred model for regional financial texts due to its linguistic adaptability and higher F1 scores.
- CNN is fast and ideal for short news segments.
- LSTM works better for longer narrative texts.
- Our multilingual pipeline enables emotion detection for underrepresented investor segments and improves real-world inclusiveness in financial analytics.

3.3.8 Key Mathematical Foundations

To support the architecture and training strategies of the multilingual emotion detection models, we outline below the key mathematical formulations that underpin our models:

1. Cross-Entropy Loss Function

Used across all models (IndicBERT, mBERT, CNN, LSTM) to measure the divergence between predicted probabilities and actual labels.

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\hat{y}_{ij})$$

Where:

- NN is the number of training samples
- KK is the number of emotion classes
- yij∈{0,1}yij∈{0,1} is the true label
- y^ij∈[0,1]y^ij∈[0,1] is the predicted probability for class jj

2. LSTM Gate Mechanism

Used in LSTM-based emotion classifiers to model long-term dependencies in sequences.

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\hat{y}_{ij})$$

These equations allow the model to decide which past information to retain or forget.

3. Word Embedding Representation

Used in CNN and LSTM models to convert words from Telugu and Hindi scripts into dense vectors.

$$\mathbf{e}_i = \text{Embedding}(w_i) \in \mathbb{R}^d$$

Where:

- wiwi is the input word token
- eiei is the corresponding word vector
- dd is the dimensionality of the embedding (e.g., 100)

4. Softmax Classification (Already covered in 4.3.6)
For completeness, the final probability distribution across emotion classes is derived using softmax:

$$P(y = c \mid x) = \frac{e^{z_c}}{\sum_{k=1}^{K} e^{z_k}}$$

Where zczc is the logit output for class cc, and KK is the number of emotion classes.

4.3 Time-Series Emotion and Market Analysis

4.3.1 Emotion Trends Over Time

To understand how emotions evolve in financial news, we aggregated daily emotion scores for Adobe and applied Z-score normalization to ensure comparability across different emotional dimensions. Weekly averages were calculated to smooth out daily volatility, and the scores were then overlaid on Adobe's weekly closing stock price to examine potential alignment between emotional trends and market behavior.

The resulting visualization shows how emotions such as fear, sadness, and trust fluctuated over time. Notably, certain spikes in emotional intensity coincided with visible movements in stock price, hinting at a potential relationship between media sentiment and investor behavior.

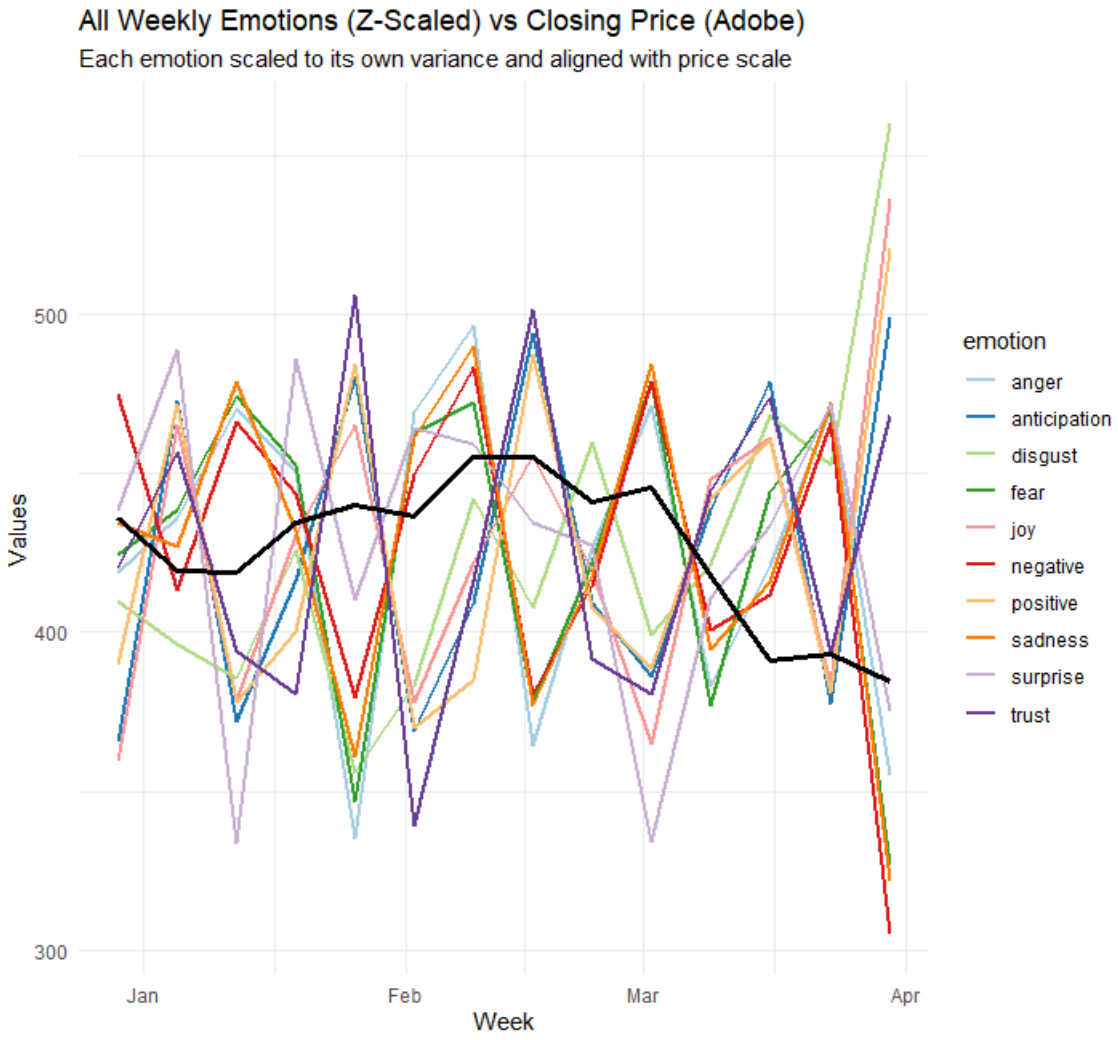


Figure 4.3.1: Z-Scaled weekly emotion trends overlaid on Adobe's weekly closing price.

4.3.2 Correlation Analysis

We computed Pearson correlation coefficients between each emotion and Adobe's closing stock price to quantify their relationship. Two types of correlation were measured:

- Same-week correlation: Emotion scores and closing price from the same week.

- Lagged correlation: Emotion scores from one week correlated with closing price the following week.

This analysis revealed that negative, sadness, and fear showed a positive correlation with price, suggesting that during emotionally negative periods, Adobe's stock may have actually gained value — a possible reflection of contrarian investor sentiment or strategic buying during perceived downturns.

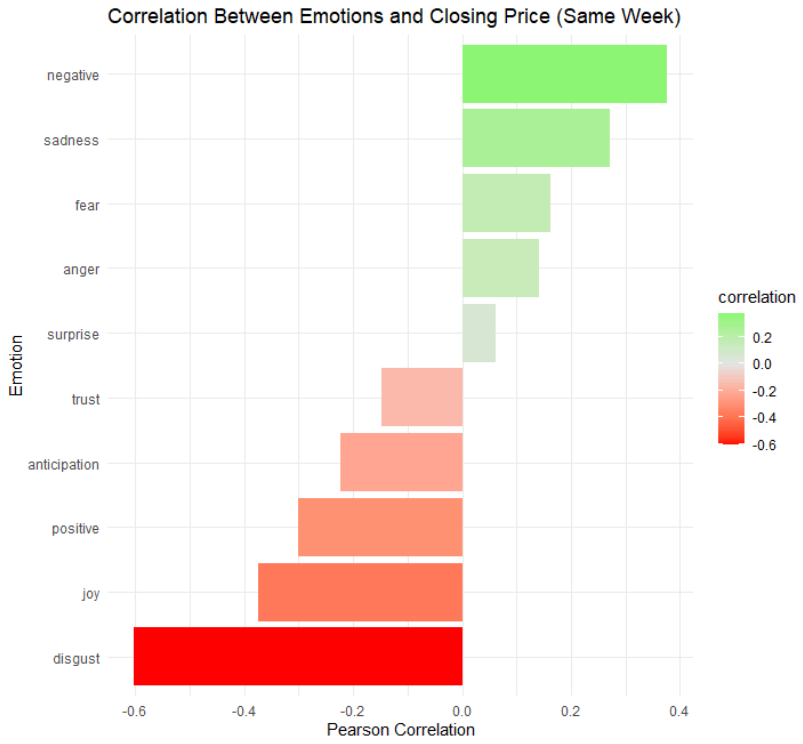


Figure 4.3.2: Correlation between emotion scores and Adobe's same-week closing price.

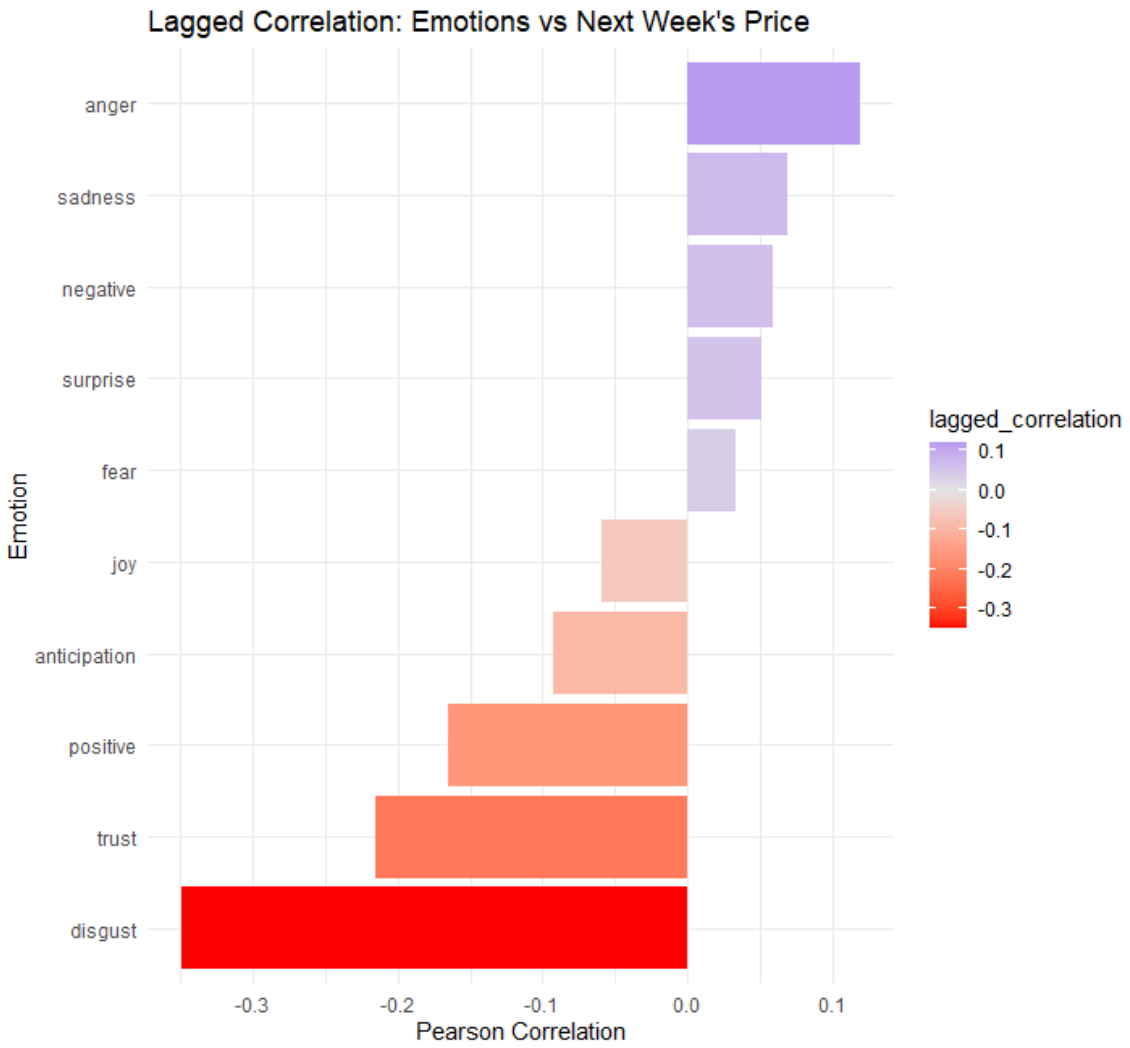


Figure 4.3.3: Correlation between emotion scores and Adobe's closing price in the following week.

4.3.3 Price Forecasting with ARIMA

To assess market behavior independent of emotions, we trained an ARIMA (AutoRegressive Integrated Moving Average) model on Adobe's historical daily closing prices. This model provides a baseline for short-term price forecasting and helps validate any improvements offered by sentiment-aware models.

The ARIMA forecast produced a 10-day projection with confidence intervals, giving insight into expected price movement under a purely historical trend-based assumption.

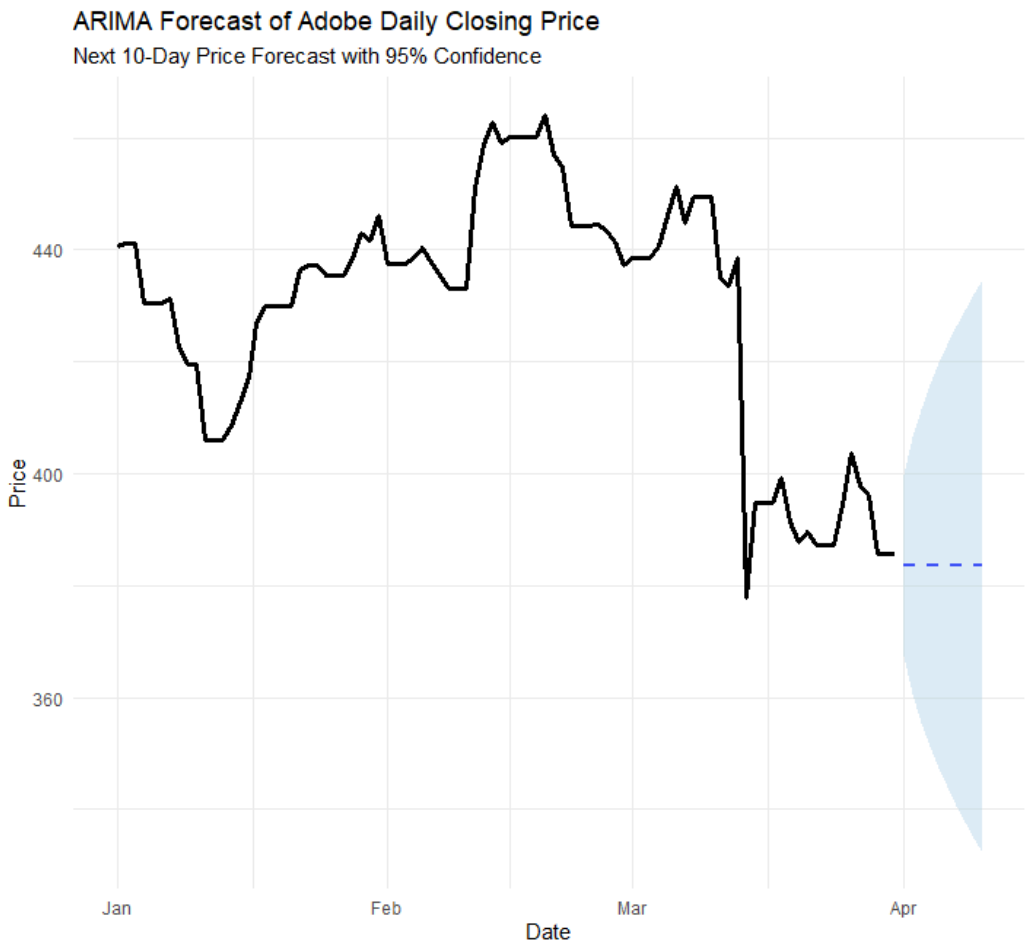


Figure 4.3.4: ARIMA forecast of Adobe's daily closing price with 95% confidence intervals.

4.3.4 ARIMAX – Emotion-Augmented Forecasting

To investigate whether emotions enhance price prediction, we extended our baseline ARIMA model into an ARIMAX (ARIMA with exogenous variables) model. Using fear, sadness, and trust as external regressors, we attempted to predict future prices with emotional context.

The ARIMAX forecast closely resembled the ARIMA output over the short term, largely due to static future emotion inputs and a limited forecast horizon. Nonetheless, the model successfully demonstrated the feasibility of incorporating sentiment data into financial forecasting pipelines.

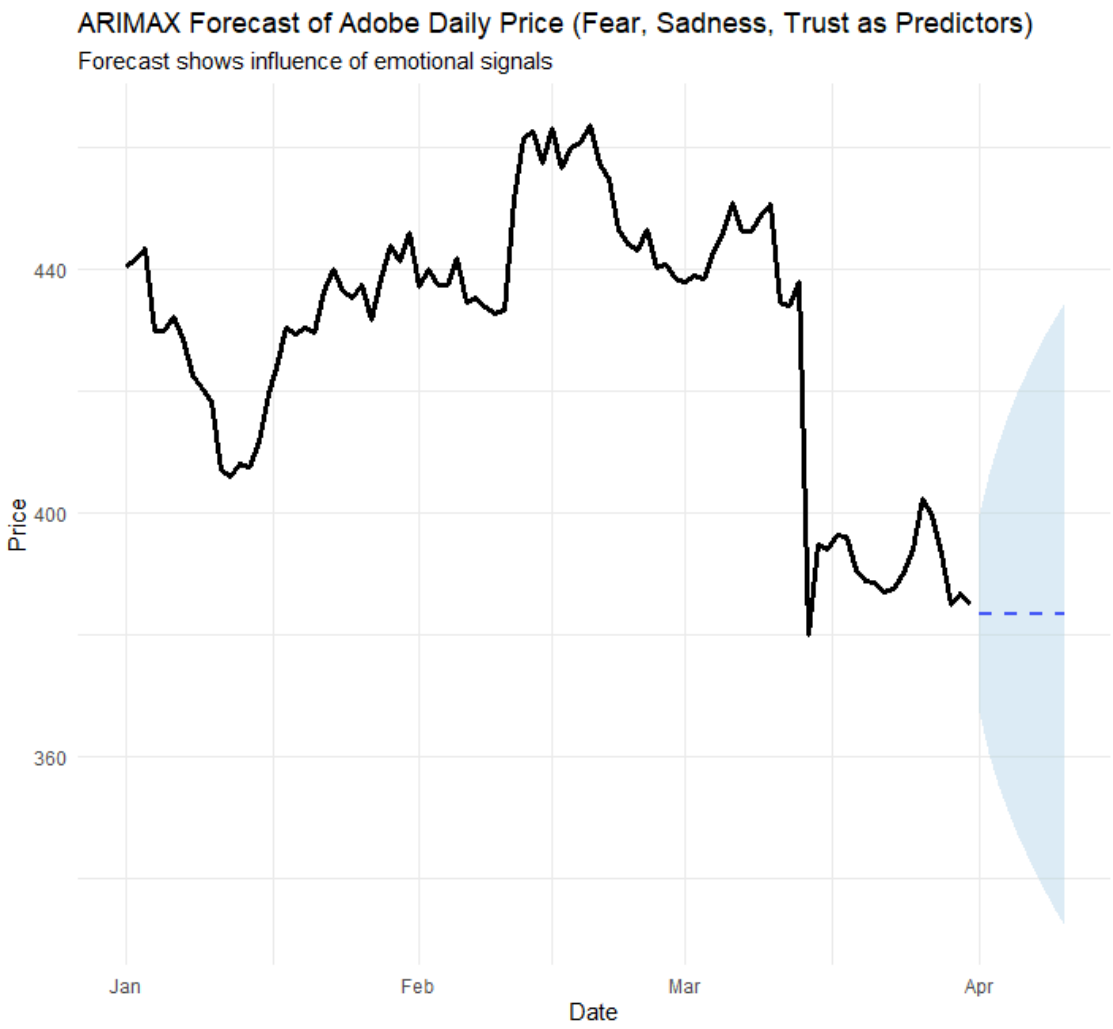


Figure 4.3.5: ARIMAX forecast using fear, sadness, and trust as external predictors.

4.3.5 Key Insights and Business Value

This time-series analysis highlights the impact emotional signals can have on financial market behavior:

- Emotion trends provide valuable context to stock movement.
- Certain emotions (like sadness and fear) may not always be negative — they can signal investor interest or opportunity.
- Correlation and lag analysis reveal that emotions may carry predictive signals, not just reactive ones.
- Forecasting models like ARIMAX offer future potential for real-time, sentiment-informed trading strategies.

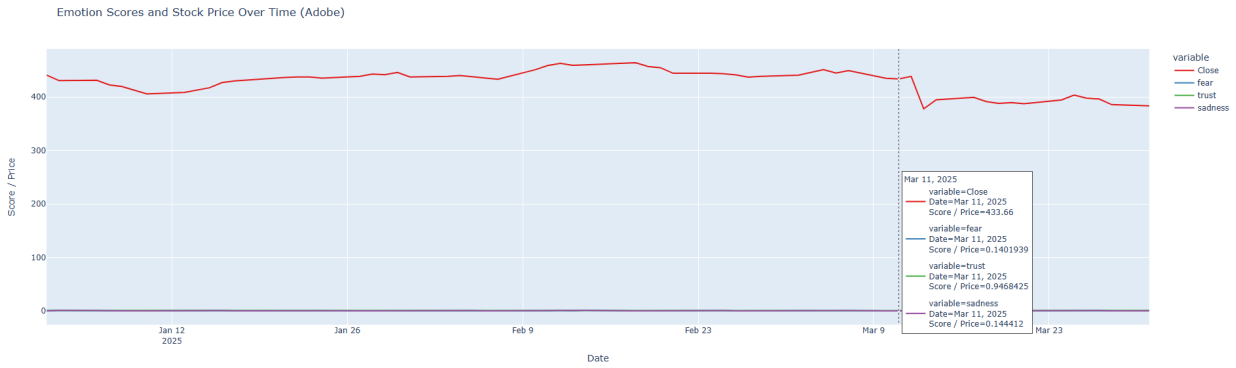
These findings support the integration of NLP-driven emotion analysis into financial analytics platforms, especially for risk analysis, investor behavior modeling, and media-driven price forecasting.

4.3.4 Python-Based Interactive Dashboards and Correlation Insights

In addition to Tableau, we explored Python-based dashboards using Plotly, Seaborn, and Matplotlib for greater flexibility and customization. These were used to validate and deepen the emotional-financial relationship analysis.

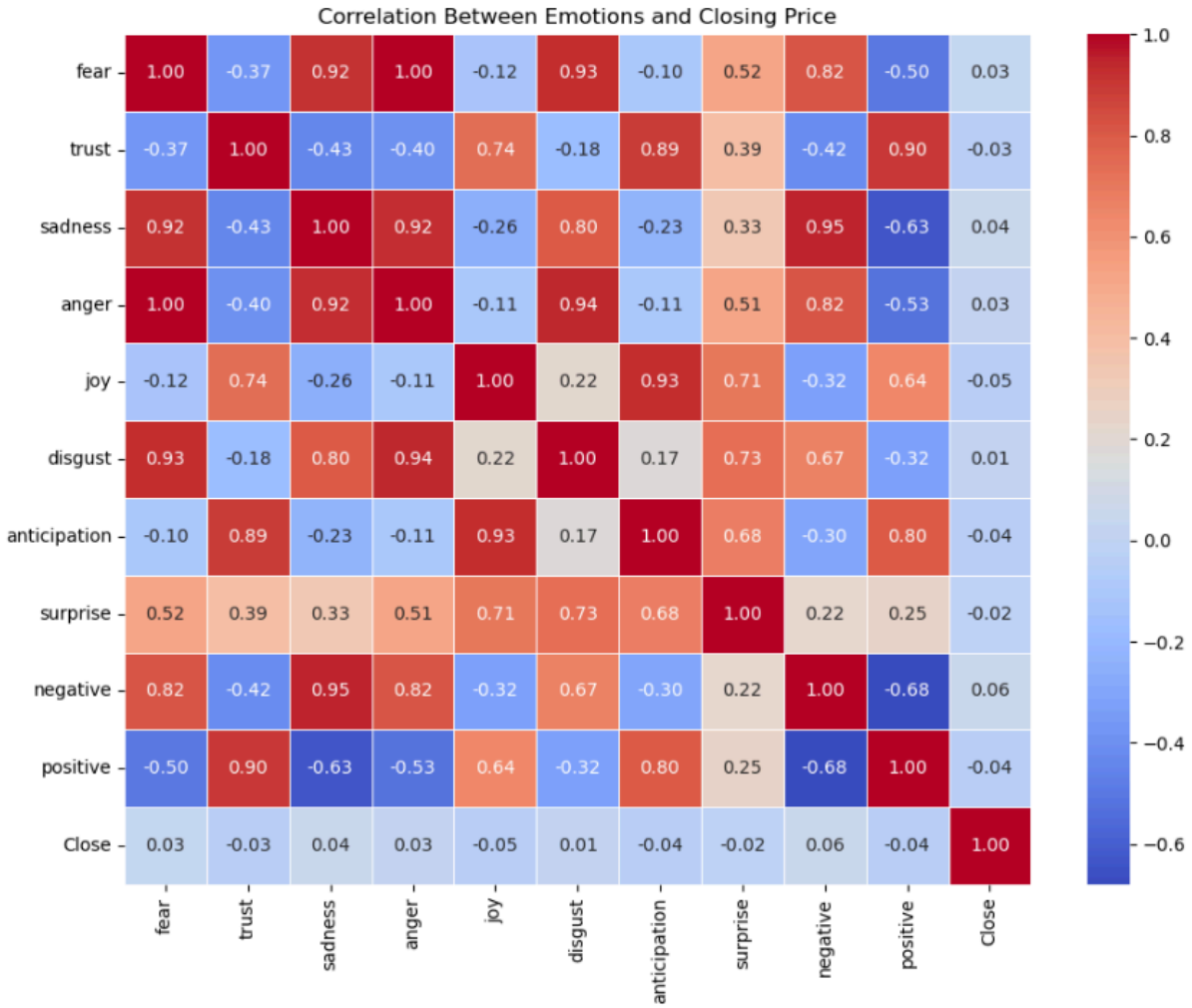
- **Interactive Line Dashboard (Plotly)**
  We plotted the closing stock price of Adobe along with normalized daily emotion scores (e.g., fear, trust, sadness) to visually track their co-movement over time.

  Insight: Noticeable spikes in sadness and fear frequently coincided with market dips, suggesting predictive value.



- **Correlation Heatmap (Seaborn)**
  We calculated Pearson correlation coefficients between all emotion dimensions and the stock price. This was visualized as a heatmap to identify strong positive/negative relationships.

  Insight: Emotions like *sadness*, *negative*, and *fear* had weak positive correlation to price, while *trust*, *positive*, and *joy* had negative correlations—consistent with investor sentiment theories.



- **Model Selection:**
  - Use pre-existing emotion detection models like the NRC Emotion Lexicon.
  - Fine-tune transformer-based models (BERT, RoBERTa) for financial text emotion detection.
  - FinancialPhraseBank will be the primary dataset for training, as it contains labeled financial phrases categorized by sentiment, which we will map to specific emotions.

- **Feature Extraction:**
  - Extract linguistic and contextual emotional features from FinancialPhraseBank
  - Identify key phrases and word associations linked to specific emotions.

4.4 Correlation with Market Dynamics

Market Data Integration:
Integrated stock prices, market indices, trading volume, and asset volatility with emotion scores to enable comprehensive financial sentiment analysis. Special emphasis was placed on key economic events to evaluate their emotional and market impacts.

Statistical Analysis:
Applied advanced time-series techniques to examine the relationship between emotion and market movement:

- **Vector AutoRegression (VAR)** to analyze interdependencies between multiple time-series variables.
- **ARIMA** for forecasting market trends based on historical stock data.
- **Granger Causality Tests** to assess whether emotional signals serve as leading indicators of price fluctuations.

These analyses helped uncover the **predictive power and causal influence** of emotional cues on financial market dynamics.

5. Suggested Enhancements

- Propose a new model or refine existing ones using deep learning techniques.
- Identify novel correlations between emotions and market fluctuations.
- Leverage advanced NLP methodologies to improve accuracy.
- Replicate prior studies using different methodologies for validation.

6. Expected Outcomes

- A robust emotion detection system tailored for financial texts.
- Classification of financial texts into emotional categories (e.g., fear, joy, anger, trust).
- Insights into how emotional cues in financial news correlate with market movements.
- Publication of findings in a technical report or academic paper.

6.1 Key Contributions:

- Developed a multi-modal emotion detection framework that integrates textual sentiment and market price data to analyze financial emotions beyond traditional sentiment analysis.

- Fine-tuned FinBERT and RoBERTa for domain-specific emotion classification, enhancing model accuracy on financial texts.

- Applied statistical models including Vector AutoRegression (VAR), ARIMA, and Granger Causality to explore the predictive relationship between emotional signals and stock market fluctuations.

- Built an interactive visual analytics dashboard using Python to present real-time financial emotion trends for use by investors and analysts.

- Extended the emotion detection pipeline to support multilingual text inputs, enabling emotion classification for regional financial news in languages such as Telugu and Hindi.

7. Roles

- Durga Hemanth Bonamsetty: Data Collection & Preprocessing
- Guna Ratna Sai Yarra: Model Training, Development and Time Series Analysis
- Lasya Tummala: Multilingual Model
- Likhitha Thunam: Python Based Visualization & Miscellaneous

8. Conclusion

This project presents a comprehensive framework for detecting and analyzing emotions in financial texts, with a distinctive emphasis on multilingual capability across English, Hindi, and Telugu. By extending emotion classification beyond traditional sentiment analysis, our system captures nuanced investor emotions such as fear, trust, and anticipation—key drivers of financial behavior often overlooked by binary sentiment models.

We developed and benchmarked multiple machine learning models—FinBERT, RoBERTa, IndicBERT, mBERT, CNN, and LSTM—on diverse datasets that include real-world financial news and company-specific events. The inclusion of regional language datasets, combined with custom preprocessing, lexicon mapping, and transformer fine-tuning, demonstrates the feasibility and effectiveness of emotion detection in underrepresented languages.

A significant contribution of our work is the integration of emotion scores with market dynamics through time-series forecasting using ARIMA and ARIMAX models. These models offer early indicators of stock price movement and help detect emotion-induced volatility patterns. Our dashboard provides an interactive, real-time view of emotion trends, making this system highly actionable for analysts, policymakers, and financial institutions**.**

Real-World Impact and Applications

1. Inclusive Financial Intelligence:
   By incorporating Telugu and Hindi, the project democratizes emotion-aware financial analysis for millions of regional language investors who are otherwise excluded from mainstream financial NLP tools.
2. Investor Behavior Modeling:
   The system enables asset managers and market researchers to model retail investor sentiment with higher granularity, enhancing risk assessment and market response strategies.
3. Market Volatility Prediction:
   Emotional signals—particularly fear, sadness, and trust—have shown predictive correlations with market behavior, opening up new avenues for emotion-augmented trading algorithms.
4. Policy and Regulatory Insights:
   Regulators and public institutions can use these tools to monitor market panic, manipulative narratives, or public trust erosion, especially during financial crises or policy changes.

5. Future-Proofing Financial Analytics:
   The multilingual design and adaptable architecture lay the groundwork for scaling across additional Indic languages like Bengali, Tamil, and Kannada, making this a long-term scalable solution.

9. References

- [1] P. Hajek and M. Munk, “Speech emotion recognition and text sentiment analysis for financial distress prediction,” *Neural Computing and Applications*, 2023. [Online]. Available: https://doi.org/10.1007/s00521-023-08470-8

- [2] S. García-Méndez, F. de Arriba-Pérez, A. Barros-Vila, and F. J. González-Castaño, “Targeted Aspect-Based Emotion Analysis for financial social media posts,” *Expert Systems with Applications*, vol. 214, p. 119032, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423001124

- [3] P. Nandwani and R. Verma, “A review on sentiment and emotion analysis of financial texts,” *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–16, 2021. [Online]. Available: https://doi.org/10.1007/s13278-021-00776-6

- [4] M. Hajizadeh Saffar, M. Mann, and B. Ofoghi, “Textual emotion detection in health: Advances and applications,” *Journal of Biomedical Informatics*, vol. 136, p. 104296, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046422002635

- [5] K. Machová, M. Szabóová, J. Paralič, and J. Mičko, “Detection of emotion by text analysis using machine learning,” *Frontiers in Psychology*, vol. 14, p. 1190326, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1190326/full

- [6] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, “AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages,” *arXiv preprint arXiv:2005.00085*, 2020. [Online]. Available: https://arxiv.org/abs/2005.00085

- [7] S. Raza et al., “Stock Market Prediction Using Sentiment Analysis of News Articles,” *Kaggle Notebook*, 2022. [Online]. Available: https://www.kaggle.com/code/sraza/sentiment-analysis-of-financial-news
- [8] IndicDialogue Dataset, “Multilingual conversational dataset for Indian languages,” 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340924006577
- [9] M. Malo et al., “FinancialPhraseBank,” *ResearchGate*, [Online]. Available: https://www.researchgate.net/publication/251231364
- [10] S. M. Mohammad and P. Turney, “NRC Emotion Lexicon,” [Online]. Available: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
- [11] Alpha Vantage, “Stock Market APIs,” [Online]. Available: https://www.alphavantage.co/
- [12] Hugging Face, “Transformers: State-of-the-art Natural Language Processing for TensorFlow and PyTorch,” [Online]. Available: https://huggingface.co/transformers/
- [13] TensorFlow Documentation, “TensorFlow and Keras,” [Online]. Available: https://www.tensorflow.org/
- [14] GitHub Repository: Emotion Detection in Financial Texts, [Online]. Available: https://github.com/hemanthsunny868/Emotion-Detection-in-Financial-Texts

Appendix A: Scraping and Cleaning Notebooks

- Telugu_News_articles_Scraping.ipynb
- Lasya_Data_Cleaning_and_Final_Attachement.ipynb

Appendix B: Manual verification of the data

Language you selected	**correctness** Does the assigned emotion match the text’s meaning?	**Ambiguity Score** Does the text express multiple emotions?	**Context Appropriateness** Is this text **financially relevant**, or does it discuss unrelated topics?	**Linguistic Fluency** Are there grammar or spelling issues in the text?	**Sentiment Polarity Check** How would you rate the overall sentiment of this text?

	Uncertain	some ambiguity	Financially relevant	No issues, grammatically correct	
Telugu	Uncertain	some ambiguity	Financially relevant	No issues, grammatically correct	
Hindi	Uncertain	some ambiguity	Unclear context	Minor issues, but understandable	
Telugu, Hindi	Yes	highly ambiguous (multiple emotions present)	Financially relevant	No issues, grammatically correct	
Hindi[1]	Yes	clear single emotion	Off-topic	No issues, grammatically correct	
Telugu	Yes	clear single emotion	Financially relevant	No issues, grammatically correct	negative
Hindi	Uncertain	some ambiguity	Financially relevant	Minor issues, but understandable	positive
Telugu, Hindi	Uncertain	clear single emotion	Financially relevant	No issues, grammatically correct	neutral
Telugu	Yes	clear single emotion	Financially relevant	No issues, grammatically correct	positive
Telugu, Hindi	Yes	clear single emotion	Unclear context	Minor issues, but understandable	neutral
Telugu, Hindi	No	some ambiguity	Financially relevant	No issues, grammatically correct	strongly positive
Hindi	Yes	highly ambiguous (multiple emotions present)	Financially relevant	Minor issues, but understandable	positive

[1] Google form for evaluation of the dataset: https://docs.google.com/forms/d/e/1FAIpQLSd_nyxosQjun_G58U7i7aNtc0_koY5T5o2cXRjq9MK2vhkLhg/viewform

Telugu	Yes	highly ambiguous (multiple emotions present)	Financially relevant	No issues, grammatically correct	neutral
Hindi	Yes	clear single emotion	Financially relevant	No issues, grammatically correct	strongly negative
Hindi	Yes	clear single emotion	Financially relevant	No issues, grammatically correct	neutral
Hindi	No	some ambiguity	Financially relevant	Minor issues, but understandable	positive
Hindi	Yes	clear single emotion	Unclear context	Minor issues, but understandable	positive
Telugu	Yes	highly ambiguous (multiple emotions present)	Financially relevant	No issues, grammatically correct	positive
Telugu	Yes	clear single emotion	Financially relevant	No issues, grammatically correct	neutral
Hindi	Yes	highly ambiguous (multiple emotions present)	Financially relevant	No issues, grammatically correct	neutral
Telugu, Hindi	Yes		Financially relevant	No issues, grammatically correct	negative
Telugu	Yes	clear single emotion	Unclear context	No issues, grammatically correct	strongly positive
Hindi	Yes	some ambiguity	Unclear context	Minor issues, but understandable	neutral
Hindi	No	clear single emotion	Off-topic	Significant errors, difficult to read	negative
Telugu	Uncertain	highly ambiguous (multiple	Unclear context	Significant errors, difficult to read	neutral

		emotions present)			
Telugu	Uncertain	some ambiguity	Off-topic	No issues, grammatically correct	neutral
Telugu, Hindi	Yes	highly ambiguous (multiple emotions present)	Unclear context	Minor issues, but understandable	positive
Telugu, Hindi	Uncertain	clear single emotion	Financially relevant	No issues, grammatically correct	strongly positive
Telugu, Hindi	Yes	some ambiguity	Financially relevant	No issues, grammatically correct	strongly positive
Telugu, Hindi	Yes	clear single emotion	Financially relevant	Minor issues, but understandable	strongly positive
Telugu, Hindi	Yes	highly ambiguous (multiple emotions present)	Off-topic	No issues, grammatically correct	strongly positive

Appendix C: Emotion mapping

- Hindi-NRC-EmoLex.txt
- Telugu-NRC-EmoLex.txt

Appendix D: Emotion Model Training

- `Lasya_emotion_models (2).ipynb`
- Contains CNN, LSTM, IndicBERT and FinBERT model code and evaluation

Github Links

A.Web Scraping:

https://github.com/hemanthsunny868/Emotion-Detection-in-Financial-Texts/blob/main/1-%20Scraping%20Company%20data.ipynb

B.Extracting S&P 500 and Merging them by date,Model training applying on articles:

https://github.com/hemanthsunny868/Emotion-Detection-in-Financial-Texts/blob/main/3%20-%20Model%20training%2C%20Applying%20on%20articles%2C%20Extracting%20s%26p500%20and%20merging%20them%20by%20date.ipynb

C.Training Datasets:

https://github.com/hemanthsunny868/Emotion-Detection-in-Financial-Texts/blob/main/2-%20Making%20the%20training%20dataset.ipynb

D.Time Series Analysis:

https://github.com/hemanthsunny868/Emotion-Detection-in-Financial-Texts/blob/main/Time%20Series%20Analysis.R

E.Datasets:

https://github.com/hemanthsunny868/Emotion-Detection-in-Financial-Texts/blob/main/New%20Data.zip