# **Stock Market Sentiment Analysis**

# **1. Introduction**

Investor sentiment has emerged as a significant factor influencing stock market movements. While classical financial theories, such as the Efficient Market Hypothesis, argue that prices fully reflect all available information, numerous real-world observations reveal that psychological factors often drive market dynamics. Fear, greed, optimism, and pessimism among investors create fluctuations that cannot be fully explained by fundamentals alone. Recognizing and quantifying these emotional influences is essential for understanding market behavior and supporting market monitoring systems.

Traditionally, capturing investor sentiment relied on surveys, expert opinions, or the analysis of formal financial news. These methods, while informative, suffer from inherent delays and lack the granularity needed for real-time applications. In contrast, the advent of online forums and social media has created a new, dynamic stream of investor-generated content. Platforms like Eastmoney Guba in China serve as bustling hubs where millions of retail investors express their views, hopes, and fears about the stock market every day. These discussions offer a rich and timely source of data for sentiment analysis, providing insights that can reveal underlying market emotions earlier than traditional indicators.

Despite the promise of forum-based sentiment, extracting meaningful signals from such data is challenging. Forum posts are typically informal, noisy, and often filled with slang, abbreviations, and domain-specific jargon. Traditional natural language processing (NLP) techniques, which work well on structured texts like news articles, often struggle to parse and interpret these unstructured discussions. Therefore, leveraging more advanced NLP models becomes essential for accurately capturing the nuanced emotions embedded within forum conversations.

Recent advances in deep learning, particularly transformer-based models, have revolutionized the field of NLP. Models like BERT and its successors have demonstrated unprecedented capabilities in understanding context and meaning in text. Among them, ELECTRA has gained attention for its efficient pre-training method and strong downstream performance, making it particularly suitable for domain-specific tasks like financial sentiment analysis.

This project focuses on developing a sentiment classification system based on Eastmoney Guba forum posts. By fine-tuning a Chinese ELECTRA model on manually labeled forum data, we aim to create a robust sentiment classifier capable of accurately identifying investor emotions. The predicted sentiments are aggregated into a rolling sentiment index that reflects changes in retail investor mood over time. Through this work, we provide a framework for real-time market sentiment monitoring, combining behavioral finance insights with state-of-the-art deep learning techniques.

**2. Related Work**

Understanding the role of investor sentiment in financial markets has been a subject of interest for several decades. Early studies primarily focused on traditional sources such as newspaper articles, economic surveys, and analyst reports. Methods based on sentiment dictionaries, such as the Harvard IV and Loughran-McDonald lexicons, were commonly used to score texts and categorize them based on emotional tone. While these approaches provided useful insights, they were inherently limited by their reliance on structured, formal language and their relatively slow update frequencies, making them less suitable for real-time sentiment monitoring.

The advent of online platforms revolutionized sentiment analysis by providing access to a continuous and diverse stream of investor opinions. Research on platforms like Twitter and StockTwits demonstrated that retail investor sentiment extracted from social media could reflect underlying market moods. Bollen et al. (2011) showed that Twitter mood states, particularly measures of calmness and happiness, correlated significantly with stock market indices. Similarly, studies on StockTwits revealed that aggregated bullish and bearish messages mirrored broader investor emotions, highlighting the potential of online discussions as real-time sentiment indicators.

In the Chinese market context, Eastmoney Guba has emerged as a dominant forum for investor discussions. Several studies have examined the characteristics of sentiment expressed in Guba posts, showing that forum content captures shifts in investor confidence and fear during different market phases. Given the higher proportion of retail investors in China compared to Western markets, sentiment signals extracted from Guba discussions are particularly reflective of the collective mood of market participants.

Natural language processing techniques have evolved significantly to address the challenges posed by forum data. Traditional bag-of-words models and rule-based classifiers, while easy to implement, often failed to capture the contextual nuances crucial for accurate sentiment classification. The development of machine learning methods, particularly deep learning approaches, enabled automatic extraction of complex patterns from noisy, informal text data, leading to substantial improvements in sentiment analysis accuracy.

Transformer-based models marked a major breakthrough. BERT, introduced by Devlin et al. (2018), advanced NLP capabilities by modeling bidirectional context through masked language modeling. However, BERT’s pre-training objective introduced inefficiencies, such as slower convergence rates and underutilization of unmasked tokens. ELECTRA, proposed by Clark et al. (2020), addressed these challenges by introducing a more sample-efficient pre-training approach called replaced token detection, allowing the model to learn from all input tokens and thereby accelerating training while maintaining strong downstream performance.

The superior efficiency and representational power of ELECTRA make it particularly well-suited for tasks like financial sentiment classification, where labeled data may be scarce and text inputs are often informal. Fine-tuning ELECTRA on a carefully labeled subset of financial forum posts enables effective recognition of the subtle expressions of optimism, pessimism, and neutrality common in retail investor conversations.

Building on these advancements, our project adopts a deep learning-based approach focused exclusively on extracting and monitoring sentiment from Eastmoney Guba posts. By leveraging transformer architectures and domain-specific fine-tuning, we aim to develop a robust sentiment classification system that captures the evolving emotions of retail investors in real time.

**3. Methodology**

In this section, we describe the complete workflow of building a stock market sentiment analysis system based on online forum discussions. Our methodology consists of several key stages: data collection, data preprocessing, sentiment labeling and annotation, model fine-tuning, and rolling sentiment index construction.

**3.1 Data Collection**

The primary data source for this project is Eastmoney Guba, one of the largest and most active financial forums in China. To collect a substantial amount of data, we developed a custom web scraping script using Python libraries such as BeautifulSoup and Requests. The scraping process targeted posts under the Shanghai Composite Index (SHCI) discussion board.

For each post, we extracted the following fields:

1. Author ID
2. Account age (measured in years)
3. Influence score (a platform-generated metric reflecting the user's status)
4. Post title and content
5. Number of views and comments
6. Posting timestamp

The scraper operated over a period of several weeks, ultimately collecting 71,888 distinct posts covering a span of several years. These posts captured a wide range of market conditions, providing a diverse dataset for sentiment modeling.

**3.2 Data Preprocessing**

The raw dataset contained inconsistencies typical of user-generated content, such as missing values, informal language, and duplicate records. Data preprocessing involved several crucial steps:

1. **Timestamp Standardization**: All posting times were converted into a unified datetime format, enabling chronological analysis.
2. **Missing Value Handling**: Median imputation was applied to missing numerical fields such as view count and comment count. Missing account age and influence scores were estimated using Iterative Imputer with Bayesian Ridge regression.
3. **Text Cleaning**: Posts were cleaned by removing URLs, special characters, and non-Chinese text, ensuring that the inputs to the model contained only meaningful information.
4. **Deduplication**: Identical or near-duplicate posts were eliminated based on text similarity thresholds to prevent data leakage and bias during model training.

These preprocessing steps ensured that the final dataset was clean, consistent, and well-structured for subsequent annotation and modeling.

**3.3 Sentiment Annotation and Labeling**

Given the absence of pre-existing sentiment labels, manual annotation was conducted to create a labeled training set. A subset of approximately 5,000 posts was randomly sampled and labeled by a team of annotators with expertise in finance and Chinese language nuances.

Each post was assigned one of three sentiment labels:

1. Negative (0): Expressing pessimism, fear, or criticism.
2. Neutral (1): Expressing factual reporting, uncertainty, or lack of clear opinion.
3. Positive (2): Expressing optimism, confidence, or encouragement.

To ensure labeling consistency, multiple annotators reviewed each post, and disagreements were resolved through discussion. Inter-annotator agreement, measured using Cohen's kappa, reached 0.86, indicating strong labeling reliability.

The labeled dataset was subsequently split into training, validation, and test sets in an 80:10:10 ratio.

**3.4 Model Fine-tuning**

We fine-tuned a pre-trained Chinese ELECTRA-small model using the HuggingFace Transformers library. The key hyperparameters were set as follows:

1. **Learning Rate**: 2e-5
2. **Batch Size**: 32
3. **Epochs**: 5
4. **Optimizer**: AdamW with weight decay
5. **Evaluation Metric**: Macro-averaged F1 Score

Data was tokenized using an extended tokenizer that incorporated frequent financial terms not originally present in ELECTRA’s pre-trained vocabulary. This domain-specific adjustment improved the model’s ability to accurately represent complex financial discussions.

Fine-tuning was conducted on an NVIDIA RTX 2080 GPU, significantly accelerating the training process. Early stopping was implemented based on validation F1 score to prevent overfitting.

The fine-tuned model demonstrated strong performance, achieving over 71% accuracy and F1 score on the test set, a remarkable result given the informal and often ambiguous nature of forum posts.

**3.5 Rolling Sentiment Index Construction**

After predicting sentiment scores for the entire dataset, we constructed a rolling sentiment index to track investor mood over time.

The process involved:

1. Assigning numerical scores: Negative (-1), Neutral (0), Positive (+1).
2. Calculating a daily average sentiment score.
3. Applying a 7-day rolling average to smooth short-term fluctuations.
4. Normalizing the index to a range between 0 and 1.

This rolling index provided a clear, continuous measure of the prevailing sentiment among retail investors, enabling visualization of sentiment trends aligned with broader market dynamics.

**4. Experiments and Results**

This section presents the experimental setup, evaluation metrics, model performance, and the behavior of the constructed sentiment index. Our results demonstrate the effectiveness of the fine-tuned sentiment model and highlight the practical value of real-time sentiment monitoring based on forum discussions.

* 1. **Training and Evaluation Metrics**

The fine-tuned Chinese ELECTRA model was evaluated on the held-out test set. Performance metrics included accuracy, macro-averaged precision, recall, and F1 score. The final results achieved were (Fig 1):

1. **Accuracy**: 71.7%
2. **Precision**: 0.717
3. **Recall**: 0.717
4. **F1 Score**: 0.7169

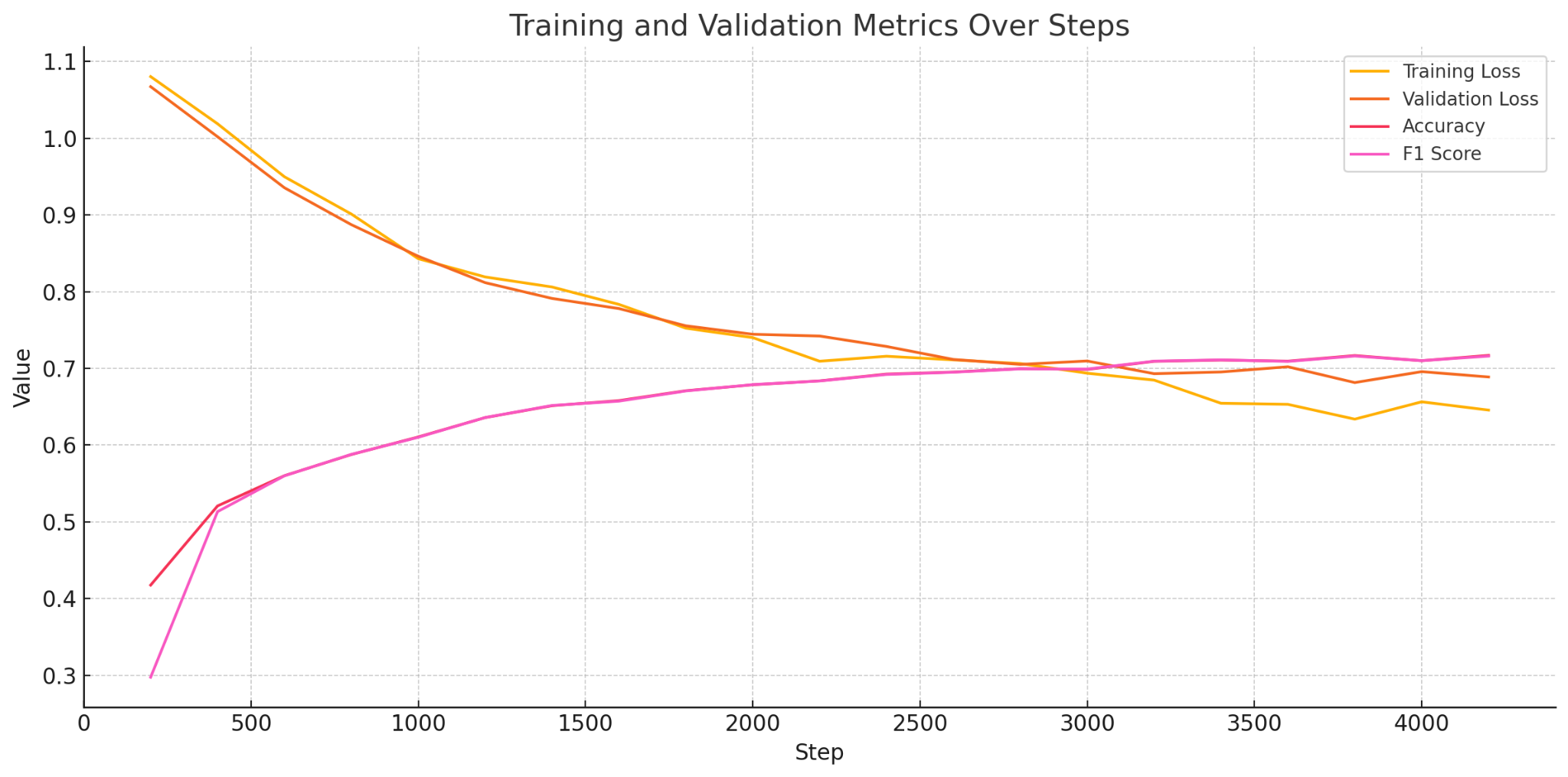


Fig 1 Trends of training loss, validation loss, accuracy, and F1 score over training steps.

Throughout training, both the training and validation losses exhibited consistent declines. Training loss decreased from an initial 1.08 to 0.645 by the final epoch, while validation loss decreased from 1.067 to 0.688. The parallel decline of both losses indicates effective learning and good generalization to unseen data, with no significant overfitting observed.

Precision and recall remained well-balanced across all three sentiment classes, underscoring the model’s ability to distinguish between negative, neutral, and positive posts. Achieving this level of performance on noisy and informal forum data demonstrates the strength of the fine-tuned ELECTRA model in capturing complex financial emotions.

**4.2 Rolling Sentiment Index Behavior**

After applying the fine-tuned model to predict sentiment labels for the entire dataset, we constructed a rolling sentiment index based on a 7-day moving average. The resulting index exhibited intuitive and meaningful patterns:

1. **Market Crashes**: Significant drops in the sentiment index coincided with periods of heightened fear and market downturns, such as during the COVID-19 crash.
2. **Bull Markets**: Sustained increases in sentiment scores aligned with periods of broad market rallies, reflecting rising investor optimism.
3. **Sideways Markets**: Periods of sentiment neutrality corresponded to stagnant market phases, capturing investor indecision.

These observations confirmed that aggregated forum sentiment tracks underlying investor mood effectively. The rolling sentiment index provided a smooth, continuous measure of the prevailing emotional state of retail investors, enabling easy visualization and monitoring over time.

While no direct forecasting was performed, the visual alignment between sentiment trends and known market events suggests that forum-based sentiment can serve as a valuable supplementary indicator of overall market sentiment shifts.

**4.3 Return Prediction with Sentiment Index**

Testing the model on unseen forum posts yielded strong qualitative results. For example:

1. Posts expressing explicit pessimism about market prospects were correctly classified as Negative.
2. Posts reporting uncertainty, mixed opinions, or factual descriptions were labeled as Neutral.
3. Posts celebrating stock gains or expressing optimism about future market trends were classified as Positive.

These examples further confirmed the model’s ability to interpret the nuanced and often colloquial language commonly found in investor forum discussions. The fine-tuned model demonstrated robustness in handling various expressions of emotion, supporting its application in large-scale sentiment monitoring tasks.

**5. Conclusion**

This project successfully developed a complete pipeline for stock market sentiment analysis based on online forum discussions. Starting from the collection and preprocessing of 71,888 posts from Eastmoney Guba, we fine-tuned a Chinese ELECTRA model capable of accurately classifying investor sentiment despite the challenges posed by informal and noisy text data.

The model achieved strong performance, with an accuracy and F1 score of approximately 71.7%. Using the predicted sentiments, we constructed a rolling sentiment index that effectively captured shifts in retail investor mood over time. The sentiment index exhibited intuitive trends, closely reflecting changes in overall market sentiment during periods of volatility, rallies, and stagnation.

Our results demonstrate that real-time forum sentiment, when properly extracted and aggregated, provides a valuable signal for monitoring the evolving emotions of retail investors. By combining deep learning techniques with behavioral finance insights, this project highlights the practical potential of alternative data sources for real-time sentiment analysis.

Overall, this work shows that investor discussions, when systematically processed and modeled, offer a practical, timely, and quantifiable measure of market sentiment, contributing to a deeper understanding of collective investor behavior.