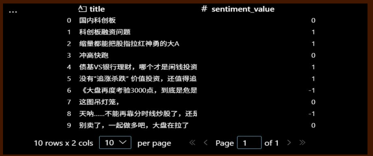
Stock Market Sentiment Analysis

Problem & Motivation

Investor sentiment is known to significantly influence stock market movements, particularly during periods of volatility.

With the growing volume of discussions on Chinese financial forums, there is a valuable opportunity to systematically quantify investor emotions from online posts.

This project aims to construct a structured sentiment index based on forum data and integrate it into a predictive model for next-day returns of the Shanghai Composite Index, bridging behavioral insights with quantitative forecasting.



```
...        title                                  # sentiment_value
        0  国内科创板                                          0
        1  科创板融资问题                                      1
        2  缩量都能把股指拉红神勇的大A                         1
        3  冲高快跑                                            0
        4  债基VS银行理财，哪个才是闲钱投资                    1
        5  没有“追涨杀跌” 价值投资，还值得追                   1
        6  《大盘再度考验3000点，到底是危是                   -1
        7  这图吊灯笼，                                        0
        8  天呐......不能再靠分时线炒股了，还是               -1
        9  别卖了，一起做多吧，大盘在拉了                      0
   10 rows x 2 cols   10 ∨   per page     « ‹ Page 1 of 1 › »
```

Dataset Description

Data Collection:

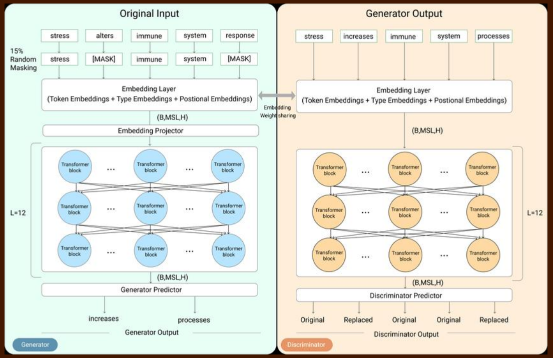
Collected from Eastmoney Guba forums using a custom scraper.

Contents: Post metadata:

1. Author
2. Forum age
3. Influence score
4. Title
5. Views
6. Comments
7. Posting date

Preprocessing: Standardized posting dates Dropped irrelevant columns Filled missing views/comments with zeros Imputed missing age/influence using IterativeImputer (Bayesian Ridge)

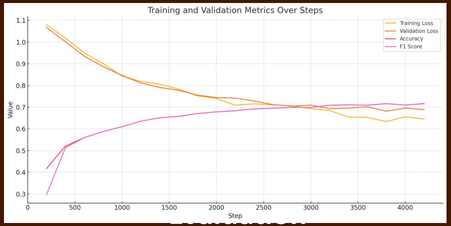
Final Output: 7 structured fields.



Model Selection

To enhance sentiment prediction performance, we selected the Chinese ELECTRA model for fine-tuning.ELECTRA offers higher pre-training efficiency, avoids the [MASK] token mismatch problem found in traditional models, and achieves strong downstream performance with fewer parameters.Its efficient training process and robust representation learning make it a powerful choice for capturing subtle and complex investor sentiment from forum posts.

Feature	BERT(MS1)	ELECTRA(MS2)
Objective	**[MASK] prediction**	**Token replacement detection**
Efficiency	**Low**	**High**
[MASK] Problem	**Yes**	**No**
Performance	**Good**	**Better**



1. Consistent Improvement in Model Performance

Throughout training, both Accuracy and F1 Score show a clear upward trend.

1. Accuracy improved from 41.7% at step 200 to about 71.7% at step 4200.

2. F1 Score improved from 0.297 to about 0.716.

This steady improvement suggests that the model consistently learned better decision boundaries with more training steps, without sudden drops or instability.

2. Smooth Loss Curve Convergence

Both Training Loss and Validation Loss decrease smoothly over time:

1. Training Loss dropped from 1.08 to 0.645.

2. Validation Loss dropped from 1.067 to 0.688.

This indicates that the model generalized well to unseen validation data, and there were no signs of overfitting during training.

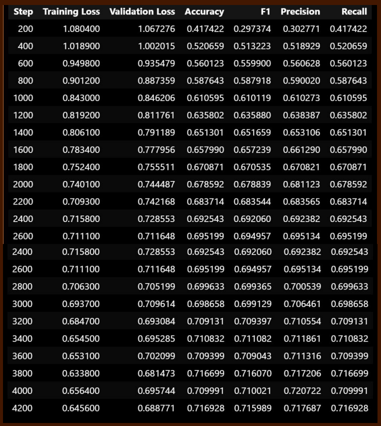
3. Strong Precision and Recall Balance

By the end of training:

1. Precision reaches about 0.717.

2. Recall reaches about 0.717 as well.

The close values between Precision and Recall demonstrate a well-balanced model that is neither too conservative nor too aggressive in making predictions — a particularly good sign for sentiment classification tasks.



Step	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
200	1.080400	1.067276	0.417422	0.297374	0.302771	0.417422
400	1.018900	1.002015	0.520659	0.513223	0.518929	0.520659
600	0.949800	0.935479	0.560123	0.559900	0.560628	0.560123
800	0.901200	0.887359	0.587643	0.587918	0.590020	0.587643
1000	0.843000	0.846206	0.610595	0.610119	0.610273	0.610595
1200	0.819200	0.811761	0.635802	0.635880	0.638387	0.635802
1400	0.806100	0.791189	0.651301	0.651659	0.653106	0.651301
1600	0.783400	0.777956	0.657990	0.657239	0.661290	0.657990
1800	0.752400	0.755511	0.670871	0.670535	0.670821	0.670871
2000	0.740100	0.744487	0.678592	0.678839	0.681123	0.678592
2200	0.709300	0.742168	0.683714	0.683544	0.683565	0.683714
2400	0.715800	0.728553	0.692543	0.692060	0.692382	0.692543
2600	0.711100	0.711648	0.695199	0.694957	0.695134	0.695199
2400	0.715800	0.728553	0.692543	0.692060	0.692382	0.692543
2600	0.711100	0.711648	0.695199	0.694957	0.695134	0.695199
2800	0.706300	0.705199	0.699633	0.699365	0.700539	0.699633
3000	0.693700	0.709614	0.698658	0.699129	0.706461	0.698658
3200	0.684700	0.693084	0.709131	0.709397	0.710554	0.709131
3400	0.654500	0.695285	0.710832	0.711082	0.711861	0.710832
3600	0.653100	0.702099	0.709399	0.709043	0.711316	0.709399
3800	0.633800	0.681473	0.716699	0.716070	0.717206	0.716699
4000	0.656400	0.695744	0.709991	0.710021	0.720722	0.709991
4200	0.645600	0.688771	0.716928	0.715989	0.717687	0.716928

4. Excellent Overall F1 Score

An F1 Score above 0.70 in a real-world noisy dataset like forum posts is a very solid result.

Especially given the complexity and informal language typical of forum discussions, achieving a 0.7169 F1 means the model effectively captures sentiment nuances.

5. Model Stability

From step 3000 onwards, metrics start to stabilize around high values without large fluctuations, reflecting a well-converged model ready for deployment or further fine-tuning.

LEHIGH UNIVERSITY