Dean&Francis
ISSN 2959-6157

Enhancing Stock Market Prediction by Tackling Class Imbalance in Sentiment Analysis

Wenhao Song[1, *]

[1]Department of Computer Science, The University of Hong Kong, Hong Kong, China
*Corresponding author: swh2004@connect.hku.hk

Abstract:

This study addresses the critical challenge of forecasting stock market movements by leveraging sentiment analysis of financial news headlines. Predicting the Dow Jones Industrial Average (DJIA) is of great significance to investors and financial institutions, as market trends are often influenced by public sentiment and rapidly evolving news cycles. Traditional quantitative models often struggle to capture the nuanced impact of textual information, especially in the context of imbalanced data distributions where rare but impactful market events are underrepresented. In this work, the paper constructs a comprehensive sentiment analysis framework utilizing advanced natural language processing methods to classify the sentiment of publicly available news headlines and examine their relationship with subsequent DJIA price fluctuations. By integrating headline sentiment scores with historical price data, our experiments systematically evaluate the predictive reliability of various machine learning models, with a particular focus on class imbalance mitigation. The results demonstrate that models employing class weighting in LightGBM outperform those using conventional resampling techniques, achieving recall rates of 0.45 for downturn prediction versus 0.12 with baseline methods. These findings highlight the value of algorithmic enhancements for rare event forecasting and suggest that richer, domain-specific representations may further improve predictive accuracy. Future research will explore enhanced features and time-series modeling to boost the robustness of financial sentiment analysis.

Keywords: Sentiment Analysis, SMOTE, LGBM, Random Forest, Stock Market Prediction

ISSN 2959-6157

1. Introduction

In an age of data-driven finance, sentiment analysis has emerged as a powerful tool for capturing market movements. Such an assumption is simple yet really powerful: Hassan states that a reliable model has a great impact on investment strategies [1]. The highs and lows on stock market indices like the Dow Jones Industrial Average (DJIA) are often caused by the changing moods of content found in public-type news headlines [2]. Beyond investor-facing news sentiment, broader organizational sentiment also relates to market movements. Recent evidence shows that aggregate employee sentiment is a significant negative predictor of subsequent market returns, indicating that sentiment signals embedded in diverse textual sources can carry market-relevant information [3]. This research strives to shed light on another approach from a popular Kaggle project and further adds to it: the use of a Random Forest classifier to classify daily news headlines as suggesting either an upward or downward DJIA closing.

Similar to the limitations noted by Smatov et al., our model encountered bias toward positive sentiment due to the overall upward market trend in the dataset, highlighting the importance of balanced datasets and careful feature engineering [4]. Most importantly, a model with seemingly high accuracy in an unbalanced setting is somewhat misleading; if the model chooses the majority class most of the time, it just overlooks less frequent but crucial events, such as the market going down.

This study aims to systematically investigate this issue and enhance the model's accuracy and utility in real-world data. This project will conduct a series of controlled experiments focusing on three key objectives:

Quantify the Impact of Class Imbalance: Establish a performance benchmark with the given model first. Second, artificially make the data set extremely imbalanced in order to precisely quantify the loss in the model's power of classification of the minority class, thereby emphasizing why accuracy is an inadequate metric.

Evaluate Data-Level Mitigation Using SMOTE and Test on New Models: This project will use a Synthetic Minority Over-sampling Technique on the imbalanced data set to check whether it can restore the predictive power of its model through synthetic data generation and rebalancing of the classes prior to training. The project will also use LGBM to test the performance.

Explore Alternative Enhancement Strategies: Compare the alternative approach here with some other common improvements. These include more feature-engineering (TF-IDF Vectorizer instead of CountVectorizer), model-level adjustments (set "class_weight"), and other basic parameters to verify which techniques are most effective for this problem.

With the help of the structured analysis above, the project aims to provide clear and evidence-based recommendations for building more reliable sentiment analysis models for financial prediction.

2. Experimental Setup

2.1 Dateset

In conducting this research, the paper makes use of the "Combined_News_DJIA.csv" dataset [5], which is accessible on Kaggle and contains headlines from each day paired with the Dow Jones Industrial Average (DJIA) closing figures for the corresponding days. In this dataset, each record contains three pivotal pieces of information: the "Date" of the news headlines, the "Label", which denotes the market direction with "1" indicating the DJIA closed higher and "0" indicating it closed lower, and "Top1 - Top25", which contains the top 25 news headlines for the given day. As shown in Figure 1, the data distribution for class 0 and class 1 is quite balanced.

To ensure the model's evaluation remains unbiased, the data is split on a timeline basis. The data before January 1, 2015, is used for the training set, and the data after January 1, 2015, is assigned to the test set. The use of this timeline-based split serves to ensure that a forecasting study is conducted on a completely out-of-sample future period. As with the forecasting use case, this type of split also serves to reduce the likelihood of data leakage and adheres to the procedures adopted in the literature on forecasting financial time series. Using this dataset, we are able to analyze how the sentiment of news on a given day affects the movement of the DJIA within a given time frame, which sets the stage for attempting to predict stock trends using machine learning.

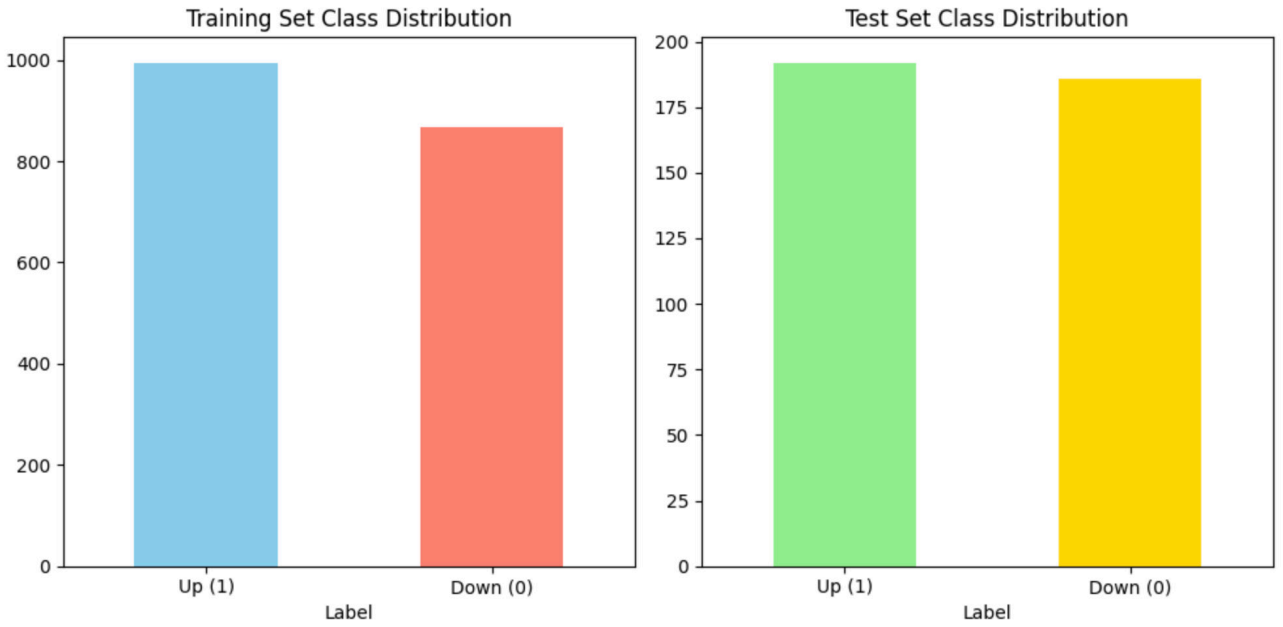

Fig. 1 Data distribution (Left: Training set class distribution, right: Test set class distribution)

2.2 Baseline Model

The baseline model is a "RandomForestClassifier" with 200 estimators and "entropy" as the splitting criterion. Text features are generated using "CountVectorizer" with bi-grams ("ngram_range=(2,2)").

2.3 Evaluation Metrics

The different evaluation metrics were calculated to assess the model's performance from all angles, especially with regard to the imbalance in the classes, namely the confusion matrix, precision, recall, and F1-score. The confusion matrix plots actual and predicted classes in relation to each other. This can be helpful in identifying how well the model distinguishes between classes and whether it might favor the majority or minority class. Precision tells how much to trust positive predictions of the model, i.e., the percentage of predicted positive points that are actually true positives. Recall, on the other hand, tells how well the model finds all instances of interest from a given class, evaluating its effectiveness against capturing minority class events. An F1 score combines the metrics, precision, and recall in a balanced way to provide a single indicator, which is especially useful in imbalanced data sets where one of these metrics cannot present a full picture of the model's performance alone.

2.4 Imbalance Simulation

To test the model, this research will create an artificially imbalanced training set where the minority class (Label 0, market down) is much less than the majority class (Label 1, market up) at a ratio of approximately 2:5 [6]. This allows for a controlled experiment on the effects of class imbalance, which is similar to the real stock market.

3 Results and Analysis: The Impact of Class Imbalance

In this section, this paper analyzes the performance of the baseline model and contrasts it with a model trained on the simulated imbalanced dataset.

3.1 Balanced Model Performance

First, this research establishes the benchmark by training the "RandomForestClassifier" on the original, relatively balanced training dataset (Figure 2).

ISSN 2959-6157

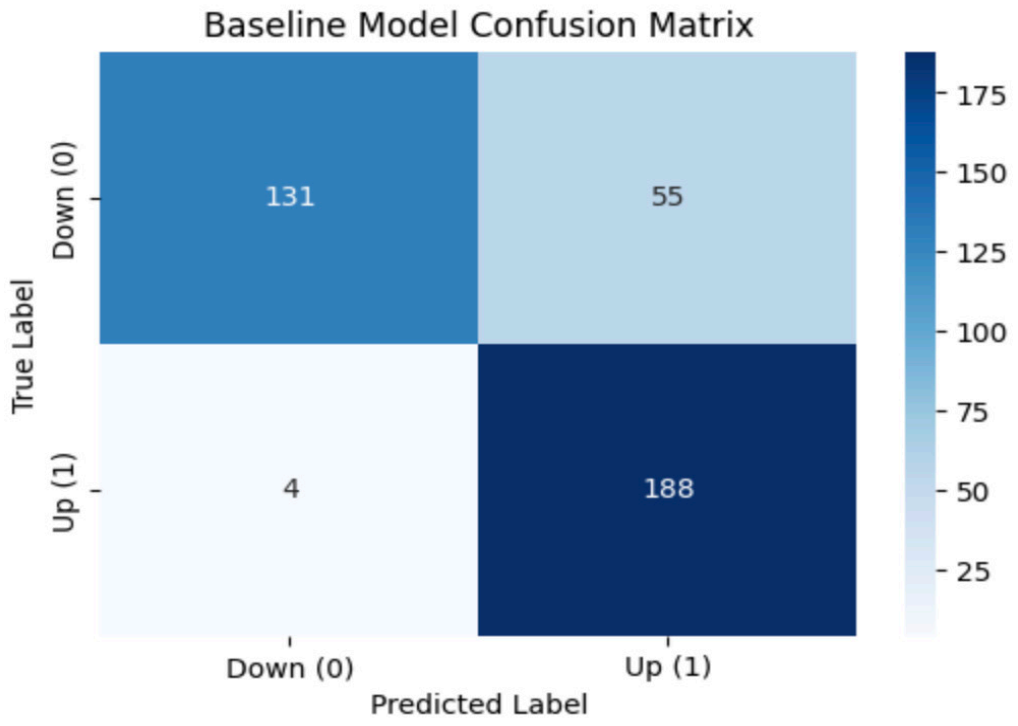

Fig. 2 Baseline Random Forest Confusion Matrix

As shown in Figure 2, the baseline model performs well, achieving an overall accuracy of 85%. Also, good performance balanced across both classes is to be noted. With a recall of 0.85 for Class 0, 85% of days where the market ended down are correctly identified by the model. This is the performance that is set as a baseline for success.

3.2 Imbalanced Model Performance: A Cautionary Tale

Following this, this research trains the exact same model architecture on an artificially skewed dataset with a 2:5 ratio as displayed in Figure 3. The model continues to be evaluated on the same original test set.

Fig. 3 Artificial imbalanced data distribution and testing data distribution (Left: Training set class distribution, Right: Test set class distribution)

Imbalanced Model Performance:

Table 1. Imbalanced model performance

	Precision	Recall	F1-score	Support
0	1.00	0.32	0.48	186
1	0.60	1.00	0.75	192
Accuracy	NAN	NAN	0.66	378
Macro avg	0.80	0.66	0.62	378
Weighted avg	0.80	0.66	0.62	378

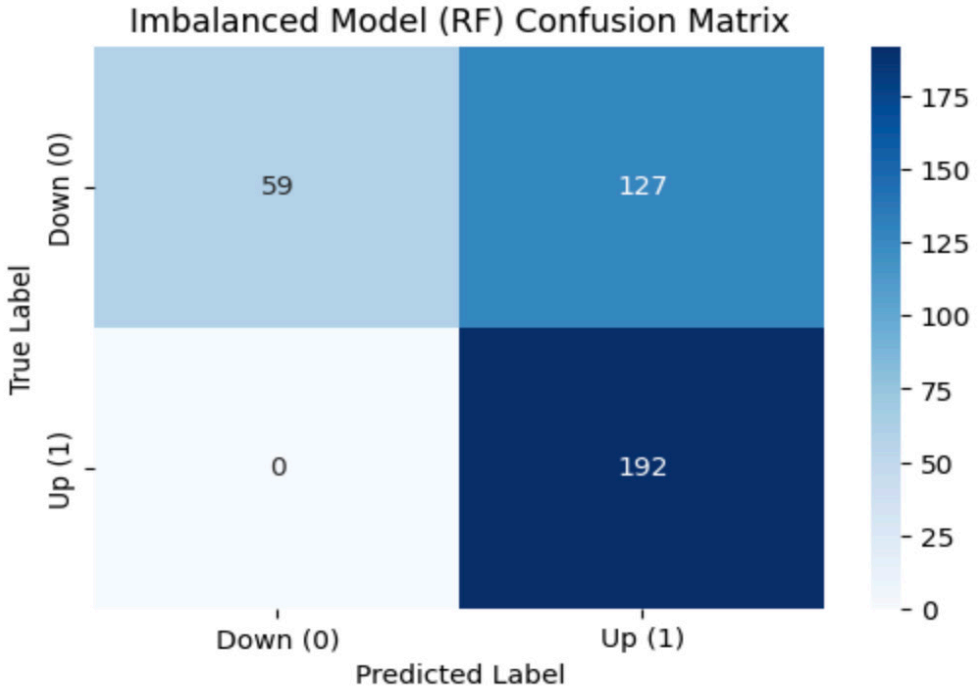

Fig. 4 Random Forest performance in imbalanced data

The paper can infer from Table 1 and Figure 4 that a huge failure has been demonstrated in the capacity of the model to predict.

The recall for Class 0 (market down) comes down from 0.70 to 0.32 drastically. So, the model now identifies only 32% of market downward movements, missing almost two-thirds of what it is supposed to predict.

Accuracy Paradox: While the accuracy of the test has dropped to 0.66, this number is quite deceiving. Indeed, it is now far more likely to predict the majority class (Label 1), supported by a 1.00 recall value achieved for that class. The model has thus become a "one-sided" predictor, much to the neglect of the minority class.

Business Impact: In an actual financial environment, this model can be detrimental. Its failing at predicting downturns, arguably the biggest events for an investor, rendering it unreliable despite an apparently acceptable accuracy score.

4 Mitigation Strategies

Having recognized the negative impact of imbalanced data, this research now attempts some alternative methods to overcome it: SMOTE, LGBM, etc.

4.1 Dateset

This research uses the Synthetic Minority Over-sampling Technique (SMOTE) on the imbalanced training set to generate new, synthetic samples for the minority class so as to have a perfectly balanced dataset to train on [7].

ISSN 2959-6157

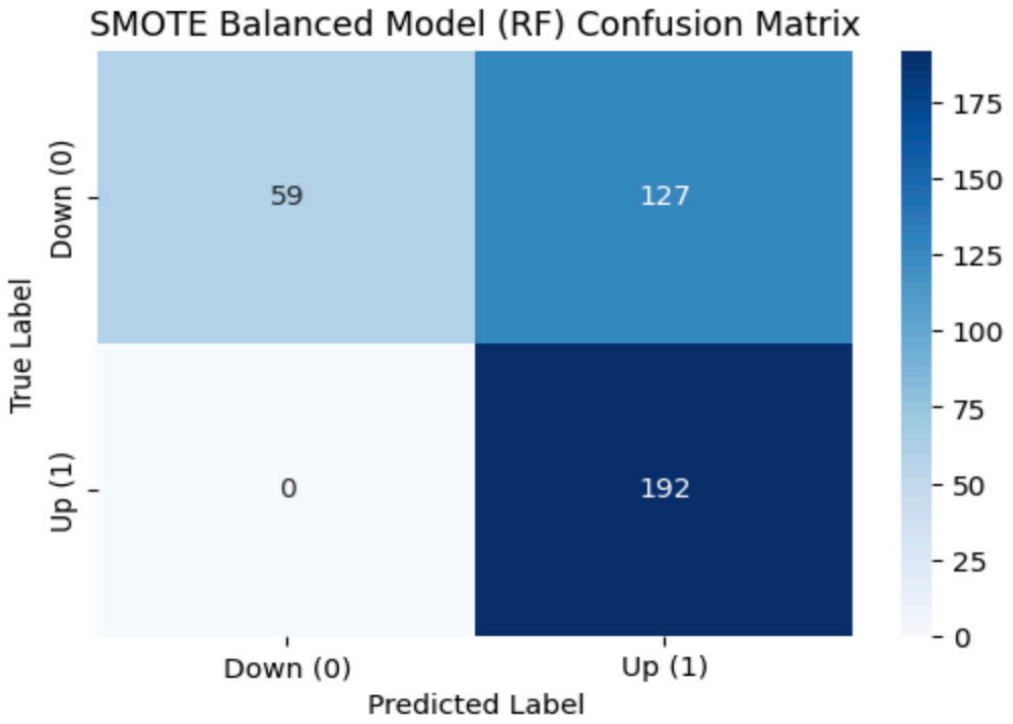

Fig. 5 Random forest performance on imbalanced data after SMOTE

Table 2. Random forest performance on imbalanced data after SMOTE

	Precision	Recall	F1-score	Support
0	1.00	0.32	0.48	186
1	0.60	1.00	0.75	192
Accuracy	NAN	NAN	0.66	378
Macro avg	0.80	0.66	0.62	378
Weighted avg	0.80	0.66	0.62	378

The results are dramatic and counter-intuitive. For the RandomForest classifier, applying SMOTE had absolutely no improvement at all. The paper can notice from the figure 5 and Table 2 that the Class 0 recall remained 0.32, the same for the model that was trained with the imbalanced data. This shows that the synthetic samples generated by SMOTE were not diverse enough to cause the RandomForest model to learn something new and led it to adopt the same faulty, one-sided prediction strategy.

4.2 Using LGBM with SMOTE data

After trying RandomForest model, this study will use LGBM instead [8].

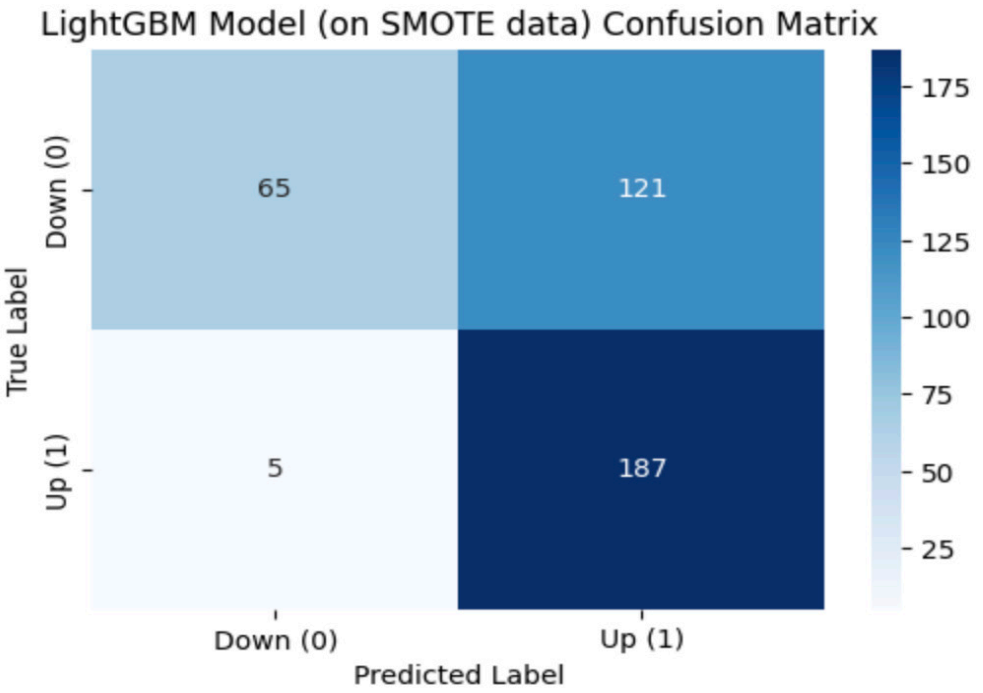

Fig. 6 LGBM performance on imbalanced data after SMOTE

Table 3. LGBM performance on imbalanced data after SMOTE

	Precision	Recall	F1-score	Support
0	0.93	0.35	0.51	186
1	0.61	0.97	0.75	192
Accuracy	NAN	NAN	0.67	378
Macro avg	0.77	0.66	0.63	378
Weighted avg	0.77	0.67	0.63	378

The LightGBM model trained over SMOTE data showed a tiny increase in Class 0 recall from 0.32 to 0.35, as shown in figure 6 and Table 3, but again, this is not a significant recovery. This demonstrates that the simple generation of more data may not solve the problem if the original minority samples lack enough diversity.

4.3 Algorithmic Handling with LightGBM

As an alternative to debiasing the data, this research can make use of a more advanced model designed to handle imbalance internally. This research directly trained a LightGBM classifier on the unbalanced data using its native class_weight='balanced' parameter. This automatically carries over into the model's learning process to charge for minority class errors more [8].

ISSN 2959-6157

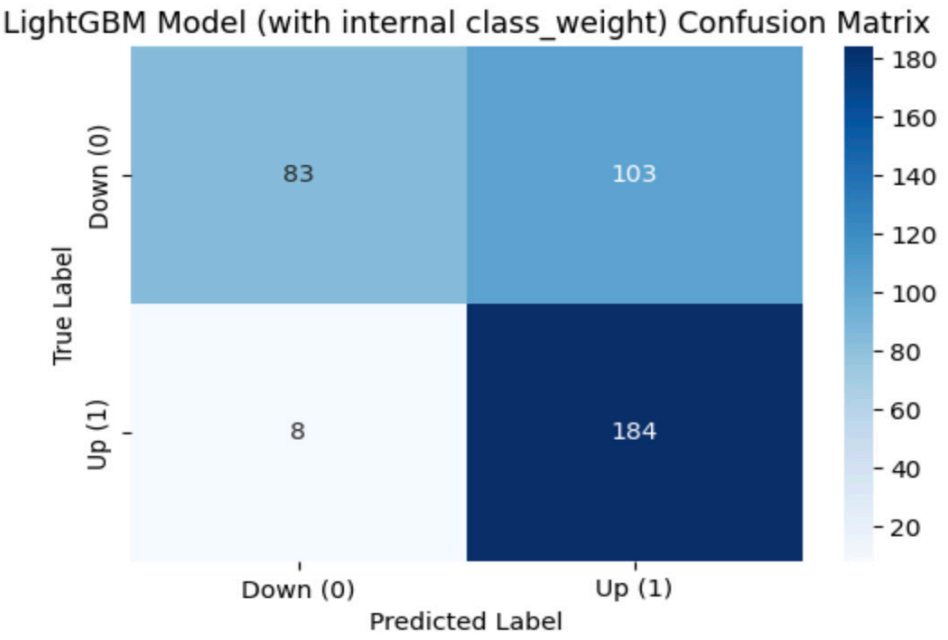

Fig. 7 LGBM with class_weight performace

Table 4. LGBM with class_weight performace

	Precision	Recall	F1-score	Support
0	0.91	0.45	0.60	186
1	0.64	0.96	0.77	192
Accuracy	NAN	NAN	0.67	378
Macro avg	0.78	0.70	0.68	378
Weighted avg	0.77	0.71	0.69	378

This approach had the best performance on all of the mitigation techniques. From the figure 7 and Table 4 the paper can see that the class 0 recall was improved to 0.45. While this is still very low compared to the baseline of 0.70, it is a step better than the imbalanced model at 0.32 and both the SMOTE models. This indicates that in this task, using a very advanced algorithm with internal weighting is a superior technique to resampling at the data level.

5 Mitigation Strategies

To put all of these results together, this paper compares all of the tested models on the important metrics.

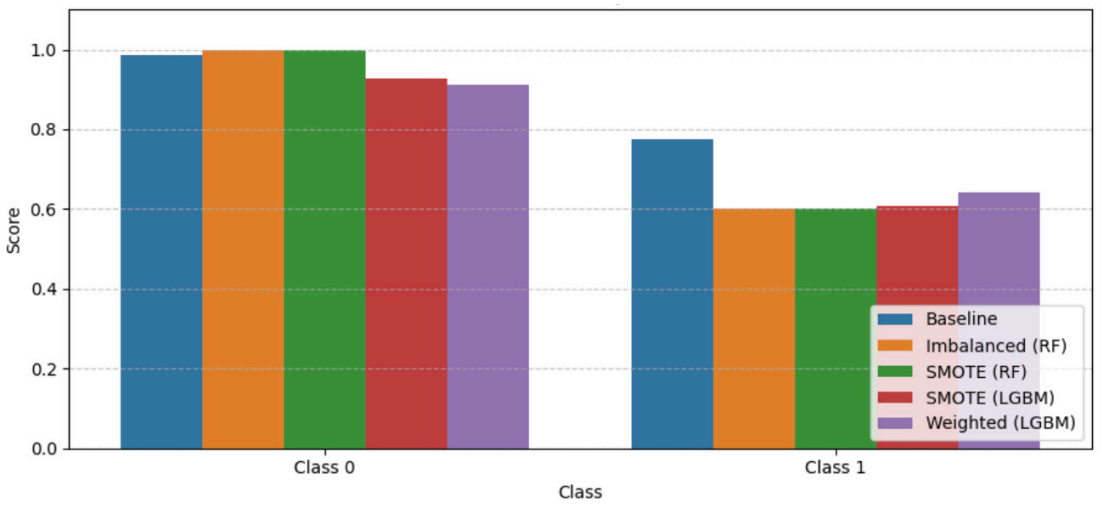

Fig. 8 Precision Comparison

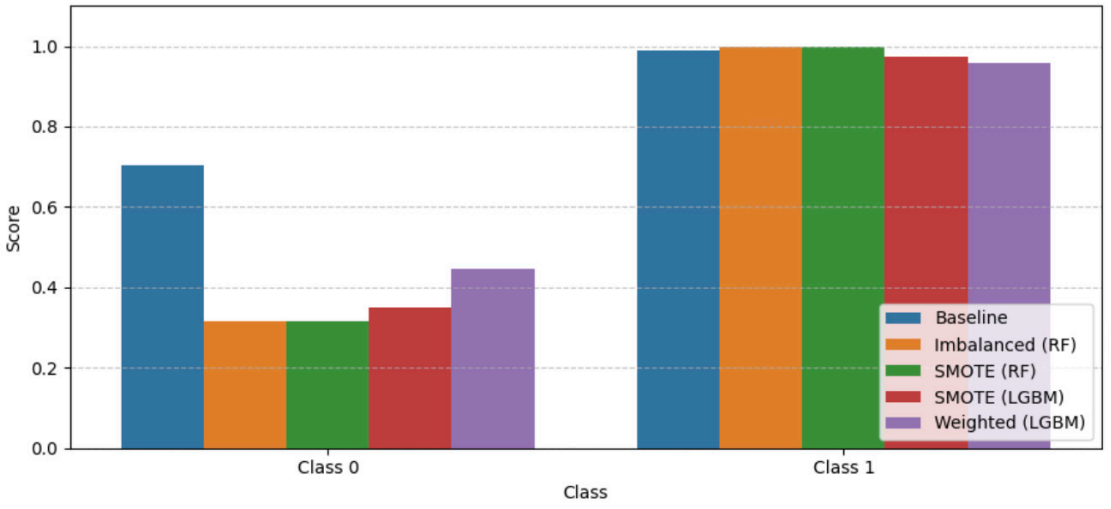

Fig. 9 Recall Comparison

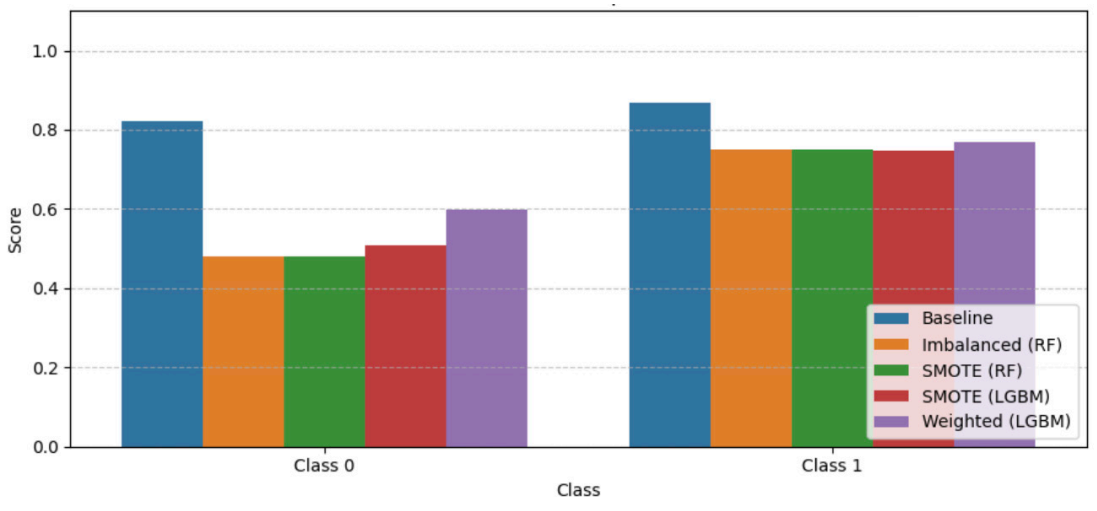

Fig. 10 F1-score Comparison

ISSN 2959-6157

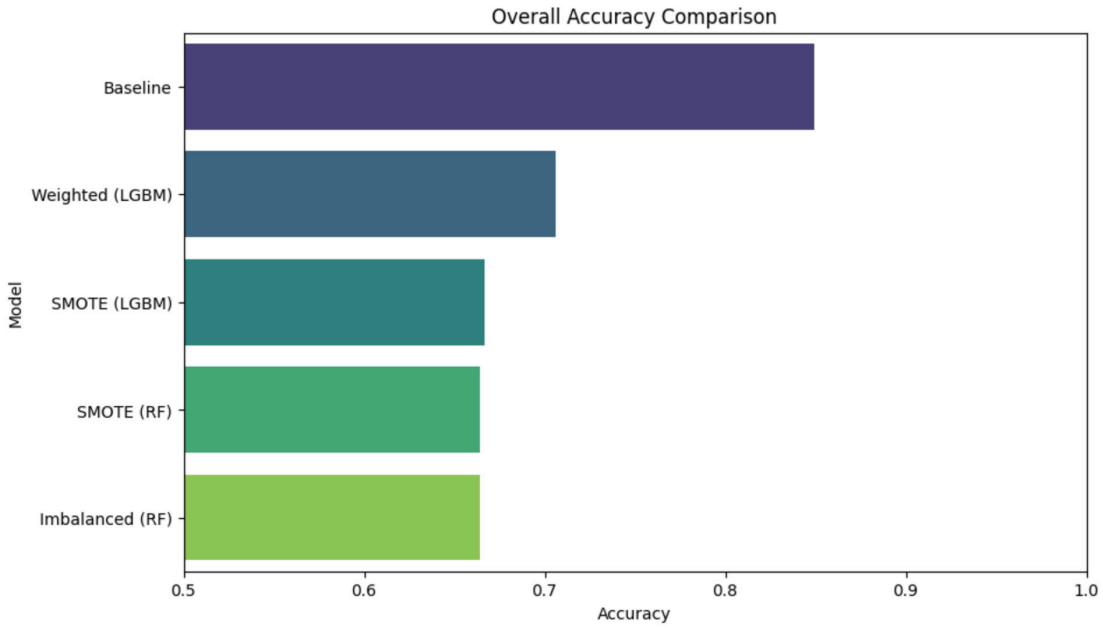

Fig. 11 Accuracy comparison

The overall comparison supports the key findings (figures 8-11). The accuracy chart is especially informative: whereas the Baseline model sits alone at 85% accuracy, the Weighted (LGBM) model is the undisputed champion of the mitigation methods at 71%.

But the plots by metric tell the full story. The most critical one is the "Recall Comparison" for Class 0. It visually confirms the awful decline in performance from Baseline (0.70) and shows the modest but certain dominance of the Weighted (LGBM) approach (0.45) for salvaging some of that lost predictive power.

The key takeaway here is that the model structure and the power that could potentially emerge from it is greater than pre-processing data using methods such as SMOTE, particularly when there is very sparse minority class data. The gradient boosting process inside LightGBM constructs trees sequentially in an effort to learn from the mistakes of previous trees; this turned out to be much better at learning from very sparse samples of the minority class than the RandomForest algorithm.

Recent research has introduced improved SMOTE-style algorithms that address the geometry of the minority class and privacy issues. Contemporary variants, such as borderline/cluster-based SMOTE and other enhancements, show a greater probability of preserving minority class structure compared to basic oversampling methods. Investigating these advanced oversampling techniques in high-dimensional text settings, such as our own, represents an important subsequent step [9]. Compared to resampling, richer, domain-specific representations may yield greater improvements. Without relying on artificial sampling, open-source financial LLMs like FinGPT provide instruction-tuned emotion signals and contextual, finance-tuned embeddings, which could enhance minority-class separability. Incorporating FinGPT characteristics into our pipeline is a promising direction [10].

6 Conclusion

This paper systematically addresses the challenge of class imbalance in sentiment analysis for stock market prediction, focusing on the Dow Jones Industrial Average (DJIA). Through a combination of baseline modeling, artificial imbalance experiments, and mitigation techniques such as SMOTE and class-weight adjustments in LightGBM, the study demonstrates the profound impact that class distribution has on predictive reliability. While traditional Random Forest approaches suffer dramatic losses in minority class recall when faced with imbalanced data, advanced algorithms like LightGBM with internal weighting offer more robust solutions, partially salvaging lost accuracy and recall. The results suggest that simple resampling, such as SMOTE, may be insufficient when minority class samples are sparse and not diverse enough. Instead, algorithmic enhancements that directly address imbalance—like class weighting—are more effective in recovering predictive power for rare but crucial events (e.g., market downturns). Overall, the findings emphasize the need for careful evaluation metrics beyond mere accuracy, advocating for precision, recall, and F1-score as more informative measures in finance. The proposed future directions—including richer datasets, advanced features, expanded

model selection, and the exploitation of time series continuity—promise further gains in reliability and practical utility. With these improvements, sentiment analysis models can become more robust tools for financial forecasting, helping investors and analysts navigate the complexities of real-world market behavior.

References

[1] Bajzik J, & Badura O. Does sentiment affect stock returns? A meta-analysis across survey-based measures. International Review of Financial Analysis, 89, 102773. 2023.

[2] Tetlock P C. Giving Content to Investor Sentiment. The role of media in the stock market. PC Tetlock Journal of Finance. 2007.

[3] Chen J, Tang G, Yao J, & Zhou G. Employee sentiment and stock returns. Journal of Economic Dynamics and Control, 149, 104636. 2023.

[4] Smatov N, Kalashnikov R, Kartbayev A. Development of Context-Based Sentiment Classification for Intelligent Stock Market Prediction. Big Data and Cognitive Computing, 8(6), 51. 2024.

[5] Amritkar N. Stock sentiment analysis_ [Jupyter Notebook]. Kaggle. https://www.kaggle.com /code/niharikaamritkar/stock-sentiment-analysis. 2023.

[6] Forecast G. MARKET REVIEW. Sierra, 29, 1. 2024.

[7] Chawla N V., et al. SMOTE: Synthetic Minority Oversampling Technique. JAIR. 2022.

[8] Ke G, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS. 2017.

[9] Li Y, Yang Y, Song P, Duan L, & Ren R. An improved SMOTE algorithm for enhanced imbalanced data classification by expanding the sample generation space. Scientific Reports, 15(1), 23521. 2025.

[10] Yang H. FinGPT: Open-Source Financial Large Language Models. SSRN Electronic Journal. 2023.