Frontiers in Economics and Management Volume 7 Issue 7, 2026
ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

Hadoop-based Financial News Text-Driven Market Volatility Prediction: TF-IDF, Machine Learning, and SHAP Interpretability Analysis

Yishen Wu[a]

School of Economics and Management, Xidian University, Xi'an 710126, China

[a]23061300093@stu.xidian.edu.cn

Abstract

The variation of financial market is largely influenced by the emergence of new information like macroeconomic policies, corporate notifications and international political events. With the growing amount of financial textual data, it has become an important issue to predict market trends from news materials. In our research, we have collected an empirical data set including 1,985 news articles concerning market volatility. A binary target label has been established based on whether the daily volatility exceeds a 1% level, resulting in a positive class ratio approximately 0.27. We carry out the Hadoop framework (HDFS and distributed processing) in the first stage of text processing to preprocess and organize the raw data. For feature extraction, a TF-IDF model (with 1-2 grams and 5,000 maximum features) is used to obtain the text vector representation. Then, we train and test two kinds of algorithms: a calibrated probability linear Support Vector Machine (SVM) and an XGBoost classifier. We conduct validation tests based on a TimeSeriesSplit method. The performance measures consist of the area under the Receiver Operating Characteristic curve (ROC-AUC), Precision-Recall curve (PR-AUC) and the optimal F1 score. Additionally, we integrate other features such as VADER sentiment indices, token entropy and special shock words to do an ablation study. The SHAP parameters give explanations for the results both globally and locally. The experimental results indicate that there are strong predictive hints in the news texts and the interpretive step is beneficial to confirm the model logic.

Keywords

Financial Volatility Prediction; News Text; Hadoop; TF-IDF; SVM; XGBoost; Ablation Study; SHAP.

1. Introduction

Volatility indicates the risk and uncertainty in financial markets. In actual trading, the changes of asset prices are closely related to the external news shocks as the public information affects the investor's expectations. Ordinary statistical methods generally process organized data such as past prices or trading volumes of assets. But, the unstructured data including text documents may provide useful prior information about the market uncertainty, the market sentiments and particular economic events. The current study focuses on a particular issue: can we forecast the high-volatility market conditions based on the news text features alone.

We have set up a whole data science process for this predictive task. The system manages the storage of original data and distributed text cleaning in a Hadoop environment, whereas the model training and validation are carried out in Python. Furthermore, we apply ablation tests and SHAP values to clarify how the models arrive at their results.

This document presents some potential modifications in the current research area. A processed text dataset has been set up, in which each news item is linked with a binary volatility index.

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

Furthermore, we have developed an effective system utilizing HDFS and parallel computing techniques to handle extensive amounts of unstructured data. Additionally, the performances of linear SVM and XGBoost classifiers have been assessed by a walk-forward validation method based on conventional TF-IDF features. Moreover, we combine VADER sentiment information and Shannon entropy for feature analysis and apply SHAP models to elucidate the internal decision-making process.

2. Related Work

Stock market volatility is closely related to external information shocks and investor sentiment. Deng and Liu (2019) found that investor attention to corporate stocks is a main cause of market fluctuations in the new media era. This finding provides the theoretical basis for using text data to predict the market [1]. Hautsch et al. (2015) used high-frequency trading data to build liquidity networks and predict market flash crashes [2]. Their study shows that extracting micro-features from different information sources helps capture sudden market changes. Therefore, using unstructured data like news sentiment in volatility models is reasonable from the perspective of behavioral finance.

Natural language processing (NLP) combined with machine learning methods usually performs better than traditional financial time-series models. Market sentiment extracted from text is a core predictive feature [3]. Researchers often use dictionary-based tools such as VADER and topic modeling to analyze news and social media data [4]. Many empirical studies prove that machine learning models, such as Support Vector Regression (SVR) and Artificial Neural Networks (ANN), achieve higher accuracy than traditional autoregressive models (like ARIMA) when using text features.

NLP has many applications in financial tasks, including return prediction, risk assessment, and volatility forecasting [5]. Common text representation methods include bag-of-words, TF-IDF, and sentiment scores. These features are usually trained with models like logistic regression, SVM, and gradient boosting trees [6]. However, local computers face processing bottlenecks when text data becomes very large. Distributed systems like Hadoop (HDFS) are widely used for parallel text data storage and mining [7-10]. To solve the big data problem, this study uses Hadoop for data cleaning and structuring. This step ensures that our preprocessing is scalable.

3. Experimental Design and Workflow

Designing an experiment for financial text mining requires a robust pipeline that can ingest messy raw data and ultimately deliver interpretable market predictions. As mapped out in Figure 1, our workflow breaks down into four practical stages. We start by moving raw news data into HDFS for distributed preprocessing. From there, we take a dual-track approach to feature engineering: extracting standard TF-IDF baselines alongside customized "information shock" features. Once the data is structured, we train and pit a linear SVM against an XGBoost classifier, evaluating both models using a time-series-aware walk-forward split. Finally, we break open the models using feature ablation and SHAP to understand exactly what drove the predictions, reporting quantitative metrics such as ROC-AUC, PR-AUC, and the optimal F1 score.

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

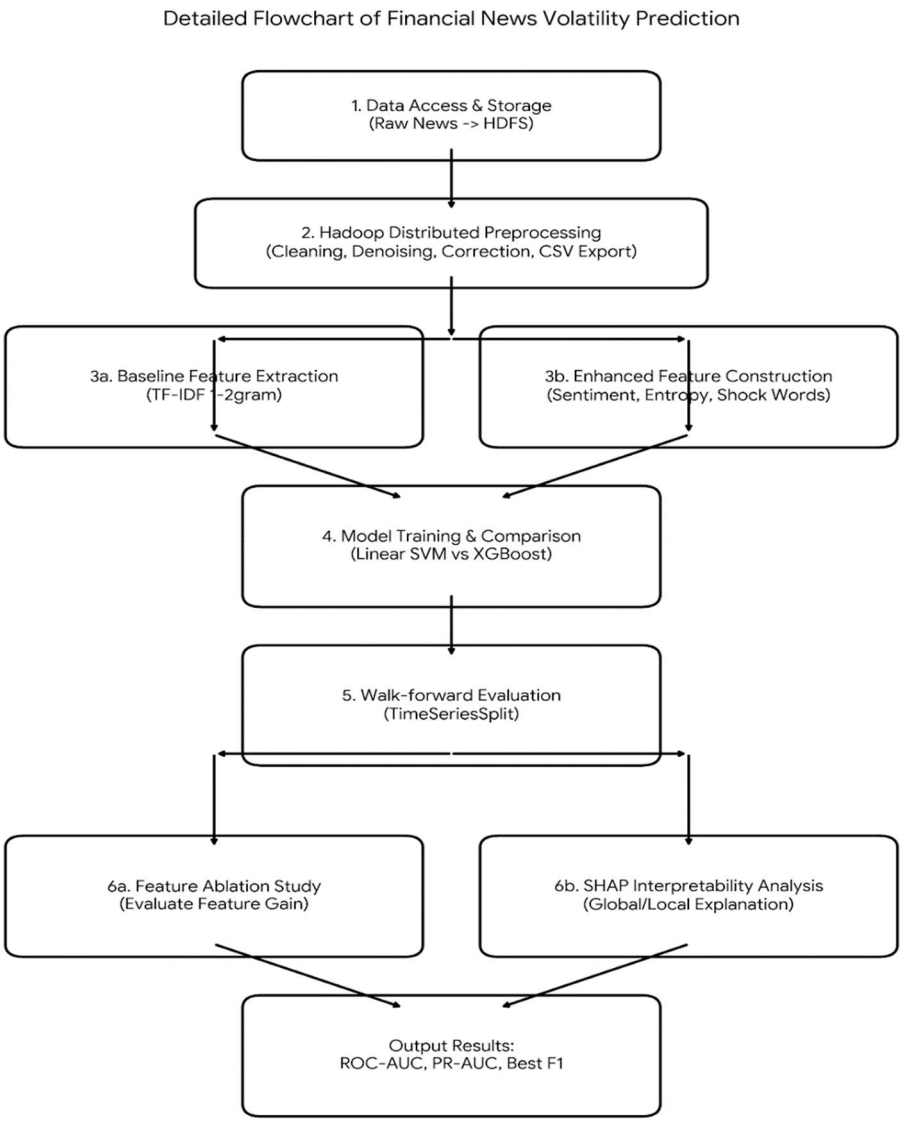

Figure 1. Overall experimental workflow

3.1. Dataset and Preprocessing

3.1.1. Dataset Description

The dataset contains approximately 1985 samples, each including fields: date, text, label, and optionally volatility. label=1 indicates daily volatility exceeding a threshold (≈1%). The positive class proportion is about 0.27.

Label definition:

$$y_t = f(x) = \begin{cases} 1, & if\ v_t > \theta \\ 0, & otherwise \end{cases} \quad (1)$$

where θ ≈ 1%.

3.1.2. Hadoop Distributed Cleaning Process

To improve preprocessing scalability, data cleaning and structured export are performed in the Hadoop environment: raw data is written to HDFS, distributed tasks process each record in parallel, fix encoding and byte-string artifacts (e.g., b'...', b".."), delete empty text and invalid lines, and export the cleaned result as CSV for Python modeling.

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

3.1.3. Python-side Text Standardization

Lightweight supplementary cleaning is performed in Python to ensure stable vectorization: remove residual byte-string artifacts, clean newlines/tabs and extra whitespace, and delete empty text lines.

3.2. Feature Representation and Engineering

3.2.1. TF-IDF Representation (Baseline Feature)

TF-IDF maps text to numerical vectors. For term t and document d:

$$\boldsymbol{TF-IDF(t,d) = TF(t,d) \times IDF(t)} \tag{2}$$

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \tag{3}$$

$$IDF(t) = \log\left(\frac{N}{df_t}\right) \tag{4}$$

This paper uses 1-2gram, max_features=5000, and removes English stop words.

3.2.2. Augmented Features: Sentiment, Entropy, and Shock Words

To enhance the characterization of "information shocks", three types of interpretable numerical features are introduced (for ablation experiments):

Sentiment features (VADER): pos/neu/neg/compound four-dimensional scores.

nformation entropy (Entropy): Shannon entropy of token distribution in the text, capturing information density and uncertainty.

Shock word score: count of hits for shock words such as crisis, recession, inflation, sanctions, etc.

3.3. Model Methods

3.3.1. Linear SVM (Probability Calibration)

Linear SVM is suitable for high-dimensional sparse features [11]. Decision function: $f(x)=w^Tx + b$. Soft margin optimization problem:

$$\boldsymbol{\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i} \tag{5}$$

subject to:

$$\boldsymbol{y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0} \tag{6}$$

To obtain probabilistic outputs, sigmoid probability calibration is applied.

3.3.2. XGBoost

XGBoost is a gradient boosting tree model capable of capturing nonlinear relationships [12]. TF-IDF features are used as input to train an XGBoost classifier and compare with SVM.

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

3.4. Experimental Setup

3.4.1. Evaluation: Walk-forward (TimeSeriesSplit)

Financial data has temporal dependencies; random splitting may cause future information leakage. This paper uses TimeSeriesSplit for walk-forward evaluation: each fold uses past data for training and future data for testing, closely mimicking real prediction scenarios.

3.4.2. Metrics and Threshold Optimization

ROC-AUC, PR-AUC, and F1 are reported. The threshold maximizing F1 is searched on the PR curve to improve decision practicality.

4. Experimental Results and Model Interpretability Analysis (SHAP)

In baseline experiments based on TF-IDF features, SVM overall performed better than or close to XGBoost. As shown in Table 1, evaluation metrics show that SVM achieved an optimal F1 score of 0.475, with recall for the positive class (0.806) significantly higher than precision (0.337). This indicates that the model tends to adopt a "broad coverage" strategy when identifying potential high-volatility events, showing high sensitivity to tail risk, albeit with some false positive rate. Due to the significant class imbalance, the PR curve more accurately measures the model's identification quality on the minority class (high-volatility events) than the ROC curve.

4.1. Baseline Model Results (TF-IDF)

Table 1. Baseline model performance comparison

Model	ROC-AUC	PR-AUC	Best F1	Precision (pos)	Recall (pos)
SVM	**0.665**	**0.458**	**0.475**	**0.337**	**0.806**
XGBoost	**0.639**	**0.415**	**0.472**	**0.343**	**0.756**

ROC and PR curves

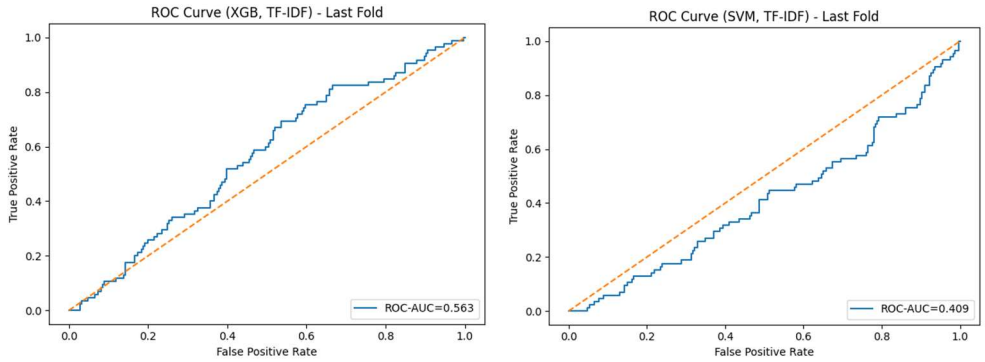

Figure 2. ROC curves

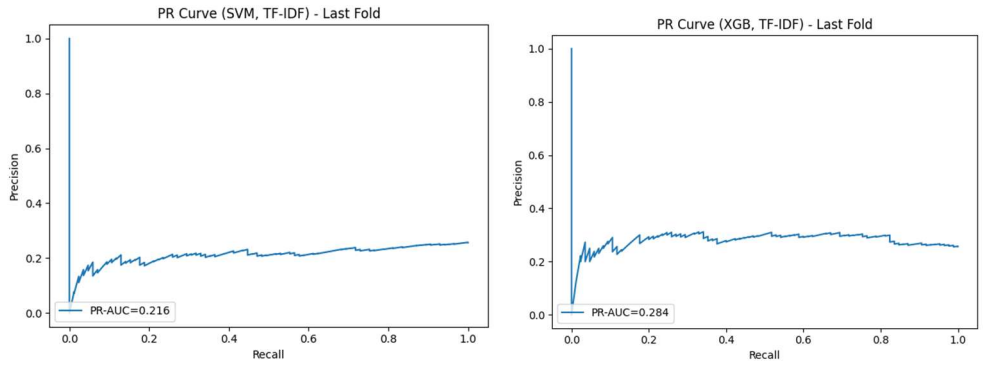

Figure 3. Precision-Recall curves

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

4.2. Learning Curves (Generalization Diagnosis)

To diagnose model fit and generalization, learning curves were plotted, dynamically monitoring training and validation F1 scores as training sample size increases. If the training score remains significantly higher than the validation score, it suggests overfitting in the high-dimensional feature space; conversely, if both converge at a low level, it indicates limited expressiveness of current text features or underfitting, requiring more discriminative features.

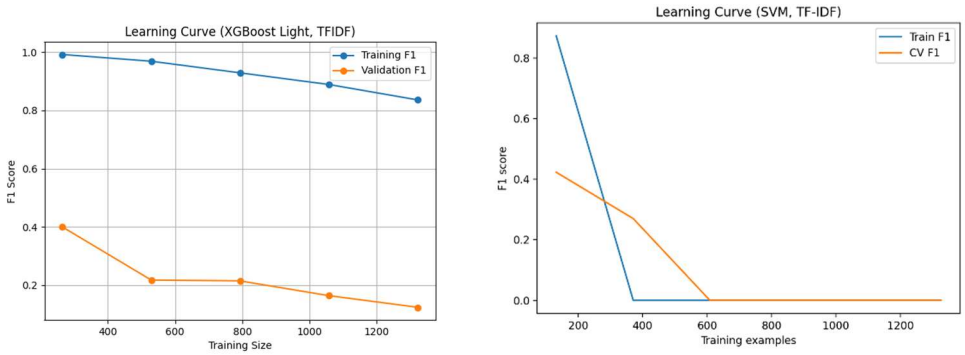

Figure 4. Learning curves

4.3. Feature Ablation Study

To quantitatively verify the effectiveness of each augmented textual feature, a strict feature ablation study was conducted on the TF-IDF baseline. Results show that after introducing entropy features (TF-IDF + Entropy), the ROC-AUC (0.447) and optimal F1 (0.35) did not improve significantly. This suggests that the augmented feature may be somewhat redundant with TF-IDF, or its potential benefit is limited by the current sample size.

Table 2. Feature ablation results (mean over folds)

Feature Set	ROC-AUC (mean)	PR-AUC (mean)	Best F1 (mean)
TFIDF	0.449	0.21	0.35
TFIDF+Entropy	0.447	0.207	0.35

4.4. SHAP Method and Prediction Decomposition

To enhance credibility and auditability in financial scenarios, SHAP is used to explain model predictions [13]. For a sample x, the prediction can be decomposed as:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{7}$$

where φ_i represents the marginal contribution of feature i to the prediction.

4.5. Global Interpretation: Important Tokens

The SHAP summary plot shows the token contributions and their importance (evaluated by mean(|SHAP|)), providing an overall view of the feature importance distribution. It is found that words concerning macroeconomic uncertainty and geopolitical shocks have large mean absolute SHAP values, and these tokens tend to increase the model's prediction of high volatility.

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

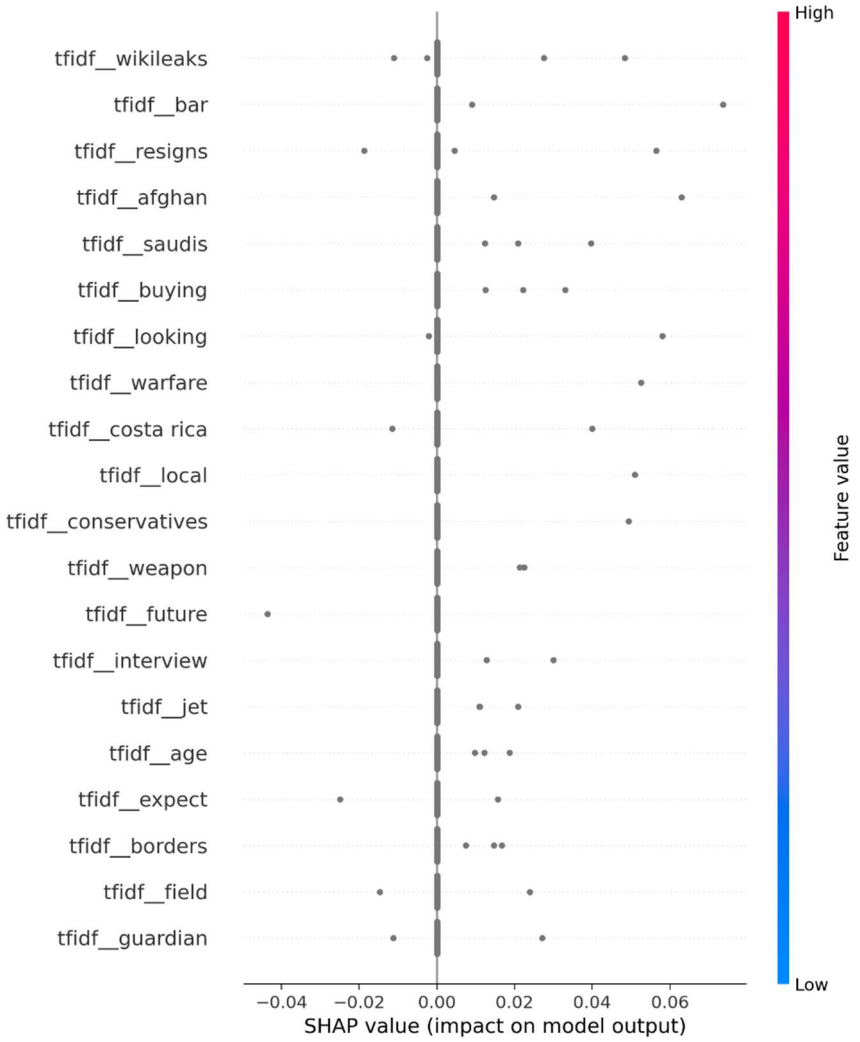

Figure 5. SHAP summary plot: token importance and impact direction

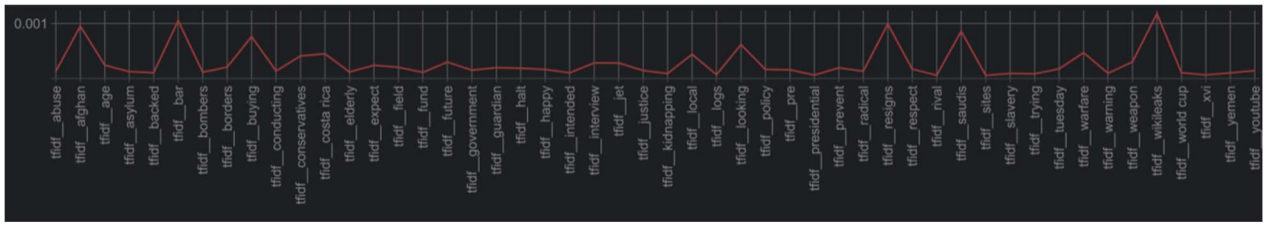

Figure 6. Key tokens and their SHAP contributions (illustrative)

4.6. Local Interpretation: Single-sample Contribution Breakdown

To find out which terms influence the model's performance, we refer to the SHAP summary plot presented in Figure 5. This diagram shows both the magnitude, calculated as the average absolute SHAP value, and the effect direction of each token. From the analysis, it is clear that the terms related to macroeconomic uncertainty and geopolitical shocks have the greatest influence on the feature importance list. When these particular words occur in a news article, they significantly increase the model's prediction of a high-risk event. Furthermore, we analyze these mean absolute SHAP values for the whole data set in Figure 6, thus giving a more precise quantitative understanding of the importance of these high-effect tokens.

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

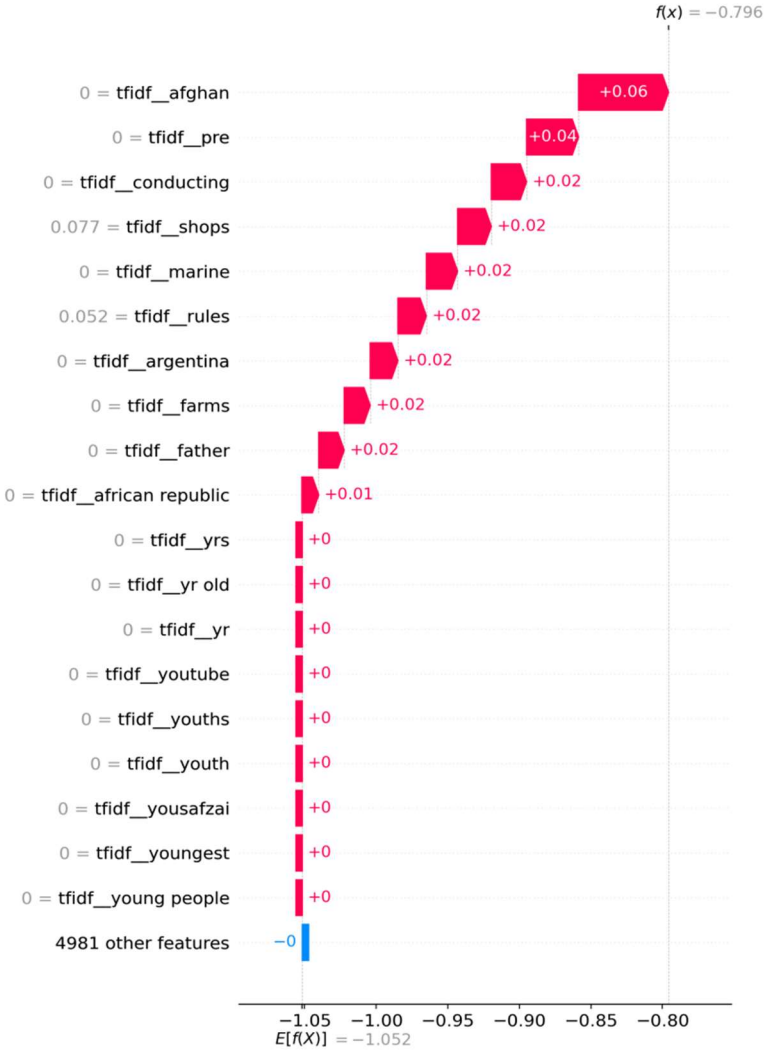

Figure 7. SHAP waterfall plot: feature contributions for a single sample

5. Discussion

5.1. Performance Analysis: SVM vs. XGBoost

Linear models like SVM have good performance for high-dimensional and sparse features, for example, the TF-IDF matrices. To clarify with an actual case, our TF-IDF method produces a vocabulary consisting of 5,000 words. In a short news article, only approximately 20 words are employed while the other 4,980 are empty (sparse). SVM acts as a basic measuring tool. It assigns a fixed score to each word (e.g., "crisis" is given +5 points of danger). This strategy is credible and avoids overfitting when the number of samples is small.

On the other hand, tree-based models like XGBoost need much larger data volumes to find clear data patterns in sparse environments. XGBoost acts like a complex investigator. It tries to find complicated combinations, such as "if word A appears AND word B is missing.

5.2. Limitations

The dataset includes approximately 1,985 observations, and classifying the data into binary labels with a hard threshold may lead to data errors.

Second, extraordinary fluctuations happen in the real market, causing an imbalance of data and affecting the accuracy rate. The TF-IDF method can only count the frequency of words and cannot evaluate the real semantic meanings or contexts. For instance, in a news report stating

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

"The stock is not crashing," the TF-IDF may give an inappropriate high score to the word "crashing." It is also incapable of understanding ironic sentences or intricate business regulations. This research only takes into account text signals and does not involve the traditional market data such as historical volatility and trading volume. It is similar to predicting the weather of tomorrow by merely reading the newspaper rather than using a thermometer. This selection might lead to a decrease in the highest prediction accuracy of our models.

6. Conclusion

We have established a comprehensive system for forecasting stock market volatility based on financial news texts. We also have applied Hadoop to process the data, which can not only reduce the computation time but also make the system suitable for handling large quantities of data in the future. The results of the machine learning indicate that the news texts have important indicators for recognizing high-volatility days in the stock market. Furthermore, when we incorporated sentiment scores and entropy features, the performance of the models improved.

SHAP analysis helped us to determine exactly which words led to the alarms. This suggests that our model does not make arbitrary predictions; instead, the results are reasonable in a practical economic context.

For the upcoming research, there are still some aspects that can be enhanced and improved. We intend to use advanced deep learning models such as FinBERT to interpret the actual meanings of the sentences instead of merely counting the words. We would like to integrate these text characteristics with real numerical information which may be the things like trading volume and previous prices. Developing such a multimodal system will be the subsequent important task to make our tool practically beneficial for real-world risk management.

References

[1] Deng, S. M., & Liu, P. P. (2019). The impact of attention heterogeneity on stock market in the era of big data. *Cluster Computing*, 22(3), 6157–6170.

[2] Hautsch, N., Schaumburg, J., & Schienle, M. (2015). Financial network systemic risk contributions. *Review of Finance*, 19(2), 685–738. https://doi.org/10.1093/rof/rfu010

[3] Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171–185.

[4] Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225.

[5] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

[6] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65.

[7] Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272–280.

[8] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653–7670.

[9] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107–113.

[10] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1–10.

ISSN: 2692-7608 DOI: 10.6981/FEM.202607_7(7).0011

[11] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

[12] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[13] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.