Can central bankers' talk predict bank stock returns? A machine learning approach

Katsafados, Apostolos G. and Leledakis, George N. and Panagiotou, Nikolaos P. and Pyrgiotakis, Emmanouil G.

Department of Accounting and Finance, School of Business, Athens University of Economics and Business, Greece, Department of Accounting and Finance, School of Business, Athens University of Economics and Business, Greece, Department of Accounting and Finance, School of Business, Athens University of Economics and Business, Greece, Essex Business School, University of Essex, U.K.

October 2024

Can central bankers' talk predict bank stock returns? A machine learning approach

by

Apostolos G. Katsafados[1], George N. Leledakis[1*], Nikolaos P. Panagiotou[1], Emmanouil G. Pyrgiotakis[2]

[1] Department of Accounting and Finance, School of Business, Athens University of Economics and Business, Greece

[2] Essex Business School, University of Essex, U.K.

Abstract

We combine machine learning algorithms (ML) with textual analysis techniques to forecast bank stock returns. Our textual features are derived from press releases of the Federal Open Market Committee (FOMC). We show that ML models produce more accurate out-of-sample predictions than OLS regressions, and that textual features can be more informative inputs than traditional financial variables. However, we achieve the highest predictive accuracy by training ML models on a combination of both financial variables and textual data. Importantly, portfolios constructed using the predictions of our best performing ML model consistently outperform their benchmarks. Our findings add to the scarce literature on bank return predictability and have important implications for investors.

JEL classification: C63, E58, G17, G21, G40

Keywords: Bank stock prediction; Trading strategies; Machine learning; Press conferences; Natural language processing; Banks

This version: November, 2024

[*] Corresponding author: Department of Accounting and Finance, School of Business, Athens University of Economics and Business, 76 Patission Str., 104 34, Athens, Greece; Tel.: +30 210 8203459. E-mail addresses: katsafados@aueb.gr (A. Katsafados), gleledak@aueb.gr (G. Leledakis), panagiotou@aueb.gr (N. Panagiotou), e.pyrgiotakis@essex.ac.uk (E. Pyrgiotakis). George Leledakis greatly acknowledges financial support received from the Research Center of the Athens University of Economics and Business (EP-3744-01). All remaining errors and omissions are our own.

1. Introduction

Forecasting equity returns is a thoroughly examined topic in the finance literature with studies focusing either on market returns (Brock et al., 1992; Campbell and Thompson, 2008; Neely et al., 2014) or individual stock returns (Lee and Swaminathan, 2000; Jegadeesh and Titman, 2002; Boudoukh et al., 2007). A key takeaway from this literature is that forecasting stock returns is particularly challenging, albeit not impossible (Rapach and Zhou, 2013). Unsurprisingly, the focus of equity forecasting literature has predominantly been on non-financial sectors, as banks are often excluded due to their unique characteristics (e.g. high leverage and heavy regulation). At the same time, there is a plethora of studies that examine the determinants of bank-stock returns (Baek and Bilson, 2015; Carmichael and Coën 2018; Venmans, 2021). This literature, however, typically focuses on in-sample statistics which are not widely regarded as reliable indicators of a model's forecasting ability (Bossaerts and Hillion, 1999). Consequently, to date, evidence of out-of-sample predictability of bank stock returns is rather limited, with the exception of a few studies (Cooper et al., 2003; Baele et al., 2015).

We investigate the predictability of bank stock returns by combining advanced machine learning (ML) algorithms with textual analysis techniques. In recent years, interest in applying machine learning techniques to financial time series forecasting has surged, due to the ML models' ability to process large datasets and capture non-linear patterns in the data (Leippold et al., 2022; Christensen et al., 2023; Jiang et al., 2024). While most studies rely on traditional financial variables to train the ML models, we enhance the approach by incorporating both financial variables and textual features. This is motivated by the capability of ML models to handle the high dimensionality of textual features. We construct textual features by applying textual analysis to the Federal Open Market Committee's (FOMC) conference calls, specifically, on the answers of the Chair (central banker) throughout the Q&A section.

We believe that the U.S. banking industry constitutes an interesting case to examine for three main reasons. First, as of October 2024, financial services firms make up 12.49% of the S&P 500 index[1], representing the second largest sector after technology firms. Second, recent literature shows that the predictability of bank classification tasks, such as failures or mergers, improves with the application of ML algorithms (Petropoulos et al., 2020), the integration of textual data (Gandhi et al., 2019), or a combination of both methods (Katsafados et al., 2024). Third, banks, unlike non-financial firms, have a common regulator (central banker) whose statements may directly influence bank stock returns.

We use a sample of 711 publicly-traded U.S. bank holding companies (BHCs) and commercial banks over the period 2011 to 2023 (25,808 bank-monthly observations). In our prediction task, we use the following machine learning models: (1) support vector regression (SVR), multilayer perceptron (MLP), and random forest (RF). We compare their performance against a traditional ordinary least squares (OLS) model, which serves as our benchmark. Our prediction horizon is one-month ahead of the FOMC press conference month.

Our main empirical approach is conducted in three steps. First, we examine the out-of-sample predictive ability of our ML models against the benchmark OLS using only financial variables as inputs. We find that MLP and RF models robustly outperform the OLS regression model. Second, we conduct the same analysis by replacing financial variables with textual features. The results show that textual features can be more informative than traditional financial variables. Third, we use both sources of data as inputs to our models. In line with our expectations, the combination of both sources of data produces the highest out-of-sample performance. RF is the best performing model, with a root mean square error of 0.0810. This figure exceeds those reported for aggregate market indices (Ferreira and Santa-Clara, 2011), However, this is expected, as market returns are less influenced by the noisy signals and the high idiosyncratic volatility associated with individual stocks.

[1] https://finance.yahoo.com/quote/SPY/holdings/

Next, we construct a portfolio using the predictions of our best performing model (RF trained with both financial variables and textual data). We do so, to examine whether the improvement in the out-of-sample performance can be translated to actual portfolio gains. We then compare this portfolio's return against both an all-bank portfolio and a portfolio based on OLS predictions. The results show that the RF portfolio significantly outperforms both benchmarks, with the return improvement ranging from 0.17% to as high as 14.42%. It is noteworthy that we report similar portfolio gains when we extend the prediction horizon from one month to three months after the FOMC press conference month.

One common criticism regarding the application of ML models in finance prediction task is that they are often viewed as "black boxes" (Zhao and Hastie, 2021). This means that the ML models' internal processes are not observable, which complicates our understanding of how input variables are transformed into output predictions. Therefore, we attempt to open the black box by shedding more light on the following two questions: (1) which variables (and in what order) contribute to improved forecasting ability, and (2) which textual features are more informative in our prediction task. To address the first question, we use the Shapley additive explanations feature importance approach (SHAP), which ranks each variable according to its contribution to the prediction task (Hansen, 2021). The rankings show that textual variables are the second most important inputs, after the federal funds rate. To address the second question, we employ the Local Interpretable Model-agnostic Explanations (LIME) method, which shows the textual features that contribute the most in improved model performance. The results are in line with our intuition. For instance, textual features relating to expectations about economic growth (inflation) have a positive (negative) impact on future bank returns.

We focus on constructing textual at the macro-level rather than the bank-level for two main reasons. First, central bankers' speeches typically set the expectations for monetary policy, interest rates changes, and the overall macroeconomic outlook. Considering that stock prices contain a forward-looking component, these macro-level insights can significantly impact market valuations and expected returns. Second, while bank-level textual features could provide valuable insights into reducing banks' inherent opacity (Katsafados et al., 2024) and evaluating their financial condition (Gandhi et al., 2019), their construction would be more computationally intensive and less replicable for the average investor. Additionally, this approach would increase the complexity and dimensionality of the data without necessarily offering more informative insights than the macro-level features. In our prediction task the most important variables are the federal funds rate and textual features from central bankers' speeches.

We contribute to several strands of the literature. First, we add to the scare literature which examines the out-of-sample predictability of U.S. bank stock returns (Cooper et al., 2003; Baele et al., 2015). We show that the use of more sophisticated ML models in conjunction with textual analysis could substantially improve our ability to forecast bank stock returns. On these grounds, we also contribute to the growing strand of banking literature that utilizes these techniques to predict bank mergers (Katsafados et al., 2024) or bank stock price crashes (Gkoumas et al., 2024). Finally, our work is related to the literature which examines how central bank tone influences investor behavior (Schmeling and Wagner, 2016; Dossani, 2021).

The remainder of the paper is structured as follows. Section 2 underlines the literature review. Section 3 outlines our data collection process. Section 4 describes our methodology, namely our employed ML models and our performance evaluation criteria. Section 5 discusses our results, and Section 6 concludes the paper.

2. Literature review

Previous literature of asset pricing studies in the banking industry focus on the effects of accounting variables on stock returns. Following Fama and French (1992), Barber and Lyon

(1997) find a significant connection between the size and the book to market, and stock returns in financial firms. They provide evidence that these results are similar across financial and non-financial firms. Schuermann and Stiroh (2006) suggest similar findings regarding the variables that could explain the stock returns for U.S. bank holding companies. They provide evidence that the market risk factor and the factors documented in Fama and French (1992, 1993) have significant effects in bank stock returns.

On the other hand, Cooper et al. (2003) could not find similar results regarding the effects of book to market and size on bank stock returns. They use bank specific variables and find that the ratios of non-interest income to net income, loan-loss reserves to total loans, earnings per share, book value to total assets and standby letters of credit to total loans can explain the cross-section of bank stock returns. In the same way, Viale et al. (2009) study a range of different models documented in the literature, such as the CAPM and the Fama-French three factor model. They highlight that the only significant factor is the shock to the yield curve in explaining the financial firms' stock returns.

Baek and Bilson (2015) support that the size and value factors are statistically significant in pricing the stock returns both for financial and non-financial companies. Baele et al. (2015) document that among 12 different independent variables, the high-minus-low factor (HML) could explain the returns of bank holding companies. Gandhi and Lustig (2015) search the tail risk of large commercial banks in the U.S. They provide evidence that large banks that are "too big to fail" have lower returns from smaller financial firms. Similarly, Venmans (2021) provides results of the effects of capital ratio on the returns of financial firms and documents that banks with lower capital tend to have lower returns.

In addition, several studies highlight the association between soft information and stock returns. Word lists have frequently been employed to gauge the pessimistic or optimistic tone in newspaper articles, earnings calls, annual reports, IPO prospectuses, and press releases

(Tetlock, 2007; Loughran and McDonald, 2011; Price et al., 2012; Doran et al., 2012; Garcia, 2013; Davis et al., 2015; Loughran and McDonald, 2016, 2020; Cohen et al., 2020; Katsafados et al., 2021; 2023b; Ardia et al., 2022; Gorodnichenko et al., 2023; Gkoumas et al., 2024).

In particular, this paper relates to the growing literature that uses textual features, individually or along with financial variables, when conducting the prediction task. More specifically, Hagenau et al. (2013) employ financial news to predict stock prices. Moreover, Tang et al. (2020) take into account a combination of financial and textual variables to predict financial distress. Next, Beaupain and Girard (2020) examine official transcripts of the press conferences of the ECB, denoting their significance when exerting monetary policy. In addition to economic variables, Zhao et al. (2022) employ sentiment scores to predict financial distress. More recently, Katsafados et al. (2023a) focus on IPO prospectuses in order to predict IPO underpricing while Katsafados et al. (2024) use annual reports in a merger prediction task.

3. Data and textual analysis

3.1. Sample selection and financial data

The dataset consists of all the entities included in Compustat Bank Fundamentals Quarterly dataset between 2011 and 2023. This database provides data for all financial companies incorporated in the U.S. Compustat provides the CUSIP number for all companies, which is a useful key, that we use to merge the firms with the Center for Research in Security Prices (CRSP). To get the maximum available number of observations, we further try to match the firms' Tickers between Compustat and CRSP.

We collect our data to form our dataset, from various sources. First, we gather all conference reports available in the Federal Reserve Board website.[2] We begin our dataset from the first available Federal Open Market Committee (FOMC) Press Conference in April 2011. Second, from the CRSP, we download the monthly stock returns from May 2011 to December 2023. Third, we gather accounting information from Compustat (Bank Fundamentals Quarterly) database, constructing bank-specific variables. Further, we download the effective federal funds rate (EFFR) from the Federal Reserve Bank of New York database.[3]

We construct eleven independent variables, which embody accounting and market information, and are usually incorporated in the related academic research (Fama and French, 1992; Barber and Lyon, 1997; Cooper et al., 2003; Mohanram et al., 2018). The data used to compute the ratios come from the databases described above.

Specifically, in our research we use the following control variables: (1) Loan loss provisions (LLP) formed as the ratio of loan loss provision to total loans, (2) equity to total assets (ETA), (3) earnings to price (EP), (4) loans to total assets (LTA), (5) non-interest income to total income (NII), (6) return on assets (ROA), (7) cost efficiency as the non-interest expenses to total income (COST_EF), (8) Tobin's Q as the ratio of the market value of the firm (obtained as total assets plus the market value of equity minus the book value of equity) to total assets (TOBINQ), (9) book value of equity to market value of equity (BM) and (10) natural logarithm of market value of equity (MVE).[4] All variables are winsorized, at 1% and 99% level to deal with any outlier effects.

Additional to the accounting variables we use the (11) effective federal funds rate (EFFR) as a control variable. We measure the monthly rate from daily data as the return between the rates of the last day every month. The speakers in FOMC Press Conferences provide evidence

[2] https://www.federalreserve.gov/default.htm

[3] https://www.newyorkfed.org/markets/reference-rates/effr

[4] We use the book value of firm's equity each quarter, divided by the market value, at the same month. We measure the market value of equity as the lagged one-month natural logarithm of market value of equity.

on monetary policy decisions and, among other, announce changes in the interest rates. To assess whether variations in interest rates could diminish the predictive power of textual features, we include the federal funds rate as an independent variable.

Consistent with Cooper et al. (2003), we assume the accounting data as public information starting two months after the Compustat quarterly report period, to confirm its availability to the market. We then match the returns from CRSP with the financial variables from Compustat. This approach ensures that accounting variables are available prior to the date of the returns they aim to explain.

For example, the quarterly data from Compustat of March 31 (Q1) would be assumed available to the public at the end of May and will be used as lagged accounting data for June, July and August returns. Our final sample includes 711 banks (25,808 bank-monthly observations) with available data in both Compustat Bank Fundamentals Quarterly and CRSP databases. Table 1 reports the summary statistics of all variables while Table 2 denotes the correlations among the variables.

*****Insert Table 1 & 2*****

3.2. Textual data

Our sample ranges from 2011 (when the first press conference was held) to September 2023. During this period, 66 press conferences were held. For each meeting, the FOMC statement and the transcript of the press conference are obtained from the Federal Reserve (Fed) website. In line with Gorodnichenko et al. (2023), our analysis focuses on the answers of the Chair throughout the Q&A section.

Next, we proceed to the pre-processing, which could potentially influence the performance of any machine learning algorithm according to the textual analysis literature (Nassirtoussi et al., 2014; Kumar and Ravi, 2016). In particular, pre-processing contains a range of sub-processes where the raw text is transformed into meaningful inputs for the machine learning models as described below.

As a first step, we remove all acronyms, abbreviations, single letter words, numbers, punctuation marks, and stop words (Gandhi et al., 2019; Katsafados et al., 2021). This filtering procedure has the benefit of reducing the informational opaqueness of the textual inputs, which contributes to superior prediction performance. Furthermore, we consider a minimum occurrence threshold to exclude words with low frequency (Schumaker and Chen, 2009; Katsafados et al., 2024). Such screening process is commonly used in textual analysis research since it limits the dimensionality of the models. Consistent with Mai et al. (2019), we take into account the 20,000 most frequent words of the press conference corpus. In essence, an excessive number of textual features could decrease the performance of any machine learning algorithm thus producing inferior results (Pestov, 2013).

Second, we convert our textual information into numerical data that a learning algorithm can understand (Mai et al., 2019). In particular, we implement the bag of words (BOW) approach to convert our unstructured textual data into inputs with explicit numerical structure (Katsafados et al., 2023a). Based on the BOW, we tokenize text into words using the Natural Language Toolkit (NLTK). In fact, we consider each unique word as a separate feature, and we generate a document-term matrix, where each row and column represent a document and a word, respectively. In such case, the value of each cell of the matrix is the value of the corresponding word feature in the particular document (Kumar and Ravi, 2016).

Finally, we compute the values of the features, where we represent each textual feature with a numeric value. To do so, we follow the popular term frequency (TF) weighting scheme normalized by document length. The TF calculates the proportion of each word in each document and assigns equal weight to each of them. The mathematical formulation for a word i in document j is as follows:

$$TF_{ij} = \frac{c_{ij}}{T_j} \tag{1}$$

where c_{ij} is the number of appearances of word i in document j, and T_j is the total number of words of document j (Katsafados and Anastasiou, 2024).

Given that we proceed to a prediction task, the independent variables should lag in time concerning the dependent variable. In this regard, we map the textual features emerging from the press conferences with the bank stock returns of the next month. For example, when the date of a press conference is within June, the textual features are linked with the monthly returns of July of the same year.

4. Models

To perform our prediction regression task, we employ the ordinary least squares (OLS) model as the benchmark for our empirical analysis. Furthermore, we use the following machine learning algorithms: (1) support vector regression, (2) multilayer perceptron, and (3) random forest. We choose these specific machine learning models, as they are not only frequently used in many prediction tasks in finance (Katsafados et al., 2023a), but also are less computationally expensive compared to complicated deep learning models. The hyper-parameters of the models are tuned via a grid search process using the 5-fold cross-validation performance of the training set. In this section, we will briefly underline the details of these predictive models.[5]

4.1. Support vector regression

Support vector machine (SVM) is a machine learning algorithm developed by Vapnik (1998). SVM has been widely-used in various tasks in finance, such as IPO underpricing prediction (Quintana et al., 2017; Katsafados et al., 2023a), merger prediction (Katsafados et al., 2024), and bankruptcy prediction (Veganzones and Severin, 2018; Mai et al., 2019). SVM typically is a supervised linear classifier, which generates a decision boundary that has the

[5] In all our models, the financial variables are standardized. Textual features are also standardized when they are combined with financial variables.

form of a hyperplane in the original feature space. The training samples at the boundaries of the margin, or (when allowing ‘slack’ in the separation) inside the margin, or on the wrong side of the hyperplane are called support vectors. Searching for the maximum margin hyperplane is conducted through quadratic programming optimization.

SVM can also be used in a regression task (Khashanah and and Shao, 2022). In our bank stock return prediction task, we adopt the support vector regression (SVR), which practically is a simple regression-based variation of SVM with some minor differences (Nassirtoussi et al., 2014). In particular, the optimization process of the SVR is similar to the SVM, with the difference that is trained on actual observed values. SVR can implement non-linear kernel function. If there are data with the non-linear structure, they are projected into a higher dimensional space representation, thus becoming more separable (Kumar and Ravi, 2016). In this regard, it is common to use the SVM with non-linear kernel functions such as the radial basis function kernel (RBF). Notably, some papers employ the SVR model to handle textual data in their prediction task (Schumaker et al., 2012; Hagenau et al., 2013).

Defining a set of data points $= \{(x_i, d_i)\}\,{}_{i}^{n}$, with x_i being the input vector, d_i representing the desired value, and n standing for the number of data patterns, SVMs approximate the relationship between the input variables and the target variable as follows:

$$y = f(x) = rm(x) + b \tag{2}$$

where $f(x)$ corresponds to the high dimensional feature space, which is non-linearly mapped from the input space x. The estimation of the coefficients r and b is achieved through the minimization of the following:

$$\mathrm{R}_{\mathrm{SVMs}}(C) = \mathrm{C}\frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{k} \mathrm{L}_{\varepsilon}(d_i, y_i) + \frac{1}{2}\|r\|^2 \tag{3}$$

$$L_{\varepsilon}(d, y) = \begin{cases} |d-y| - \varepsilon |d-y| \geq \varepsilon \\ \quad 0 \quad otherwise \end{cases} \tag{4}$$

The first term $\mathrm{C}\frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{L}_{\varepsilon}(d_i, y_i)$ of the regularized function (3) denotes the empirical risk.

C is the regularized constant and it influences the trade-off between the regularization term and the empirical risk. ε stands for the tube size, which is analogous to the approximation accuracy of the training data points. The next equation (4) represents the loss function, which allows the utilization of sparse data points for the decision function (2). On the other hand, $\frac{1}{2}\|r\|^2$ corresponds to the regularization term.

Then, equation (3) is converted into a primal function, while positive slack variables z_i and z_i^* are also introduced below:

$$\text{Minimize} \qquad \mathrm{R_{SVMs}}(C) = \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{n}}(z_i + z_i^*) + \frac{1}{2}\|r\|^2 \tag{5}$$

$$\text{Subjected to} \qquad \mathrm{d_i} - rf(x_i) - b_i \leq \varepsilon + z_i,$$

$$rf(x_i) + b_i - d_i \leq \varepsilon + z_i^*,$$

$$z^* \geq 0, \tag{6}$$

Finally, taking advantage of the Lagrange multipliers and the optimality constraints, as Tay and Cao (2001) exhibit, equation (3) transforms:

$$g(x, a_i, a_i^*) = \sum_{i=1}^{n}(a_i - a_i^*)K(x, x_i) + b \tag{7}$$

where a_i and a_i^* constitute the Lagrange multipliers, which are estimated by maximizing equation (6) as follows:

$$R(a_i, a_i^*) = \sum_{i=1}^{n} d_i(a_i - a_i^*) - \varepsilon\sum_{i=1}^{n}(a_i + a_i^*) - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}(a_i - a_i^*)\left(a_j - a_j^*\right)K(x_i, x_j) \tag{8}$$

subject to the constraints:

$$\sum_{i=1}^{n}(a_i - a_i^*) = 0,$$

$$0 \leq a_i \leq C, i = 1,2, \dots, n,$$

$$0 \leq a_i^* \leq C, i = 1,2, \dots, n.$$

To deal with overfitting problem, we should find the proper hyper-parameter value of the regularization parameter (C). In our empirical setting, we define C equal to 0.1. Moreover, given that we use the non-linear kernel RBF, there is another hyper-parameter known as gamma that controls for the curvature of the decision boundary. In our study, gamma is defined as 0.1; exception is the case where we use only financial variables where gamma is equal to 1.[6]

4.2. Multilayer perceptron

Artificial neural networks are a widely-used category of machine learning models especially in the domain of Natural Language Processing (Goldberg, 2017). In finance, one of the most popular artificial neural networks is the multilayer perceptron (MLP) model (Kumar and Ravi, 2016). In particular, MLP is frequently used for various prediction tasks (Mai et al., 2019; Ibrahim et al., 2022; Jiang et al., 2022; Katsafados et al., 2024). In addition to its popularity, we focus on the MLP models because they are more directly applicable to BOW text representations and at the same time more directly comparable to the other models.

In MLP architecture, there is initially an input layer of neurons, where our variables are fed as inputs into the network. Next, there are one or more hidden layers. When the hidden layer receives the content from the input layer, non-linear functions are activated before transferring the estimated values to the next hidden or lastly to the output layer. In the end, the output layer produces the predictive outcome based on the received input from the hidden layers.[7] Notably, all the hyper-parameters are tuned based on a grid search process using the 5-fold cross-validation performance of the training set. In our empirical setting, the optimal MLP model has 3 hidden layers, each of which has 200 neurons. Also, we use a rectified linear unit (ReLU) as the activation function of each hidden layer. In general, this kind of activation function among the various hidden layers can effectively handle non-linear relationships between independent and dependent variables.

[6] The hyper-parameters of our SVR models are tuned based on the 5-fold cross-validation performance of the training set.

[7] For training purposes, we apply a backpropagation algorithm. In addition, we employ Adam as the optimizer algorithm, and cross-entropy as the loss function. ReLU is defined as f(x) = max (0, x). Finally, we use early stopping to mitigate overfitting (Mai et al., 2019). To do so, we set aside 10% of training data as validation or development set.

4.3. Random forest

Random forest (RF) practically is an ensemble learning algorithm that is suitable for both regression and classification purposes. RF was introduced by Breiman (2001), which virtually is a variant of the Bagging ensemble learning method (Breiman, 1996). It is common that RF frequently outperforms the classical decision trees including CART, as it has slightly less possibility to over-fit to the training sample. RF generates plenty of uncorrelated Decision Trees (DTs) trained on bootstrap copies of original samples by randomly choosing a subset of features (Mai et al., 2019). Notably, some papers use RF models to incorporate textual features in their prediction tasks with superior performance (Mai et al., 2019; Katsafados and Anastasiou, 2024; Katsafados et al., 2023a; 2024).

To deal with overfitting, we optimize the five key hyper-parameters of the RF model: (1) the number of decision trees, (2) the number of features randomly chosen to grow each decision tree when searching for the best split (max_features), (3) the minimum number of samples required to be at a leaf node (min_samples_leaf), (4) the maximum depth of the trees (max_depth), and (5) the minimum required number of observations in any given node in order to split it (min_samples_split). Notably, all the hyper-parameters are tuned using a grid search process based on the 5-fold cross-validation performance of the training set. Apart from the randomization of training samples (bootstrap), the tuning optimization of max_features benefits the proper randomization of feature space, leading to a reduced variance (low overfitting).

In case of only textual and both textual and financial variables, the optimal amount of each hyper-parameter is 200 for the number of trees, 2 for max_features, 3 for min_samples_leaf, 90 for max_depth, and 10 for min_samples_split. However, when we use only financial variables, max_features and min_samples_leaf optimal numbers are 3 and 4 respectively.

4.4. Evaluation

Our approach is to evaluate the models based on their out-of-sample performance. Consistent with pertinent literature, we split our data by selecting 80% of our sample as the training set and the remaining 20% as the out-of-sample set (Doumpos et al., 2017; Mai et al., 2019). Noteworthy, we select the out-of-sample set from a future period rather than at random (Pasiouras and Tanna, 2010; Katsafados et al., 2024). In fact, the usefulness of a predictive model depends on its ability to correctly predict future values (Espahbodi and Espahbodi, 2003). In our framework, the evaluation methodology requires both an out-of-sample and an out-of-time. That is, we first sort the data by date, and only then we consider the first 80% as the training set while the rest as the testing set.

To evaluate the performance of the models, we first use the root mean square error (RMSE), which is commonly-used in similar prediction tasks (Quintana et al., 2017; Katsafados et al., 2023a). RMSE represents the square root of the average squared differences between predicted and observed values. The lower the value of RMSE, the higher the predictive ability of our models. In line with Birim et al. (2022), RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{v}\sum_{i=1}^{v}(\hat{y}_i - y_i)^2} \tag{9}$$

where i defines each bank-level observation in the testing set, v represents the overall amount of predictions (equal to the number of observations in the testing set), $\hat{y_i}$ is the vector of predicted values of bank stock returns, and y_i is the vector of observed values of bank stock returns.

Second, we use the mean square error (MSE) as an additional evaluation measure. It offers a measure of the average magnitude of errors, which increases its sensitivity to both overestimation and underestimation. MSE squares the differences between predicted and actual values, giving more weight to larger errors. In a sense, MSE is practically the RMSE after being raised to the square:

$$MSE = RMSE^2 = \frac{1}{v}\sum_{i=1}^{v}(\hat{y}_i - y_i)^2 \tag{10}$$

Finally, we employ the mean absolute error measure (MAE). Similarly, MAE measures the average absolute differences between predicted and observed values, leading to a clear picture of the average magnitude of errors. The mathematical formula behind MAE is presented:

$$MAE = \frac{1}{v}\sum_{i=1}^{v}|\hat{y}_i - y_i| \tag{11}$$

These three specific metrics are widely used in the related literature since numerous studies conduct their forecasting exercise using them (see among others, Pai and Lin, 2005; Anastasiou and Drakos, 2021; Anastasiou et al., 2022). On the one hand, the use of RMSE aligns with our objective of capturing both the magnitude and directionality of errors in predicting bank stock returns. Also, RMSE provides the advantage that is in the same unit as the dependent variable, making it more interpretable. On the other hand, the choice of MSE offers added value since it is quite sensitive to the magnitude of errors via squaring the differences between predicted and actual values. In the context of predicting bank stock returns, it is crucial to effectively capture the magnitude of errors. Otherwise, the consequences would be severe either for financial stability or for investors when shaping their portfolios. Finally, we also chose MAE as it offers a straightforward interpretation of the model's accuracy (Ftiti and Jawadi, 2019).

5. Results

5.1. Prediction using textual features and financial variables separately

Table 3 presents the predictive ability of our models regarding the bank stock returns using only financial variables whereas Table 4 denotes the performance of the same models including only textual features extracted from the press conferences of central bankers. As we observe when comparing the tables, each of the models perform better in case of using textual features (Table 4) than using financial variables (Table 3). Notably, this finding is robust across the three alternative evaluation measures of models' performance. That is, all models with only text predict with lower errors, thus achieving more accurate estimations. In this regard, we prove that the textual disclosures of press conferences have predictive power, indicating their high informational value.

When our attention shifts to the comparison among the models, we conclude that MLP and RF outperform the benchmarking OLS in the models using only financial variables. Nonetheless, the three models achieve quite similar performance according to the results in Table 4 where only textual features are employed. In general, SVR appears the worst performance in each case. Although we have proved the high informational value of text, next the vital question is whether this information is unique, and if so, to what extent it can complement the financial variables.

*****Insert Table 3 & 4*****

5.2. Prediction with both SVD-100 textual features and financial variables

In Table 5, we show the predictive ability of our models including both financial variables and textual features. In the empirical setting, we apply the singular value decomposition (SVD) dimensionality reduction technique (Degiannakis et al., 2018; Katsafados et al., 2023a). In practice, we use SVD to decrease the dimensions of the textual features from 20,000 to just 100. As a result, the 100 textual features concentrate all the available information derived from the text of press conferences. This process has the advantage of further limiting the curse of dimensionality as well as the plethora of textual features does not overrule the role of financial variables.

Yet, according to Table 5, we observe that the OLS and SVR models do not present improved performance compared with previous results. However, the interesting point here is that the RF and MLP achieve better out-of-sample performance including the combined input of textual features and financial variables than the models using a single type of input. Moreover, the RF exhibits the best predictive accuracy in comparison with all the other models. To sum up, the machine learning models could effectively combine textual data with financial information to improve the accuracy in the bank stock returns prediction task.

*****Insert Table 5*****

5.3. Portfolio strategies

In this section, we design an investment strategy based on the predictions of the best prediction model. Consistent with previous results, we consider the RF with both financial variables and textual features as the model with the best performance. The investment approach in this case is to identify bank stocks with a high probability of price increases, which translates to positive returns, thereby suggesting them as favorable investment choices. As a result, the percentage profit for each stock is equal to the magnitude of each return, assuming that we will sell the stock at the end of the first month after the press conference.

Similar to Dal Pra et al. (2018) and Katsafados et al. (2023a), we focus only on the out-of-sample predictions when evaluating the portfolio strategies. To begin with, we create 3 portfolio strategies; we invest in a stock if the predicted return: (1) is positive, (2) is over 5%, and (3) is over 10%. In fact, each portfolio includes all bank stocks that cover the aforementioned criterion with equal shares. Given that the return of the created portfolio is computed as the average returns of all bank stocks included in the portfolio. The mathematical formula is given as follows:

$$portofolio\ return = \sum_{i=1}^{N} stock\ return_i \tag{12}$$

where *i* is the bank stock included in the portfolio and *N* is the total number of chosen stocks.

According to our findings in Figure 1, the portfolios with 0%, 5%, and 10% thresholds correspond to 4.07%, 8.81%, and 15.69% average performance respectively. As a first benchmark, we consider the portfolio that contains the entire amount of bank stocks in the out-of-sample with equal weights. Such a portfolio has 1.27% performance on average. Consequently, our machine learning portfolio manages to substantially improve the average return by 2.80%, 7.54%, and 14.42% respectively (see Figure 1).

*****Insert Figure 1*****

Moreover, we use a second benchmark portfolio based on OLS predictions. According to the empirical findings, the OLS portfolios with 0%, 5%, and 10% thresholds correspond to 3.90%, 7.46%, and 13.04% average performance respectively. Hence, we observe that again the machine learning portfolios outperform the OLS with improved return by 0.17%, 1.35%, and 2.65% respectively (see Figure 2).

*****Insert Figure 2*****

These findings are in line with Cerniglia and Fabozzi (2020) and Katsafados et al. (2023a) since the authors highlight that the machine learning models frequently generate more accurate predictions compared to standard econometric methods. They do so because they can capture nonlinearities in data, grasp complex interactions among variables and allow the use of large, unstructured datasets.

5.4. Longer prediction horizons

It is crucial to examine the persistence of textual information over time. Therefore, we extend our study to explore how the models' prediction performance change as we increase the prediction horizon. Instead of connecting the textual features with the returns of next month, now we let two months as a gap between the month of the release of press conference and the stock returns. For example, when the date of a press conference is within June, then the textual features are linked with the monthly returns of September. Consistent with the previous results from one-month-ahead prediction horizon, the machine learning models continue to outperform the OLS and manage to improve their performance by combining textual features with financial variables (see Table 6 & 7).

*****Insert Table 6 & 7*****

To provide further evidence regarding the superiority of our machine learning models, we extend the time horizon of our portfolios as well. We implement such an empirical analysis for two reasons. First, it can prove the robustness of the results documented, and second, we can investigate how persistent is the improved performance of the portfolios over time. Figure 3 presents how the results are altered based on this modification.

According to our empirical findings, the OLS portfolios manage to 3.87%, 7.08%, and 12.17% average performance. On the other side, the RF portfolios achieve 4.01%, 8.10%, and 13.49% average performance, indicating an improvement by 0.14%, 1.03%, and 1.32% respectively. Based on the results, we make three inferences. First, the machine learning portfolios still perform better in the longer horizon than OLS. Second, the improved performance of RF over OLS is now limited, which means that there is a convergence over time. Finally, both RF and OLS portfolio have lower returns compared to the case of one-month window prediction task. The rationale behind this is that the predictions are severely influenced by the textual features. In longer periods, the information hidden in the texts has already incorporated into the market. In conclusion, machine learning portfolios perform better in all cases, but more significantly, in the short-term period.

*****Insert Figure 3*****

5.5. Textual transparency

5.5.1. SHAP methodology

Shapley additive explanations feature importance approach (SHAP) substantially uses the Shapley values from game theory to gauge the extent of contribution of each variable in the prediction process (Hansen, 2021). In this paper, we adopt the SHAP method in order to compare the comparative value of textual information compared to the financial variables. In essence, we implement SHAP to the RF model with both textual and financial variables as inputs. In fact, we employ the singular value decomposition (SVD) dimensionality reduction technique (Degiannakis et al., 2018; Katsafados et al., 2023a). More specifically, we use SVD to decrease the dimensions of our textual features from 20,000 to just 1. That is, the variable (TEXT) concentrates all available information derived from the press conferences of central bankers.

In Figure 4, we illustrate the SHAP importance scores for each variable in the RF model in order of importance. As a result, we show that the most influential variable is the EFFR. Nevertheless, the most interesting fact is that TEXT is the second-best variable in the bank stock return prediction task. To sum up, we provide evidence of high informative value of central bankers' press conferences, which should complement financial variables.

*****Insert Figure 4*****

In Figure 5, we report a bees-warm plot that sheds additional light on the relationship between the dependent variable (stock returns) and the independent variables. In fact, the independent variables in this plot are ranked by mean absolute SHAP values as in Figure 4. If the red color is on the left (right) side of the graph, it recommends that the variable has a negative (positive) impact on the dependent variable. Unsurprisingly, the direction of the relationship between textual information and bank stock returns depends on the content of the textual information. Textual features that convey a positive outlook are positively associated with returns (red values on the right side), while features that convey a negative outlook are negatively related with returns (red values on the left side).

*****Insert Figure 5*****

5.5.2. LIME approach

To further evaluate the importance of our textual features in our bank stock returns prediction task, we use the novel LIME method. This method has the advantage to be agnostic that practically explains the predictions of any predictive model (Ribeiro et al., 2016). Moreover, it provides swift results compared with SHAP method, fact that is beneficial in case of extremely large datasets. As in the case of SHAP, we apply LIME to the

RF model as it is the best-performing model. However, we now use bigram textual features as inputs into the RF model to achieve a better transparency of the results.

Table 8 reports the results of the LIME analysis. In particular, positive impact leads to higher bank stock returns whereas negative impact drives to the opposite effect. Notably, the results are consistent with our expectations. On the one hand, the bigrams with the positive impact are mainly related to growth, progress, and improvement ("growth going", "conditions improving", "job creation", "economic growth", "substantial improvement". In addition to phrases that describe positive prospects of economy, there are some phrases that imply action from central bank such as "use monetary", "provide additional", and "asset purchases". On the other hand, in the case of bigrams with negative impact, we observe phrases regarding inflation ("expect inflation", "inflation continues", "inflation pressures", "projected inflation") and others such as "risks associated", "bring unemployment", and "Ukraine Russia".

*****Insert Table 8*****

Table 9 demonstrates suggestive evidence on this point, by providing a number of parallel passages. Words with orange colour translate to positive impact and high returns, whereas those with blue colour imply negative impact and low returns. Note that the more intense the background shade, the stronger the effect.

For instance, you can see the following sentences: ("HIGHER COSTS BROADER PRICES CREATING BROADER INFLATION ECONOMY LONG INFLATION EXPECTATIONS" and "HOWEVER PERCENT UNEMPLOYMENT RATE REMAINS ELEVATED LOOKING AHEAD COMMITTEE ANTICIPATES UNEMPLOYMENT RATE WILL DECLINE GRADUALLY"). Obviously, these reflect negative impact, stemmed mainly by "inflation" and unemployment" words. On the other side, there are sentences with positive impact including words such as "growth" and "developments"

("GROWTH POSITIVE LASTED FIVE YEARS" and "PLANS WARRANTED ECONOMIC FINANCIAL DEVELOPMENTS").

*****Insert Table 9*****

6. Conclusions

We examine the predictability of U.S. bank stock returns using several machine learning algorithms. To improve the forecasting ability, we train our models using a combination of traditional financial variables and textual features constructed from FOMC's press releases. Our results provide several insightful conclusions. First, by benchmarking our ML models against OLS regressions, we show that both MLP and RF consistently outperform traditional econometric techniques. Second, we find that textual features can be more meaningful inputs that financial variables. Nevertheless, we show that the combination of both sources of data produces the best out-of-sample performance, with RF achieving the lowest prediction errors. In terms of predictive power, macro-level variables outperform bank-specific fundamentals. Specifically, the federal funds rate is the most influential variable in forecasting bank stock returns, followed by textual features from central bankers' press conferences. The textual features that contribute to this improved performance are also in line with intuition; language suggesting a positive (negative) economic outlook is associated with higher (lower) bank returns.

Our findings are economically meaningful for investors. Portfolios constructed using predictions from our best-performing ML model achieve higher substantially returns than both all-bank portfolios and those based on OLS predictions. Notably, this outperformance persists even over longer prediction horizons. Collectively, our study highlights the value of incorporating textual features in bank stock return predictions, as well as the advantage of using machine learning models that can effectively manage the complexity of such data.

References

Anastasiou, D., Ballis, A., & Drakos, K. (2022). Constructing a positive sentiment index for COVID-19: Evidence from G20 stock markets. *International Review of Financial Analysis*, 81, 102111.

Anastasiou, D., & Drakos, K. (2021). European depositors' behavior and crisis sentiment. *Journal of Economic Behavior & Organization*, 184, 117-136.

Ardia, D., Bluteau, K., & Boudt, K. (2022). Media abnormal tone, earnings announcements, and the stock market. *Journal of Financial Markets*, 61, 100683.

Baek, S., & Bilson, J. F. O. (2015). Size and value risk in financial firms. *Journal of Banking and Finance*, 55, 295-326.

Baele, L., De Bruyckere, V., De Jonghe, O., & Vander Vennet, R. (2015). Model uncertainty and systematic risk in US banking. *Journal of Banking and Finance*, 53, 49-66.

Barber, B. M., & Lyon, J. D. (1997). Firm size, book-to-market ratio, and security returns: A holdout sample of financial firms. *Journal of Finance*, 52, 875-883.

Beaupain, R., & Girard, A. (2020). The value of understanding central bank communication. *Economic Modelling*, 85, 154-165.

Birim, S., Kazancoglu, I., Mangla, S. K., Kahraman, A., & Kazancoglu, Y. (2022). The derived demand for advertising expenses and implications on sustainability: A comparative study using deep learning and traditional machine learning methods. *Annals of Operations Research*,1-31.

Bossaerts, P., & Hillion, P. (1999). Implementing statistical criteria to select return forecasting models: what do we learn? *Review of Financial Studies*, 12, 405-428.

Boudoukh, J., Michaely, R., Richardson, M., & Roberts, M. R. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing. *Journal of Finance*, 62, 877-915.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance*, 47, 1731-1764.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21, 1509-1531.

Carmichael, B., & Coën, A. (2018). Real estate as a common risk factor in bank stock returns. *Journal of Banking and Finance*, 94, 118-130.

Cerniglia, J. A., & Fabozzi., F. J. (2020). Selecting computational models for asset management: Financial econometrics versus machine learning—Is there a conflict? *Journal of Portfolio Management*, 47, 107-118.

Christensen, K., Siggaard, M., & Veliyev, B. (2023). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics*, 21, 1680-1727.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *Journal of Finance*, 75, 1371-1415.

Cooper, M. J., Jackson III, W. E., & Patterson, G. A. (2003). Evidence of predictability in the cross-section of bank stock returns. *Journal of Banking and Finance*, 27, 817-850.

Dal Pra, G., Guidolin, M., Pedio, M., & Vasile., F. (2018). Regime Shifts in Excess Stock Return Predictability: An Out-of-Sample Portfolio Analysis. *Journal of Portfolio* Management, 44, 10-24.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20, 639-673.

Degiannakis, S., Filis, G., & Hassani, H. (2018). Forecasting global stock market implied volatility indices. *Journal of Empirical Finance*, 46, 111-129.

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings conference call content and stock price: The case of REITs. *Journal of Real Estate Finance and Economics*, 45, 402-434.

Dossani, A. (2021). Central bank tone and currency risk premia. *Journal of International Money and Finance*, 117, 102424.

Doumpos, M., Andriosopoulos, K., Galariotis, E., Makridou, G., & Zopounidis, C. (2017). Corporate failure prediction in the European energy sector: A multicriteria approach and the effect of country characteristics. *European Journal of Operational Research*, 262, 347-360.

Espahbodi, H., & Espahbodi, P. (2003). Binary choice models for corporate takeover. *Journal of Banking and Finance*, 27, 549-574.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47, 427-465.

Fama, E. F., & French, K. R. (1993). Common risk factors in returns on stocks and bonds. *Journal of Financial Economics*, 33, 3-56.

Ferreira, M. A., & Santa-Clara, P. (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics*, 100, 514-537.

Ftiti, Z., & Jawadi, F. (2019). On the oil price uncertainty. *Energy Journal*, 40, 19-40.

Gandhi, P., & Lustig, H. (2015). Size anomalies in US bank stock returns. *Journal of Finance*, 70, 733-768.

Gandhi, P., Loughran, T., & McDonald, B. (2019). Using annual report sentiment as a proxy for financial distress in US banks. *Journal of Behavioral Finance*, 20, 424-436.

Garcia, D. (2013). Sentiment during recessions. *Journal of Finance*, 68, 1267-1300.

Gkoumas, N. C., Leledakis, G. N., Pyrgiotakis, E. G., & Androutsopoulos, I. (2024). Bank competition, loan portfolio concentration and stock price crash risk: The role of tone ambiguity. *British Journal of Management*, forthcoming.

Goldberg, Y. (2017). Neural network methods for natural language processing. Morgan & Claypool Publishers.

Gorodnichenko, Y., Pham, T., & Talavera, O. (2023). The voice of monetary policy. *American Economic Review*, 113, 548-584.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55, 685-697.

Hansen, J. (2021). Diabetic risk prognosis with tree ensembles integrating feature attribution methods. *Evolutionary Intelligence*, 1-10.

Ibrahim, B. A., Elamer, A. A., & Abdou, H. A. (2022). The role of cryptocurrencies in predicting oil prices pre and during COVID-19 pandemic using machine learning. *Annals of Operations Research*, 1-44.

Jegadeesh, N., & Titman, S. (2002). Cross-sectional and time-series determinants of momentum returns. *Review of Financial Studies*, 15, 143-157.

Jiang, M., Jia, L., Chen, Z., & Chen, W. (2022). The two-stage machine learning ensemble models for stock price prediction by combining mode decomposition, extreme learning machine and improved harmony search algorithm. *Annals of Operations Research*, 309, 553-585.

Jiang, F., Ma, T, & Zhu, F. (2024). Fundamental characteristics, machine learning, and stock price crash risk. *Journal of Financial Markets*, 69, 100908.

Katsafados, A. G., & Anastasiou, D. (2024). Short-term prediction of bank deposit flows: Do textual features matter? *Annals of Operations Research*, 1-26.

Katsafados, A. G., Androutsopoulos, I., Chalkidis, I., Fergadiotis, E., Leledakis, G. N., & Pyrgiotakis, E. G. (2021). Using textual analysis to identify merger participants: Evidence from the US banking industry. *Finance Research Letters*, 42, 101949.

Katsafados, A. G., Androutsopoulos, I., Chalkidis, I., Fergadiotis, E., Leledakis, G. N., & Pyrgiotakis, E. G. (2023a). Textual information and IPO underpricing: A machine learning approach. *Journal of Financial Data Science*, 5, 100-135.

Katsafados, A. G., Leledakis, G. N., Pyrgiotakis, E. G., Androutsopoulos, I., & Fergadiotis, M. (2024). Machine learning in bank merger prediction: A text-based approach. *European Journal of Operational Research*, 312, 783-797.

Katsafados, A. G., Nikoloutsopoulos, S., & Leledakis, G. N. (2023b). Twitter sentiment and stock market: a COVID-19 analysis. *Journal of Economic Studies*, 50, 1866-1888.

Khashanah, K., & Shao, C. (2022). Short-term volatility forecasting with kernel support vector regression and Markov switching multifractal model. *Quantitative Finance*, 22, 241-253.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114, 128-147.

Lee, C. M., & Swaminathan, B. (2000). Price momentum and trading volume. *Journal of Finance*, 55, 2017-2069.

Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, 145, 64-82.

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

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54, 1187-1230.

Loughran, T., & McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, 12, 357-375.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274, 743-758.

Mohanram, P., Saiy, S. & Vyas, D. (2018). Fundamental analysis of banks: The use of financial statement information to screen winners from losers. *Review of Accounting*

Studies, 23, 200-233.

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

Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G. (2014). Forecasting the equity risk premium: the role of technical indicators. *Management Science*, 60, 1772-1791.

Pai, P. F., & Lin, C. S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, 33, 497-505.

Pasiouras, F., & Tanna, S. (2010). The prediction of bank acquisition targets with discriminant and logit analyses: Methodological issues and empirical evidence. *Research in International Business and Finance*, 24, 39-61.

Pestov, V. (2013). Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Computers and Mathematics with Applications*, 65, 1427-1437.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, 36, 1092-1113.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking and Finance*, 36, 992-1011.

Rapach, D., & Zhou, G. (2013). Forecasting stock returns. In Handbook of Economic Forecasting (Vol. 2, pp. 328-383). Elsevier.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Quintana, D., Sáez, Y., & Isasi, P. (2017). Random forest prediction of IPO underpricing. *Applied Sciences*, 7, 636.

Schmeling, M., & Wagner, C. (2016). Does central bank tone move asset prices? *Journal of Financial and Quantitative Analysis*, 1-48.

Schuermann, T., & Stiroh, K. (2006). Visible and hidden risk factors for banks, Staff Reports 252, Federal Reserve Bank of New York.

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, 53, 458-464.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions of Information Systems*, 27, 1-19.

Tang, X., Li, S., Tan, M., & Shi, W. (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting*, 39, 769-787.

Tay, F. E. H., & Cao, L. J. (2001). Application of support vector machines in financial time series forecasting. *Omega*, 29, 309-317.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62, 1139-1168.

Vapnik, V. (1998). Statistical learning theory. (1st ed.). Wiley.

Veganzones, D., & Severin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, 112, 111-124.

Venmans, F. (2021). The leverage anomaly in U.S. bank stock returns. *Journal of International Financial Markets, Institutions and Money*, 75, 101425.

Viale, A. M., Kolari, J. W., & Fraser, D. R. (2009). Common risk factors in bank stocks. *Journal of Banking and Finance*, 33, 464-472.

Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business and Economic Statistics*, 39, 272-281.

Zhao, S., Xu, K., Wang, Z., Liang, C., Lu, W., & Chen, B. (2022). Financial distress prediction by combining sentiment tone features. *Economic Modelling*, 106, 105709.

Figure 1: RF portfolio returns vs Average portfolio returns

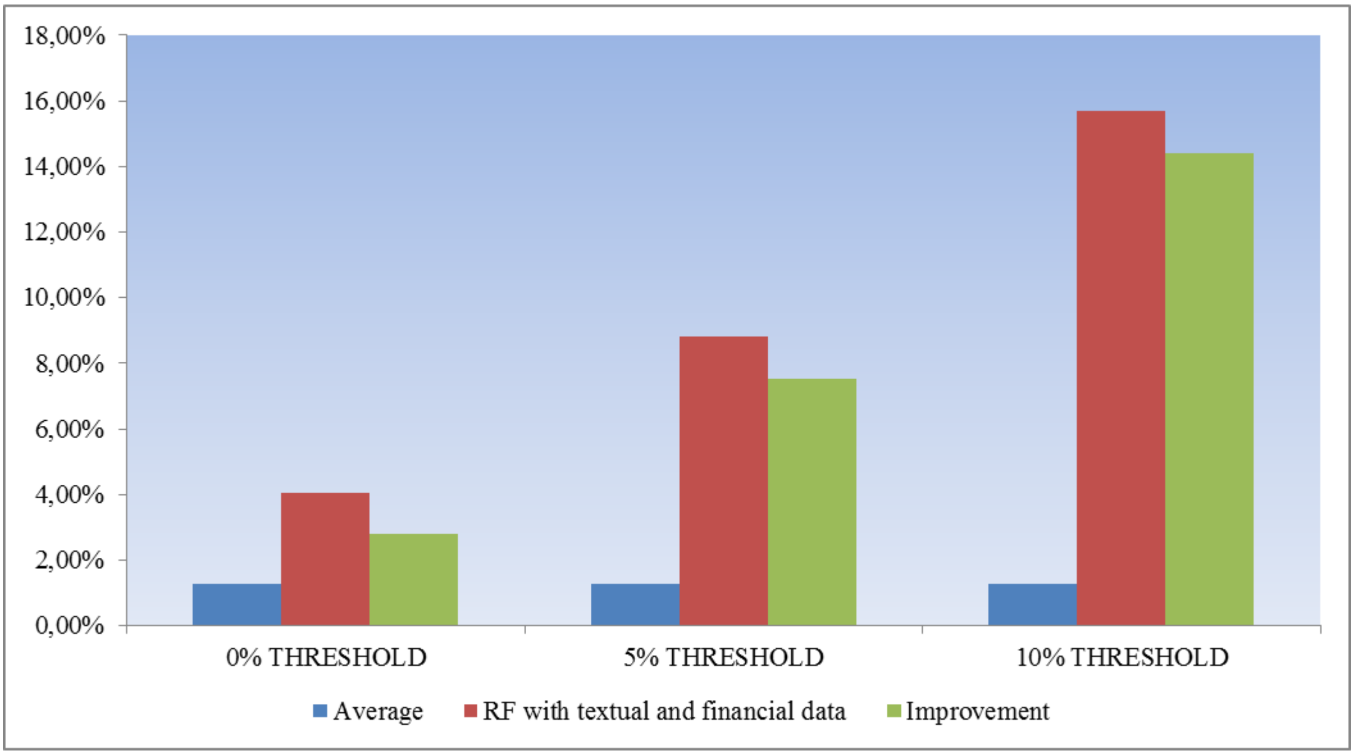

Figure 2: RF portfolio returns vs OLS portfolio returns

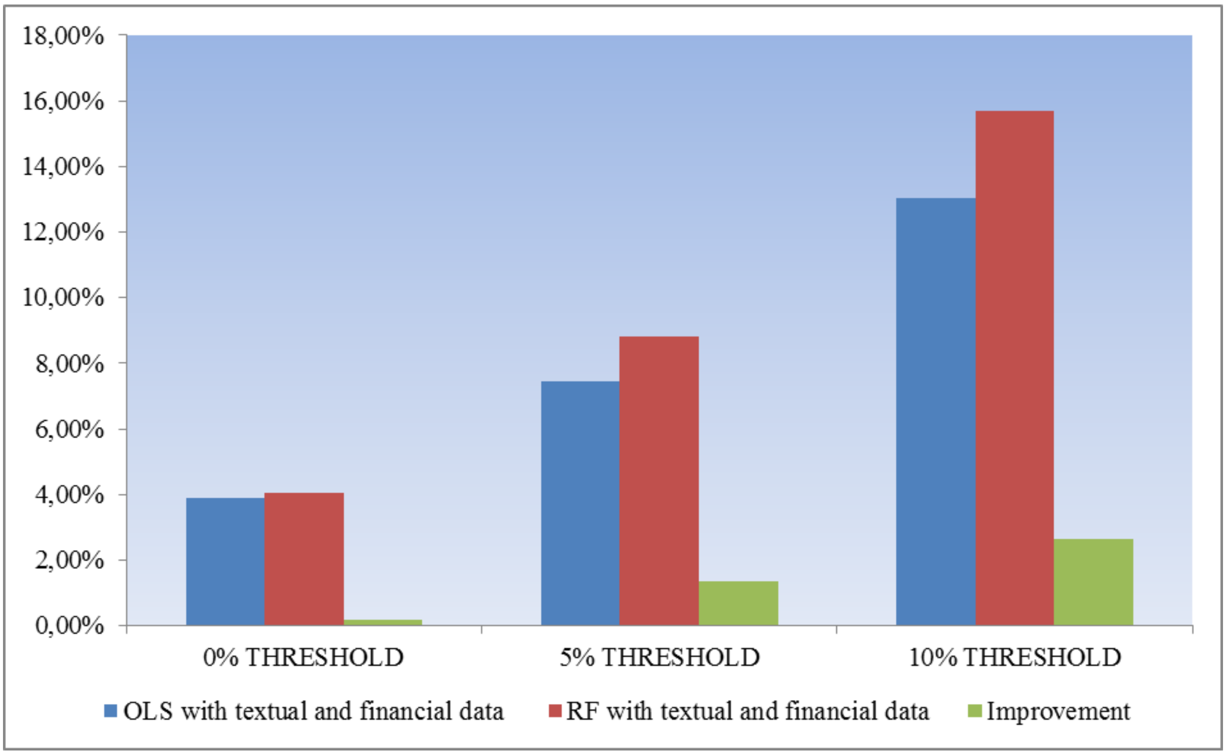

Figure 3: RF portfolio returns vs OLS portfolio returns (3-month window)

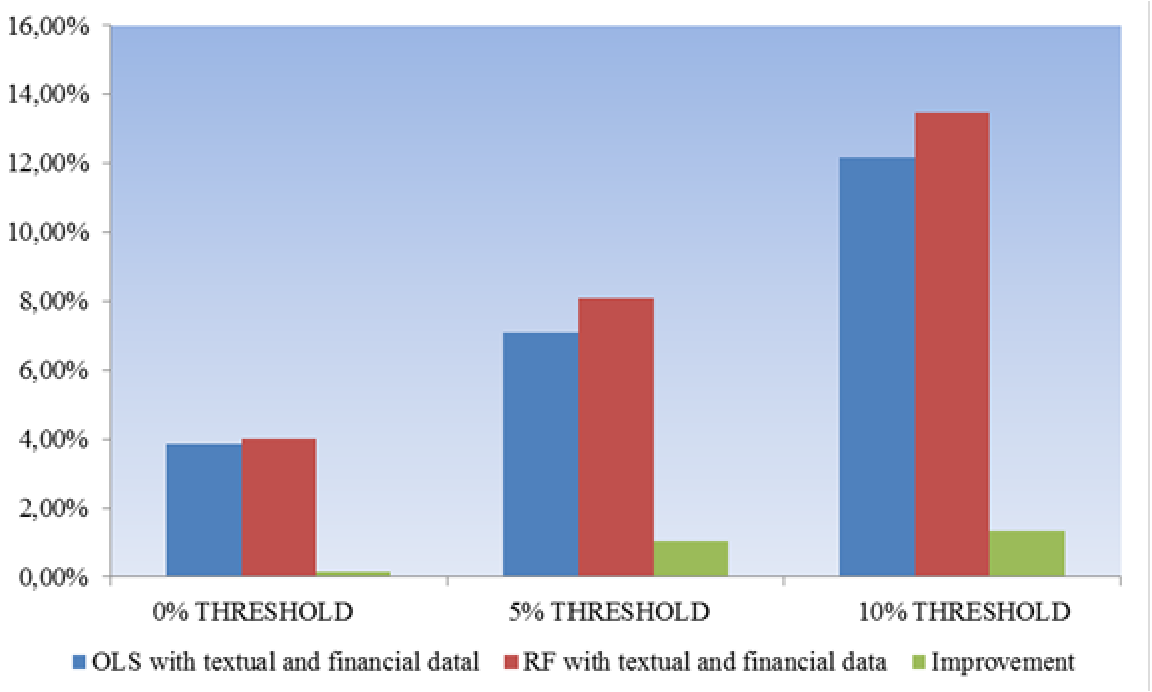

Figure 4: SHAP feature importance scores

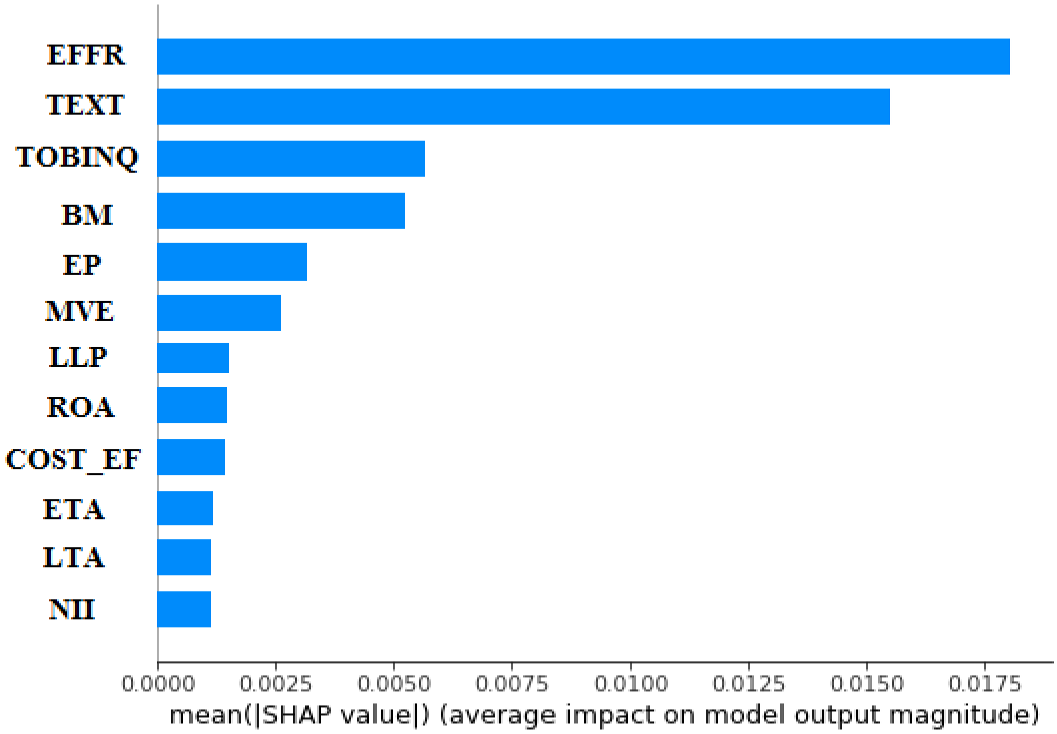

Figure 5: Bees-warm plot

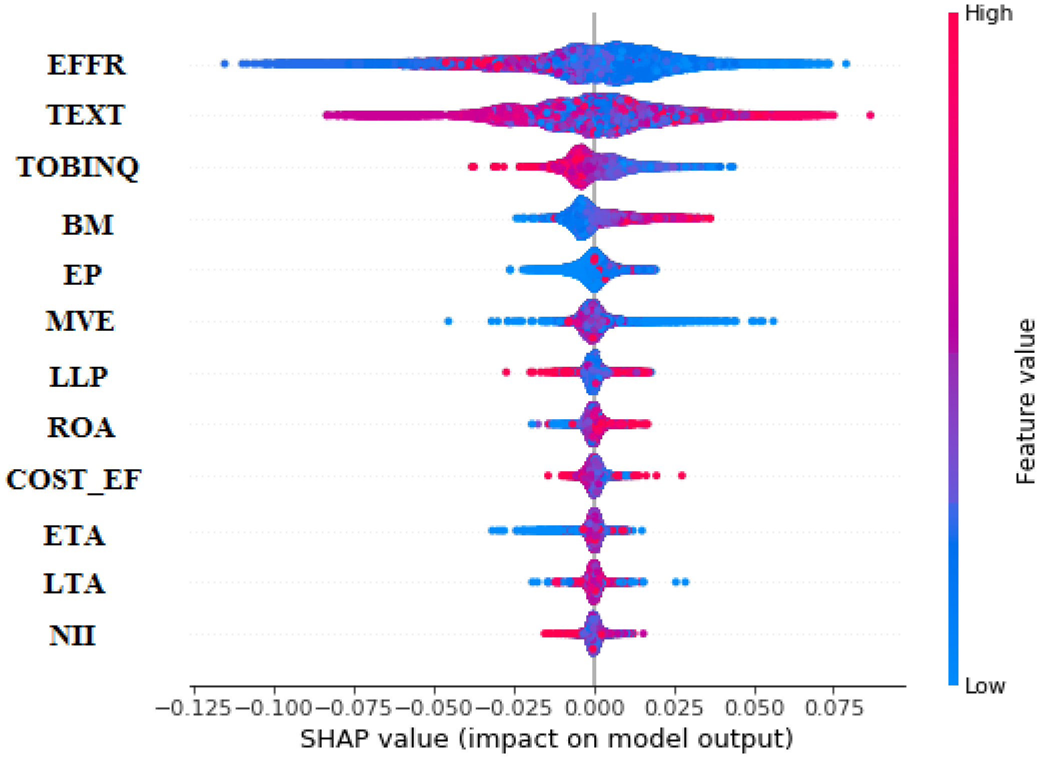

Table 1: Summary statistics

	N	Mean	Std. Dev.	p25	Median	p75
LLP	25808	0.001	0.002	0	0	0.001
ETA	25808	0.106	0.032	0.086	0.103	0.121
EP	25808	0.018	0.021	0.013	0.019	0.025
LTA	25808	0.660	0.127	0.597	0.682	0.750
NII	25808	0.230	0.143	0.130	0.207	0.291
ROA	25808	0.002	0.002	0.002	0.002	0.003
COST_EF	25808	0.654	0.150	0.565	0.638	0.723
TOBINQ	25808	1.011	0.059	0.985	1.011	1.040
MVE	25808	6.199	1.778	4.937	5.961	7.311
BM	25808	2.400	9.211	0.705	0.894	1.151
EFFR	25808	0.172	0.559	-0.020	0	0.243

Notes: This table reports the summary statistics of the control variables in our sample. All variables are winsorized at 1% and 99% level.

Table 2: Correlation matrix

	LLP	ETA	EP	LTA	NII	ROA	COST_EF	TOBINQ	MVE	BM	EFFR
LLP	1.000										
ETA	-0.028	1.000									
EP	-0.282	-0.023	1.000								
LTA	-0.075	0.088	0.000	1.000							
NII	0.123	-0.162	0.087	-0.456	1.000						
ROA	-0.287	0.148	0.697	0.017	0.105	1.000					
COST_EF	0.043	-0.029	-0.494	-0.103	0.117	-0.682	1.000				
TOBINQ	-0.180	-0.066	0.075	0.010	0.032	0.265	-0.262	1.000			
MVE	-0.069	-0.002	0.167	-0.290	0.326	0.280	-0.366	0.337	1.000		
BM	0.078	-0.182	0.086	-0.171	0.116	-0.092	-0.009	-0.236	-0.021	1.000	
EFFR	-0.020	-0.021	0.002	-0.056	-0.001	0.002	-0.016	0.020	0.033	0.008	1.000

Notes: This table reports Pearson correlations of the control variables in our sample.

Table 3: Out-of-sample performance of bank stock returns prediction using only financial variables

Only financial	**OLS**	**SVR**	**MLP**	**RF**
RMSE	0.0974	0.1009	0.0956	0.0900
MSE	0.0095	0.0102	0.0091	0.0081
MAE	0.0617	0.0682	0.0603	0.0558

Notes: This table reports the out-of-sample RMSE, MAE and MSE scores for our machine learning models, using only financial variables as inputs. We use 80% of our sample as the training dataset and the remaining 20% as the out-of-sample (testing) dataset. Beyond the baseline OLS model, we use the following machine learning models: support vector regression (SVR), multilayer perceptron (MLP), and random forest (RF).

Table 4: Out-of-sample performance of bank stock returns prediction using only textual features from press conferences of central bankers

Only textual	**OLS**	**SVR**	**MLP**	**RF**
RMSE	0.0831	0.0923	0.0835	0.0831
MSE	0.0069	0.0085	0.0070	0.0069
MAE	0.0522	0.0652	0.0527	0.0522

Notes: This table reports the out-of-sample RMSE, MAE and MSE scores for our machine learning models, using only textual features as inputs. To construct the textual features, we use the 20,000 most frequent words of the central bankers' press conferences. We use 80% of our sample as the training dataset and the remaining 20% as the out-of-sample (testing) dataset. Beyond the baseline OLS model, we use the following machine learning models: support vector regression (SVR), multilayer perceptron (MLP), and random forest (RF).

Table 5: Out-of-sample performance of bank stock returns prediction, using both SVD-100 textual features from press conferences of central bankers and financial variables

Both	**OLS**	**SVR**	**MLP**	**RF**
RMSE	0.0860	0.0935	0.0821	0.0810
MSE	0.0074	0.0087	0.0067	0.0066
MAE	0.0523	0.0639	0.0509	0.0489

Notes: This table reports the out-of-sample RMSE, MAE and MSE scores for our machine learning models, using both textual features and financial variables as inputs. To construct the textual features, we use the 20,000 most frequent words of the central bankers' press conferences. However, the dimensions of textual features are further reduced to 100 using the singular value decomposition dimensionality reduction technique (SVD100). We use 80% of our sample as the training dataset and the remaining 20% as the out-of-sample (testing) dataset. Beyond the baseline OLS model, we use the following machine learning models: support vector regression (SVR), multilayer perceptron (MLP), and random forest (RF).

Table 6: 3-month window prediction of bank stock returns using only textual features from press conferences of central bankers

Both	**OLS**	**SVR**	**MLP**	**RF**
RMSE	0.0738	0.0823	0.0752	0.0738
MSE	0.0054	0.0068	0.0057	0.0054
MAE	0.0488	0.0598	0.0507	0.0488

Notes: This table reports the out-of-sample RMSE, MAE and MSE scores for our regression machine learning models using textual features as inputs. To construct the textual features, we use the 20,000 most frequent words of the central bankers' press conferences. In fact, now there is a 2-month gap between the textual features and the bank stock returns. We use 80% of our sample as the training dataset and the remaining 20% as the out-of-sample (testing) dataset. Beyond the baseline OLS model, we use the following machine learning models: support vector regression (SVR), multilayer perceptron (MLP), and random forest (RF).

Table 7: 3-month window prediction of bank stock returns, using both SVD-100 textual features from press conferences of central bankers and financial variables

Both	**OLS**	**SVR**	**MLP**	**RF**
RMSE	0.0743	0.0813	0.0728	0.0705
MSE	0.0550	0.0066	0.0530	0.0050
MAE	0.0494	0.0583	0.0480	0.0460

Notes: This table reports the out-of-sample RMSE, MAE and MSE scores for our regression machine learning models, using both textual features and financial variables as inputs. To construct the textual features, we use the 20,000 most frequent words of the central bankers' press conferences. However, the dimensions of textual features are further reduced to 100 using the singular value decomposition dimensionality reduction technique (SVD100). In fact, now there is a 2-month gap between the textual features and the bank stock returns. We use 80% of our sample as the training dataset and the remaining 20% as the out-of-sample (testing) dataset. Beyond the baseline OLS model, we use the following machine learning models: support vector regression (SVR), multilayer perceptron (MLP), and random forest (RF).

Table 8: LIME results with bigrams

POSITIVE	**NEGATIVE**
Economic growth	Risks associated
Use monetary	Take additional
Job creation	Bring unemployment
Conditions improving	Expect inflation
Growth going	Projected inflation
Provide additional	Inflation continues
Asset purchases	Ukraine Russia
Substantial progress	Inflation pressures

Notes: This table denotes the most significant bigrams used as inputs in the RF model via the LIME methodology. Positive (negative) impact implies higher (lower) bank stock returns.

Table 9: LIME results with coloured text

TEXT
"SHORTTERM INCREASE INFLATION PROMPTED COMMITTEE TIGHTEN POLICY"
"PLANS WARRANTED ECONOMIC FINANCIAL DEVELOPMENTS"
"HIGHER COSTS BROADER PRICES CREATING BROADER INFLATION ECONOMY LONG INFLATION EXPECTATIONS"
"GROWTH POSITIVE LASTED FIVE YEARS"
"HOWEVER PERCENT UNEMPLOYMENT RATE REMAINS ELEVATED LOOKING AHEAD COMMITTEE ANTICIPATES UNEMPLOYMENT RATE WILL DECLINE GRADUALLY"
"ULTIMATELY WANT EARN MONEY INVESTMENTS INVEST ECONOMY GROWING"
"SIGNIFICANT MOVE INFLATION ALSO PERSISTENT RAISING RATES ADDRESS INFLATION CONCERNS VIEW"
"USE MONETARY POLICY TOOLS HELP KEEP JOBS MARKET STRONG SHOWING HEALTHIER WAGE GAINS PROMPTING MANY PEOPLE JOIN REMAIN WORKFORCE"

Notes: This table denotes the most significant words inside the text through LIME methodology. In this empirical setting, we use a RF as classifier that predicts whether there are positive or non-positive returns with only textual features as inputs. Words with orange colour translate to positive impact and high returns, whereas those with blue colour imply negative impact and low returns. Note that the more intense the background shade, the stronger the effect.