Short-term Prediction of Bank Deposit Flows: Do Textual Features matter?

Katsafados, Apostolos and Anastasiou, Dimitris

Athens University of Economics and Business, Alpha Bank

January 2022

Short-term Prediction of Bank Deposit Flows: Do Textual Features matter?

Apostolos G. Katsafados[†,*,1]
[†]Corresponding author: [*]Department of Statistics, Bank of Greece, Athens, Greece.
[1]Department of Accounting and Finance, Athens University of Economics and Business.
Email: katsafados@aueb.gr

Dimitris Anastasiou[1,2]
[2]Economic Research Division, Alpha Bank, Athens, Greece.
[1]Department of Accounting and Finance, Athens University of Economics and Business.
Email: anastasioud@aueb.gr

Abstract

The purpose of this study is twofold. First, to construct short-term prediction models for bank deposit flows in the Euro area peripheral countries, employing machine learning techniques. Second, to examine whether textual features enhance the predictive ability of our models. We find that Random Forest models including both textual features and macroeconomic variables outperform those that include only macro factors or textual features. Monetary policy authorities or macroprudential regulators could adopt our approach to timely predict potential excessive bank deposit outflows and assess the resilience of the whole banking sector in the Euro area peripheral countries.

Keywords**:** Bank deposit flows; European banks; textual analysis; short-term prediction; machine learning

***JEL classification**:* C22, C51, G10, E44.

Disclaimer: The views and opinions expressed herein are those of the authors and do not reflect their respective institutions.

Funding: This research received no external funding.

Data Availability Statement: Data are not publicly available due to ethical reasons, though the data may be made available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

1. Introduction

Bank deposit (out) flows are clearly significant since any extreme fluctuations can disrupt aggregate consumption and aggregate investment, thus bringing about considerable adverse effects in the macroeconomic environment (Demirguç-Kunt and Detragiache, 1998; Anastasiou and Katsafados, 2020). Furthermore, given that bank assets are usually illiquid assets, excessive deposit outflows can trigger banking insolvency, or even worse, a banking panic. Consequently, this would disturb the credit flows both to households and enterprises, decreasing investment and consumption, hence forcing even sustainable firms into bankruptcy (Demirguç-Kunt and Detragiache, 1998). Therefore, it becomes apparent that predicting bank deposit (out) flows is imperative for policymakers and regulators.

Even though there is an extended background theory on the determinants of bank deposits (Martinez-Peria and Schmukler, 2001; Hondroyiannis, 2004; Finger and Hesse, 2009; Oliveira *et al.,* 2014; Nys *et al.,* 2015; Anastasiou and Katsafados, 2020; Anastasiou and Drakos, 2021a; Anastasiou and Drakos, 2021b) the literature on forecasting deposit flows is still sparse. Especially, as far as we know, the only studies that conduct a forecasting exercise for bank deposits are these of Piscopo (2010), Petropoulos *et al.,* (2018), and Anastasiou and Petralias (2021). In more detail, Piscopo (2010) developed a functional data model with ARIMA terms for forecasting the evolution of Italian bank deposits. Petropoulos *et al.,* (2018) developed a Markov-regime switching autoregressive model to forecast Greek private sector bank deposits. Finally, Anastasiou and Petralias (2021), after constructing a novel leading indicator based on Bloomberg news headlines, examined its forecasting ability on Greek bank deposit flows employing a Markov Regime Switching Regression model.

While previous studies have contributed to forecasting bank deposit flows, optimizing the predictive factors and prediction models is still arguably in need of improvement.

Therefore, this study combines traditional macroeconomic fundamentals with textual features derived from the ECB President's speeches to expand the predictive factors. In the same spirit, Hagenau *et al.* (2013) use financial news to predict stock price. Moreover, Tang *et al.* (2020) use a combination of financial and textual variables to predict financial distress. More particularly, in this study, we attempt to answer the following research questions:

Q1. Does textual information from ECB speeches influence the bank deposit flows of Euro area peripheral countries?

Q2. Which models perform better than others for short-term prediction of bank deposit flows of Euro area peripheral countries?

This study examines several one-month ahead prediction models' predictive ability to engage with these research objectives, revealing changes in the key predictive factors. We examine five classification algorithms that have had a principal role in the finance prediction literature. Particularly, we employ Logistic Regression, Support Vector Machine, Random Forest, Naive Bayes, and Multilayer Perceptron. We infer that the best machine learning model for a short-term prediction of bank deposit flows is the Random Forest with both TF-IDF features and macroeconomic variables as inputs. It is hoped that the results of this research can provide a template for early warning mechanisms for relevant economic agents to take the corresponding efforts to avoid bank runs and bank losses.

In some more detail, our study makes several significant contributions to the related literature. First, we attempt to predict the deposit outflows in four peripheral Eurozone countries (Portugal, Italy, Greece, and Spain). Such an attempt was never tried before due to the prolonged macroeconomic deterioration faced by these countries after the 2008 financial crisis. Thus, it becomes apparent that any attempt for prediction in this country group is a challenging task. Second, as far as we know, this is the first study of bank deposit flows prediction employing textual features. In other words, our study complements the relevant literature by adding fresh insights on how textual features can signal early warning signs for bank deposit outflow events. Third, we demonstrate that machine learning constitutes a promising framework for predicting financial outcomes.

The remainder of this study is structured as follows. In Section 2, we describe the data and variables used. In Section 3, we describe the methodology we followed and the architecture of the models used to predict bank deposit flows. In Section 4, we present the empirical results, and we assess the forecasting power of our proposed forecasting model. Finally, Section 5 presents the conclusions of this study.

2. Data

We utilize monthly data spanning the period 2008-2018 for the Euro area peripheral countries. Following Goretti and Souto (2013), Angelopoulou *et al.,* (2014), Bijsterbosch and Falagiarda (2015) and Anastasiou *et al.,* (2019), we define as Euro area peripheral countries Portugal, Italy, Greece, and Spain. Attempting to predict bank deposit flows in these countries and during a crisis period is challenging since, during this period, bank deposit flows demonstrated high volatility. The selection of this country group and this period under scrutiny makes our study even more important, especially if we consider that during the last sovereign debt crisis, macroeconomic fundamentals along with financial markets in the peripheral countries collapsed amid a deepening loss of confidence in the ability of governments to tackle their severe economic problems (Anastasiou *et al.,* 2022a). This loss of confidence, in turn, led economic agents (depositors in our case) not to fully trust the macroeconomic fundamentals in these countries anymore, therefore making other non-fundamental factors the main driving force affecting their decisions.

2.1. Dependent variable and matching

We obtain the deposit transaction flows from domestic households and non-profit institutions from the ECB Statistical Data Warehouse. When deposit flows in a country attain positive (negative) values, this country witnesses deposit inflows (outflows). Thus, as a dependent variable, we construct a dummy variable (DF) attaining 0 if a country of our sample witnesses deposit inflows and 1 when it witnesses deposit outflows, respectively.

However, the number of outflows is smaller than inflows, which means our dataset is imbalanced. Imbalanced datasets are a common issue in classification tasks in finance (Pasiouras *et al.,* 2007, 2010; Katsafados *et al.,* 2020). For that reason, we apply the undersampling technique of Veganzones and Severin (2018) to deal with this issue. This technique creates a balanced subsample from our original sample by excluding observations from the majority category (in this case, the inflows).

2.2. Textual methodology

After creating a web-crawling algorithm, we gather all the speeches of the ECB president from February 2008 to February 2017.[1] All the retrieved speeches are encoded in a hypertext markup language (HTML). We adopt the parsing process for each retrieved speech as described in Loughran and McDonald (2013). In particular, we remove HTML formatting and any other non-textual information (Bodnaruk *et al.,* 2015; Katsafados *et al.,* 2021). As a result, we end up with speeches that include merely words.

Given that knowledge retrieval from a text is a highly delicate process, it is essential to perform high-quality pre-processing. Notably, pre-processing is vital in analyzing textual information, thereby influencing the overall performance of any classification algorithm (Nassirtousi *et al.,* 2014; Kumar and Ravi, 2016). It practically contains a variety of sub-

[1] We have collected the speeches directly from the website of the ECB. In our sample, there are merely those speeches that are in English language.

processes. The purpose is to convert the raw format of our texts into meaningful inputs for our predictive models.

First of all, we exclude from our analysis all non-germane characters such as single-letter words, numbers, punctuation marks, and stop words (Gandhi *et al.,* 2019). The high quality of the purging process retains only the inputs that contain valuable information regarding our prediction task. Thus, it contributes to superior prediction performance. In addition, there is another advantage through this process: eradicating the curse of dimensionality problem (Nassirtoussi *et al.,* 2014). If we have many textual features, this can adequately decrease the effectiveness of any learning algorithm (Pestov, 2013).

Although we now have purified the speeches, they cannot be used as inputs in our models. This is because any learning algorithm or mathematical model is unable to understand the unstructured format of textual data and any natural language unless we convert our data into inputs with numerical form (Mai *et al.,* 2019). This challenging process is called feature selection. By far, the most popular method is arguably the bag of words (BOW) model.

As a first step, this model proceeds to tokenization. This implies that our speeches are parsed into the words included within. To do so, we use Natural Language Toolkit (NLTK) Python library, as mentioned in Mai *et al.* (2019). To be more in-depth, the BOW model considers each unique word as a separate textual feature and generates a document-term matrix, where each column and row assigns to a word and a document, respectively (Kumar and Ravi, 2016). Although BOW naively ignores word sequence, it is widely used in many tasks in the textual finance literature (Loughran and McDonald, 2016).

Finally, given that we select our textual features, we proceed to feature representation. We practically use a numeric value to represent each feature throughout the feature representation procedure. However, in the textual analysis realm, raw counts of textual features are not considered the best measure of a text's information content because this is apparently strongly bound to document length. For that reason, one solution to the problem is to adopt simple proportions, or we may choose to adjust a word's weight in the analysis considering how unusual the word is in the corpus. In our empirical setting, we employ two widely used term weighting schemes: (1) the term frequency (TF), and (2) the term frequency-inverse document frequency (TF-IDF).

The former measure considers all words to be equivalent. Substantially it computes the raw count of each word in each document divided by the document length for normalization purposes. The mathematical formulation for a word *i* in document *j* is:

$$\mathrm{TF}(\mathrm{w_{ij}}) = \frac{c_{ij}}{Tj} \tag{1}$$

where c_{ij} is the raw count of word *i* in document *j* and T_j is the total number of words of document *j*.

On the other hand, *TF-IDF* down weights the *TF* scores based on how frequently each word appears in our sample of speeches in overall (Kearney and Liu, 2014; Nassirtoussi *et al.,* 2014). We define our *TF-IDF* measure of word *i* in the j^{th} document as follows:

$$\mathrm{TF\text{ - }IDF}\ (\mathrm{t_{ij}}) = \mathrm{TF}(\mathrm{t_{ij}}) \times [-\log(\frac{n_i}{N})] \tag{2}$$

where *N* represents the number of speeches in our entire dataset, n_i the number of speeches that include at least one occurrence of the i^{th} word. *TF-IDF* weighting scheme is a common approach, widely used by the literature due to its merit of providing more considerable attention to rarer words across our entire speech sample collection (Loughran and

McDonald, 2016). So far, plenty of studies have employed it (Balakrishnan *et al.,* 2010; Brown and Tucker, 2011; Kumar *et al.,* 2012; Mai *et al.,* 2019; Katsafados *et al.,* 2021).

2.3. Macroeconomic variables

The level of private sector deposit transaction flows in a country is directly related to its macroeconomic conditions (Petropoulos *et al.,* 2018). Therefore, we take into consideration several additional macroeconomic and financial variables that reflect both the data availability and the background literature (see among others, Martinez-Peria and Schmukler, 2001; Finger and Hesse, 2009; Nys *et al.,* 2015; Petropoulos *et al.,* 2018; Anastasiou and Katsafados, 2020; Anastasiou and Drakos, 2021a; Anastasiou and Drakos, 2021b).

Specially, we employ the following set of macroeconomic factors as additional explanatory variables:

- 10GBY: Long Term 10-Year Government Bond Yields.
- IPI: Industrial Production Index.
- DEPRATE: Average Deposits Interest Rate that each country sets.
- UNMP: Unemployment rate (as % of the active population)
- ESI: Economic Sentiment Indicator.

Finally, we include the one-period lag of DF as a possible determinant to forecast future bank deposit flows. Table 1 provides the main descriptive statistics for each under examination variable by country.

Insert **Table 1** here

3. Machine learning models

In this study, we set out to investigate whether the machine learning models can accurately predict one-month ahead European bank deposit flows. In what follows, we examine five classification algorithms that have had a principal role in the finance prediction literature. Particularly, we use Logistic Regression, Support Vector Machine, Random Forest, Naive Bayes, and Multilayer Perceptron.

3.1. Logistic regression (Logit)

Among alternative classification algorithms used in finance prediction tasks, the most common is the Logit model (Palepu, 1986; Ambrose and Megginson, 1992; Papoulias and Theodossiou, 1992; Espahbodi and Espahbodi, 2003; Pasiouras and Tanna, 2010; Boehm and DeGennaro, 2011; Mai *et al.,* 2019). The logit model estimates a non-linear sigmoid function between our binary variable DF and the independent variables (i.e., textual and macroeconomic). The estimation is achieved through the maximum likelihood method (MLE). The mathematical framework behind the Logit model is denoted as follows:

$$P(Y_{t+1}=1 \mid X_{i,t}) = \frac{\exp(b_0 + \sum_{i=1}^{n} b_i X_{i,t})}{1+\exp(b_0 + \sum_{i=1}^{n} b_i X_{i,t})} \tag{3}$$

where Y_{t+1} defines the dichotomy deposit flow event, $X_{i,t}$ is a vector that includes n variables at time t, b_i denotes the parameters of the model, and at last, b_0 is a bias term.

3.2. Support vector machine (SVM)

Another well-established method in the literature is that of SVM. The SVM, first introduced by Vapnik and Vapnik (1998), has been used quite frequently in a plethora of forecasting tasks in finance, such as merger prediction (Pasiouras *et al.,* 2008), time-series provision (Cao, 2003; Huang *et al.,* 2005; Pai and Lin, 2005), and bankruptcy forecasting (Min and Lee, 2005; Shin *et al.,* 2005; Wu *et al.,* 2007). In practice, SVM aims to find the best hyperplane that separates two clashes of observations with a maximum margin (Kumar and Ravi, 2016). The only training samples used to fulfil the classification task are called support vectors and those near the hyperplane. To handle non-linear separable data, the employment of a non-linear kernel mapping is vital (Nassirtoussi *et al.,* 2014). In our case, we apply the radial kernel function (RBF), consistent with Mai *et al.,* (2019).

3.3. Random forest (RF)

RF is an ensemble learning method that generates numerous decision trees at training time. Breiman (1996) introduces Bagging, an early version of RF. In general, RF produces superior results than the classical decision trees. The rationale behind this is that RF models do not suffer from an over-fitting problem. That is, RF can generalize more efficiently. In our research, RF uses some uncorrelated decision tree classifiers. After the random selection of a subset of features, the training is achieved based on bootstrap copies of original samples (Mai *et al.,* 2019; Iworiso and Vrontos, 2020). Finally, each tree decides to support a class. The class with the most votes automatically becomes the predictive output. Some other papers also use RF to handle textual information for their predictions are those of Moniz and Jong (2014) and Katsafados *et al.,* (2020).

3.4. Naive Bayes (NB)

The NB classifier belongs to the family of probabilistic learning algorithms, and it is based upon implementing Bayes's theorem. It assumes that there is complete independence among the features set. Given its predicting capability, NB is commonly used so far for binary problems and multi-class classifications (Kumar and Ravi, 2016). Given that the class variable y and dependent feature vector x_1 through $x_{n'}$, then the mathematical formula is:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)} \tag{4}$$

Under the naive conditional independence assumption:

$$P(x_i | y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i | y), \tag{5}$$

for all i, the relationship is simplified to:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} \quad P(x_i \mid y)}{P(x_1, \dots, x_n)} \tag{6}$$

Considering that $P(x_1, \dots, x_n)$ is constant given the input, we employ the following classification rule:

$$\begin{gathered} P(y \mid x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} \quad P(x_i \mid y) \\ \Downarrow \\ \hat{y} = \arg \max_{y} P(y) \prod_{i=1}^{n} \quad P(x_i \mid y), \end{gathered} \tag{7}$$

As mentioned in Iworiso and Vrontos (2020), we can use maximum posterior estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

In our study, we implement the Gaussian Naive Bayes algorithm for the classification task. The likelihood of the features is assumed to be Gaussian, and it reads as follows:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{8}$$

where the parameters σ_y and μ_y are estimated using maximum likelihood.

3.5. Multilayer perceptron (MLP)

Considering their ability to efficiently deal with textual information due to the non-linearity they offer, artificial neural networks (ANN) are used in a broad spectrum of tasks in the Natural Language Processing (NLP) domain (Goldberg, 2016). The most famous representative is the MLP models, which belong to the feed-forward network category.[2] Their advantage is that they are so powerful, and at the same time, easy to implement. MLP models include three separate layers, as explained by Kumar and Ravi (2016). First, the input

[2] Feed-forward neural networks are networks with fully-connected layers. Namely, each neuron is linked to all of the neurons in the next layer.

layer is the first stage in the network structure, whereby the variables are injected into the network. Second, there are one or more hidden layers.[3] When the hidden layers receive the content from the input layer, they use non-linear functions to process it. Afterward, they transfer the computed values to the output layer. Finally, the output layer applies a softmax or sigmoid function upon the received output from the last hidden layer deciding the predictive class. Mai *et al.,* (2019) document that the back-propagation algorithm upgrades the weights of the model throughout the training process. Based on the mathematics, each value of an input pattern $A \in R^N$ is linked with weight value $W \in R^N$ which takes values between 0 and 1 (Dosdogru, 2019). Given that *F(x)* is the function that computes the output from the neurons, this output could be represented with the following mathematical formula:

$$Y = F\left(\sum_{i=1}^{N} a_i * w_i + u\right)$$

where w_i denotes the synaptic weights, and *u* is the bias levels. Figure 1 shows the MLP architecture, where as in our study, we use three hidden layers.

Insert **Figure 1** here

4. Empirical Results and Evaluation

4.1. Evaluation measures

It is necessary to ensure that our deposit flow predictions are properly evaluated concerning their out-of-sample predictive ability (Mai *et al.,* 2019; Katsafados *et al.,* 2020). Espahbodi and Espahbodi (2003) suggest that a realistic assessment must provide an out-of-time perspective in addition to the out-of-sample. However, an accurate assessment of

[3] The networks with two or more layers of hidden neurons are known as deep networks, thus leading to the terminology of deep learning (Goldberg, 2016). According to Sun *et al.,* (2017), the existence of many hidden layers benefits us with higher learning capacity.

learning algorithms' ability to classify objects is only established if they get tested in a future period (Pasiouras *et al.,* 2008). Such a superior method considers the possibility of a population drifting over time. As a result, we follow the approach with the two distinct samples in the present study. In line with the prior literature, we choose 80% of our data as the training set for model fitting (Geng *et al.,* 2015; Doumpos *et al.,* 2017; Routledge *et al.,* 2017) and the rest 20% of them (that are not employed throughout the training procedure) is defined as our testing set.[4] For all reasons above, we apply the partitioning method of defining the testing set from a future period rather than randomly (Pasiouras *et al.,* 2008; Pasiouras and Tanna, 2010; Mai *et al.,* 2019).

When the models are trained, we need to evaluate their out-of-sample performance. As a first evaluation criterion, we utilize the accuracy metric. Plenty of past papers in finance prediction literature have used the accuracy metric to assess their models (see among others Palepu, 1986; Pasiouras *et al.,* 2007; Pasiouras and Tanna, 2010; Pasiouras *et al.,* 2010; Boehm and DeGennaro, 2011; Mai *et al.,* 2019). Accuracy results range from 0 to 1. A higher accuracy score implies a better out-of-sample performance of the model. Generally, the accuracy metric can be defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{9}$$

where *TP* is the number of observations labeled adequately as deposit outflows by the model, *TN* is the number of observations correctly decided as deposit inflows by the classifier, *FP* the number of observations erroneously identified as deposit outflows by the classifier and *FN* is the number of observations incorrectly labeled as deposit inflows by the model.

[4] This set is practically used to assess the out-of-sample and out-of-time performance of our classifiers.

To ensure the stability of our results, we also adopt some widely-used prediction performance measures, such as Precision and Recall. Notably, there is a measure, called F1-score, that harmonically combines Precision and Recall. First, we practically compute the measures for each category (inflows and outflows), and next, we apply the macro average approach to estimate the general performance.[5] As follows, we provide the mathematical formulas of these measures:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{12}$$

Alternatively, we use receiver operating characteristic (ROC) curves to provide some robustness check to our predictive outcome. The prior literature broadly utilizes ROC in many tasks, such as bankruptcy prediction (Gaganis *et al.,* 2005; Mai *et al.,* 2019) and bank merger prediction (Pasiouras *et al.,* 2008; Pasiouras and Tanna, 2010). In practice, the ROC curve plots the true-positive rate of the model on the vertical axis and the false-positive rate on the horizontal axis as cut-off points variegates. The basic concept is that models closer to the upper and left corner of the diagram suggest a better out-of-sample classification power. Apart from the curves of each model, we also plot a 45-degree line, which indicates a random assignment of class labels. Based on the ROC curve, we can compute the area under the

[5] The macro average approach just sums the measure scores for inflows and outflows, and finally, it divides the result by two.

curve (AUC) measure that ranges between 0 to 1. An uninformative classifier performs an AUC score of 0.5, while 1 demonstrates a perfect predictive ability.

4.2. Support vector machine results

Table 2 presents the results of the support vector machine model (SVM). We observe that the model with the best predictive performance has only TF features as inputs approaching 69% approximately. Surprisingly, when combining textual features and macroeconomic variables, we find that the predictive performance deteriorates.

Insert **Table 2** here

To get further insights into how the SVM tries to separate the two classes based on textual data, we present the decision boundary of the model. We first implement the singular value decomposition (SVD) dimensionality decrease technique. It can practically project high-dimensional data into a low-dimensional space. We apply SVD to project the textual features to 2 dimensions to visually present the decision boundary at the Cartesian level. Figure 2 shows the decision boundary of the model.

Insert **Figure 2** here

4.3. Multilayer perceptron results

Table 3 reports the results from the MLP models, according to which the models with the TF-IDF textual features provide the best predictive outcome. More precisely, this model achieves a precision score equal to 70%, which is slightly better than the best SVM model. However, we find again that the combination of textual features and macroeconomic factors does not improve the predictive performance of the models.

Insert **Table 3** here

4.4. Naïve Bayes results

Table 4 shows the results from the Naïve Bayes models. Notably, we find that Naïve Bayes with TF textual features as inputs has the highest scores, achieving 71% predictive performance. This model seems to produce marginally better scores than the previously noted models. Once again, the combination of textual features and macroeconomic variables does not seem to augment the predictive performance of the models.

Insert **Table 4** here

To shed lights on how the NB attempts to separate the two classes based on textual data, we present the decision boundary of the model. Again, similar to the SVM model, we apply SVD to project the textual features into two dimensions. Figure 3 shows the decision boundary of the model.

Insert **Figure 3** here

4.5. Logistic regression results

Table 5 illustrates the results from the logistic regression models. Particularly, we find that these models with TF textual features as inputs yield the best performance (71%), which is similar to this of the Naïve Bayes model. Once more, the mixture of macroeconomic variables and textual features seems to lessen the predictive performance of the models.

Insert **Table 5** here

In Figure 4, we visually report the decision boundary of the logistic regression model based on the textual information. As previously, we employ SVD to project the textual features into two dimensions.

Insert **Figure 4** here

4.6. Random forest

4.6.1. Random forest results

Table 6 demonstrates the results from the random forest models. We find for the first time that textual features can efficiently complement macroeconomic variables in terms of prediction efficacy, as models that utilize both sources of data produce more accurate estimates. Specifically, RF with both TF and macroeconomic factors achieves 73% accuracy, while interestingly, RF with both TF-IDF and macroeconomic factors as inputs achieve 76% performance, which is the highest score compared to all models under-scrutiny. In line with past literature (Loughran and McDonald, 2016; Mai *et al.,* 2019; Katsafados *et al.,* 2021), the TF-IDF weighting scheme is considered more effective than the TF approach.

Insert **Table 6** here

Figure 5 shows the decision boundary created by the RF model based on textual information. As before, we project the textual features into two dimensions with the SVD technique; thus, we can visually present the decision boundary.

Insert **Figure 5** here

4.6.2. Gini Impurity

To further prove the high importance of textual features in our RF model, we now present more quantitative evidence. When TF-IDF textual features and macroeconomic variables are jointly used as inputs in the RF model, we attempt to find the most important features. We practically use the Gini importance methodology. Essentially this technique provides an internal insight into the mathematical mechanisms behind the structure's model. In each internal node of each decision tree, the RF selects a feature to decide how to divide the datasets into two separate sets. The feature selection is based on some criteria, such as Gini Impurity in classification tasks. The mathematical formula is:

$$Gini\ Impurity = 1 - Gini \tag{13}$$

where Gini is computed as:

$$Gini = \sum_{i=1}^{n} p^2(c_i) \tag{14}$$

where n is the number of the classes and $p(c_i)$ denotes the percentage of class c_i in the node. Therefore, the mathematical framework is expressed:

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} p^2(c_i) \tag{15}$$

In our case, we have a binary problem where the classifiers try to separate deposit inflows from deposit outflows. We present the Gini Impurity formulas for each node in the trees for both of our tasks:

$$Gini\ Impurity_{outflows} = 1 - (percentage\ of\ outflows)^2 - (percentange\ of\ inflows)^2$$

For each leaf node, the feature with the highest decrease of impurity is selected for the node as the most appropriate. Finally, given that we use RF instead of a single decision tree model, we compute the average impurity decrease of each feature across all decision trees in the forest.

We next define the textual (or macro) Gini Impurity score as the sum of Gini Impurity scores of all textual (or macro) variables. The mathematical framework could be expressed as follows:

$$Textual\ Gini\ Impurity = \sum_{i=1}^{n} Gini\ Impurity_i \tag{16}$$

where i represents each textual feature.

Similarly, we compute the macro–Gini Impurity as follows:

$$Macro\ Gini\ Impurity = \sum_{j=1}^{k} Gini\ Impurity_j \tag{17}$$

where j represents each macro feature.

As a result, we interestingly find that textual Gini Impurity is larger (0.53) than macro Gini Impurity (0.47). That finding supports the statement of the high importance of textual information in our task. To conclude, using textual features vitally comes to supplement macro variables, thus leading to a much better predictive outcome overall.

4.7. ROC curves and AUC scores

Figure 6 depicts the ROC curves of our four best machine learning algorithms (i.e., MLP, NB, Logit, and RF), when we use a combination of TF textual features and macroeconomic factors. We observe that AUC values are steadily above 0.7, with the RF model producing the best AUC score (0.75). Also, when we compare the second and third best models, we find that MLP and NB compete as each one prevails across a particular spectrum of cut-off probabilities.

Finally, Figure 7 illustrates the ROC curves of our four best machine learning algorithms (i.e., MLP, NB, Logit, and RF), when we employ a blend of TF-IDF textual features and macroeconomic variables. In general, as we may well observe, all models yield AUC scores consistently above 0.7. In fact, RF is the model with the best AUC score (0.75). In addition, comparing the second and third best models (i.e., MLP and NB), we conclude that they demonstrate an equal performance (0.72).

Insert **Figures 6 and 7** here

Overall, we find that Random Forest models including both textual features and macroeconomic variables outperform those that include only macro factors or textual features. Textual features capture an aspect of the so-called non-fundamental variables that may affect bank deposits. Economic agents in peripheral countries, which have been hit at a higher degree by the sovereign debt crisis, may rely more on non-fundamental factors, such as textual sentiment rather than macroeconomic fundamentals.[6] This is in line with prior literature showing that non-fundamental factors exert a more significant impact on peripheral countries (see, among others, Gómez-Puig *et al.,* 2014; Galariotis *et al.,* 2016; Anastasiou *et al.,* 2022a; Anastasiou *et al.,* 2022b).

5. Conclusions

Motivated by the successful usage of machine learning in the area of computer science and its wide acceptance from the economic literature (Li *et al.,* 2020; Huo and Chaudhry, 2021; Kamble *et al.,* 2021), we introduce machine learning models for predicting bank deposit flows in the Euro area peripheral countries. We infer that for a short-term prediction of bank deposit flows, the best machine learning models are the random forest with TF-IDF features combined with macroeconomic fundamentals.

[6] Anastasiou and Drakos (2021b) found that depositors have lower confidence in the peripheral countries' banking systems, making the latter suffer from larger deposit outflows (especially in crisis periods), leading to more frequent panics in bank deposits and thus financial instability in the periphery. All these, in turn, further deteriorate agents' trust in the domestic banking system, which may lead them to rely more on sentiment than macro-financial factors (fundamentals).

Our study prompts future further investigations. First, a micro-level dataset could be employed, where instead of having bank deposit flows at a country-level, bank deposit flows at a bank level could be examined. Thus, bank-specific variables could also be employed as possible factors to forecast bank deposit flows, such as return on equity, leverage, and non-performing loans. Second, other machine learning techniques could be examined. For example, more advanced deep learning models such as Recurrent Neural Networks (RNNs) and Transformer-based Models could be examined.

References

Ambrose, B.W., and Megginson, W.L., (1992). "The role of asset structure, ownership structure, and takeover defenses in determining acquisition likelihood". *Journal of Financial and Quantitative Analysis*, 27, 575-589.

Anastasiou, D., and Drakos, K., (2021a). "Nowcasting the Greek (semi-) deposit run: Hidden uncertainty about the future currency in a Google search". *International Journal of Finance and Economics*, 26, 1133-1150.

Anastasiou, D., Kapopoulos, P., and Zekente, KM., (2022a), "Sentimental Shocks and House Prices". *Journal of Real Estate Finance and Economics*.

Anastasiou, D., Kallandranis, C., and Drakos, K., (2022b), "Borrower Discouragement Prevalence for Eurozone SMEs: Investigating the Impact of Economic Sentiment". *Journal of Economic Behavior and Organization*, 194, 161-171.

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

Anastasiou, D., and Katsafados, A., (2020). "Bank deposits flows and textual sentiment: When an ECB president's speech is not just a speech". MPRA Working Paper No. 99729.

Anastasiou, D., and Petralias, A., (2021). "On the construction of a leading indicator based on news headlines for predicting Greek deposit outflows", MPRA Working Paper No. 107602.

Anastasiou, D., Louri, H., and Tsionas, M., (2019). "Non-performing loan in the Euro area: Are Core-Periphery Banking Markets Fragmented?". *International Journal of Finance and Economics*, 24, 97-112.

Angelopoulou, E., Balfoussia, H., and Gibson, H.D., (2014). "Building a financial conditions index for the euro area and selected euro area countries: What does it tell us about the crisis?". *Economic Modelling*, 38, 392-403.

Balakrishnan R., Qiu X.Y., and Srinivasan P., (2010). "On the predictive ability of narrative disclosures in annual reports". *European Journal of Operational Research*, 202, 789-801.

Bijsterbosch, M., and Falagiarda, M., (2015). "The macroeconomic impact of financial fragmentation in the euro area: Which role for credit supply?". *Journal of International Money and Finance*, 54, 93-115.

Bodnaruk, A., Loughran, T., and McDonald, B., (2015). "Using 10-K text to gauge financial constraints". *Journal of Financial and Quantitative* Analysis, 50, 623-646.

Boehm, T.P., and DeGennaro, R.P., (2011). “A discrete choice model of dividend reinvestment plans: Classification and prediction”. *Managerial and Decision Economics*, 32, 215-229.

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

Brown, S.V, and Tucker, J.W., (2011). “Large-sample evidence on firms’ firms’ year-over-year MD&A modifications”. *Journal of Accounting Research*, 49, 309-346.

Cao, L., (2003). “Support vector machines experts for time series forecasting”. *Neurocomputing*, 51, 321-339.

Demirguç-Kunt, A., and Detragiache, E., (1998). “The determinants of banking crises in developing and developed countries”. *International Monetary Fund Staff Papers*, 45, 81-109.

Dosdogru, A.T., (2019). “Comparative study of hybrid artificial neural network methods under stationary and nonstationary data in stock market”. *Managerial and Decision Economics*, 40, 460-471.

Doumpos, M., Andriosopoulos, K., Galariotis, E., Makridou, G., and 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., and Espahbodi, P., (2003). “Binary choice models for corporate takeover”. *Journal of Banking and Finance*, 27, 549-574.

Finger, M.H., and Hesse, H., (2009). “Lebanon-determinants of commercial bank deposits in a regional financial center”. *International Monetary Fund*, No. 9-195.

Gaganis, C., Pasiouras, F., and Tzanetoulakos, A., (2005). “A comparison and integration of classification techniques for the prediction of small UK firms failure”. *Journal of Financial Decision Making*, 1, 55-69.

Galariotis, E.C., Makrichoriti, P., & Spyrou, S., (2016). “Sovereign CDS spread determinants and spill-over effects during financial crisis: A panel VAR approach”. *Journal of Financial Stability*, 26, 62-77.

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

Geng, R., Bose, I., and Chen, X., (2015). “Prediction of financial distress: An empirical study of listed Chinese companies using data mining”. *European Journal of Operational Research*, 241, 236-247.

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

Gómez-Puig, M., Sosvilla-Rivero, S., and del Carmen Ramos-Herrera, M., (2014). "An update on EMU sovereign yield spread drivers in times of crisis: A panel data analysis". *The North American Journal of Economics and Finance*, 30, 133-153.

Gorton, G., (1988). "Bank panics and business cycles". *Oxford Economic Papers*, 40, 751-781.

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

Hondroyiannis, G., (2004). "Estimating private savings behaviour in Greece". *Journal of Economic Studies*, 31, 457-476.

Huang, W., Nakamori, Y., and Wang, S.Y., (2005). "Forecasting stock market movement direction with support vector machine". *Computers and Operations Research*, 32, 2513-2522.

Huo, D., and Chaudhry, H.R., (2021). Using machine learning for evaluating global expansion location decisions: An analysis of Chinese manufacturing sector. *Technological Forecasting and Social Change*, 163, 120436.

Iworiso, J., and Vrontos, S., (2020). "On the directional predictability of equity premium using machine learning techniques". *Journal of Forecasting*, 39, 449-469.

Kamble, S.S., Gunasekaran, A., Kumar, V., Belhadi, A., and Foropon, C., (2021). "A machine learning based approach for predicting blockchain adoption in supply Chain". *Technological Forecasting and Social Change*, 163, 120465.

Katsafados, A.G., Androutsopoulos, I., Chalkidis, I., Fergadiotis, E., Leledakis, G.N., and Pyrgiotakis, E.G., (2020). "Textual information and IPO underpricing: A machine learning approach". Working Paper SSRN: ssrn.com/sol3/papers.cfm?abstract_id=3720213

Katsafados, A.G., Androutsopoulos, I., Chalkidis, I., Fergadiotis, E., Leledakis, G.N., and Pyrgiotakis, E.G., (2021). "Using textual analysis to identify merger participants: Evidence from U.S. banking industry". *Finance Research Letters*, forthcoming.

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

Kumar R.B., Kumar B.S., and Prasad C.S.S., (2012). "Financial news classification using SVM". *International Journal of Scientific and Research Publications*, 2, 2250-3153.

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

Li, J.P., Mirza, N., Rahat, B., and Xiong, D., (2020). "Machine learning and credit ratings prediction in the age of fourth industrial revolution". *Technological Forecasting and Social Change*, 161, 120309.

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

Loughran, T., and McDonald, B., (2013). “IPO first-day returns, offer price revisions, volatility, and form S-1 language). *Journal of Financial Economics*, 109, 307-326.

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

Martinez-Peria, M.S., and Schmukler, S.L., (2001). “Do depositors punish banks for bad behavior? Market discipline, deposit insurance, and banking crisis”. *Journal of Finance*, 56, 1029-1051.

Min, J.H., and Lee, Y.C., (2005). “Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters”. *Expert Systems with Applications*, 28, 603-614.

Moniz, A., and Jong, F.D., (2014). “Classifying the influence of negative affect expressed by the financial media on investor behavior”. 5th Information Interaction in Context Symposium (IIiX), 275-278.

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

Nys, E., Tarazi, A., and Trinugroho, I., (2015). “Political connections, bank deposits, and formal deposit insurance”. *Journal of Financial Stability*, 19, 83-104.

Oliveira, R., Schiozer, R.F. and Barros L., (2014). “Depositors’ perception of too-big-to-fail”. *Review of Finance*, 18, 1-37.

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

Palepu, K.G., (1986). “Predicting takeover targets: A methodological and empirical analysis”. *Journal of Accounting and Economics*, 8, 3-35.

Papoulias, C., and Theodossiou, P., (1992). “Analysis and modeling of recent business failures in Greece”. *Managerial and Decision Economics*, 13, 163-169.

Pasiouras, F., and 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.

Pasiouras, F., Gaganis, C., Tanna, S., and Zopounidis, C., (2008). “An application of support vector machines in the prediction of acquisition targets: Evidence from the EU banking sector). In: Zopounidis, C., Doumpos, M., Pardalos, P. (Eds.), Handbook of Financial Engineering. Springer.

Pasiouras, F., Gaganis, S., and Zopounidis, C., (2010). “Multicriteria classification models for the identification of targets and acquirers in the Asian banking sector”. *European Journal of Operational Research*, 204, 328-335.

Pasiouras, F., Tanna, S., and Zopounidis, C., (2007). “The identification of acquisition targets in the EU banking industry: An application of multicriteria approaches”. *International Review of Financial Analysis*, 16, 262-281.

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., Vlachogiannakis, N. E., and Mylonas, D., (2018). “Forecasting private sector bank deposits in Greece: determinants for trend and shock effects”. *International Journal of Banking, Accounting and Finance*, 9, 141-169.

Piscopo, G., (2010). “Italian deposits time series forecasting via functional data analysis”. *Banks and Bank Systems*, 5, 12-19.

Routledge, B.R., Sacchetto, S., and Smith, N.A., (2017). “Predicting merger targets and acquirers from text”. Working Paper, Carnegie Mellon University.

Shin, K.S., Lee, T.S., and Kim, H., (2005). “An application of support vector machines in bankruptcy prediction model”. *Expert Systems with Applications*, 28, 127-135.

Sun, S., Luo, C., and Chen, J., (2017). “A review of natural language processing techniques for opinion mining systems”. *Information Fusion*, 36, 10-25.

Tang, X., Li, S., Tan, M., and 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.

Vapnik, V.N., and Vapnik, V., (1998). “Statistical learning theory: 1”. New York: Wiley.

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

Wu, C.H., Tzeng, G.H., Goo, Y.J., and Fang, W.C., (2007). “A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy”. *Expert Systems with Applications*, 32, 397-408.

Tables

Table 1: Main descriptive statistics						
	DF	**UNMP**	**10GBY**	**IPI**	**ESI**	**DEPRATE**
N	436	430	436	436	431	436
min	-23,840.75	4.80	1.99	79.44	72.20	0.08
mean	1,499.39	9.91	6.04	104.52	98.12	2.23
max	23,773.50	27.90	29.24	135.80	112.90	5.37

Note: This table reports the main descriptive statistics for the sample countries.

Table 2: One-month ahead out-of-sample performance using ECB president speech with SVM model				
Variables/Features used	**Accuracy**	**Precision**	**Recall**	**F1-score**
Only TF features	0.689	0.690	0.690	0.690
Only TF-IDF features	0.667	0.670	0.670	0.660
Only Macro variables	0.667	0.670	0.670	0.670
TF *and* Macro variables	0.667	0.670	0.670	0.660
TF-IDF *and* Macro variables	0.667	0.670	0.670	0.660

Note: This table reports the accuracy scores for our first classification machine learning model, Support Vector Machine (SVM), using either textual information or macroeconomic variables as separate inputs, as well as in combination. The final (imbalanced) sample consists of 246 observations from 2008-2018. The analysis is based on a balanced sample of inflows and outflows. We employ 80% of our sample as the training set and the remaining 20% as the testing set. TF and TF-IDF are the two term weighting schemes for our textual features. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table 3: One-month ahead out-of-sample performance using ECB president speech with MLP model				
Variables/Features used	**Accuracy**	**Precision**	**Recall**	**F1-score**
Only TF features	0.622	0.650	0.620	0.590
Only TF-IDF features	0.689	0.700	0.690	0.680
Only Macro variables	0.644	0.660	0.640	0.640
TF *and* Macro variables	0.667	0.670	0.670	0.660
TF-IDF *and* Macro variables	0.689	0.690	0.690	0.680

Note: This table reports the accuracy scores for our second classification machine learning model, Multilayer Perceptron (MLP), using either textual information or macroeconomic variables as separate inputs, as well as in combination. The final (imbalanced) sample consists of 246 observations from 2008-2018. The analysis is based on a balanced sample of inflows and outflows. We employ 80% of our sample as the training set and the remaining 20% as the testing set. TF and TF-IDF are the two term weighting schemes for our textual features. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table 4: One-month ahead out-of-sample performance using ECB president speech with NB model				
Variables/Features used	**Accuracy**	**Precision**	**Recall**	**F1-score**
Only TF features	0.711	0.710	0.710	0.710
Only TF-IDF features	0.689	0.690	0.690	0.680
Only Macro variables	0.667	0.730	0.670	0.650
TF *and* Macro variables	0.667	0.680	0.670	0.670
TF-IDF *and* Macro variables	0.689	0.700	0.690	0.690

Note: This table reports the accuracy scores for our third classification machine learning model, Naïve Bayes (NB), using either textual information or macroeconomic variables as separate inputs, as well as in combination. The final (imbalanced) sample consists of 246 observations from 2008-2018. The analysis is based on a balanced sample of inflows and outflows. We employ 80% of our sample as the training set and the remaining 20% as the testing set. TF and TF-IDF are the two term weighting schemes for our textual features. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table 5: One-month ahead out-of-sample performance using ECB president speech with logit model				
Variables/Features used	**Accuracy**	**Precision**	**Recall**	**F1-score**
Only TF features	0.711	0.710	0.710	0.710
Only TF-IDF features	0.689	0.690	0.690	0.680
Only Macro variables	0.667	0.670	0.670	0.670
TF *and* Macro variables	0.689	0.690	0.690	0.690
TF-IDF *and* Macro variables	0.689	0.690	0.690	0.690

Note: This table reports the accuracy scores for our fourth classification machine learning model, Logistic regression (Logit), using either textual information or macroeconomic variables as separate inputs, as well as in combination. The final (imbalanced) sample consists of 246 observations from 2008-2018. The analysis is based on a balanced sample of inflows and outflows. We employ 80% of our sample as the training set and the remaining 20% as the testing set. TF and TF-IDF are the two term weighting schemes for our textual features. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Table 6: One-month ahead out-of-sample performance using ECB president speech with RF model				
Variables/Features used	**Accuracy**	**Precision**	**Recall**	**F1-score**
Only TF features	0.689	0.690	0.690	0.690
Only TF-IDF features	0.711	0.710	0.710	0.710
Only Macro variables	0.533	0.540	0.530	0.530
TF *and* Macro variables	0.733	0.740	0.730	0.730
TF-IDF *and* Macro variables	0.756	0.760	0.760	0.750

Note: This table reports the accuracy scores for our final classification machine learning model, Random Forest (RF), using either textual information or macroeconomic variables as separate inputs, as well as in combination. The final (imbalanced) sample consists of 246 observations from 2008-2018. The analysis is based on a balanced sample of inflows and outflows. We employ 80% of our sample as the training set and the remaining 20% as the testing set. TF and TF-IDF are the two term weighting schemes for our textual features. TF stands for the term frequency scheme normalized by document length, and TF-IDF for the term frequency-inverse document frequency scheme.

Figures

Figure 1: MLP architecture

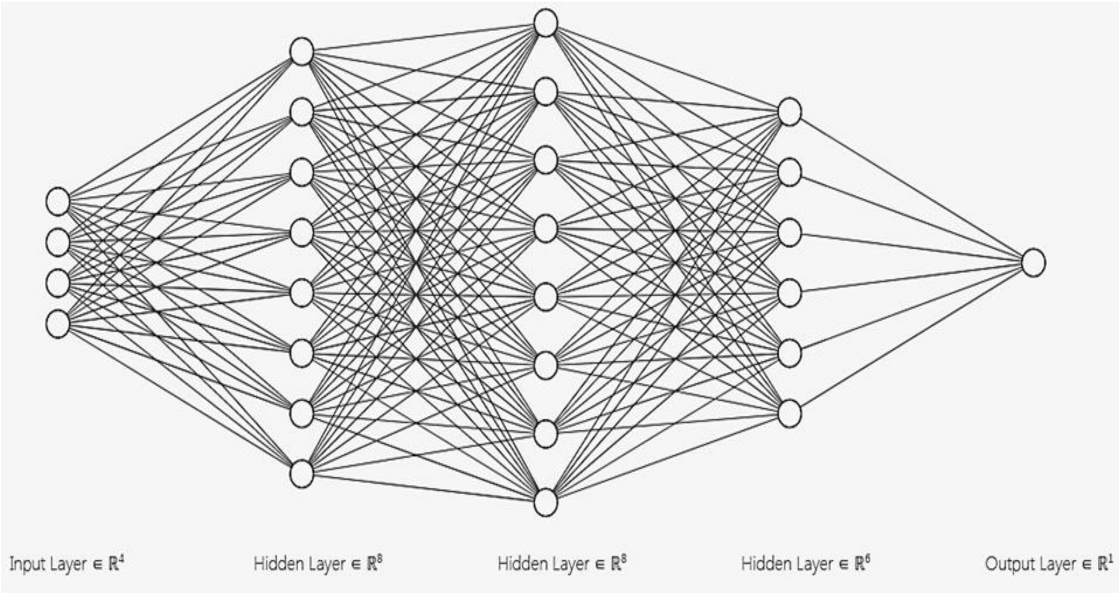

Figure 2: Decision boundary of SVM model

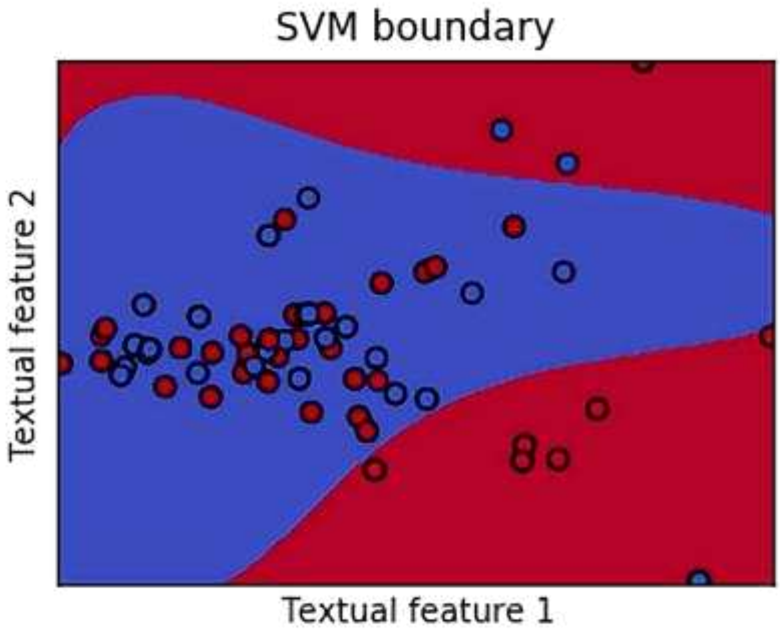

Figure 3: Decision boundary of NB model

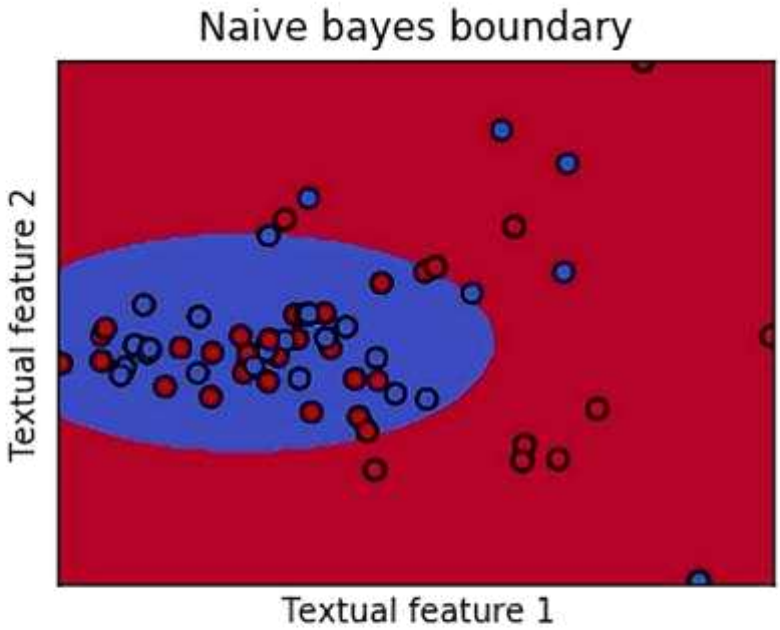

Figure 4: Decision boundary of Logit model

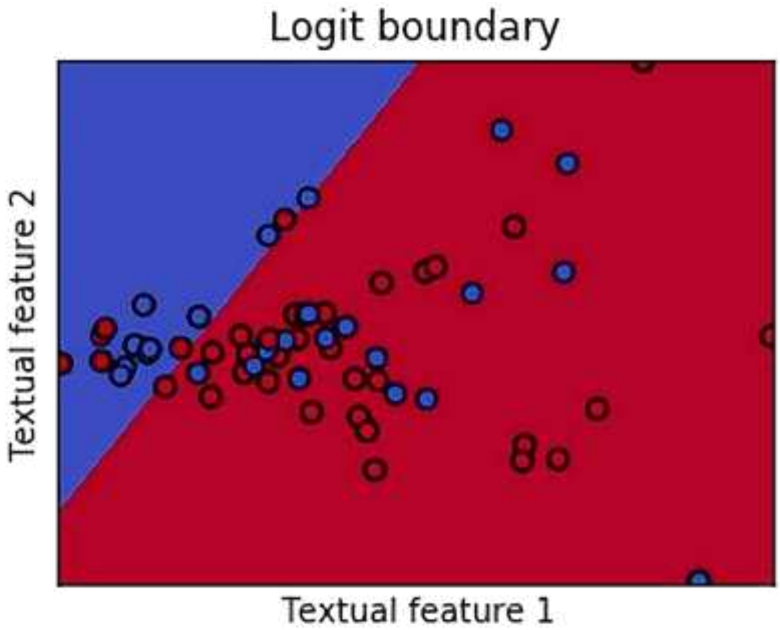

Figure 5: Decision boundary of RF model

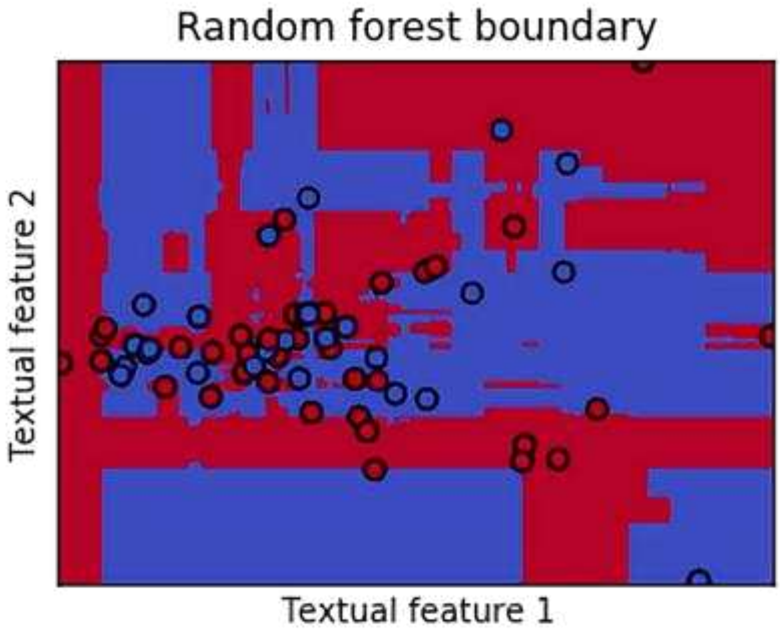

Figure 6: ROC curve with both TF textual features and macro variables

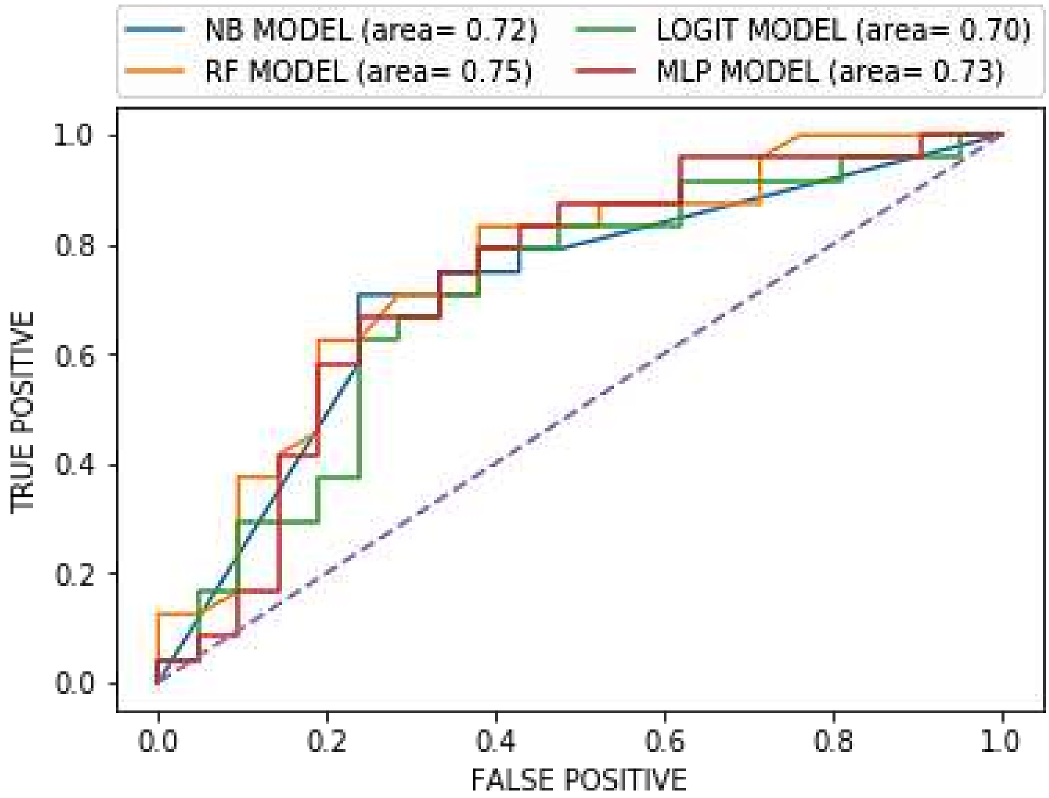

Figure 7: ROC curve with both TF-IDF textual features and macro variables