BTRM

The Certificate of Bank Treasury Risk Management

Social media and liquidity risk management: applying a “sentiment analysis” early warning indicator

Thought Leadership Series #21

Authors:
Gerardo Salazar PhD
Abraham M Izquierdo FRM, BTRM
Mariana Grajales
Saul Echazarreta

February 2026

Abstract

The purpose of this paper is to analyse the relationship between social media opinions about a bank and the sensitivity of deposits. The aim is to develop a framework to anticipate accelerated deposit withdrawals in response to events that create perceptions of fragility or stress within the sector or the bank itself. This approach will support the implementation of preventive balance-sheet measures, but also, highlights the relevance of considering social dynamics on digital platforms as a new dimension of liquidity risk assessment. We conclude that sentiment analysis applied to social media information represents a valuable complementary tool for anticipating changes in bank deposit behaviour. By transforming unstructured data into quantitative indicators, the proposed model enables the incorporation of an additional dimension of perception and reputational risk into traditional liquidity monitoring.

INTRODUCTION

In recent years, social media has played an important role in shaping public perceptions of banks' financial health. Events such as bank runs at Silicon Valley Bank (SVB), Signature Bank, and First Republic Bank in 2023 have shown how digital platforms can amplify systemic risk through the rapid spread of information and misinformation, influencing depositor behavior during periods of financial stress.

Recent studies (Cookson et al., 2023; Gam et al., 2024; ECB, 2023) show that banks' exposure to social media—particularly X (formerly Twitter)—can accelerate deposit withdrawals when there are signs of deterioration, even before traditional financial indicators justify such movements. This occurs because depositors react not only to direct information received from the bank but also to their expectations about the behavior of other depositors. In this context, social media enables real-time observation of overall user sentiment, amplifying contagion and coordination dynamics.

BACKGROUND

The expansion of internet access and the rise of new digital channels—particularly online news media and social networks—have led to a rapid increase in data production, especially unstructured data. This environment has driven the development of advanced technologies and methodologies capable of processing large volumes of information in diverse formats, enabling the extraction of relevant signals more quickly.

Banco de México has documented this structural shift in its Financial Stability Report (December 2024), highlighting that text analysis of news and social media offers a privileged window into monitoring expectations, risk perceptions, and narratives that may influence financial stability. In particular, the use of Natural Language Processing (NLP) techniques has enabled the construction of high-frequency indicators such as the Economic Policy Uncertainty (EPU) Index and the Banking Risk Index, both designed to detect early signs of economic or financial stress based on the dynamics of digital information.

These developments are supported by a growing body of literature. Baker et al. (2016) demonstrated the value of quantifying economic uncertainty through systematic analysis of news, while more recent studies, such as Cookson et al. (2023), have shown the critical role of social media in accelerating episodes of banking stress. International initiatives, such as the BIS Ellipse Project, have reinforced

the use of unstructured data within early-warning systems for the financial sector, emphasizing its usefulness in identifying emerging vulnerabilities.

These precedents show that integrating digital data and NLP methodologies is an essential component of modern risk supervision. They also provide the technical and conceptual foundation for developing machine-learning models capable of anticipating deposit stress episodes, particularly when such stress may originate from or be amplified within the information environment of social networks and digital media.

WHAT ARE SENTIMENT MODELS?

One of the most widely used approaches in Natural Language Processing (NLP) is sentiment analysis. These computational tools are designed to identify, classify, and quantify the opinions, emotions, or attitudes expressed in natural-language texts. Their main purpose is to determine content polarity —typically classified as positive, negative, or neutral— although more advanced approaches measure sentiment intensity or identify nuanced emotional attributes.

These models use statistical techniques, machine learning, and deep learning to extract structured information from large volumes of unstructured text, such as social media posts, forums, reviews, or news articles (Jurafsky & Martin, 2023).

Because of their ability to capture perceptions in a timely manner, sentiment models have become essential in applications where user attitudes can have immediate impacts, such as financial markets, institutional reputation, risk analysis, or early detection of stress episodes. In these contexts, sentiment analysis provides a systematic approach to studying how public perception evolves and, importantly, how it can translate into behaviors that affect the stability or functioning of a system.

HOW ARE SENTIMENT ANALYSIS MODELS CALCULATED?

Computing sentiment analysis models involves methodological stages that transform unstructured text into quantitative metrics representing sentiment polarity or intensity. Although specific procedures depend on the chosen approach, the literature agrees that the general process includes: (i) text preprocessing, (ii) numerical representation of language, and (iii) sentiment estimation or inference through a formal model (Liu, 2012; Jurafsky & Martin, 2023).

Text Preprocessing

Preprocessing is a fundamental stage aimed at cleaning and normalizing text to reduce noise and improve model performance. Common techniques include removing punctuation, URLs and special characters; converting text to lowercase; removing stopwords; and applying lemmatization or stemming to normalize words.

For text from social media, preprocessing is particularly important due to the frequent use of abbreviations, emoticons, hashtags, and informal language, all of which may contain relevant sentiment information and therefore must be handled carefully (Taboada et al., 2011).

Sentiment Calculation in Lexicon-Based Models

In lexicon-based models, sentiment is calculated by assigning each word in the text a numerical value previously defined in a sentiment dictionary. These values typically represent polarity (positive or

negative) or emotional intensity. The overall sentiment of a text is obtained by aggregating the values of its component words, generally through sums or averages (Liu, 2012).

Formally, the sentiment score of a text S can be expressed as:

$$S = \sum_{i=1}^{N} w_i$$

where w_i represents the sentiment score associated with word i, and N is the total number of words with emotional weight in the text. In financial applications, it is common to use specialized dictionaries that adjust semantics to the economic context, since words with a negative connotation in general language may not carry the same meaning in financial documents (Loughran & McDonald, 2011).

Sentiment Calculation in Supervised Machine Learning Models

In supervised models, sentiment calculation is based on training a classification algorithm using a set of previously labeled texts. In the first stage, the texts are transformed into numerical vectors through representation techniques such as bag-of-words, TF-IDF, or n-grams (Pang et al., 2002).

Subsequently, the model estimates a function that assigns a probability to each sentiment class. In the case of a binary classification model, the sentiment of a text is determined as:

$$P(y = 1 \mid x) = f(x; \theta)$$

where x represents the feature vector of the text, θ the model's estimated parameters, and y the sentiment category. The result can be expressed either as a discrete label (positive, negative, or neutral) or as a continuous probability, which is particularly useful for constructing aggregated sentiment indicators.

Sentiment Calculation in Deep Learning Models

Deep learning models employ distributed representations of language, known as *embeddings*, which capture the semantic meaning and context of words. Models based on transformer architecture process the entire text bidirectionally, allowing the meaning of each word to depend on the context in which it appears (Devlin et al., 2019).

In these models, sentiment is obtained from the output of the neural network, typically a classification layer that produces a probability distribution over sentiment categories. The final score can be interpreted as a probability, a continuous value, or a discrete classification, depending on the specific application.

Aggregation and Construction of Sentiment Indicators

Once sentiment has been calculated at the level of individual comments, the results can be aggregated to construct temporal or institutional indicators. Examples include daily average sentiment, the percentage of negative comments, or a volume-weighted interaction index. This type of aggregation allows sentiment extracted from social media to be linked with observable economic and financial variables, facilitating its use in statistical and econometric analysis (Tetlock, 2007).

In the banking context, these indicators can be interpreted as indirect measures of perception, trust, or reputational risk, which justifies their joint analysis with variables such as deposit liquidity.

MACHINE LEARNING MODELS FOR CLASSIFICATION

General Machine Learning Approach

Machine Learning models are used to identify complex relationships between a set of explanatory variables and a target variable, especially when such relationships cannot be adequately represented through simple parametric assumptions. These approaches allow patterns to be extracted from data in a flexible manner, making them particularly useful in economic and financial applications, where interactions among variables are often nonlinear and high-dimensional.

Supervised Classification

Within the set of Machine Learning techniques, supervised classification focuses on assigning observations to discrete categories based on previously labeled historical information. The model learns a function that relates the explanatory variables to a categorical target variable, adjusting its parameters during the training process to minimize classification errors observed in historical data. This approach has been widely documented in the statistical learning literature as a flexible alternative to traditional models when working with complex datasets (Hastie, Tibshirani & Friedman, 2009).

In the financial domain, this type of model is used to anticipate various economic behaviors, such as changes in credit quality, agent decision-making, or variations in balance-sheet variables. In this study, the problem is formulated as a binary classification task, where the objective is to predict the direction of change in bank deposit balances over a given period.

Sentiment Indicators as Explanatory Variables

As previously mentioned, social media data are inherently unstructured. Through sentiment analysis, this information is transformed into quantitative measures that summarize the overall tone of user comments within a specific period. The sentiment indicators constructed from this process capture the public's perception of a financial institution in an aggregated and time-comparable manner.

Several studies have shown that such indicators contain relevant information for explaining and anticipating the behavior of financial variables, supporting their use as explanatory variables in predictive models (Tetlock, 2007). In this sense, sentiment indicators function as a bridge between qualitative information from social media and the quantitative models used in financial analysis.

Classification Models in Financial Applications

The Financial Machine Learning literature has explored a wide variety of classification algorithms, including logistic regression, decision trees, ensemble methods, neural networks, and support vector machines. The selection of the most suitable model depends on the specific characteristics of the dataset, including sample size, the degree of correlation between explanatory variables, and the possible presence of nonlinear relationships with the target variable.

Advantages of the Machine Learning Approach in This Study

Machine Learning models offer important advantages over traditional parametric approaches, as they do not require strict assumptions about data distribution and allow for the flexible capture of complex patterns. These features are especially valuable when using indicators derived from social media, which often exhibit high volatility and strong interdependence. In this context, supervised classification models provide an appropriate methodological framework for integrating sentiment information with financial variables, serving as the conceptual foundation for the model proposed in the following section.

METHODOLOGY

General Approach

To estimate the sentiment contained in social media comments, a deep learning model based on transformer-type architectures is employed (a neural network model that uses attention mechanisms to capture semantic relationships in both short and long texts). The model is specifically fine-tuned for the task of sentiment classification in short text. This approach allows the capture of complex semantic relations and contextual dependencies that cannot be adequately modeled using traditional word-count-based approaches or linear classifiers.

The model output is then used to construct quantitative sentiment indicators, which are subsequently analyzed jointly with bank deposit balances.

Deep Learning Model Architecture

The model used is based on a pretrained transformer architecture, which represents each input text as a sequence of tokens $(\boldsymbol{t_1, t_2, \dots, t_n})$. Each token is initially transformed into a vector representation through an embedding layer (numerical vectors representing words or tokens that capture semantic similarities and contextual relationships), incorporating both semantic and positional information.

Formally, the model input can be represented as:

$$\boldsymbol{X} = [\boldsymbol{x_1, x_2, \dots, x_n}], \quad \boldsymbol{x_i} \in \mathbb{R}^{\boldsymbol{d}}$$

Where $\boldsymbol{d}$ is the dimensionality of the embedding space. These representations are processed through multiple self-attention layers, which allow each token to incorporate information from the full context of the text.

In each attention layer, the **scaled dot-product attention** mechanism—which weights the importance of each token based on its similarity to other tokens in the sequence—is defined as:

$$\boldsymbol{Attention(Q, K, V) = softmax}\left(\frac{\boldsymbol{QK^T}}{\sqrt{\boldsymbol{d_K}}}\right)\boldsymbol{V}$$

where Q, K and V represent the query, key, and value matrices, respectively, and d_K is the dimensionality of the keys. This mechanism allows the model to dynamically weigh the importance of each word based on its contextual relevance.

Sentiment Classification Layer

For the sentiment analysis task, the aggregated contextual representation of the text—typically associated with a special classification token—is used. This final representation $h \in \mathbb{R}^d$ is fed into a fully connected (feed-forward) layer that acts as the classifier:

$$\boldsymbol{z = Wh + b}$$

where $W \in \mathbb{R}^{K \times d}$ is the weight matrix, $b \in \mathbb{R}^{K}$ is the bias vector, and K is the number of sentiment classes considered (positive, neutral, and negative).

The probability associated with each sentiment class is obtained by applying a softmax function (an activation function that converts model scores into normalized probabilities over a set of classes):

$$\boldsymbol{P(S = k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}}$$

The sentiment assigned to each comment corresponds to the class with the highest estimated probability, although the continuous probabilities can be retained for the construction of aggregated indicators.

Loss Function and Model Training

During the model fine-tuning stage, the parameters are estimated by minimizing a categorical cross-entropy loss function:

$$\boldsymbol{\mathcal{L} = -\sum_{k=1}^{K} y_k \log\left(P(S = k \mid x)\right)}$$

where y_k represents the true sentiment label. The optimization process is carried out using adaptive stochastic gradient descent algorithms, updating the network's weights to maximize the model's predictive performance on unseen data.

Construction of the Sentiment Score

Once the model has been trained, each comment $\boldsymbol{i}$ yields a probability vector $(\boldsymbol{p_i^+, p_i^0, p_i^-})$. Based on these probabilities, a continuous sentiment score is defined as:

$$\boldsymbol{S_i = p_i^+ - p_i^-}$$

This score makes it possible to capture not only the direction of sentiment but also its relative intensity, facilitating its temporal aggregation.

Temporal Aggregation of Sentiment

Individual scores are aggregated at the temporal level to construct institutional sentiment indicators. For a given period t, the aggregated indicator is defined as:

$$\boldsymbol{\bar{S}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} S_{t,i}}$$

Where $\boldsymbol{N_t}$ is the number of comments observed in period $\boldsymbol{t}$. This indicator summarizes the average perception expressed on social media and is used as an explanatory variable in the analysis of its relationship with bank deposit liquidity.

Construction of Sentiment Indicators

The deep learning model described in the previous section allows each individual comment to be assigned both a discrete sentiment classification (positive, neutral, or negative) and a continuous

score associated with sentiment intensity. While these results are informative at the individual level, their joint analysis with aggregated financial variables—such as the behavior of bank deposits—requires transforming them into quantitative indicators that summarize the perception expressed on social media over time.

Let $\boldsymbol{t}$ be a temporal aggregation period (for example, daily or monthly) and let $\boldsymbol{N_t}$ be the total number of comments observed during that period. For each comment i, the model produces a discrete sentiment classification and a continuous score $\boldsymbol{S_i}$

Based on the discrete classification generated by the model, the number of comments classified as positive, neutral, and negative is obtained for each period $\boldsymbol{t}$. These counts provide a preliminary description of the distribution of sentiment expressed on social media and constitute the basic input for the construction of aggregated indicators.

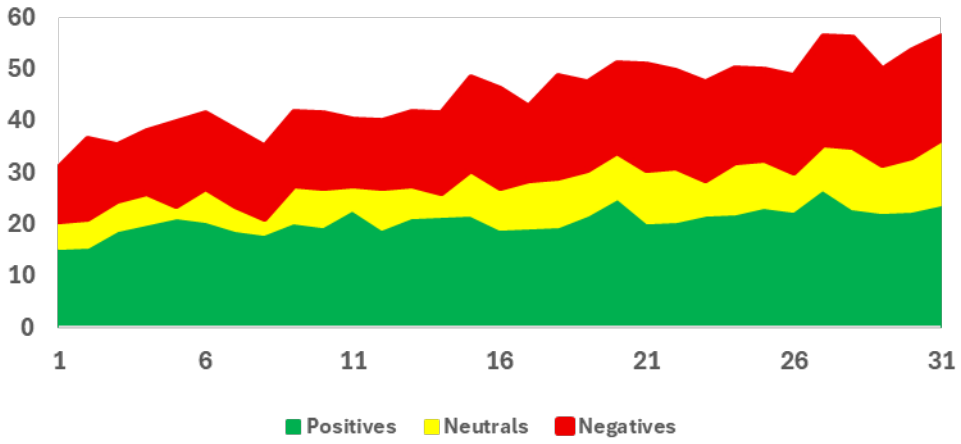

Figure 1: Count of Comments by Sentiment Category

The figure shows the temporal evolution of the number of comments classified as positive, neutral, and negative in each period, as identified by the sentiment analysis model.

Based on the previous results, aggregated indicators are constructed to summarize different dimensions of the sentiment expressed on social media and to facilitate its quantitative analysis.

Proportion of Negative Comments

The first indicator corresponds to the proportion of comments classified as negative relative to the total number of comments in period $\boldsymbol{t}$. This indicator helps capture episodes of heightened adverse perception or discontent expressed on social media.

Average Sentiment

The average sentiment summarizes the overall tone of the comments observed in period $\boldsymbol{t}$, using the continuous score estimated by the deep learning model. It is defined as:

$$\bar{S}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} S_{t,i}$$

where $\bar{S}_t$ represents the average sentiment score for period t. Higher values of this indicator reflect a more positive tone, while lower values indicate a more negative perception.

Net Sentiment Index

This index measures the balance between positive and negative opinions, providing a synthetic measure of the perception expressed on social media. It is defined as:

$$Net\ Index_t = \frac{Positives_t - Negatives_t}{Positives_t + Negatives_t}$$

This indicator takes values in the interval $[-1, 1]$, where positive values indicate a predominance of favorable comments and negative values reflect predominance of unfavorable opinions.

Figure 2 shows the evolution of the indicators constructed from social media comments: the proportion of negative comments, the average sentiment, and the net sentiment index. These series allow for analyzing the dynamics of sentiment throughout the study period and identifying changes in the overall tone of the opinions expressed by users.

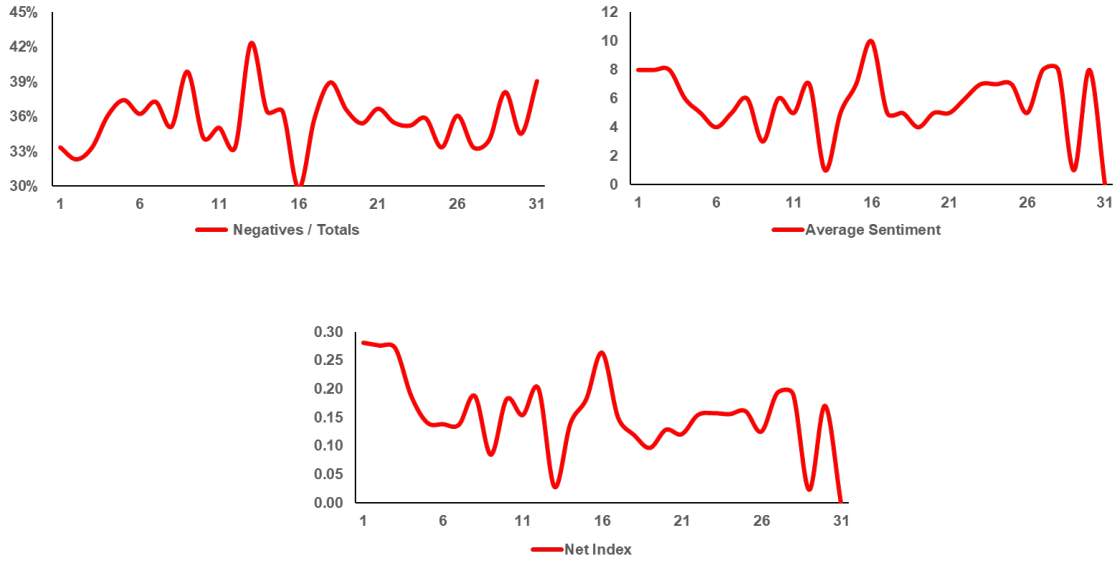

Figure 2: Temporal evolution of sentiment indicators

The indicators defined in this section are used as explanatory variables in the classification model based on Support Vector Machines, described in the following section. This model does not process text directly; instead, it uses the aggregated indicators derived from the sentiment analysis. By being aligned with the same temporal frequency as the financial variables, these indicators make it possible to assess whether the information contained in sentiment expressed on social media helps distinguish between periods of increase or decrease in bank deposit balances.

Figure 3 summarizes the proposed methodology for sentiment analysis:

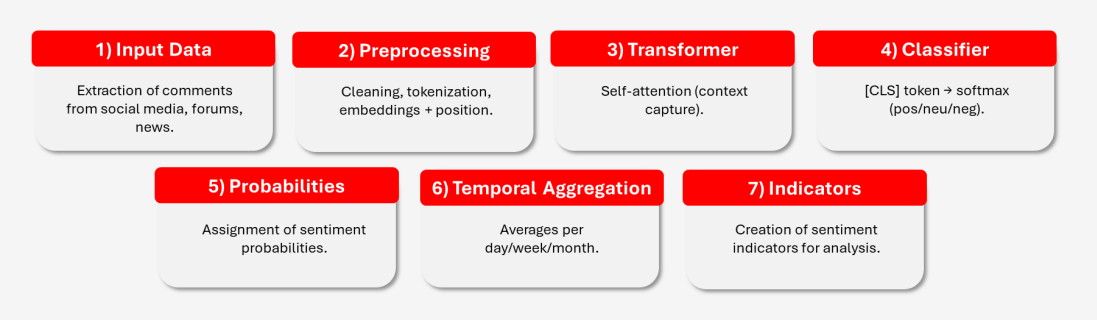

Figure 3: General flow of sentiment analysis methodology

The general flow of the sentiment analysis model is shown, starting with the collection of comments from social media, forums, and news sources, and continuing through the estimation of sentiment probabilities using a transformer-based architecture. These probabilities are then aggregated over time to construct institutional perception indicators, which are used as explanatory variables in the analysis of bank deposit sensitivity.
Just to clarify, deep learning is used for feature extraction; SVM is used for deposit classification.

Classification Model Based on Support Vector Machines
To evaluate the relationship between the sentiment indicators constructed in the previous section and changes in bank deposit balances, a classification model based on Support Vector Machines (SVM) is employed. This approach makes it possible to assess whether the information contained in social media indicators helps distinguish between periods of increase or decrease in deposit balances.

Definition of Variables
Let t be a time observation period. The categorical dependent variable is defined as:

$$y_t = \begin{cases} 1, if\ the\ deposit\ balance\ increases\ in\ period\ t \\ 0, if\ the\ deposit\ balance\ decreases\ or\ does\ not\ increase \end{cases}$$

This variable captures the direction of the change in the deposit balance and allows the problem to be formulated as a binary classification task.

To illustrate the process of transforming the continuous deposit balance series into a categorical variable, a graphical representation of this transformation is presented below.

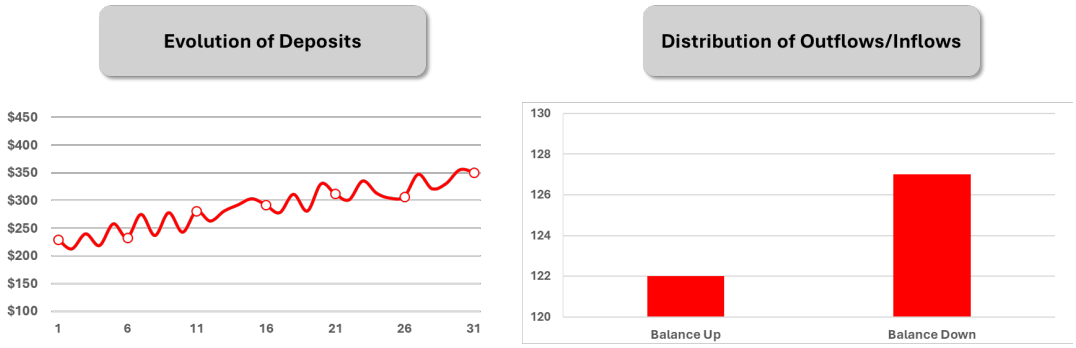

Figure 4: Transformation of the deposit balance into a binary variable

Figure 4 shows the temporal evolution of the deposit balance and the distribution of periods classified as increases or decreases. The classification is performed based on the change in the balance between consecutive periods, assigning the category of increase when the variation is positive and decrease when it is negative.

The explanatory variable vector $\boldsymbol{x_t} \in \mathbb{R}^{\boldsymbol{p}}$ is composed of the sentiment indicators (average sentiment of the period, proportion of negative comments relative to the total, net sentiment index, etc.) previously described. These indicators are formally represented as:

$$\boldsymbol{x_t} = \left(\boldsymbol{S_t^{(1)}}, \boldsymbol{S_t^{(2)}}, \ldots, \boldsymbol{S_t^{(p)}}\right)$$

where $\boldsymbol{p}$ is the number of indicators constructed from the sentiment analysis.

Mathematical Formulation of the SVM Model

The SVM model seeks to find a hyperplane that optimally separates the observations belonging to the two classes defined by $\boldsymbol{y_t}$. In its linear form, the classifier is expressed as:

$$\boldsymbol{f(x) = w^T x + b}$$

where $\boldsymbol{w}$ is the weight vector and $\boldsymbol{b}$ is the bias term. The classification rule is:

$$\hat{\boldsymbol{y}} = \begin{cases} \boldsymbol{1, si\ f(x) \geq 0} \\ \boldsymbol{0, si\ f(x) < 0} \end{cases}$$

The optimization problem associated with the soft-margin SVM is defined as:

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

Subject to:

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

where:

- ξ_i are slack variables that allow classification errors.
- $C > 0$ is the regularization parameter that controls the trade-off between maximizing the margin and penalizing errors.

The model in its linear version can be visualized as follows:

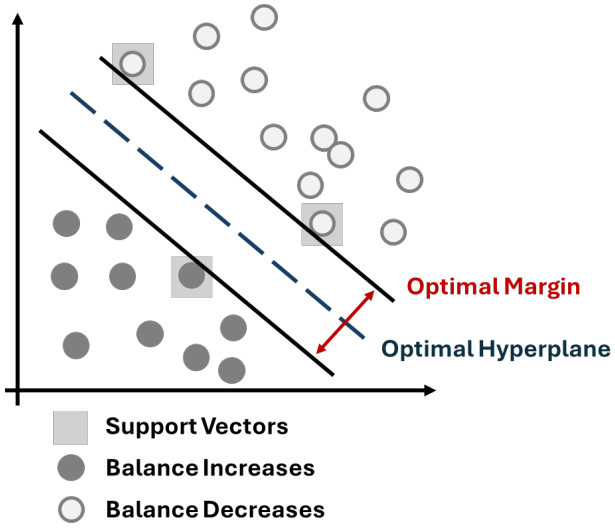

Figure 5: Linear SVM classification model

In Figure 5, we can observe a two-dimensional linear SVM model. The axes represent a set of two variables that have a direct relationship with increases or decreases in balances. The goal of the model is to find a hyperplane that correctly separates the observations. The support vectors, represented by the gray squares, help define the widest possible margin between the two classes.

Non-linear extension using kernel functions

Given that the relationship between sentiment indicators and deposit variation may be non-linear, the model is extended using a kernel function $K(x_i, x_j)$ which measures the similarity between observations in a transformed space. This approach allows projecting the data into a higher-dimensional space without explicitly computing that transformation.

A commonly used kernel function is the radial basis function (RBF) kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

where γ controls the relative influence of each observation.

Model Training and Validation

The dataset is divided into training and test subsets, preserving the temporal structure to avoid information leakage. The model is trained using the training set and evaluated on the test set. The selection of hyperparameters (external values that regulate the model's performance and are chosen during the validation process), (C, γ) is carried out through cross-validation, maximizing the model's predictive performance.

Model Performance Evaluation (Confusion Matrix)
The classifier's performance is evaluated using the confusion matrix, which summarizes the model's predictions compared with the observed values:

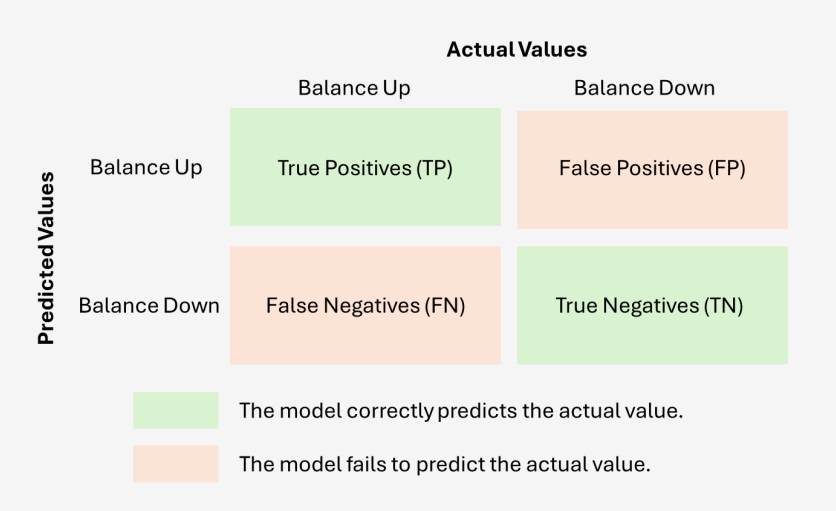

Figure 6: Confusion matrix of the classification model

This matrix (Figure 6) makes it possible to identify asymmetric classification errors, which are particularly relevant in financial risk analysis, where the consequences of misclassification may differ depending on the type of error.

Evaluation Metrics
Based on the confusion matrix, the following metrics are calculated:

- Accuracy

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

- Precision

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

- Recall

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

- F1-score (a metric that combines precision and recall into a single performance indicator)

$$\boldsymbol{F1} = 2 \cdot \frac{\boldsymbol{Precision} \cdot \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

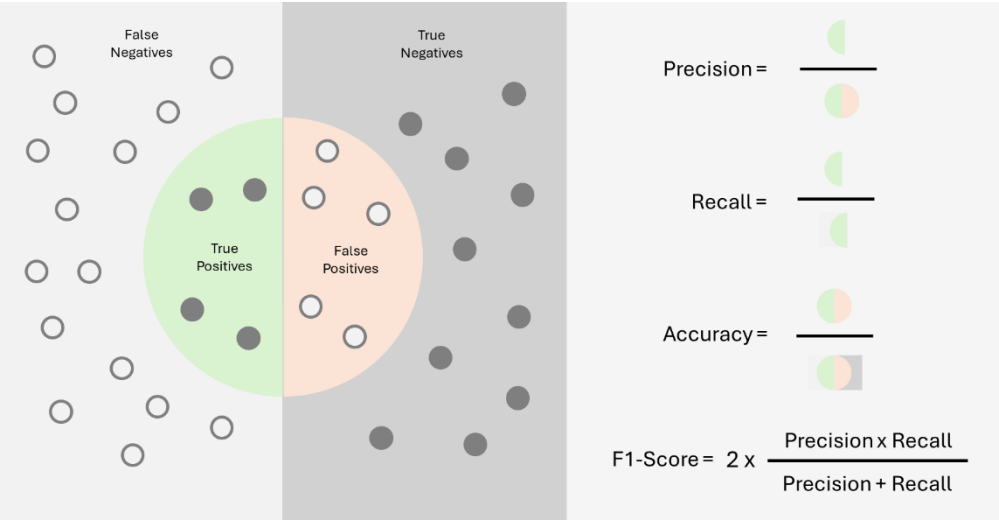

Figure 7: Evaluation metrics for the classification model

These metrics make it possible to evaluate not only the model's overall accuracy, but also its ability to correctly identify periods of increasing deposits. This is particularly relevant from a liquidity monitoring and management perspective, where the costs associated with classification errors can be asymmetric.

Model Interpretation

If the SVM model shows predictive performance superior to that of a random classifier, it is interpreted as evidence that sentiment indicators contain relevant information for anticipating changes in deposit balances. This result suggests that sentiment analysis can function as an early-warning signal that complements traditional financial indicators.

APPLICATIONS AND USE

The proposed model can be used as a complementary tool for analyzing and monitoring the behavior of bank deposits, integrating alternative information from social media with Machine Learning techniques. Its main value lies in its ability to transform qualitative information into quantitative signals that support decision-making across different areas of financial and risk management.

Early Monitoring of Deposit Changes

The model enables continuous monitoring of social media sentiment as a leading indicator of changes in deposit balances. Early detection of sentiment deterioration generates alerts regarding potential deposit outflows, facilitating preventive liquidity management.

Activation of the Contingency Funding Plan

The model can serve as a mechanism for activating a bank's contingency funding plan. By providing signals of worsening sentiment and anticipating changes in deposits, it allows institutions to implement appropriate funding measures in critical situations.

Support for Liquidity Risk Management
This model can be integrated as an additional source of information within liquidity risk management frameworks. It provides signals about the expected direction of deposits, complementing traditional and regulatory analyses.

Assessment of Reputational Risk Impact
Sentiment indicators reflect public perception of the bank and can be used to assess the impact of reputational events, communication campaigns, or strategic changes. Tracking sentiment before and after these events provides valuable insights for decision-making by governance bodies and senior management.

Exploratory and Analytical Use
The model can also be employed for exploratory purposes to analyze the relationship between social media sentiment and financial variables over time. This approach helps identify patterns and evaluate hypotheses, laying the groundwork for developing more advanced models not only in liquidity risk, but also in other areas of financial institutions or for integrating new sources of alternative data into financial analysis.

TECHNICAL ASSUMPTIONS

The model is based on several assumptions that simplify the complexity of the phenomenon and allow its implementation. These assumptions do not aim to fully describe all factors influencing deposit balances, but rather to establish a coherent framework for using social media information as a predictive signal.

1. **Representativeness of comments:** It is assumed that social media comments are a representative sample of public sentiment toward the bank, reflecting relevant perceptions, especially during periods of high digital interaction.
2. **Stability of sentiment:** It is assumed that the relationship between comment content and sentiment classification remains stable over time, without significant changes in language or classification rules.
3. **Validity of sentiment indicators:** It is assumed that the constructed indicators adequately capture the prevailing tone of social media conversations and remain comparable over time.
4. **Relationship between sentiment and deposits:** It is taken as a starting point that there is a meaningful relationship between sentiment indicators and the direction of changes in deposit balances, interpreted as predictive.
5. **Temporal stability of the model:** It is assumed that the relationship learned by the model during training remains valid during evaluation and application, without abrupt structural changes.
6. **Independence from external factors:** It is assumed that the effects of macroeconomic or regulatory factors do not overshadow the relationship between sentiment and deposits, allowing sentiment indicators to contribute additional information.

RISKS AND LIMITATIONS

The model presents certain risks and limitations that must be considered when interpreting its results. These limitations arise from both the nature of the data and the methodological decisions involved.

1. **Social media bias:** Comments may not adequately represent all depositors, leading to an over-representation of extreme opinions.
2. **Language noise:** The informal and ambiguous nature of social media language can introduce errors in sentiment classification, affecting the model's performance.
3. **Limitations of sentiment analysis:** The effectiveness of the model depends on the quality of the initial sentiment analysis. Changes in user language or behavior may reduce its predictive capacity.
4. **Risk of overfitting:** There is a risk that the model becomes overly fitted to historical data, capturing patterns that may not repeat over time.
5. **Limited interpretation:** The model predicts the direction of changes in deposit balances but does not estimate the magnitude (volume) of the effect nor establish causal relationships.
6. **Sensitivity to exceptional events:** Extraordinary events may alter depositor behavior and social media dynamics, limiting the model's predictive ability.

CONCLUSIONS

Sentiment analysis applied to social media information represents a valuable complementary tool for anticipating changes in bank deposit behavior. By transforming unstructured data into quantitative indicators, the proposed model enables the incorporation of an additional dimension of perception and reputational risk into traditional liquidity monitoring. Although it presents limitations inherent to the nature of the data and the methodological assumptions, its use as an early-warning signal can strengthen the preventive management of liquidity risk and support decision-making in contexts of financial stress.

References

(1) Baker, S. R., Bloom, N., & Davis, S. J. (2016). *Measuring economic policy uncertainty*. The Quarterly Journal of Economics, 131(4), 1593–1636.

(2) Banco de México. (2024, diciembre). *Reporte de estabilidad financiera: Diciembre 2024*. Banco de México.

(3) Bank for International Settlements. (2022). *Project Ellipse: Regulatory reporting and data analytics platform*. BIS Innovation Hub.

(4) Cookson, J. A., Fox, C., Gil-Bazo, J., Imbet, J. F., & Schiller, C. (2023). *Social media as a bank run catalyst*. FDIC Bank Research Conference.

(5) Del Sarto, N., Bocchialini, E., Gai, L., & Ielasi, F. (2024). *Digital banking: how social media is shaping the game*. Qualitative Research in Financial Markets.

(6) Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)* (pp. 4171–4186). Association for Computational Linguistics.

(7) Giuliana, R., Panfilo, M., & Peltonen, T. (2024). *Deposit flows during monetary tightening: The role of digital banking and social media*. European Central Bank / ESRB Workshop.

(8) Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

(9) Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed., draft). Online manuscript.

(10) Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

(11) Loughran, T., & McDonald, B. (2011). *When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks*. Journal of Finance, 66(1), 35–65.
(12) Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using Machine Learning techniques*. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 79–86). Association for Computational Linguistics.
(13) Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). *Lexicon-based methods for sentiment analysis*. Computational Linguistics, 37(2), 267–307.
(14) Tetlock, P. C. (2007). *Giving content to investor sentiment: The role of media in the stock market*. The Journal of Finance, 62(3), 1139–1168.