JOURNAL
OF EUROPEAN ECONOMY
Vol. 25. № 1 (96). January–March 2026
Publication of West Ukrainian National University

Economic Theory

Leonidas THEODORAKOPOULOS,
Alexandra THEODOROPOULOU,
Evangelos SISKOS,
Yevhen SAVELYEV

THE ECONOMIC COST OF FINANCIAL FAKE NEWS IN EUROPEAN CAPITAL MARKETS

Abstract

Financial news misinformation is a systemic risk that distorts price discovery and capital allocation by widening information asymmetry and weakening investor trust. This study develops a scalable NLP and machine-learning framework to detect deceptive financial narratives in large digital corpora through Big Data. After text normalization, lemmatization, and stopword elimination, the framework contrasts TF-IDF with Word2Vec embeddings and trains Logistic Regression,

© Leonidas Theodorakopoulos, Alexandra Theodoropoulou, Evangelos Siskos, Yevhen Savelyev 2026.

Theodorakopoulos Leonidas, PhD (Big Data in Management and Economics), Adjunct Professor, Department of Management Science and Technology, University of Patras, Patras, Greece. ORCID: 0000-0002-0891-6780 Email: theodleo@upatras.gr
Theodoropoulou Alexandra, MSc (Digital Innovation and Management), PhD Candidate, Department of Management Science and Technology, University of Patras, Patras, Greece. ORCID: 0009-0004-6314-7795 Email: theodoropouloua@upatras.gr
Siskos Evangelos, DSc (Economics), Professor of International, European and Black Sea Economic Relations, Department of International and European Economic Studies, University of Western Macedonia, Kozani, Greece. ORCID: 0000-0002-5221-4444 Email: esiskos@uowm.gr
Savelyev Yevhen, DSc (Economics), Professor, Department of International Economics, West Ukrainian National University, Ternopil, Ukraine. ORCID: 0000-0003-0137-2263 Email: savelyev@wunu.edu.ua

Random Forest, and Gradient Boosting classifiers. Performance is assessed with Accuracy, Precision, Recall, F1-score, and ROC-AUC. Across models, TF-IDF provides stronger discrimination than Word2Vec; the TF-IDF Random Forest reaches near-perfect results (ROC-AUC 0.9999; Precision 0.9977). The emphasis on transparent, feature-based models supports auditability (for example, via feature importance) and helps limit harmful false positives that could suppress legitimate signals. The results indicate that high-precision, interpretable pipelines can reduce the verification gap in fast-moving information environments, mitigate macroeconomic costs of deceptive narratives, and inform DSA- and ESMA-aligned market surveillance workflows. The framework is designed for deployment on streaming news feeds and large-scale platform archives.

Key Words:

Big Data, economic cost, European capital markets, financial decision-making, financial misinformation, market manipulation, Natural Language Processing.

JEL: G14, D82, C55.

2 figures, 4 tables, 14 formulas, 20 references.

Problem Statement

In the contemporary digital economic landscape, the efficiency of financial markets is increasingly dependent on the quality and integrity of information flows. While the rapid digitalization of media has reduced transaction costs, it has also facilitated the emergence of financial misinformation as a significant systemic risk. From an economic perspective, fake financial news represents a severe form of information asymmetry that distorts the price discovery process and leads to suboptimal capital allocation.

The macroeconomic consequences of deceptive narratives are not merely theoretical. A notable instance occurred in 2013 when a single fraudulent report regarding a

White House explosion caused an instantaneous $130 billion decrease in market value (Selyukh, 2013). Such events underscore how «noise» in the information ecosystem can trigger irrational market behavior and widespread economic instability.

The challenge for modern European economies is that the volume of data produced by online media, social networks, and algorithmic trading platforms now outpaces traditional human verification procedures. This creates a «verification gap» that can be exploited for market manipulation. Consequently, developing automated, scalable systems for detecting untrustworthy content is no longer just a technical objective; it is a requirement for maintaining market integrity and protecting investor trust.

The purpose of this study is to identify economic approaches for quantifying the economic cost of financial fake news in European capital markets. The study proposes a framework that utilizes Natural Language Processing (NLP) and Machine Learning (ML) to identify trustworthy and deceptive reporting characteristics within extensive financial datasets (Du et al., 2024). Unlike general-domain detection systems, our approach accounts for the specific technical terminology and market-based emotional indicators inherent in financial discourse.

We evaluate three supervised learning models – Logistic Regression, Random Forest, and Gradient Boosting – across two feature representation methods: TF-IDF and Word2Vec. By focusing on model interpretability and statistical robustness, this research aims to provide a reliable monitoring tool capable of identifying the subtle linguistic markers of financial fraud, thereby safeguarding the foundational pillars of a stable digital economy. Within the European Union, the regulatory landscape has evolved rapidly with the implementation of the Digital Services Act (DSA), which mandates that systemic platforms take proactive measures against the spread of harmful misinformation. From a policy perspective, the challenge lies in developing «auditable» and «interpretable» systems that can distinguish between healthy market speculation and coordinated deceptive narratives that threaten market integrity. By focusing on high-accuracy machine learning models that remain transparent to auditors, this research provides a methodology for operationalizing these European regulatory goals.

The remainder of this paper is organized as follows: The Literature Review section establishes the theoretical framework and reviews relevant literature concerning information asymmetry and computational fake news detection. The Methodology section details the research methodology, encompassing the dataset characteristics, the multi-stage preprocessing pipeline, and the mathematical foundations of the feature extraction and machine learning architectures employed. The Research Results section presents a comprehensive analysis of the experimental results, utilizing diverse performance metrics and visual representations to evaluate model efficacy. The Discussion section discusses the practical implementation of the framework within the European regulatory context and outlines strategic directions for future research in the field of financial informatics. Finally, the Conclusions section synthesizes the key findings and highlights the broader economic implications for market integrity and financial news credibility assessment.

Literature Review

Research about fake news detection through computational methods continues to expand, but financial misinformation remains insufficiently studied. Our initial research phase focused on fundamental fake news detection methods that used text analysis and stylistic and psycholinguistic signal detection techniques. Ahmad et al. (2020) demonstrated the effectiveness of the ensemble method through bagging and boosting tests on different datasets, proving that model combination improves classification reliability.

The detection of market rumors and manipulative reports using NLP/ML has been demonstrated in finance-specific settings (Alshuwaier & Alsulaiman, 2025). For instance, Majumdar and Bose (2018) present a big-data framework to flag financial rumors around the Bombay Stock Exchange, while Cheng et al. (2023) use social-media rumor signals and machine-learning classifiers to predict abnormal trading. Manipulative campaigns such as cryptocurrency pump-and-dump schemes have also been detected using text signals from messaging platforms (Nghiem et al., 2021). Linguistic regularities in financial discourse e.g., event-phrasing, entity repetition, and extremity in wording are consistent with deception markers documented in financial communications (Larcker & Zakolyukina, 2012) and with domain-specific lexicons needed to model financial language (Loughran & McDonald, 2011). Finally, the field still lacks widely accepted, finance-specific misinformation benchmarks; recent efforts such as Fin-Fact were proposed precisely because of this gap (Rangapur et al., 2025), and broader surveys note dataset fragmentation more generally (D'Ulizia et al., 2021).

The research of Alghamdi et al. (2024) and Mishra et al. (2022) established three categories for fake news detection which include content-based, context-based and hybrid approaches. The authors demonstrate that classical ML models achieve both high interpretability and equivalent performance through proper NLP preprocessing and feature engineering techniques. The research findings lead to the current study which uses content-based analysis through structured textual features because these methods match the requirements for financial applications that need transparency and scalability.

The research extends previous studies through its application of financial text analysis to classical NLP and ML systems which examine how different feature types and classifier architectures affect domain-specific text patterns. The results of the study help identify both the advantages and disadvantages of using conventional models to identify financial misinformation.

The Economic Cost of Financial Information Asymmetry

From an economic perspective, fake financial news represents a severe form of information asymmetry that distorts market efficiency. When untrustworthy content is disseminated, it creates «noise» that interferes with the price discovery process, leading to suboptimal capital allocation. As evidenced by the 2013 Associated Press hack (Selyukh, 2013), a single piece of misinformation can cause instantaneous wealth destruction on a macroeconomic scale. Automated detection systems, therefore, serve as critical infrastructure for maintaining market integrity and protecting investor trust, which are foundational pillars of a stable digital economy. The literature increasingly recognizes that financial misinformation is not merely a social problem but a structural economic inefficiency (Lyzun et al., 2019; Lyzun et al., 2023). While early studies focused on simple keyword filtering, the shift toward Big Data analytics allows for a more nuanced understanding of how deceptive narratives circumvent market filters. In the European context, where market fragmentation across different nations is high, the ability of NLP models to maintain high precision across diverse financial news sources is critical for preventing cross-border information arbitrage.

Methodology

The research methodology is divided into four main stages: (1) dataset preparation, (2) text preprocessing, (3) feature extraction, and (4) model classification. The study also implements two parallel NLP pipelines using TF-IDF and Word2Vec representations.

Dataset and Labeling

The dataset used in this study comes from the publicly available «Fake and Real News Dataset» on Kaggle (kaggle.com, n.d.), with filters applied to it to include articles related to financial or economic topics. It consists of two separate CSV files:

- Fake.csv: contains fabricated news articles (23,481 entries)
- True.csv: contains legitimate news articles (21,417 entries)

Each file includes the following columns:

- Title: The headline of the news article
- Text: The full body of the article
- Subject: A broad category label
- Date: The publication date of the article

For this study, a binary classification label was manually assigned:

- Label 0: Fake news (from Fake.csv)
- Label 1: True news (from True.csv)

The two datasets were combined into a single corpus containing 44,898 news articles for our analysis. The title and text fields were combined into a single input feature, «content», for each instance. The «subject» and «date» columns received no modeling treatment, but can be used by researchers for future exploratory work and temporal analysis. The combined dataset was randomly arranged to prevent order-based bias from affecting model training.

Text Preprocessing

The text data required preprocessing to obtain useful features because multiple normalization and cleaning operations were performed. These operations served as an important tool that eliminated unneeded data points and inconsistent information, while minimizing language differences, which resulted in better machine learning model input quality.

- **Lowercasing**: The entire text was changed to lowercase in order to ensure uniformity. This step prevents the model from treating words like «Trump» and «trump» as different entities, reducing redundancy in the vocabulary.
- **Digit and Punctuation Removal**: Regular expressions were used, together with string translation methods, to eliminate all numeric digits and punctuation marks from the text. Such elements in this context lack meaningful semantic value so their removal simplifies the text.
- **Tokenization**: The text was then split into individual words, or tokens. This process enables downstream tasks, such as stopword removal and lemmatization, to be applied to each word. The TF-IDF pipeline used basic string splitting, while the Word2Vec-based pipeline used gensim.simple_preprocess, which proved sufficient due to its non-contextual nature.

- **Stopword Removal**: Common English stopwords such as «the», «is», and «and» were removed using NLTK's stopword list. These high-frequency words contribute little to distinguishing fake from real news and often introduce noise into the model.
- **Lemmatization**: Each word was lemmatized using the WordNetLemmatizer from NLTK. This tool reduces words to their base or dictionary form (e.g., «running» becomes «run»). This process helps in grouping different forms of a word under a single representation, thereby reducing dimensionality and enhancing semantic consistency.

These preprocessing techniques were applied uniformly to both pipelines (Word2Vec and TF-IDF) so as to make sure that feature extraction was performed on clean and semantically meaningful text.

Feature Extraction

Two separate numerical transformation methods were applied to the text, to prepare it for machine learning analysis, which included Word2Vec embeddings and TF-IDF vectorization. The two methods represent opposing text representation approaches because they use contextual meaning and statistical word frequency to create their representations.

Word2Vec Embeddings

Word2Vec is a NLP technique that uses a shallow, two-layered neural network model and was developed by Google; it learns distributed representations of words in a continuous vector space, such that semantically similar words are mapped to nearby points. In this experiment, we trained a Skip-gram Word2Vec model on the corpus with a vector size $d = 100$, window size $w = 5$, and minimum word frequency threshold of 2.

The Skip-gram model maximizes the probability of a word's surrounding context given the word itself. For a sequence of training words $w_1, w_2, \ldots, w_T$, the objective is:

$$\max \frac{1}{T} \sum_{t=1}^{T} \sum_{-w \le j \le w,\, j \neq 0} \log P(w_{t+j} \mid w_t), \tag{1}$$

where the conditional probability $P(w_{t+j} \mid w_t)$ is computed using softmax:

$$P(w_0 \mid w_1) = \frac{\exp(u_{w_0}^T u_{w_1})}{\sum_{w=1}^{|V|} \exp(u_w^T u_{w_1})}. \tag{2}$$

Here, u_{w_1} is the input vector of the center word; u_{w_0} is the output vector of the context word; and $|V|$ is the vocabulary size.

Once word embeddings are learned, each document is transformed into a fixed-length vector by averaging its constituent word vectors. For a document represented by a set of n valid tokens $\{w_1, w_2, \ldots, w_n\}$, the document vector $\vec{d}$ computed as:

$$\vec{d} = \frac{1}{n}\sum_{i=1}^{n} \vec{w}_i. \quad (3)$$

TF-IDF Vectorization

The second feature extraction technique employed was Term Frequency–Inverse Document Frequency (TF-IDF), a widely used statistical measure that reflects how important a word is to a document relative to a corpus.

The TF-IDF weight for a term t in document d is defined as:

$$TF - IDF(t,d) = TF(t,d) * IDF(t), \quad (4)$$

where:

- **Term Frequency (TF)** is the normalized frequency of the term in the document:

$$TF(t,d) = \frac{f_{t,g}}{\sum_{t' \in d} f_{t',d}}, \quad (5)$$

with $f_{t,d}$ being the raw count of term t in document d.

- **Inverse Document Frequency (IDF)** measures how rare the term is across all documents:

$$IDF(t) = \log\left(\frac{N}{1+n_t}\right), \quad (6)$$

where N is the total number of documents; n_t is the number of documents containing the term t.

TF-IDF emphasizes terms that are frequent in a given document but infrequent across the corpus, helping to reduce the impact of common but less informative words.

In our implementation, we used the TfidfVectorizer from scikit-learn with a vocabulary limited to 5,000 most frequent terms. This resulted in a sparse matrix

of dimensionality N×5000, where each row corresponds to a document vector and each column corresponds to a term in the vocabulary.

Model Training and Evaluation

To classify news articles as either fake or true, three classical machine learning classifiers were employed: Logistic Regression, Random Forest, and Gradient Boosting. Each model was applied using both feature representations (Word2Vec and TF-IDF), and their performance was evaluated using a consistent set of quantitative metrics.

Logistic Regression

Logistic Regression is a linear probabilistic classifier that models the probability of the binary label using a logistic (sigmoid) function. For an input vector $\vec{x} \in R^d$, the probability that the output label $y \in \{0,1\}$ is true (i.e., news is real) is given by:

$$P(y = 1 \mid \vec{x}) = \sigma(\vec{w}^T \vec{x} + b) = \frac{1}{1 + e^{-(\vec{w}^T \vec{x} + b)}}, \quad (7)$$

where $\vec{w}$ is the weight vector; b is the bias term; and $\sigma(.)$ is the sigmoid activation function.

Model parameters $\vec{w}$ and b are learned by minimizing the log-loss (cross-entropy) function over the training set:

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]. \quad (8)$$

Random Forest

Random Forest is an ensemble learning algorithm that constructs a large number of decision trees during training and outputs the mode (majority vote) of the individual trees' predictions for classification tasks.

Each decision tree is trained on a random bootstrap sample of the training data, and at each split in a tree, a random subset of features is considered. This promotes diversity among trees and reduces overfitting.

The prediction for input $\vec{x}$ is given by:

$$y = majority\_vote(h_1(\vec{x}), h_2(\vec{x}), ..., h_T(\vec{x})), \quad (9)$$

where $h_T(\vec{x})$ is the prediction from the t^{th} decision tree, and T is the total number of trees in the forest.

Gradient Boosting Classifier

Gradient Boosting is another ensemble method that builds a sequence of weak learners (typically shallow decision trees) in a stage-wise fashion. Each new learner focuses on correcting the residuals (errors) of the previous learners by minimizing a differentiable loss function (typically log-loss for binary classification).

Let $F_m(\vec{x})$ denote the model after mmm iterations. It is updated as:

$$F_{m+1}(\vec{x}) = F_m(\vec{x}) + \eta * h_m(\vec{x}), \quad (10)$$

where $h_m(\vec{x})$ is the base learner fitted to the negative gradient of the loss; η is the learning rate (shrinkage parameter); $F_0(\vec{x})$ is initialized to a constant (e.g., mean log-odds).

Evaluation Metrics

To comprehensively evaluate the performance of each model, we employed several standard classification metrics:

- **Accuracy**: The proportion of correctly predicted instances.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + \\ + False\ Positives + False\ Negatives}. \quad (11)$$

- **Precision**: The proportion of positive predictions that are actually positive (useful in minimizing false positives).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}. \quad (12)$$

- **Recall (Sensitivity)**: The proportion of actual positives correctly identified (useful in minimizing false negatives).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}. \quad (13)$$

- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two.

$$F1\_Score = 2 * \frac{Precision * Recall}{Precision + Recall}. \quad (14)$$

- **ROC-AUC (Area Under the Curve)**: Measures the area under the Receiver Operating Characteristic curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. AUC values closer to 1.0 indicate better discrimination between classes.
- **ROC Curve**: A graphical tool that illustrates the trade-off between sensitivity and specificity across different decision thresholds.

Research Results

This section shows the experimental results from financial fake news detection tasks using classification models before presenting an extensive analysis of their performance. From a regulatory perspective, the high precision (0.9977) of the Random Forest model is more significant than its overall accuracy. In financial markets, a «false positive» (mislabeling legitimate news as fake) can be as damaging as a «false negative», as it may lead to the suppression of valid market signals. The statistical robustness shown in the ROC-AUC curves (0.9999) suggests that this framework provides a reliable «safety net» that can be deployed without introducing new distortions into the price discovery process. The research compares three popular machine learning classifiers that also applied TF-IDF and Word2Vec feature extraction methods; (a) Logistic Regression, (b) Random Forest and (c) Gradient Boosting. The assessment of each model used standard performance metrics that included accuracy, precision, recall, F1 score, and ROC-AUC to evaluate their overall performance. This research aims to determine which model works best under different representations, all the while analyzing how different feature types affect classifier performance. The research results are presented through quantitative tables and visual ROC curves that provide straightforward analysis of financial fake news detection approach performance.

Logistic Regression Comparison

The Logistic Regression classifier showed almost perfect results with both feature types, but the TF-IDF version slightly outperformed the Word2Vec version (see Table 1). More specifically, TF-IDF yielded an accuracy of 0.9895 compared to Word2Vec's 0.9636, and F1 score improved from 0.9624 to 0.9892. The TF-IDF model also demonstrated better balance between precision and recall, indicating superior generalization. The ROC-AUC values for both were above 0.99, but again TF-IDF noted a bit higher rates at 0.9992.

Table 1

Logistic Regression Results

Metric	TF-IDF	Word2Vec
Accuracy	0.9895	0.9636
Precision	0.9844	0.9592
Recall	0.9940	0.9656
F1 Score	0.9892	0.9624
ROC-AUC	0.9992	0.9942

Source: authors' calculations.

This performance gap illustrates that for linear models like Logistic Regression, high-dimensional sparse representations like TF-IDF provide greater discriminative power, especially when the text patterns are distinctive between classes, as is often the case in fake vs. real news.

Random Forest Comparison

Random Forest models showed the largest performance differential between the two feature representations. When paired with TF-IDF, Random Forest achieved near-perfect results: accuracy and F1 score of 0.9978, and ROC-AUC of 0.9999. In contrast, the Word2Vec variant performed significantly lower, with accuracy of 0.9548 and F1 score of 0.9527 (see Table 2).

Table 2

Random Forest Results

Metric	TF-IDF	Word2Vec
Accuracy	0.9978	0.9548
Precision	0.9977	0.9610
Recall	0.9977	0.9446
F1 Score	0.9977	0.9527
ROC-AUC	0.9999	0.9916

Source: authors' calculations.

This gap can be attributed to how Random Forests operate. Since decision trees in the ensemble rely on distinct thresholds and splits in the input space, the sparse and feature-rich TF-IDF vectors allow for more precise partitioning, whereas the dense and averaged Word2Vec vectors smooth out many informative distinctions between documents.

Gradient Boosting Comparison

Gradient Boosting models also favored TF-IDF. With TF-IDF, the model reached an accuracy of 0.9953 and an F1 score of 0.9952, while the Word2Vec counterpart dropped to 0.9359 in accuracy and 0.9337 in F1 score (see Table 3). Although Gradient Boosting still performed well with Word2Vec, the performance decline was consistent with the pattern seen in the other models.

Table 3

Gradient Boosting results

Metric	TF-IDF	Word2Vec
Accuracy	0.9953	0.9359
Precision	0.9940	0.9301
Recall	0.9963	0.9374
F1 Score	0.9952	0.9337
ROC-AUC	0.9987	0.9858

Source: authors' calculations.

This further confirms that context-independent vector averaging in Word2Vec may discard critical cues, which TF-IDF is able to retain by directly leveraging term frequencies and document rarity.

ROC curves for TF-IDF

The ROC curves for the Logistic Regression, Random Forest, and Gradient Boosting classifiers using TF-IDF feature representation indicate exceptional classification performance across all three models (see Figure 1).

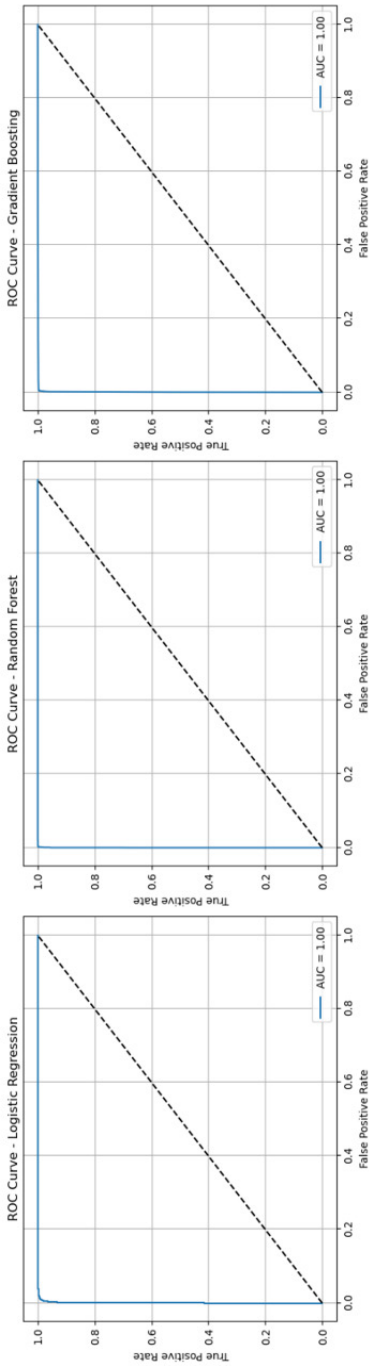

Figure 1
ROC curves for TF-IDF

Source: made by the authors.

The curves maintain perfect alignment with the top-left border of the ROC space because they achieve perfect equilibrium between true positive rate (sensitivity) and false positive rate. The shape of the curves alone suggests that the classifiers are highly effective at distinguishing between fake and real news.

As we can see, the models produce AUC (Area Under the Curve) scores that approach 1.00. This shows their exceptional performance. The Random Forest model shows an almost perfect vertical increase, followed by a horizontal line at the top, which indicates that it maintains high true positive rates with no false positives at all thresholds. The Logistic Regression model produces a curve that is both smooth and steep because it generates strong probabilistic output results. Lastly, the Gradient Boosting model produces results that are reliable, but its initial curve segment shows a gentler slope than the other models.

The above results confirm that the TF-IDF tool provides a robust and discriminative feature space for all three classifiers. The near-perfect ROC curves indicate that the models used in these experiments are not only accurate in their final predictions, but are also well-calibrated in ranking instances by how probable it is for news to be real or fake. The ROC analysis supports the finding that classical models using Random Forest and Gradient Boosting, in combination with TF-IDF features, produce highly effective fake news detection solutions.

ROC curves for Word2Vec

The ROC curves for the Logistic Regression, Random Forest, and Gradient Boosting classifiers, under the use of Word2Vec embeddings, show that all three models achieve high classification performance (see Figure 2). The ROC space curves show steep upward movement toward their top-left corner, something that shows high true positive rates at various threshold settings. All three models demonstrate successful identification of authentic news content and fabricated news through their low rate of incorrect positive results.

The Logistic Regression model shows a smooth curve that rises sharply because it operates as a linear model that produces probability values. The Random Forest curve, on the other hand, shows a fast initial increase because it keeps delivering good results, while splitting the feature space effectively when input vectors become dense. Finally, the Gradient Boosting model shows a curve that rises steadily to enhance recall performance while keeping false positive rates at a minimum. This demonstrates its ability to improve classification accuracy through sequential error correction.

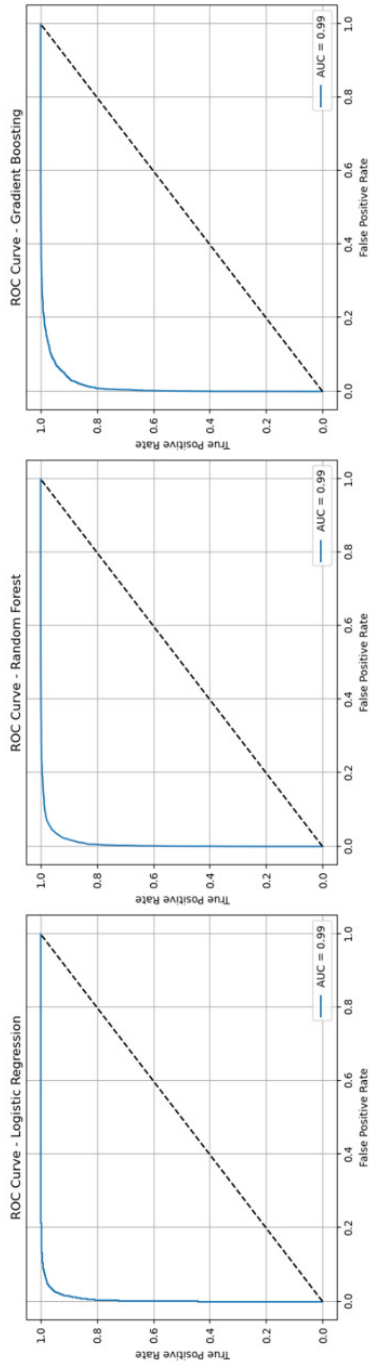

Figure 2
ROC curves for Word2Vec

Source: made by the authors.

The three models produce AUC values of 0.99; this indicates that they perform at an excellent, almost perfect, level. The shape and placement of these curves confirm that the Word2Vec embeddings are effective in capturing relevant semantic features for the binary classification task of these experiments. The two classifiers generate dependable results through dense, vector-based representations, despite the fact that each model uses different learning strategies.

Comparison of ROC curves

The ROC curves for the three classifiers used in this study, using Word2Vec embeddings, show that all three models achieve high discriminative power although their performance is a little lower than their TF-IDF counterparts. The curves demonstrate superior performance because they exist above the random guess baseline, which represents the diagonal line. However, there is a visible softening of the curve shape compared to the TF-IDF models, suggesting a slightly reduced ability to separate the two classes at all thresholds.

As stated above, the three models produce AUC values that approach 0.99, which indicates their excellent performance. The Logistic Regression model produces a smoothly increasing curve because it applies probabilistic scoring to produce its results. The Random Forest and Gradient Boosting models show high TPR values at low FPR points, but their performance curves have a gentler slope and more rounded shape during the initial stages compared to TF-IDF models.

This performance difference is likely due to the averaging nature of Word2Vec representations, which can dilute the specific lexical cues important for distinguishing fake news from real news. While Word2Vec captures semantic relationships well, the loss of word-level specificity appears to hinder its ability to support tree-based models, such as Random Forest and Gradient Boosting, as much as TF-IDF does. Overall, while the ROC curves affirm that Word2Vec still enables effective classification, they also reinforce the earlier conclusion: TF-IDF offers superior performance in this task due to its ability to preserve word frequency and importance, which are critical for detecting subtle stylistic and lexical patterns inherent in fake news content.

Discussion

Practical Implementation and Economic Policy Implications

The transition from a theoretical detection framework to a functional market-oversight tool requires a clear implementation strategy that aligns with the European digital agenda. For the European economy, the practical value of the proposed Natural Language Processing (NLP) model lies in its capacity to reduce the systemic «noise» that hinders efficient capital allocation and disrupts the price discovery mechanism.

Integration into European Market Surveillance

The high statistical reliability of the Random Forest and TF-IDF pipeline facilitates its direct integration into the surveillance infrastructures of national and EU-level financial authorities, such as the European Securities and Markets Authority (ESMA). By deploying this model as an automated, pre-emptive filter for high-frequency news feeds and social media data streams, regulators can achieve a more proactive stance against market manipulation. Furthermore, the inherent interpretability of the Random Forest architecture ensures that the feature importance metrics—such as specific linguistic markers or fabricated sentiment triggers—can be documented. This transparency is essential for legal enforcement and regulatory compliance under the Market Abuse Regulation (MAR) (European Parliament & Council of the European Union, 2014), providing an auditable trail that «black-box» deep-learning models lack.

Institutional Guardrails and Price Discovery Protection

Beyond regulatory use, the framework offers significant protections for both institutional and retail investors. Information asymmetry typically penalizes participants who lack the resources for instantaneous fact-checking. By implementing this framework, financial news aggregators and trading terminals can provide a «Linguistic Authenticity Rating» for digital content, allowing market participants to distinguish between evidence-based reporting and speculative misinformation. Furthermore, in the high-frequency trading (HFT) environment, the model's output can serve as a critical risk-management parameter. If a news event triggers a high misinformation probability flag, automated systems can be programmed to lower their sensitivity to that specific news stream, effectively acting as a digital circuit breaker to prevent cascading sell-offs based on fabricated data.

Operationalizing the EU Digital Services Act (DSA)

Under current European regulations, systemic platforms are increasingly responsible for the financial and societal risks hosted on their networks. This study provides a technical benchmark for the «proactive mitigation of systemic risks» as mandated by Article 35 of the Digital Services Act (DSA) (European Parliament & Council of the European Union, 2022). It offers a standardized methodology for platforms to identify market-manipulating content through a process that is visible to independent auditors and vetted researchers. By adopting these NLP benchmarks, European news outlets and social platforms can foster a more transparent digital ecosystem, ensuring that the velocity of digital information does not compromise the accuracy required for a stable and resilient economy.

Resource Efficiency and Cross-Border Scalability

Traditional manual fact-checking is insufficient in a digital market where news travels in milliseconds. The computational efficiency of the TF-IDF-based approach allows for the simultaneous processing of vast data volumes across the 27 EU member states. Such scalability is essential for preventing information arbitrage, where misinformation is utilized to exploit price differences between fragmented national markets before human oversight can intervene. Ultimately, the deployment of this framework supports the European Union's goal of creating a unified, secure, and transparent Single Digital Market.

Future Work

Our results show that standard machine-learning models with a careful NLP pipeline can flag misleading financial articles with strong accuracy on our test sets. Still, a practical system for newsrooms or regulators needs more than accuracy on a benchmark. It should work across outlets and languages, stay reliable when topics shift, and explain its flags in ways a human can check. To move in that direction, future work could concentrate on the following 5 areas.

First, it is possible to build richer, finance-specific datasets that reflect real publishing patterns. Second, the use of language models especially tuned to financial text so the system understands reporting styles and entity names. Third, connecting text with simple market signals (prices, volume, volatility) to reduce false alarms. Fourth, the possibility of handling timing and spread: stories move across sources and platforms, and models should account for that and remain stable under change. Fifth, the deployment of guardrails human review, drift monitoring, and clear logs so decisions can be audited and improved over time (Hu et al., 2022; Theodorakopoulos et al., 2025). Table 4 below summarizes these 5 possible future directions.

Table 4

Future research directions

Category	Research focus	Data/signals needed	Methods (sketch)	Risks & ethics
Data enrichment & diversification	Build finance-specific, representative datasets for misinformation detection.	Labeled news and filings, verified fact-checks, social streams (X/Reddit/ StockTwits), multilingual sources.	Clear labeling guidelines; weak supervision to scale; publish datasheets.	Label bias, privacy/ToS issues, uneven market coverage.
Domain models & adaptation	Use finance-tuned language models with short, evidence-grounded outputs.	Your corpus + Financial PhraseBank/earnings calls; entity dictionaries.	Fine-tune FinBERT/FinancialBERT or adapters; rationale extraction.	Hallucinated rationales, leakage of sensitive text.
Multimodal market linkage	Combine text with market data to reduce false positives/ negatives.	News text plus OHLCV, volatility, options activity, timestamps.	Late/attention fusion; event linking between articles and prices.	Chasing normal price moves by mistake; making swings worse by reacting to them.
Temporal / graph robustness	Model spread over time and remain stable under shifts.	Time-stamped source–article–entity links; platform interaction graphs.	Temporal GNNs/ Hawkes processes; adversarial/ paraphrase tests.	Platform overfitting; brittleness to missing links.
Deployment, monitoring & governance	Operate a real-time, auditable system with humans in the loop.	Streaming ingestion, analyst feedback, drift signals, explanation logs.	Triage dashboard; drift detection/ recalibration; SHAP/attention views.	Over/under-blocking, outlet fairness, audit trails and model cards.

Source: authors' elaboration.

Conclusions

This study has presented a robust framework for the identification of deceptive financial narratives, addressing a fundamental challenge to the stability and efficiency of the modern digital economy. By shifting the analytical lens from a purely technical classification problem to a mechanism for safeguarding market integrity, this research demonstrates that Natural Language Processing (NLP) is a critical tool for reducing the information asymmetry that historically leads to suboptimal capital allocation and systemic wealth destruction.

The empirical evidence established in this study confirms that classical ensemble methods, particularly Random Forest when integrated with TF-IDF vectorization, provide the necessary precision and interpretability required for institutional deployment. From a regulatory perspective, these findings are significant as they offer a scalable methodology for operationalizing the requirements of the European Digital Services Act (DSA). By providing a transparent and auditable means of filtering «noise» from the financial information stream, such models help preserve the foundational pillars of the price discovery process, ensuring that asset valuations remain grounded in fundamental economic data rather than fabricated sentiment.

Furthermore, the research highlights that, as the velocity of digital information increases, the «cost of verification» becomes a primary barrier to market efficiency. The proposed NLP pipeline effectively lowers this barrier, offering a public-good utility for both retail and institutional investors. By mitigating the impact of market-manipulating narratives, this framework fosters a more resilient financial ecosystem capable of withstanding the rapid shifts of the digital age.

Looking ahead, the integration of these models into real-time regulatory oversight remains a priority for the European Union's Single Digital Market. Future research should expand upon these findings by exploring the cross-linguistic capabilities of these models within fragmented markets and investigating the recursive relationship between automated news verification and algorithmic trading volatility. Ultimately, the adoption of interpretable, high-accuracy misinformation detection is not merely a technical advancement but a necessary evolution in the governance of contemporary financial markets.

References

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, Article 8885861. https://doi.org/10.1155/2020/8885861

Alghamdi, J., Luo, S., & Lin, Y. (2024). A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications, 83*, 51009–51067. https://doi.org/10.1007/s11042-023-17470-8

Alshuwaier, F. A., & Alsulaiman, F. A. (2025). Fake news detection using machine learning and deep learning algorithms: A comprehensive review and future perspectives. *Computers*, *14*(9), Article 394. https://doi.org/10.3390/computers14090394

Cheng, L.-C., Lu, W.-T. & Yeo, B. (2023). Predicting abnormal trading behavior from internet rumor propagation: A machine learning approach. *Financial Innovation, 9*, Article 3. https://doi.org/10.1186/s40854-022-00423-9

Du, K., Xing, F., Mao, R., & Cambria, E. (2024). Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, *56*(9), 1–42. https://doi.org/10.1145/3649451

D'Ulizia, A., Caschera, M. C., Ferri, F., & Grifoni, P. (2021). Fake news detection: A survey of evaluation datasets. *PeerJ Computer Science 7*, Article e518. https://doi.org/10.7717/peerj-cs.518

European Parliament & Council of the European Union. (2014). Regulation (EU) No. 596/2014 of the European Parliament and of the Council of 16 April 2014 on market abuse (market abuse regulation) and repealing Directive 2003/6/EC of the European Parliament and of the Council and Commission Directives 2003/124/EC, 2003/125/EC and 2004/72/EC and Commission Decision 2004/693/EC. *Official Journal of the European Union, L 173,* 1–142. https://eur-lex.europa.eu/eli/reg/2014/596/oj

European Parliament & Council of the European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union, L 277*, 1–102. https://eur-lex.europa.eu/eli/reg/2022/2065/oj

Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. *AI Open*, *3*, 133–155. https://doi.org/10.1016/j.aiopen.2022.09.001

kaggle.com. (n.d.). *Fake and real news dataset* [Data set]. Retrieved July 17, 2025, from https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research, 50*(2), 495–540. https://doi.org/10.1111/j.1475-679X.2012.00450.x

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10 - Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Lyzun, M., Desyatnyuk, O., Savelyev, Y., Kuryliak, V., Sachenko, S., Lishchynskyy, I. (2023). Architectonics of the European Currency Integration: Cluster and Gravity Modeling. *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Dortmund, Germany, 2023, pp. 661–664. https://doi.org/10.1109/IDAACS58523.2023.10348652

Lyzun, M., Lishchynskyy, I., Savelyev, Y., Kuryliak, V. and Kurylyak, Y. (2019). Modeling Evaluation of Dollarization Economic Efficiency. *International Conference on Advanced Computer Information Technologies (ACIT)*. Ceske Budejovice, Czech Republic: 366–370.

Majumdar, A., & Bose, I. (2018). Detection of financial rumors using big data analytics: The case of the Bombay Stock Exchange. *Journal of Organizational Computing and Electronic Commerce*, 28(2), 79–97. https://doi.org/10.1080/10919392.2018.1444337

Mishra, S., Shukla, P., & Agarwal, R. (2022). Analyzing machine learning enabled fake news detection techniques for diversified datasets. *Wireless Communications and Mobile Computing*, Article 1575365. https://doi.org/10.1155/2022/1575365

Nghiem, H., Muric, G., Morstatter, F., & Ferrara, E. (2021, November). Detecting cryptocurrency pump-and-dump frauds using market and social signals. *Expert Systems with Applications, 182,* Article 115284. https://doi.org/10.1016/j.eswa.2021.115284

Rangapur, A., Wang, H., Jian, L., & Shu, K. (2025, April 28-May 2). Fin-Fact: A benchmark dataset for multimodal financial fact-checking and explanation generation. In *WWW Companion'25: Companion Proceedings of the ACM Web Conference 2025* (pp. 785–788). ACM. https://doi.org/10.1145/3701716.3715292

Selyukh, A. (2013, April 24). *Hackers send fake market-moving AP tweet on White House explosions*. Reuters. https://www.reuters.com/article/technology/hackers-send-fake-market-moving-ap-tweet-on-white-house-explosions-idUSBRE93M12Y/

Theodorakopoulos, L., Theodoropoulou, A., Tsimakis, A., & Halkiopoulos, C. (2025). Big Data-driven distributed machine learning for scalable credit card fraud detection using PySpark, XGBoost, and CatBoost. *Electronics*, *14*(9), Article 1754. https://doi.org/10.3390/electronics14091754

Received: November 7, 2025.
Reviewed: March 10, 2026.
Accepted: March 20, 2026.