Machine learning in bank merger prediction:

A text-based approach

by

Apostolos G. Katsafados[1], George N. Leledakis[1*], Emmanouil G. Pyrgiotakis[2], Ion Androutsopoulos[3], Manos Fergadiotis[3]

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

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

[3] Department of Informatics, School of Information Sciences and Technology, Athens University of Economics and Business, Greece

Abstract

This paper investigates the role of textual information in a U.S. bank merger prediction task. Our intuition behind this approach is that text could reduce bank opacity and allow us to understand better the strategic options of banking firms. We retrieve textual information from bank annual reports using a sample of 9,207 U.S. bank-year observations during the period 1994-2016. To predict bidders and targets, we use textual information along with financial variables as inputs to several machine learning models. We find that when we jointly use textual information and financial variables as inputs, the performance of our models is substantially improved compared to models using a single type of input. Furthermore, we find that the performance improvement due to the inclusion of text is more noticeable in predicting future bidders, a task which is less explored in the relevant literature. Therefore, our findings highlight the importance of textual information in a bank merger prediction task.

JEL classification: C63, G14, G21, G34, G40
Keywords: *Finance; Bank merger prediction; Textual analysis; Natural language processing; Machine learning*

This version: June, 2023

[*] Corresponding author: Department of Accounting and Finance, School of Business, Athens University of Economics and Business, 76 Patission Str., 104 34, Athens, Greece; Tel.: +30 210 8203459. E-mail addresses: katsafados@aueb.gr (A. Katsafados), gleledak@aueb.gr (G. Leledakis), e.pyrgiotakis@essex.ac.uk (E. Pyrgiotakis), ion@aueb.gr (I. Androutsopoulos), fergadiotis@aueb.gr (M. Fergadiotis). We would like to thank Ilias Chalkidis, Nikolaos Gkoumas, Prodromos Malakasiotis, Thanos Verousis, and the participants at the Annual Event of Finance Research Letters 2022 CEMLA Conference: New Advances in International Finance for their valuable comments and suggestions. Apostolos Katsafados acknowledges financial support co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY). George Leledakis greatly acknowledges financial support received from the Research Center of the Athens University of Economics and Business (EP-2256-01). All remaining errors and omissions are our own.

1. Introduction

Over the last decades, the U.S. banking industry has experienced a severe wave of consolidation through mergers and acquisitions (M&A). Aligned with this trend, the academic literature has given increased attention to the topic of bank M&As. The vast majority of the literature focuses on investigating the shareholder wealth effects around the announcement of bank mergers (Houston et al., 2001; DeLong and DeYoung, 2007; Filson and Olfati, 2014; Leledakis and Pyrgiotakis, 2022), while other studies analyze the merger-related performance changes (Cornett and Tehranian, 1992; Cornett et al., 2006), or the efficiency effects (Rhoades 1993; 1998).

Another strand of the literature attempts to identify the characteristics of merging U.S. banks, especially from the perspective of the target (Prasad and Melnyk, 1991; Wheelock and Wilson, 2000). These studies report that smaller, less profitable, and poorly-managed banks are more attractive acquisition targets. In this respect, Katsafados et al. (2021) find that banks with more positive (negative) tone in their annual reports have a higher probability of becoming bidders (targets). However, the latter study focuses on the determinants of merger likelihood, and not in the prediction of future merger participants. Up to date, therefore, there is a gap in the literature regarding the development of classification models in a U.S. bank merger prediction task.

In the non-financial sector, there is a plethora of studies that utilize classification models to predict M&As (Palepu, 1986; Slowinski et al., 1997; Espahbodi and Espahbodi, 2003; Edmans et al., 2012; Routledge et al., 2017; Delis et al., 2023). One possible explanation on why there is no substantial empirical work on this issue for U.S. banks could be that the banking industry is inherently more opaque than other industries (Flannery et al., 2004; Blau et al., 2017). Opacity means that banking assets are hard-to-value due to their financial nature which distinguishes banks from non-bank firms (Morgan, 2002). In other words, banks hold very few physically-fixed assets compared to other types of firms. Instead, banks primarily hold loans, which are privately negotiated transactions with their borrowers. The opaqueness of these types of assets limits the ability of investors to properly evaluate the financial condition of a bank (Huizinga and Laeven, 2012; Jones et al., 2013). Researchers in merger prediction for non-financial firms use accounting measures to evaluate the financial condition of the firm. Potential bidders are perceived to be in sound financial position, whereas potential targets may face financial constraints (Espahbodi and Espahbodi, 2003). Taken altogether, it is likely that bank opacity could be one possible reason for the lack of empirical work on bank merger prediction.

Bank opacity is inversely related to disclosure of information, as the level of bank opacity decreases with the quality of disclosure (Flannery et al., 2013; Jiang et al., 2016; Zheng, 2020). Banks disclose information to the public mainly through their financial statements and annual reports. On the one hand, financial statements may not effectively reduce opacity, as banks manage their statements to smooth their earnings and circumvent the capital requirements (Ahmed et al., 1999; Beatty et al., 2002; Bushman and Williams, 2012; Gandhi et al., 2019). On the other hand, bank annual reports contain one other important source of information besides balance sheet data: textual information.

There is a growing literature on how textual information can reduce firms’ valuation uncertainty and the asymmetry of information on the initial public offerings (IPOs) in the U.S. (Hanley and Hoberg, 2010; Loughran and McDonald, 2013; Jegadeesh and Wu, 2013). Collectively, these studies find that the textual information of the IPO prospectuses can mitigate the uncertain valuation of IPO firms, a fact that leads to a more accurate pricing of the newly issued shares. In a similar manner, Gandhi et al. (2019) use the sentiment of banks' annual reports to gauge financial distress. The authors argue that text is more informative than simple accounting measures, as the latter source of information could be influenced by bank managers. This happens because over-optimism in annual reports by managers increases litigation risk (Rogers et al., 2011; Loughran and McDonald, 2013). Building on these arguments, it is reasonable to assume that the use of textual information could improve our ability to evaluate the financial condition of banks by reducing bank opacity. Hence, if this assumption is valid, textual information may also enable us to more accurately identify bidders and targets in the U.S. banking industry.

Apart from reducing bank opacity, textual information could have an additional benefit in a merger prediction task. In most cases, the choice to engage in a merger is a strategic decision for the bank, especially on the part of the bidder (Ramaswamy, 1997). Potential bidders have different characteristics from potential targets, as they are usually larger and more profitable (Becher, 2009). However, the fact that a bank is financially healthier (according to its financial statements), does not necessarily imply that its strategy is to engage in M&As. For this reason, annual reports may be more insightful regarding the bank's strategic options, as managers disclose information regarding the future prospects of their bank in these reports. In fact, managerial motives consist one of the main explanations behind M&As (Gregoriou and Renneboog, 2007).

Therefore, the primary aim of this paper is to investigate whether the use of textual information from bank annual reports is meaningful in a merger prediction task. More precisely, we develop classification models to identify bidders and targets in the U.S. banking industry, and we use both textual information and financial variables as inputs in these models. In our classification task, we choose to employ several machine learning algorithms instead of traditional econometric techniques. We do so, because machine learning algorithms make fewer assumptions about the data and can often produce more accurate estimates (Mai et al., 2019).

To address our research question, we collect annual reports from banks that filed the reports over the period 1994-2016. By doing so, we obtain a large sample of 9,207 U.S. bank-year observations, which includes bidders, non-bidders, targets, and non-targets. For each year and for each bank, we retrieve the annual reports from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website. For the purpose of our analysis, we extract textual information by creating textual features from these reports using the bag of words approach. In fact, we use the following textual features as inputs: term frequency (TF) features, and term frequency-inverse document frequency (TF-IDF) features corresponding to words (or combination of words and bigrams). Finally, we use these textual features along with financial variables in the classification machine learning models. More precisely, we use the following models: (1) logistic regression, (2) linear support vector machine, (3) support vector machine with radial basis function kernel, (4) random forest, and (5) multilayer perceptron.

A key innovation of our study is that apart from the aforementioned textual features, we create textual features based on word embeddings. In addition to the frequently-used generic word embeddings, we also create our own finance word embeddings. Textual features based on word embeddings are used as inputs to the multilayer perceptron model.

As a first step in our empirical analysis, we run our models using only financial variables as inputs, and we consider these models our benchmark. Then, we re-run our models using only textual features as inputs. Finally, we augment our benchmark models by combining both financial and textual data. Our results indicate that in both prediction tasks, textual features are meaningful inputs in our models. More precisely, we achieve the highest prediction scores when we augment our benchmark models with textual features. It is noteworthy that textual features have higher incremental value in the bidder prediction task. This is expected to some extent, as the language used by managers in the annual reports may be more informative regarding the bidding banks' strategic decisions, compared to financial ratios. By contrast, target banks' prediction relies more on weak fundamentals (Becher, 2009). These findings are also consistent with Cornett et al. (2011), which suggest bidders can be predicted more accurately than targets.

We conduct a series of robustness tests. First, we employ the bootstrap resampling method of Berg-Kirkpatrick et al. (2012) to validate the performance of our models. More precisely, in both tasks, we compare the best-performing models that use both financial and textual data as inputs with the benchmark models. We find that the former models statistically outperform the latter in both tasks, a fact that provides further support to our conjecture regarding the importance of textual information in a merger prediction task. Second, we compute the importance score for each variable in our best-performing models using the Gini impurity technique of Kurt et al. (2008). The results of this analysis indicate that in the bidder prediction task, textual features are more important inputs than traditional financial variables. In the target prediction task, the scores of financial variables are slightly higher than the ones of textual information, a finding which is also consistent with our main results, as we have documented that weak fundamentals are the main driver behind target prediction. Third, we use the Local Interpretable Model-agnostic Explanations (LIME) method as in Ribeiro et al. (2016), to find out which specific textual features increase the predictive ability of our best-performing models. The results for both tasks were in line with our intuition. Fourth, we use word embeddings (both the generic and the finance-specific ones) as inputs in Bi-directional Long Short-Term Memory Recurrent Neural Network (BLSTM-RNN) models. We do so, as BLSTM-RNNs are very effective for modelling sequential data like documents (word sequences). The results indicate that our finance-specific word embeddings are the most informative textual features when combined with sophisticated deep learning models (BLSTM-RNN in our case).

Our findings could benefit all key parties of a bank merger transaction. From the regulators' perspective, identifying future acquirers may be more beneficial than identifying future targets. When acquirers grow large through M&As, they can become too-big-to-fail (TBTF) and enjoy oligopolistic market power (O'hara and Shaw, 1990; Demirgüç-Kunt and Huizinga, 2013). Furthermore, TBTF banks usually enjoy special treatment from policy-makers, as they are more likely to receive government bailouts in case of insolvency, compared to their smaller and less systemically important peers (Bernanke, 2010). This special treatment creates moral hazard issues, because TBTF banks are more inclined to engage in excessive risk-taking, which poses a threat to the stability of the banking industry (Brewer and Jagtiani, 2013). Therefore, the development of an accurate classification model could enable regulatory authorities to *a priori* evaluate any merger-related anticompetitive effects and ensure the stability of the banking industry. These regulatory actions are of major importance for social welfare, since changes in the structure of the banking industry may have a detrimental effect on competition (Koskela and Stenbacka, 2000). From the investors' perspective, identifying future bank merger participants could be a profitable strategy, because bidders (targets) typically realize negative (positive) abnormal returns around the announcement of the merger (DeLong, 2001). In this case, an investor can apply a merger arbitrage strategy by purchasing shares of the target firm and short-selling shares of the acquiring firm (Buehlmaier and Zechner, 2021). Finally, the development of an accurate classification model could also be of use to bank managers. Managers of banks who want to expand via M&As can use such a tool to identify potential targets. At the same time, financially constrained banks that have to be acquired may use such classification models to identify and attract potential bidders (Pasiouras et al., 2010).

We contribute to the literature in four main aspects. First, instead of focusing merely on predicting future acquisition targets, we also attempt to predict future acquiring banks, as this task is more important to regulators and depositors. Notably, we provide evidence that textual information can significantly improve the prediction of bank bidders, a fact that might provide fertile ground for future research on this topic. Second, instead of using econometric techniques to perform our task, we utilize several machine learning models, which have several advantages over traditional econometric methodologies (Mai et al., 2019). This argument is supported by our findings, since more sophisticated machine learning algorithms, such as the random forest or the BLTSM-RNNs, yield more accurate estimates compared to the traditional logistic regressions. Hence, any improvement in prediction accuracy is important, considering that there is still no consensus regarding the development of an accurate method to predict M&A activity (Very et al., 2012). Third, we create our own finance word embeddings, which appear to be the most meaningful textual inputs in the bidder prediction task. Finally, we provide evidence that textual information can effectively complement traditional financial variables in bank merger prediction. Our study contributes to the textual analysis literature by adding fresh insights into how textual features can significantly improve the predictive performance of classification models (Stevenson et al., 2021; Kriebel and Stitz, 2022; Nguyen and Huynh, 2022). Our interpretation of this result is that textual information reduces bank opacity, since the language used by managers in the annual reports provides a clearer picture of the financial condition of the bank and its future strategic options.

The rest of the paper is organized as follows. Section 2 describes our sample collection and our textual analysis procedure. Section 3 discusses our classification models, and Section 4 reports our empirical findings. Finally, Section 5 concludes the paper.

2. Data and textual analysis

2.1. Sample selection

To construct our dataset, we follow a three-step approach. The first step is to collect bank annual reports (10-Ks, 10-K405s, 10-KSBs, and 10-KSB40s) from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website. To do so, we use a web-crawling algorithm, which gathers the reports and excludes all amended documents. In our primary sample, we require that banks' filing dates are between 1994 and 2016. Furthermore, we exclude 97 observations from our sample because the filing contained fewer than 2,000 words (Loughran and McDonald, 2011). Further, we also exclude 2 observations from our sample, due to the fact that 2 banks had more than one filing in the same fiscal year (we include only the first filing). By applying these criteria, our initial sample consists of 18,031 bank-year observations.

The second step is to gather bank-specific characteristics from the Federal Reserve Bank of Chicago (FRBC), as in Katsafados et al. (2021). More precisely, we acquire financial information of bank holding companies (BHCs) from the FR Y-9C reports and of commercial banks and savings institutions from Call Reports. Then, we collect banks RSSD

IDs using the Federal Reserve Bank of New York's CRSP-FRB link. Next, we use the bank names and locations (state and/or city) to merge our initial sample from EDGAR with FRBC data. By doing so, we are able to link the banks' RSSD IDs with their corresponding Central Index Keys (CIK). To ensure the maximum number of observations, we manually match banks' RSSD IDs with their CIKs using the National Information Centre (NIC) database. This matching process leaves us with a final sample of 9,207 bank-year observations consisting of 1,160 unique banks.

As a third step, we obtain our bank merger sample from the Thomson ONE database. We focus on deals announced between February, 1994 and December, 2017.[1] To filter our sample, we use the following criteria similar to Leledakis et al. (2021):

1. Both bidders and targets are commercial banks with a three-digit primary SIC code of 602, savings institutions with a three-digit primary SIC code of 603, or bank holding companies with a four-digit primary SIC code of 6712.
2. The bidder is publicly-traded. The target can be a public firm, a private firm, or an unlisted subsidiary of a publicly-traded firm.
3. All public firms are listed on NYSE, AMEX, or NASDAQ.
4. The bidder acquired an interest of more than 50% of the target firm after the merger. Before the merger, its interest was below 50%.
5. Failed bank acquisitions are excluded.

The above selection process results in a sample of 966 bank M&As. As described in the filter criteria, all bidders are publicly-traded. However, in the case of targets, 389 are publicly-traded, and the remaining ones are either private-owned banks or subsidiaries of listed banks. Since the sample also includes unlisted targets, our sample selection process ensures that our subsample of bidding banks includes all listed banks that had acquired another bank during our examination period. Hence, from the final sample of 9,207 bank-year observations, 8,241 refer to non-bidder banks and 8,818 to non-target banks. Table 1 reports the number of bidders (non-bidders) and targets (non-targets) on an annual basis over our examination period.

2.2. Financial variables

We choose to use a set of financial variables as inputs in our predictive models that satisfy the following two criteria: (i) they are likely to influence acquisition decisions (Wheelock and Wilson, 2000; Pasiouras et al., 2010), and (ii) these variables are limited in number to avoid overfitting of our models (Palepu, 1986). In what follows, we briefly describe the twelve financial variables used in this study. Eight of these variables are variables from financial statements and the remaining four are market variables.[2]

The first two financial variables relate to the inefficient management hypothesis. According to this hypothesis, the motive behind M&As is to replace the inefficient management of the target firm (Manne, 1965). Hence, following Pasiouras et al. (2010), we employ two bank efficiency measures: the cost-to-income ratio (*Cost efficiency*), and the return on total assets (*ROA*). Capital strength is also an important determinant of bank acquisition behavior, as weaker-capitalized banks are more likely to be acquired (Hannan and Rhoades, 1987; Pasiouras et al., 2007). For this reason, we use the ratio of common equity to total assets (*Capital strength*). Moreover, we control for the impact of loan

[1] To be included in our merger sample, a bank should be a bidder or a target in a twelve-month period after the filing date (Routledge et al., 2017). The earliest filing date of our sample is in the end of January, 1994 and the latest is in the end of December, 2016.

[2] All variables from financial statements are measured at the most recent fiscal year end prior to the filing date.

activity on bank acquisition likelihood using the ratio of loans to total assets (*Loans*), as in Pasiouras et al. (2010).

Market power is a commonly-stated motive behind bank M&As (Hankir et al., 2011). Hence, we also use in our models the ratio of each bank's deposits to the total deposits of the U.S. banking industry in a given year (*Market power*). Further, acquired banks tend to have higher amounts of loan loss reserves relative to non-acquired banks (Wheelock and Wilson, 2000; Pasiouras et al., 2010). In line with these results, we include the ratio of loan loss provisions to total loans (*Asset quality*). Further, we proxy for the banks' dependence on off-balance sheet activities using the ratio of non-interest income to total income (*Non-interest income*), as in Ellul and Yerramilli (2013). Finally, in the spirit of Cornett et al. (2006), we use the ratio of deposits to total assets (*Deposits*) as a measure of liquidity. Table A1 in the Appendices provides a detailed list of these eight variables, along with the corresponding codes from the FR Y-9C reports for bank holding companies and the Call reports for commercial banks and savings institutions.

Table 1
Yearly distribution of observations

Filing year	Bidders	Non-bidders	Targets	Non-targets
1994	39	81	5	115
1995	39	128	11	156
1996	55	308	14	349
1997	84	410	33	461
1998	81	423	30	474
1999	49	464	24	489
2000	43	466	29	480
2001	35	451	21	465
2002	35	446	17	464
2003	66	430	32	464
2004	48	464	16	496
2005	57	435	18	474
2006	57	414	25	446
2007	37	366	13	390
2008	15	368	8	375
2009	4	369	3	370
2010	10	350	11	349
2011	11	330	7	334
2012	29	321	10	340
2013	40	335	16	359
2014	48	316	16	348
2015	52	300	17	335
2016	32	266	13	285
Total	966	8,241	389	8,818

This table summarizes the yearly distribution of our sample based on the 10-K filing year. *Bidders* is the number of banks that participate in a merger with the role of the bidder within a twelve-month period after the 10-K filing date, while *Non-bidders* is the number of banks that do not participate in a merger with the role of the bidder. Similarly, *Targets* is the number of banks that participate in a merger with the role of the target within a twelve-month period after the 10-K filing date, and *Non-targets* is the number of banks that do not participate in a merger with the role of the target.

Furthermore, we include the following market variables that are frequently-used in the merger literature: (1) *MTB* the ratio of the market value of equity to book value of equity (Palepu, 1986; Ambrose and Megginson, 1992; Espahbodi and Espahbodi, 2003; Cremers et al., 2009; Cocco and Volpin, 2013), (2) *LnMV* the natural logarithm of the market value of equity (Cremers et al., 2009) and it is a proxy for bank size since smaller banks are more likely to become acquisition targets (Wheelock and Wilson, 2000), (3) *ERet* the excess returns over the cumulative value-weighted return of a portfolio with NYSE/AMEX or NASDAQ stocks within a 12-month period before the 10-K filing date similar to Espahbodi and Espahbodi (2003), and (4) *DVOL* the daily return volatility over the 12-month period prior to the 10-K filing date (Cocco and Volpin, 2013). The forward-looking nature of market variables might contain important information for our merger classification task, as they capture the investors' perception regarding the future prospects of the banks. Table 2 reports the summary statistics of all financial variables. Particularly, we split the sample into the following four categories: bidders (Panel A), targets (Panel B), non-bidders (Panel C), and non-targets (Panel D).

2.3. Textual analysis and parsing methodology

2.3.1. Textual sources

All bank annual reports are encoded in the hypertext markup language (HTML). Hence, as in most studies using textual analysis in finance, we follow the parsing process of Loughran and McDonald (2011). Through this process, we remove HTML formatting and any other non-textual information, such as embedded images or spreadsheets that might be present in the text (Bodnaruk et al., 2015). Moreover, we exclude all identified HTML tables, if their numeric character content is higher than 10%, as effectively documented by Loughran and McDonald (2014).

2.3.2. Pre-processing and bag of words

After the parsing procedure, we have to transform the textual information into numerical features before we insert them as inputs to our models. To do so, we follow the pre-processing procedure, which consists of several steps (Jegadeesh and Wu, 2013; Loughran and McDonald, 2014; Nassirtoussi et al., 2014).

First, we eliminate single letter words, abbreviations, numbers, punctuation marks, and stop words (Gandhi et al., 2019; Anastasiou and Katsafados, 2023). Second, we impose a minimum occurrence threshold in order to remove words with low frequency (Chen et al., 2020). Bernabé-Moreno et al. (2020) highlight that such a filtering process is important as it helps to avoid sparsity and noise. Following Mai et al. (2019), we consider the 20,000 most frequent words of the bank annual reports of the remaining text. Third, we use the bag of words (BOW) approach to transform our unstructured textual information into inputs with explicit numerical structure. More precisely, we use the Natural Language Toolkit (NLTK) to tokenize text into individual words. As a matter of fact, this approach treats each unique word as a separate textual feature, and constructs a document-term matrix, where each row and column represent a document and a word, respectively (Loughran and McDonald, 2011).

In the textual analysis literature, raw counts of textual features are not considered the best measure of a text's information content. Therefore, we represent each textual feature using the two most widely-used term weighting schemes: (1) the term frequency (TF) normalized by document length, and (2) the term frequency-inverse document frequency (TF-IDF). *TF* is calculated as the proportion of each textual feature in each document, so it assigns an equal weight for each feature. *TF-IDF* adjusts the TF scores by putting a lower weight on features that appear more frequently in our sample of bank annual reports (Jegadeesh and Wu, 2013; Loughran and McDonald, 2016; Katsafados et al., 2021). Prior studies suggest that *TF-IDF* is a more effective weighting scheme compared to *TF*, as it assigns lighter weights to common words, which have a less meaningful impact on textual analysis tasks (Balakrishnan et al., 2010; Brown and Tucker, 2011; Loughran and McDonald, 2011; Loughran and McDonald, 2016; Mai et al, 2019). We calculate the *TF-IDF* weight of word *i* in the j^{th} document as reported in the equation below:

Table 2
Summary statistics

Variables	**N**	**Mean**	**Median**	**Std. Dev.**
Panel A: Bidders				
Cost efficiency %	966	63.15	63.09	10.37
ROA %	966	1.07	1.08	0.48
Capital strength %	966	9.58	9.24	2.37
Loans %	966	65.97	66.93	10.17
Market power %	966	0.35	0.06	1.08
Asset quality %	966	0.34	0.28	0.47
Non-interest income %	966	24.90	23.58	11.99
Deposits %	966	76.22	77.87	8.66
MTB	966	1.92	1.78	0.86
LnMV	966	13.43	13.29	1.66
ERet %	966	7.22	2.51	24.00
DVOL %	966	1.69	1.59	0.89
Panel B: Targets				
Cost efficiency %	389	67.88	66.42	15.66
ROA %	389	0.82	0.91	0.79
Capital strength %	389	9.28	8.69	2.94
Loans %	389	66.56	67.17	10.23
Market power %	389	0.17	0.02	0.65
Asset quality %	389	0.43	0.25	0.86
Non-interest income %	389	20.74	18.50	11.82
Deposits %	389	77.31	78.61	9.64
MTB	389	1.61	1.52	0.72
LnMV	389	12.21	11.95	1.64
ERet %	389	4.42	1.29	30.51
DVOL %	389	2.02	1.77	1.01
Panel C: Non-bidders				
Cost efficiency %	8,241	68.15	65.35	21.38
ROA %	8,241	0.74	0.91	1.07
Capital strength %	8,241	9.34	8.99	2.88
Loans %	8,241	66.27	67.36	12.09
Market power %	8,241	0.21	0.02	1.20
Asset quality %	8,241	0.57	0.29	0.99
Non-interest income %	8,241	22.15	20.11	13.52
Deposits %	8,241	77.12	79.19	10.27
MTB	8,241	1.56	1.43	0.95
LnMV	8,241	12.25	11.92	1.71
ERet %	8,241	4.93	1.36	32.79
DVOL %	8,241	2.33	1.95	1.52
Panel D: Non-targets				
Cost efficiency %	8,818	67.62	64.97	20.76
ROA %	8,818	0.77	0.94	1.04
Capital strength %	8,818	9.37	9.04	2.83
Loans %	8,818	66.23	67.31	11.97
Market power %	8,818	0.23	0.02	1.21
Asset quality %	8,818	0.55	0.29	0.96
Non-interest income %	8,818	22.52	20.49	13.46
Deposits %	8,818	77.01	79.05	10.14
MTB	8,818	1.60	1.46	0.97
LnMV	8,818	12.40	12.09	1.75
ERet %	8,818	5.23	1.50	31.99

DVOL %	8,818	2.27	1.92	1.49

$$TF\text{-}IDF(t_{ij}) = TF(t_{ij}) \times \left[-\log\left(\frac{n_i}{N}\right)\right] \quad (1)$$

where *TF(t*$_{ij}$*)* is the number of times a term *i* appears in a document *j*, divided by the total word count of the same document for normalization purposes, *N* represents the number of documents in our entire dataset, and n_i is the total number of documents including at least one occurrence of the i^{th} word.

At this point, it is worth mentioning that one limitation of the BOW is that it does not control for the presence of polysemous words (words with multiple meanings) in the text. To control for this issue, we also use bigrams in our textual analysis. Bigrams are essentially word pairs, obtained using the word n-gram features (n equal to 2). The use of bigrams may improve the ability of our models to disambiguate the meaning of a polysemous word. Note that the BOW approach is also based on word n-gram features, when n equals 1 (unigrams).

2.3.3. Word embeddings

The aforementioned BOW approach has a prevalent role in studies that employ textual analysis in finance. As mentioned before, a main drawback of this approach is that it does not account for polysemous words, an issue which can be partially resolved with the use of bigrams. However, another drawback of the BOW approach is that it is not able to capture well the morpho-syntactic and semantic properties of the words of the text (Manning and Schutze, 1999; Loughran and McDonald, 2016). This happens because conventional BOW models rely on the frequency of words under the assumption that each word occurs independently of all others. In this regard, it is likely that models that use conventional BOW representations as textual inputs are not fully capable of understanding the underlying semantics of the text (Loughran and McDonald, 2016). To alleviate this concern, we also employ word embedding features to represent textual information.

The word embedding approach is a relatively new representation of textual data in natural language processing (NLP). The fundamental concept behind this model is that words with similar properties co-occur with similar neighbors (Mai et al., 2019). In other words, a word embedding is a type of word representation which allows words with similar properties to have a similar representation. More precisely, this model represents each word as a vector in a low-dimensional space (Goldberg, 2017). The word embedding vector includes real values, which reflect the morpho-syntactic and semantic properties of the word.

Mikolov et al. (2013) develop the word2vec technique, where word embeddings can be produced either through the continuous bag of words (CBOW) model, or the skip-gram model. Both models use shallow neural networks to learn word representations for each unique word. The CBOW model combines the embeddings of surrounding words to predict the word in the middle of a window of text, whereas the skip-gram model tries to predict the context words in a window of text for a given word in the middle of the window.

Pennington et al. (2014) introduce an alternative method for producing word embeddings, known as global vectors for word representation (GloVe). GloVe embeddings typically lead to similar performance in NLP tasks as word2vec embeddings, but GloVe embeddings are more readily available in different dimensionalities, and pre-trained on diverse corpora. Therefore, in our paper, we employ the available 200-dimension generic embeddings created by Pennington et al. (2014). These embeddings are obtained from 6 billion tokens from Wikipedia 2014 and Gigaword 5, and have a vocabulary size of 400K words.[3]

In our empirical setting, one possible concern with the generic word embeddings is that they are not trained on (obtained from) a finance-specific corpus. To account for this issue, we also employ domain-specific (DS) word embeddings. DS word embeddings are trained on data from a specific domain of interest. For this reason, they may be able to represent better the semantics of the text compared to generic word embeddings. In particular, we use word2vec to create our 200-dimension finance word embeddings (FWE) induced from textual disclosure in the finance domain.[4] In particular, our finance word embeddings are derived from 4.9 billion tokens of EDGAR financial disclosures from 1994 to 2016 (including all 10-K, 10-Q, and S-1 filings), and have a vocabulary size of 2.3M words.

In more detail, we employ the skip-gram model to produce our finance word embeddings. As noted earlier, the skip-gram model learns word vector representations aiming to predict the context (surrounding words in a window) from the central word of each (sliding) text window (Mikolov et al., 2013). In this regard, if we have a corpus of *T* words w_1, w_2,.., w_T, skip-gram aims to maximize the following log-likelihood objective:

$$\sum_{t=1+m}^{T-m} \sum_{-m \le i \le +m, i \ne 0} \log P(w_{t+i} \mid w_t) \qquad (2)$$

where w_t is the central word of the (sliding) window at location *t* in the corpus, w_{t+i} is the context word at location $t+i$, and *m* defines the window size $(2 \times m - 1)$ of the window around w_t. Our FWE are created with window size equal to 5.

Each word has two embeddings (vectors of real numbers), an input (w^{in}) and an output (w^{out}) one, which are randomly initialized, and learned by minimizing the objective. For every token w_t at position *t* of the corpus and every position $t+i$ (i≠0) within a window [t-m, t+m] around position t, we aim to be capable of predicting which vocabulary word occurs at position $t+i$ by multiplying (dot product) w_t^{in} and w_{t+i}^{out}. The basic form of skip-gram employs the softmax function to calculate the likelihood of a surrounding word w_{t+i} given a center word w_t:

$$P(w_{t+i} \mid w_t) = \text{softmax}\left(\mathrm{w}_{t+i}^{out} \times w_t^{in}\right) = \frac{\exp\left(\mathrm{w}_{t+i}^{out} \times w_t^{in}\right)}{\sum_{w \in V} \exp\left(\mathrm{w}^{out} \times w_t^{in}\right)} \qquad (3)$$

where V is the vocabulary. We learn the w_t^{in} and w_{t+i}^{out} by maximizing the probability we assign to the word w_{t+i} that actually occurs at each position t+i of each window. In fact, we obtain the word embeddings as follows:

$$\left\langle E^{in},\ E^{out} \right\rangle = \underset{\left\langle E^{in},\ E^{out} \right\rangle}{\arg\max} \sum_{t=1+m}^{T-m} \sum_{-m \le i \le +m, i \ne 0} \log P(w_{t+i} \mid w_t) \qquad (4)$$

where E^{in} and E^{out} are matrices that include in their columns all the in (w_t^{in}) and out (w_{t+i}^{out}) vectors of all words in the

[3] These word embeddings have been proved to be efficient to many tasks. Also, they are publicly available https://nlp.stanford.edu/projects/glove/

[4] To do so, we use the free available Python library of gensim (https://radimrehurek.com/gensim/).

vocabulary. We maximize the objective by stochastic gradient ascent. However, in practice the softmax of $P(w_{t+i}|w_t)$ is computationally expensive, because of the large size of the vocabulary V. We, therefore, use the negative sampling version of the skip-gram model. Instead of predicting the context word w_{t+i} from the central word w_t, we now aim to be able to identify the true context word w_{t+i}, when given the true context word w_{t+i} and a randomly sampled word r (multiple randomly sampled words are used in practice, instead of just one). In effect, instead of aiming to produce a probability distribution over the vocabulary V for position $t+i$, we now have a binary classification problem, where we need to classify w_{t+i} in the true (positive) class, and r to the false (negative) class. The objective now becomes:

$$\left\langle E^{in},\ E^{out} \right\rangle = \underset{\left\langle E^{in},\ E^{out} \right\rangle}{\arg\max} \sum_{t=1+m}^{T-m} \sum_{-m \le i \le +m, i \ne 0} \log \sigma\left(\mathrm{w}_{t+i}^{out} \times w_t^{in}\right) + \log\left[1 - \sigma\left(r^{out} \times w_t^{in}\right)\right] \quad (5)$$

where σ is the sigmoid (logistic) function, and $\sigma(w_{t+i}^{out} \times w_t^{in})$ is the probability estimate that word w is the true context word. After maximizing the objective, we keep the vectors of E^{in} as word embeddings, though the vectors of E^{out} can also be used alternatively.

For visualization purposes, we project our finance word embeddings (with 200 dimensions) into 2 dimensions using the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique. We plot the various words from our financial word embeddings in a 2-dimensional vector space (see Figure A1 in the Appendices). Words with similar properties are located in close proximity to each other in the word embedding space. For instance, words close to the upper right corner of the embedding space relate to merger events, such as acquired, acquire, target, purchase, merger, and acquisition. This finding is in line with our conjecture that annual reports contain information regarding the banks' strategic choices, and particularly their M&As strategies. Furthermore, at the bottom of the embedding space, there is a set of words that express negativity, such as crisis, distress, weak, recession, and turmoil among others. Considering the previous facts, we can infer that our FWE serve their purpose of being specialized in financial texts.

3. Methodology

In this section, we describe the three parts of our methodological approach. First, how we split our datasets into training set and out-of-sample (testing) set, and how we match our training sets to control for the imbalance issue. Second, we analyze the machine learning models we employ in our study. Third, we describe the measure we use to evaluate the performance of our models.

3.1. Splitting datasets and matching training sets

To address our research question, we specify two binary models that are capable of distinguishing between: (1) bidders and non-bidders and (2) targets and non-targets. To do so, we have to construct our two datasets in a proper way. The first dataset will include only bidders and non-bidders, and the second dataset will include only targets and non-targets. Then, we split our two datasets into training and out-of-sample datasets. Following Geng et al. (2015), Doumpos et al. (2017), and Routledge et al. (2017), we select 80% of each dataset as the training set, and the remaining

20% as the out-of-sample. The out-of-sample is selected from a future period, as the usefulness of a classification model is evaluated according to its ability to correctly predict observations that occur in the future (Espahbodi and Espahbodi, 2003). More precisely, for both prediction tasks, the out-of-sample covers the period from March, 2012 to December, 2016 in terms of 10-K filing date.

It is obvious from Table 1 that the number of bidders and targets is disproportionally smaller compared to non-bidders and non-targets. This suggests that our datasets are imbalanced. Imbalanced datasets are a common issue in finance classification tasks, a fact which might jeopardize the training of our models (Neophytou and Mar Molinero, 2004; Pasiouras et al., 2007, 2010). To address this issue, we adopt the undersampling approach of Veganzones and Severin (2018), and we balance our training samples by excluding observations from the majority category. We use the filing year of the banks' annual reports as the matching criterion. This matching criterion has two main benefits: (1) it helps us control for any time effects in our analysis, (2) it allows us to include all the other variables as inputs in our models (Hasbrouck, 1985).[5] However, we leave the out-of-sample datasets imbalanced, because the out-of-sample dataset should be representative of the whole sample. We do so, to test whether textual features can enhance the performance of our models in a real-time setting.

3.2. Machine learning models

To perform our merger classification task, we use our machine learning models.[6] The machine learning models we use are: (1) logistic regression (LOGIT), (2) linear support vector machine (SVM-linear), (3) support vector machine with radial basis function kernel (SVM-RBF), (4) random forest (RF), and (5) multilayer perceptron (MLP). These models use as textual inputs the features obtained by the BOW approach. We select these models for our merger prediction task, as they are widely-used in several finance classification tasks due to their prediction efficacy in such tasks (Mai et al., 2019; Stevenson et al., 2021; Yildirim et al., 2021). Therefore, we rely on this literature, and we choose the machine learnings that all these studies have in common. By doing so, our results can be comparable to findings from other finance classification tasks such as bankruptcy prediction. Furthermore, in the MLP, we use the textual features obtained by word embeddings (generic or finance).[7] Figure 1 illustrates this process step by step.

We note that using centroids of word embeddings is still, in effect, a bag of words approach, since word order is discarded. More powerful deep learning models, like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can be applied to text (Goldberg, 2017), using word embeddings as inputs, in ways that consider word order. Also, more recent deep learning models for text, mostly Transformer-based models (Vaswani et al., 2017)

[5] Size is also frequently-used as a matching criterion. However, if we use size to match our datasets, then we have to exclude it from our classification models. In line with previous studies, we prefer to use size as a control variable rather than as a matching criterion, because it is an important factor in explaining merger behavior (Espahbodi and Espahbodi, 2003; Pasiouras et al., 2007, 2010).

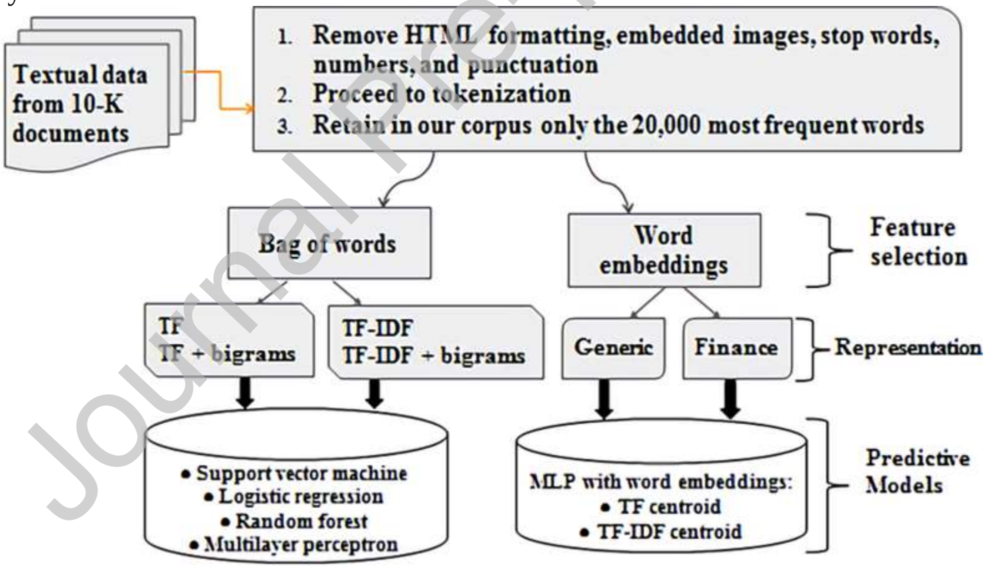
[6] In all our models, all financial variables are standardized. Textual features are also standardized when they are combined with financial variables. The hyper-parameters of the models are tuned using grid search based on the 5-fold cross-validation performance of the training set.

[7] The python algorithms were run through Google Colab, a product of Google Research, which allows anyone to write and run arbitrary python code. Since Colab randomly assigns GPU cards with different features each time, the estimated time for each experiment is approximately as follows: the average time needed to train a model with BOW textual features was 1.5 hours for bidders and 1 hour for targets. Word embeddings are more time consuming, especially with TF-IDF weights. More precisely, the equivalent average time for TF word embeddings was 2 hours for bidders and 1.5 hours for targets. Strikingly, for TF-IDF word embeddings, the average time for was 15 hours for bidders and 12 hours for targets, respectively.

can be pre-trained on gigantic corpora (Wikipedia and book collections) of unlabelled documents and then fine-tuned (further trained) on much fewer (compared to a gigantic corpus) task-specific labelled training instances, achieving better performance than when using only the task-specific labelled training instances. However, models of this kind can so far cope only with very short documents. For example, the commonly used BERT models (Devlin et al., 2018), which employ Transformers, can typically process up to 512 sub-word tokens (sub-word tokenizers break words into smaller units). Even very recently proposed variants of Transformer-based models for "long" text (Zaheer et al., 2020) can only process text input of up to 4,096 sub-word tokens, whereas the documents we consider are much longer. By contrast, centroids of word embeddings have no input length limitation.

One possible methodological concern with our machine learning models is the risk of overfitting. More precisely, overfitting is a serious issue where the models perform much better on the training set than on the out-of-sample (low bias, but high variance). In other words, in such a case the models learn the peculiarities of the training data to an excessive degree. We address this issue in several ways. First, we decrease the feature space by using dimensionality reduction techniques. Second, we optimize the hyper-parameters of our machine learning models to reduce the likelihood of overfitting. Third, in the MLP model, we employ regularization terms, dropout techniques, and early stopping. More details will be provided next.

Figure 1
Flow chart of analysis



3.2.1. Machine learning models with bag of words approach

3.2.1.1. Logistic regression

Logistic regression (LOGIT) is also one of the most commonly-used models in merger prediction task (Hasbrouck, 1985; Palepu, 1986; Ambrose and Megginson, 1992; Barnes, 1998, 1999; Powel, 2001; Espahbodi and Espahbodi, 2003; Cremers et al., 2009; Routledge et al., 2017). LOGIT belongs to the generalized linear models and uses a sigmoid function to convert the log-odds to probability as the predictive output of the model. LOGIT's rationale is to maximize the conditional log-likelihood of training samples in order to learn the parameters of the model. In fact, it typically uses stochastic gradient ascent or variants. To deal with overfitting the training dataset, regularization terms could be added to the log-likelihood. In our empirical setting, we employ L2 regularization, which subtracts the squared L2 norm of the weights vector (multiplied by a hyper-parameter), from the log-likelihood. The mathematical formula behind this model is described as follows:

$$P(Y_{t+1} = 1 | X_{i,t}) = \frac{\exp\left(\mathrm{bo} + \sum_{i=1}^{\mathrm{n}} \mathrm{b}_i \mathrm{X}_{i,t}\right)}{1 + \exp\left(\mathrm{bo} + \sum_{i=1}^{\mathrm{n}} \mathrm{b}_i \mathrm{X}_{i,t}\right)} \qquad (6)$$

where *Y* is the binary output. In the bidder prediction task *Y* equals if the firm i is a bidding bank at year *t*, and 0 otherwise. In the target prediction task *Y* equals if the firm *i* is a target bank at year *t*, and 0 otherwise. $X_{i,t}$ is a vector of *n* control variables at time *t*, b_i are parameters of the model, and b_0 is a bias term.

3.2.1.2. Support vector machine

Support vector machine (SVM) is a non-probabilistic supervised learning algorithm, first introduced by Vapnik (1998). So far, several studies have used SVMs in finance tasks, such as bankruptcy forecasting (Min and Lee, 2005; Shin et al., 2005; Wu et al., 2007; Manthoulis et al., 2020), stock price forecasting (Cao, 2003; Pai and Lin, 2005). Given a set of training instances that explicitly belong to various pre-defined categories, the SVM learns a decision boundary that defines the predicted identity of each instance. This decision boundary is practically a hyperplane in the feature space. The aim is to find the optimal hyperplane that maximizes the width of the gap (margin) among the instances of different categories (Kumar and Ravi, 2016). Notably, only the training samples near the hyperplane, either at the boundaries of the margin or inside the margin in case of letting “slack” in the separation, matter when creating the hyperplane. It is worth mentioning that finding the maximum margin hyperplane belongs to the general quadratic programming optimization problems. Interestingly, SVM has the advantage that being able to handle non-linearly separable data. In such a case, it can employ non-linear kernel functions such as radial basis function (RBF) kernel. As a result, our training data are projected into a higher dimensional space so that our data become more separable. (Nassirtoussi et al., 2014). Hence, in our empirical analysis, we use: (i) a linear SVM, and (ii) an SVM with RBF kernel. To prevent the overfitting issue, we choose the appropriate hyper-parameter value of the regularization parameter (C). In the case of SVM-linear, C equals 1 for both tasks. For SVM-RBF, C equals 0.5 (5) in the bidder (target) prediction task. As in all non-linear SVM models, there is another hyper-parameter known as gamma that controls for the curvature of the decision boundary. In our study, gamma is set to 0.01 for both tasks.

3.2.1.3. Random forest

Random forest (RF) is an ensemble machine learning algorithm, initially designed by Breiman (2001) as a variant of Bagging (Breiman, 1996). We employ RF by creating several uncorrelated decision tree classifiers. These decision trees are typically trained on bootstrap copies of original samples by randomly selecting a subset of features (Mai et al., 2019). The prediction process is then performed with each individual tree predicting a class. Based on majority voting, the class with the most votes becomes the output of our model. In general, RF outperforms the classical decision trees (DT), since it addresses the DT issue of overfitting to the training sample. To address overfitting, we optimize the two key hyper-parameters of the RF model: (1) the number of decision trees, and (2) the number of features randomly chosen to grow each decision tree when searching for the best split (max_features). First, the number of decision trees is defined as 200 for bidders and 100 for targets. Second, in addition to the randomization of training samples (bootstrap), the tuning optimization of max_features allows the proper randomization of feature space, leading to a decreased variance (low overfitting). We choose max_features to be equal to 10.

3.2.1.4. Multilayer perceptron

Artificial neural networks (ANNs) have widely been used in several prediction tasks in the area of finance (Kumar and Ravi, 2016). Among them, one of the simplest kinds of neural networks, and at the same time very popular is the multilayer perceptron (MLP) model. Not only for these reasons but also because MLP is able to handle all the text representations we use (TF-IDF-based or embedding-based) makes it an ideal choice for our analysis along with the rest of the machine learning models we use. Given that MLP is a feed-forward model, it maps inputs (financial variables and textual features) to a binary outcome. In a typical MLP model, there is an input layer of neurons, where our variables, textual or financial, are used as inputs (Goldberg, 2017). Next, there are one or more hidden layers. Each neuron computes a weighted sum of its inputs, applies a non-linear activation function to the resulting sum, and passes its output to the neurons of the next layer. The weights are learned by minimizing a loss function via back-propagation, a version of stochastic gradient descent for networks with hidden layers. In a classification task, the non-linear activation functions allow the model to cope with non-linearly separable data. In binary classification, as in our case, the output layer contains a single neuron with a sigmoid activation function, which provides the probability the model assigns to the positive class. The loss function is typically binary cross-entropy, in effect minimizing the divergence of the predicted probability distribution over the two classes from the correct (one-hot) distribution, for each training example. In the bidder prediction task, we use 3 hidden layers of 100 neurons in each.[8] In the target prediction task, the MLP model has 3 hidden layers, each of which has 200 neurons.

3.2.2. MLP model with word embedding approach

In all the previous models, we use the BOW text representations as textual inputs. Word embeddings are very uncommon in non-neural models, such as LOGIT, SVM, and RF, firstly for historical reasons; they were developed in the realm of neural network research. There are also theoretical reasons. In particular, each dimension (feature) of the word embeddings provides latent information, not easily interpretable on its own, and typically multiple dimensions of the word embeddings need to be non-linearly combined, by several hidden layers in MLPs, to obtain useful features. Linear models (e.g., LOGIT, linear SVMs) do not form such non-linear feature combinations. Also, in deep learning multiple stacked layers of approximately the same number of neurons lead to better performance compared to fewer stacked layers with many more neurons. Non-linear SVMs can be viewed as belonging to the latter kind as they project the input feature space only once to a higher-dimensional space. Furthermore, the dimensions of the word embeddings are real-valued with unknown and possibly skewed distributions, which requires considering multiple alternative

[8] We use Adam (a version of stochastic gradient descent) as the optimizer algorithm, and rectified linear unit (ReLU) as the activation function of each hidden layer. ReLU is defined as f(x) = max (0, x). Finally, we use early stopping to mitigate overfitting (Mai et al., 2019). To do so, we set aside 10% of training data as validation or development set.

discretizations or multiple alternative value-splits, which increases the computational cost of decision tree-based learners, including RF.

Kriebel and Stitz (2022) demonstrate that the MLP models with word embeddings as inputs provide comparable performance with more complicated deep learning networks in credit default prediction. To utilize textual information based on word embeddings, we have a vector (embedding) for each vocabulary word (obtained using tools like word2vec, see Section 2.3.3). To obtain a vector representation of an entire text, we can average the word embeddings of its words, called "centroid". Equivalently, we can obtain the centroid of a text by summing the word embeddings of all the words of the vocabulary, but weighting each word embedding in the sum as many times as the TF of the corresponding word in the text. As follows, we first provide the mathematical formula of the TF centroid textual feature:

$$\overrightarrow{TFcentroid_i} = \frac{\sum_j^V (TF_{ij} \times \overrightarrow{w_j})}{\sum_j^V TF_{ij}} \qquad (7)$$

where i represents each text in the sample, j represents each word in the vocabulary (V), $\overrightarrow{w_j}$ represents the 200-dimensional word embedding of each word j, and TF_{ij} represents the term frequency of the word j in the text i.

Alternatively, TF-IDF scores can also be combined with word embeddings. Notably, we can obtain the centroid of a text by weighting each word embedding using the TF-IDF weighting scheme. We multiply each embedding by the TF-IDF score. This has the effect that words of the text that are very common in the language (low IDF, e.g., articles) are in effect ignored when forming the centroid of word embeddings of a text, even if their TF is high in the particular text. Moreover, we present the mathematical formula for TF-IDF centroid textual feature:

$$\overrightarrow{TF-IDFcentroid_i} = \frac{\sum_j^V (TF_{ij} \times IDF_j \times \overrightarrow{w_j})}{\sum_j^V (TF_{ij} \times IDF_j)} \qquad (8)$$

where IDF_j represents the inverse document frequency of each word j.

With both kinds of centroids, a text can initially be viewed as a (d x n) matrix, where d is the dimensionality of the word embeddings and n is the length of the text in words (word occurrences). By computing the centroid of the text, the text representation becomes a (d x 1) matrix (vector), as illustrated in Figure 2. In particular, the word embeddings could be either pre-trained, such as the generic word embeddings of GloVe, or be our financial word embeddings trained on the EDGAR documents. Finally, we use these word embedding textual features as inputs to a MLP model.

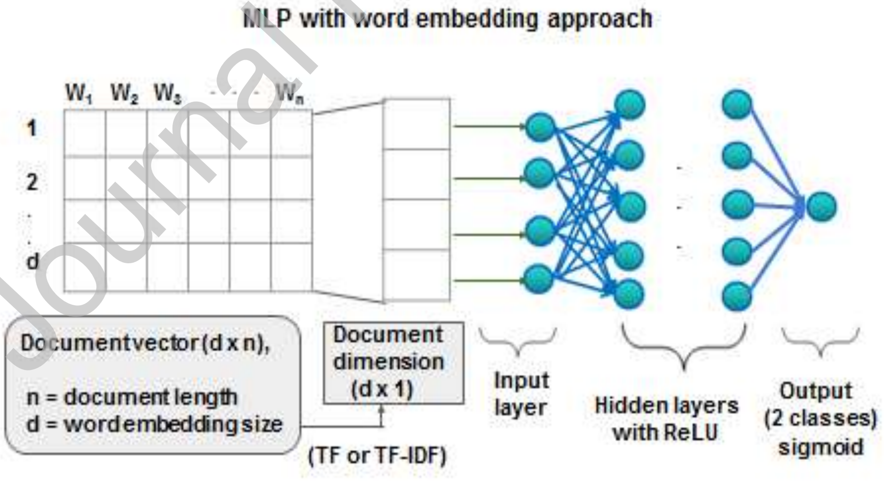
Figure 2 illustrates the architecture of the MLP models with the word embedding approach. First, we use a 200-dimensional vector to represent each document, as the size of the pre-trained word embeddings is 200. Second, these vectors are inserted as inputs in the model, and then they are processed by some hidden layers with the rectified linear unit (ReLU) activation function. Finally, there is the output layer where a sigmoid function provides the probability of the positive class.

We create our models using the Keras library with a TensorFlow backend (Chollet, 2017). We employ a batch size of 16, and the models take less than 12 epochs to converge. In the bidder prediction task, the models have 2 hidden layers without any linear projection in the input layer. In the target prediction task, the input layer linearly projects into 50 dimensions, and afterwards, there are 2 hidden layers. It is worth mentioning that linear projection is a beneficial mechanism, as it can limit the overfitting problem by reducing the number of features used as inputs in the model. Further, our models use the Glorot weight initialization scheme and each layer contains 512 neurons. Finally, to control for the issue of overfitting, we use the dropout technique and the early stopping strategy, as in Mai et al. (2019). The dropout method randomly omits a subset of hidden neurons at every step of the training process. The dropout rate in our models is defined as 10%. On the other hand, early stopping requires monitoring the performance of the validation set, a subset of the training set, so that we stop the training process when there is no more improvement.

3.3. Evaluation measure

We evaluate the out-of-sample performance of our classification models using the Area Under the Curve (AUC) score, which is frequently-used in studies with imbalanced datasets (Yildirim et al., 2021). The AUC score is computed from the receiver operating characteristic (ROC) curves, which are typically used in finance prediction tasks, such as bankruptcy prediction (Chava and Jarrow, 2004; Mai et al., 2019; Manthoulis et al., 2020; Borchert et al., 2023; Korangi et al., 2023), and credit risk prediction (Stevenson et al., 2021; Dumitrescu et al., 2022). The ROC curve plots the true-positive rate of the classifier on the vertical axis, and the false-positive rate on the horizontal axis, as the classification threshold varies. The area under this curve is called AUC, and its values are in the range of [0, 1]. Higher AUC values imply better out-of-sample classification ability of our models.

Figure 2
Architecture of the MLP models with the word embedding approach



4. Empirical results and discussion

4.1. Prediction with financial variables

As the first step in our empirical analysis, we examine the predictive power of our models when we use only financial variables as inputs. In fact, we investigate whether financial variables alone can distinguish between bidders and non-bidders or targets and non-targets. The results are reported in Table 3. First, we present the AUC scores of our classification models for the bidding banks. AUC scores range from 0.500 (SVM-RBF) to 0.604 (MLP). Furthermore, LOGIT, which is the most frequently-used model in the literature, performs relatively well with an AUC score of 0.595.

Second, we present the results for the target firms. In general, our benchmark models perform better compared to what is reported by the bidding banks. In fact, with the exception of SVM-linear, all other models yield higher AUC scores. More precisely, AUC scores range from 0.512 (SVM-linear) to 0.669 (RF). These results suggest that traditional financial variables are more informative in a target prediction task. This is expected to a large extent, since the target prediction usually focuses on weak bank fundamentals. However, in the case of bidding banks, financial variables do not account for strategic managerial decisions. To this end, we expect textual features to be more informative in the case of bidder prediction.

Table 3
Out-of-sample performance using only financial variables

	LOGIT	SVM-linear	SVM-RBF	RF	MLP
Bidders	0.595	0.598	0.500	0.581	0.604
Targets	0.655	0.512	0.636	0.669	0.666

This table reports the AUC scores for our machine learning models, using financial variables as inputs.

4.2. Prediction with textual features

In this section, we investigate whether the language used by managers in the bank annual reports has any predictive power in our merger classification task. To be consistent with our empirical setting, we will first analyze results based on the BOW approach, and then, we will report the results of the word embedding approach.

Table 4 presents out-of-sample AUC scores of our prediction models, using only textual data as inputs based on the BOW approach. We use four different types of textual features: (1) term frequency (TF), (2) term frequency-inverse document frequency (TF-IDF), (3) term frequency with bigrams (TF + bigrams), and (4) term frequency-inverse document frequency with bigrams (TF-IDF + bigrams). Types 1 and 2 use only unigrams, and types 3 and 4 use a combination of unigrams and bigrams.

Table 4
Out-of-sample performance using only textual features based on bag of words approach

	LOGIT	SVM-linear	SVM-RBF	RF	MLP
Panel A: Bidders					
TF	0.658	0.626	0.616	0.511	0.645
TF-IDF	0.595	0.657	0.636	0.638	0.643
TF + bigrams	0.658	0.645	0.621	0.581	0.650
TF-IDF + bigrams	0.657	0.666	0.599	0.620	0.651
Panel B: Targets					
TF	0.615	0.500	0.573	0.587	0.579
TF-IDF	0.596	0.579	0.599	0.623	0.587
TF + bigrams	0.520	0.542	0.546	0.575	0.562
TF-IDF + bigrams	0.590	0.561	0.549	0.605	0.604

This table reports the AUC scores for our machine learning models, using textual features based on the bag of words approach. The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams.

Panel A of Table 4 shows the results for our first dataset (bidders and non-bidders). Overall, our models using only textual data as inputs perform better than our benchmark models. SVM-linear yields the highest AUC score (0.666) followed by LOGIT (0.650) and MLP (0.650). Notably, almost all AUC scores (except the RF model with TF textual features) are equal to or higher than the ones reported in Table 3. This fact indicates that textual information of the 10-

K filings contains vital information for predicting future acquirers in the U.S. banking industry, a finding which is in line with our research question.

Panel B of Table 4 shows the results for our second dataset (targets and non-targets). In this task, the results are mixed. For instance, SVM-linear performs better with textual features compared to our benchmark models. For the remaining models, traditional financial variables are more meaningful inputs compared to textual features. However, this is legitimate considering that target prediction relies on weak fundamentals, as measured by the financial variables that we also use in our benchmark models. Taken altogether, our results might explain why the relevant literature, which almost exclusively uses financial variables, focuses primarily on target prediction rather than bidder prediction.

Table 5 reports the results when we employ textual features based on the word embedding approach. We examine the performance of two different models, the TF centroid embedding model and the TF-IDF centroid embedding model. In each model, we use as inputs either the generic word embeddings based on GloVe, or our finance word embeddings. More precisely, we use the MLP model with four different word embedding features: (1) TF Centroid with generic word embeddings as inputs (TF Generic centroid), (2) TF-IDF Centroid with generic word embeddings as inputs (TF-IDF Generic centroid), (3) TF Centroid with finance word embeddings as inputs (TF Finance centroid), and (4) TF-IDF Centroid with finance word embeddings as inputs (TF-IDF Finance centroid).

Panel A presents the results for the bidders and Panel B presents the results for the targets. In predicting future bidders, the TF-IDF Finance centroid embedding model has the best performance (0.568). Similarly, in predicting future targets, the TF-IDF Finance centroid embedding has the best performance, with an AUC score of 0.579. Collectively, these results highlight two important issues. First, our FWE embeddings are more informative textual features compared to generic word embeddings since they are trained on a finance-specific corpus. Second, in all cases, our MLP models with word embeddings perform better with the use of the TF-IDF weighting scheme. However, the results obtained are inferior to the ones obtained by the BOW approach and the benchmark models.

4.3. Prediction with both financial variables and textual features

In this section, we jointly use both financial variables and textual features as inputs in our classification models. We do so, in order to investigate whether and to what extent textual information can effectively be combined with financial variables in our merger classification task.

Table 5
Out-of-sample performance of the MLP model using only textual features based on word embedding approach

	Generic centroid	**Finance centroid**
Panel A: Bidders		
TF	0.481	0.528
TF-IDF	0.521	0.568
Panel B: Targets		
TF	0.560	0.577
TF-IDF	0.562	0.579

This table reports the AUC scores for the MLP model using textual features as inputs.

4.3.1. Combination of financial variables with bag of words textual features

We now investigate the prediction performance when both financial variables and textual features based on BOW are utilized. One issue that emerges here is that textual features dramatically outnumber financial variables, and as a result, the plethora of textual data may overrule the role of financial variables. Such a model may suffer from the "curse of dimensionality" (Mai et al., 2019). To alleviate this concern, we decrease the dimensionality of our textual features (number of words in the vocabulary).

We project our high-dimensional document vectors into a low-dimensional space using the singular value decomposition (SVD) dimensionality reduction technique as in Kim et al. (2005), and Degiannakis et al. (2018), among others. In our empirical analysis, we use SVD to project the original feature vectors to 100 dimensions (SVD100). We consider only the 100 first SVD components, as they were found to explain almost 80% of the joint variance of the 20,000 most frequent textual features in the 10-K filings. Hence, this method reduces the dimensions of our textual features from 20,000 to 100. By using such a low level of textual representation, we are able to deal with the curse of dimensionality, while preserving the meaningful information of the 10-K filings.

Table 6 presents the results of this analysis. Panel A reports the AUC scores for our first dataset (bidders and non-bidders). First, RF is the best-performing model with an AUC score of 0.706. This score is achieved with the combination of financial variables and TF-IDF$_{\text{SVD100}}$ textual features. Next, SVM-linear produces the second-best AUC score (0.681), followed by SVM-RBF (0.664) and MLP (0.662). Finally, LOGIT produces the lowest AUC score, which equals to 0.659. These findings could imply that machine learning models such as the RF can handle textual data more efficiently compared to traditional techniques such as logistic regression.

Table 6
Out-of-sample performance using both SVD100 textual features based on the bag of word approach and financial variables as inputs

	LOGIT	**SVM-linear**	**SVM-RBF**	**RF**	**MLP**
Panel A: Bidders					
TF$_{\text{SVD100}}$	0.658	0.669	0.616	0.638	0.647
TF-IDF$_{\text{SVD100}}$	0.640	0.654	0.664	0.706	0.662
(TF + bigrams)$_{\text{SVD100}}$	0.659	0.681	0.621	0.669	0.659
(TF-IDF + bigrams)$_{\text{SVD100}}$	0.638	0.659	0.662	0.673	0.658
Panel B: Targets					
TF$_{\text{SVD100}}$	0.517	0.552	0.570	0.632	0.582
TF-IDF$_{\text{SVD100}}$	0.569	0.518	0.618	0.695	0.637
(TF + bigrams)$_{\text{SVD100}}$	0.519	0.541	0.588	0.647	0.583
(TF-IDF + bigrams)$_{\text{SVD100}}$	0.568	0.533	0.599	0.689	0.623

This table reports the AUC scores for our machine learning models, using both textual features based on the bag of word approach and financial variables. The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams.

Two inferences are obtained when we compare the results of the models using both types of inputs with the models using a single type of input. On the one hand, the performance of our models is substantially improved when we use both textual features and financial variables instead of a single type of input. On the other hand, in some cases, the performance of the augmented models is comparable to the performance of the models using only textual features as inputs. Collectively, these findings may indicate that textual information from the bank annual reports is more informative relative to financial variables in the bidder prediction task.

Panel B of Table 6 reports the AUC scores for our second dataset (targets and non-targets). Interestingly, we achieve the highest AUC score when we augment our benchmark models with textual features. More precisely, RF yields an AUC score of 0.695 when we use TF-IDF$_{\text{SVD100}}$ as textual features, which is the highest score in the target prediction task so far. Furthermore, in all cases, the SVM-linear outperforms the benchmark SVM-linear model of Table 3. However, the scores for the three remaining models (LOGIT, SMV-RBF, and MLP) are inferior to the ones produced by the benchmark models. Overall, our results indicate that textual features can also have some incremental value in the target prediction task. Nevertheless, bank fundamentals, as proxied by our financial variables, are strong predictors of future targets, a fact which is consistent with the relevant literature.

4.3.2. Combination of financial variables with word embedding textual features

In this section, we examine the out-of-sample performance of the MLP model using a combination of word embedding textual features and financial variables as inputs. Table 7 presents our findings for bidder classification (Panel A) and target classification (Panel B).

The results of Panel A suggest that the combination of word embeddings with financial variables can improve the performance of the MLP model. More specifically, the TF-IDF Finance centroid produces an AUC score of 0.663 which is substantially higher than the score of the benchmark MLP model in Table 3. The TF-IDF Generic centroid also slightly outperforms the benchmark with a score of 0.607. The results of Panel B of Table 7 indicate that the combination of financial variables with word embeddings performs well also in the target classification task. More precisely, all models (except the TF Generic centroid) outperform the benchmark MLP model of Table 3. Similar to the bidder classification task, the TF-IDF Finance centroid yields the highest AUC score (0.690), followed by the TF-IDF Generic centroid (0.680), and the TF Finance centroid (0.670).

Table 7
Out-of-sample performance of the MLP model using both textual features based on the word embedding approach and financial variables as inputs

	Generic centroid	**Finance centroid**
Panel A: Bidders		
TF	0.559	0.590
TF-IDF	0.607	0.633
Panel B: Targets		
TF	0.660	0.670
TF-IDF	0.680	0.690

This table reports the AUC scores for the MLP model using both textual features and financial variables as inputs.

Further, our results provide two additional important findings. First, the TF-IDF centroid embedding model outperforms the TF centroid embedding model in all cases. This means that the TF-IDF weighting scheme produces a set of weights for our textual features that enhance the classification ability of our models. This result is consistent with previous findings, as the TF-IDF approach tends to perform better in many NLP tasks compared to simple proportional weighting (Loughran and McDonald, 2011; Loughran and McDonald, 2016; Katsafados et al., 2023). Second, the use of our finance word embeddings yields more accurate estimates, compared to using generic word embeddings. Therefore, we argue that the finance word embeddings are more meaningful inputs than generic word embeddings in both classification tasks. This is expected to some extent, because FWEs take into account the most likely meaning of a word in a business context, and as such, they are able to understand better the semantics of the text.

4.4. Robustness tests

4.4.1. Bootstrap statistical significance test

So far, our findings provide supportive evidence that the inclusion of textual features substantially improves the performance of our benchmark models, especially in the bidder classification task. In both tasks, the best-performing model is the RF with TF-IDF$_{SVD100}$ as textual features. However, it is important to test the consistency of these results, by including statistical significance tests to validate metric gains. To do so, we employ the bootstrap resampling method of Berg-Kirkpatrick et al. (2012), and we examine whether the best-augmented model outperforms the best benchmark model in both tasks. We provide a more detailed description of this technique in Appendix B.

Table 8 presents the results of this analysis. Initially, we compare the RF(TF-IDF$_{SVD100}$) model of Table 6 with the MLP model of Table 3, which was the best-performing benchmark model in the bidder classification task. The comparison suggests that the RF(TF-IDF$_{SVD100}$) significantly outperforms the benchmark MLP model (*p*=0.000). For the target classification task, we compare the performance of the RF(TF-IDF$_{SVD100}$) of Table 6 with the benchmark RF of Table 3, which was the best-performing benchmark model in the target classification task. Similar to the bidder classification task, the RF(TF-IDF$_{SVD100}$) significantly outperforms the benchmark RF model (*p*=0.006). Hence, our findings provide further support to our argument that textual features can significantly complement traditional financial variables in merger prediction tasks.

Table 8
Bootstrap randomization and statistical significance

Comparisons	Task	Winner	p-value
RF(TF-IDF$_{SVD100}$) *vs* benchmark MLP	Bidders	RF(TF-IDF$_{SVD100}$)	0.000***
RF(TF-IDF$_{SVD100}$) *vs* benchmark RF	Targets	RF(TF-IDF$_{SVD100}$)	0.006***

This table reports the p-values of our results based on bootstrap statistical significance tests. In each task (bidders or targets), we compare the best performing model with textual features and financial variables as inputs with the best benchmark model of Table 3. P-vales are calculated using the bootstrap resampling method of Berg-Kirkpatrick et al. (2012).

4.4.2. Importance of textual features

To further illustrate the high importance of textual features, we adopt the Gini impurity technique (Kurt et al., 2008). Practically, this technique computes the importance score for each variable in the model, and it is applied to the RF models. Hence, we compute the Gini importance scores for the 25 most important features of our RF models, which are the best-performing models in both tasks. We limit the analysis to the 25 most important features, due to the fact that our textual features substantially outnumber our financial variables. Then, we compute the sum of these scores separately for textual features and for financial variables.

Panel A of Table 9 presents the Gini importance scores for our bidder classification task. By comparing those sums, we observe that textual features are more important inputs than financial variables in all cases and by a large margin. This result is in line with our baseline findings, since we have documented the importance of textual features in predicting bidders in U.S. bank M&As. Furthermore, Panel B of Table 9 reports the Gini importance scores for our target classification task. In this task, the scores for financial variables are slightly higher compared to textual features in all cases except the (TF-IDF + bigrams)$_{SVD100}$ textual features. Again, this finding is consistent with what we have reported so far. In fact, identifying future targets in the banking industry is usually conditional upon the weak bank fundamentals measured by financial variables. However, textual features are still important inputs in such a task.

4.4.3. Identification of important textual features

To understand which are the most important features in our prediction tasks, we adopt the novel LIME method which practically explains the predictions of any classifier (Ribeiro et al., 2016). More precisely, we employ LIME to visualize the important bigrams of the RF model, as it is the best-performing model with textual features in both tasks.

Table 9
Sum of Gini impurity scores

	Financial variables Gini	**Textual variables Gini**
Panel A: Bidders		
TF$_{SVD100}$	0.054	0.232
TF-IDF$_{SVD100}$	0.063	0.229
(TF + bigrams)$_{SVD100}$	0.077	0.210
(TF-IDF + bigrams)$_{SVD100}$	0.069	0.221
Panel B: Targets		
TF$_{SVD100}$	0.246	0.194
TF-IDF$_{SVD100}$	0.236	0.199
(TF + bigrams)$_{SVD100}$	0.258	0.191
(TF-IDF + bigrams)$_{SVD100}$	0.229	0.236

This table reports the Gini impurity scores when both SVD100 textual features and financial variables are used as inputs in the RF model. We provide the sum of Gini scores separately for financial variables and textual features. However, in our calculations we consider only the 25 most important features. The first two lines of each panel report results using only unigrams, while the last two lines report results using combinations of unigrams and bigrams.

Table 10 presents the results of this analysis. Positive (negative) impact translates to higher (lower) probability to be classified as bidder or target. By looking at the results, we observe that they are in line with our intuition. For instance, bigrams that relate to the banks' organizational structure can explain merger activity. Typically, bidding banks operate as bank holding companies (BHCs). BHCs file consolidated financial statements because they own a controlling stake in one or more banks and non-bank financial institutions. By contrast, stand-alone commercial banks are rarely bidders and more likely targets in the U.S. bank merger market. These banks do not operate as holding companies and they do not file consolidated financial statements. Therefore, it is reasonable why terms like "Consolidated financial" or "Holding company" are positively related to bidder prediction and negatively related to target prediction. Furthermore, in the case of targets, the positive impact relates to words that describe weak bank fundamentals, such as loan losses and credit losses, a result that was consistent with our expectations.

Table 10
LIME visualization of textual features

Panel A: Bidders		**Panel B: Targets**	
Positive Impact	*Negative Impact*	*Positive Impact*	*Negative Impact*
Financial group	Subsidiary banks	Credit losses	Internal control
Bank holding	Loan losses	Loan losses	Federal reserve
Consolidated financial	Affiliate banks	Fair value	Mortgage loans
Internal control	Federal reserve	Assets liabilities	Loan portfolio
Interest rates	Investment securities	Interest revenue	Investment securities
Preferred securities	Mortgage loans	Financial reporting	Increased million
Fair value	Depository institutions	Net interest	Allowance loan
Financial statements	Deferral plan	Results operations	Loan leases
Loan portfolio	Loss share	Cash flows	Consolidated financial
Interest income	Subordinated debentures	Interest income	Holding company

This table reports the most important bigrams used as inputs in the RF model in both tasks. Their importance is computed using the LIME methodology. Positive (negative) impact translates to higher (lower) probability to be classified as bidder or target.

4.4.4. Recurrent neural network analysis

One final concern regarding our empirical setup is the degree to which word embeddings, and particularly our finance-specific ones, are more meaningful inputs than the ones produced with the traditional BOW approach. In the target prediction task, MLP models with word embeddings outperform the vast majority of models with BOW textual features. However, in the bidder prediction task, evidence along these lines is not yet conclusive.

To alleviate this concern, we employ another, more sophisticated (but also computationally more expensive) deep learning model, namely the Bi-directional Long Short-Term Memory Recurrent Neural Network (BLSTM-RNN). In the computer science literature, BLSTM-RNNs with word embeddings have been proved to be quite effective in producing accurate predictions due to their capability in modelling sequential data (Wang et al., 2015). Furthermore, in their recent survey paper, Doumpos et al. (2023) suggest that BLSTM-RNNs should be considered for further research in banking prediction tasks. Hence, we run both prediction tasks using word embeddings and financial variables as inputs to BLSTM-RNN models.

In the bidder prediction task with GloVe word embeddings, the model has 2 Bi-directional LSTM layers with 256 neurons in each layer (128 neurons in each direction) and linear projection to 64 dimensions where the outcome is transmitted to an MLP model with 1 hidden layer of 256 neurons. For regularization reasons, we employ a drop-out rate equal to 0.3. When we use our finance word embeddings, the model has 1 Bi-directional LSTM layer with 256 neurons (128 neurons in each direction), and linear projection to 128 dimensions where the outcome is transmitted to an MLP model with 2 hidden layers of 256 neurons in each layer. The drop-out rate equals 0.1. In the target prediction task with GloVe word embeddings, the model has 2 Bi-directional LSTM layers with 256 neurons in each layer (128 neurons in each direction) and linear projection to 64 dimensions where the outcome is transmitted to an MLP model with 2 hidden layers of 128 neurons in each layer. The drop-out rate equals 0.2. When we use our finance word embeddings, the model has 1 Bi-directional LSTM layer with 256 neurons (128 neurons in each direction) and linear projection to 128 dimensions where the outcome is transmitted to an MLP model with 2 hidden layers of 256 neurons in each layer. The drop-out rate equals 0.1. In all experiments, we use class weighting to deal with the class imbalance which proved to be more efficient than undersampling in the case of BLTSM-RNNs.[9]

Table 11 reports the results of the BLSTM-RNN models. By comparing our results with the ones in Table 7 (MLP models with word embeddings), we see that the AUC scores are substantially improved in the bidder prediction task. More importantly, when we compare our results with Table 6, the AUC score for bidders using FWE is the highest one reported in our study (0.725). Hence, this evidence suggests that our finance word embeddings can be the most informative textual features when used as inputs in the appropriate deep learning model. For targets, AUC scores range from 0.661 to 0.669, scores which are higher than most scores reported in Table 6. Furthermore, in both tasks, AUC scores are higher with FWE compared to GloVe word embeddings. The importance of these findings is twofold. First, BLSTM-RNNs are more efficient in handling textual information from word embeddings, especially in the bidder prediction task. Second, domain-specific word embeddings, in conjunction with the appropriate model, should be

[9] The hyperparameters of the models are tuned using a development set containing 15% of the training set selected at random.

considered as inputs in such prediction tasks.

Table 11
Bi-directional LSTM predictions

	GloVe	**FWE**
Bidders	0.720	0.725
Targets	0.661	0.669

This table reports the AUC scores for the Bi-directional LSTM model using both word embedding textual features and financial variables as inputs.

5. Conclusions

In this study, we utilize several machine learning models to predict bank mergers in the U.S. Our key innovation is that we investigate the role of textual disclosure of bank annual reports in our merger prediction task. More precisely, we examine whether the language used by bank managers in the annual reports has any additional predictive power in our classification models beyond the traditional financial variables. The intuition behind this text-based approach is that textual information could reduce the opaqueness of bank assets and provide some important insights regarding the strategic options of the banking firms. Hence, our study contributes to the recent body of research that utilizes textual analysis in various finance tasks.

We create a comprehensive dataset of 9,207 U.S. bank-year observations during the period 1994-2016. To create our textual features, we use the bag of words and the word embedding approaches. One important aspect of our empirical approach is that we go beyond the frequently-used generic word embeddings, and we create our own word embeddings specialized in the finance sector. Then, we use our textual features (with or without financial variables) as inputs in our classification models, and we examine whether the inclusion of textual data can improve the performance of our benchmark models.

Our findings provide strong evidence for the importance of textual information in a bank merger classification task. In fact, when we augment our benchmark models with textual data, we achieve the highest AUC scores. In the bidder classification task, textual data alone are in many cases more informative than financial variables. When we combine both types of inputs, our models significantly outperform all our benchmark models. By using the bootstrap resampling method of Berg-Kirkpatrick et al. (2012), we find that this outperformance is also statistically significant. Furthermore, we employ additional robustness tests to quantify the importance of textual features in our merger prediction task, and to examine which textual features indeed contribute to the enhanced performance of our best models. Finally, to illustrate the predictive ability of our finance word embeddings, we use them as inputs (along with financial variables) in BLSTM-RNN models, which are more capable of exploiting such information. Our findings suggest that in the bidder prediction task, our finance word embeddings are the most informative textual sources since we achieve the highest AUC score. This is in line with our intuition that textual information is more meaningful in this task, since the choice to become a bidder is a strategic decision for the bank. To conclude, we hope that our study will provide fertile ground for future research in the fast-growing literature of textual analysis in finance.

References

Ahmed, A. S., Takeda, C., & Thomas, S. (1999). Bank loan loss provisions: A reexamination of capital management, earnings management and signaling effects. *Journal of Accounting and Economics*, 28, 1-25.

Ambrose, B. W., & 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., & Katsafados, A. (2023). Bank deposits and textual sentiment: When an European Central Bank president's speech is not just a speech. *The Manchester School*, 91, 55-87.

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

Barnes, P. (1998). Can takeover targets be identified by statistical techniques? Some UK evidence. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 573-591.

Barnes, P. (1999). Predicting UK takeover targets: Some methodological issues and an empirical study. *Review of Quantitative Finance and Accounting*, 12, 283-302.

Beatty, A. L., Ke, B., & Petroni, K. R. (2002). Earnings management to avoid earnings declines across publicly and privately held banks. *Accounting Review*, 77, 547-570.

Becher, D. A. (2009). Bidder returns and merger anticipation: Evidence from banking deregulation. *Journal of Corporate Finance*, 15, 85-98.

Berg-Kirkpatrick, T., Burkett, D., & Klein, D. (2012). An empirical investigation of statistical significance in nlp. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 995-1005).

Bernabé-Moreno, J., Tejeda-Lorente, A., Herce-Zelaya, J., Porcel, C., & Herrera-Viedma, E. (2020). A context-aware embeddings supported method to extract a fuzzy sentiment polarity dictionary. *Knowledge-Based Systems*, 190, 105236.

Bernanke, B. C. (2010) Causes of the recent financial and economic crisis. Statement before the Financial Crisis Inquiry Commission, Washington, D.C., September 2.

Blau, B. M., Brough, T. J., & Griffith, T. G. (2017). Bank opacity and the efficiency of stock prices. *Journal of Banking and Finance*, 76, 32-47.

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

Borchert, P., Coussement, K., De Caigny, A., & De Weerdt, J. (2023). Extending business failure prediction models with textual website content using deep learning. European Journal of Operational Research, 306, 348-357.

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

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

Brewer, E., & Jagtiani, J. (2013). How much did banks pay to become too-big-to-fail and to become systemically important? Journal of Financial Services Research, 43, 1-35.

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

Buehlmaier, M. M., & Zechner, J. (2021). Financial media, price discovery, and merger arbitrage. *Review of Finance*, 25, 997-1046.

Bushman, R. M., & Williams, C. D. (2012). Accounting discretion, loan loss provisioning, and discipline of banks' risk-taking. *Journal of Accounting and Economics*, 54, 1-18.

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

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, 8, 537-569.

Chen, J., Yan, S., & Wong, K. C. (2020). Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications*, 32, 10809-10818.

Chollet, F. (2017). Deep learning with python. Manning Publications.

Cocco, J. F., & Volpin, P. F. (2013). Corporate pension plans as takeover deterrents. *Journal of Financial and Quantitative Analysis*, 48, 1119-1144.

Cornett, M. M., & Tehranian, H. (1992). Changes in corporate performance associated with bank acquisitions. *Journal of Financial Economics*, 31, 211-234.

Cornett, M. M., McNutt, J. J., & Tehranian, H. (2006). Performance changes around bank mergers: Revenue enhancements versus cost reductions. *Journal of Money, Credit and Banking*, 38, 1013-1050.

Cornett, M. M., Tanyeri, B., & Tehranian, H. (2011). The effect of merger anticipation on bidder and target firm announcement period returns. *Journal of Corporate Finance*, 17, 595-611.

Cremers, K. J. M., Nair, V. B., & John, K. (2009). Takeovers and the cross-section of returns. *Review of Financial Studies*, 22, 1409-1445.

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

Delis, M. D., Kazakis, P., & Zopounidis, C. (2023). Management and takeover decisions. *European Journal of Operational Research*, 304, 1256-1268.

DeLong, G. L. (2001). Stockholder gains from focusing versus diversifying bank mergers. *Journal of Financial Economics*, 59, 221-252.

DeLong, G. L., & DeYoung, R. (2007). Learning by observing: information spillovers in the execution and valuation of commercial bank M&As. *Journal of Finance*, 62,181-216.

Demirgüç-Kunt, A., & Huizinga, H. (2013). Are banks too big to fail or too big to save? International evidence from equity prices and CDS spreads. *Journal of Banking and Finance*, 37, 875-894.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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

Doumpos, M., Zopounidis, C., Gounopoulos, D., Platanakis, E., & Zhang, W. (2023). Operational research and artificial intelligence methods in banking. European Journal of Operational Research, 306, 1-16.

Dumitrescu, E., Hue, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297, 1178-1192.

Edmans, A., Goldstein, I., & Jiang, W. (2012). The real effects of financial markets: The impact of prices on takeovers. *Journal of Finance*, 67, 933-971.

Ellul, A., & Yerramilli, V. (2013). Stronger risk controls, lower risk: Evidence from US bank holding companies. *Journal of Finance*, 68, 1757-1803.

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

Filson, D., & Olfati, S. (2014). The impacts of Gramm–Leach–Bliley bank diversification on value and risk. *Journal of Banking and Finance*, 41, 209-221.

Flannery, M. J., Kwan, S. H., & Nimalendran, M. (2004). Market evidence on the opaqueness of banking firms' assets. *Journal of Financial Economics*, 71, 419-460.

Flannery, M. J., Kwan, S. H., & Nimalendran, M. (2013). The 2007–2009 financial crisis and bank opaqueness. *Journal of Financial Intermediation*, 22, 55-84.

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

Geng, R., Bose, I., & 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.

Gregoriou, G. N., & Renneboog, L. (2007). Understanding mergers and acquisitions: Activity since 1990. In G. N. Gregoriou, & L. Renneboog (Eds.), *International mergers and acquisitions activity since 1990* (pp. 1-20), Academic Press.

Hankir, Y., Rauch, C., & Umber, M. P. (2011). Bank M&A: A market power story? *Journal of Banking and Finance*, 35, 2341-2354.

Hanley, K. W., & Hoberg, G. (2010). The information content of IPO prospectuses. *Review of Financial Studies*, 23, 2821-2864.

Hannan, T. H., & Rhoades, S. A. (1987). Acquisition targets and motives: The case of the banking industry. *Review of Economics and Statistics*, 69, 67-74.

Hasbrouck, J. (1985). The characteristics of takeover targets q and other measures. *Journal of Banking and Finance*, 9, 351-362.

Houston, J. F., James, C. M., & Ryngaert, M. D. (2001). Where do merger gains come from? Bank mergers from the perspective of insiders and outsiders. *Journal of Financial Economics*, 60, 285-331.

Huizinga, H., & Laeven, L. (2012). Bank valuation and accounting discretion during a financial crisis. *Journal of Financial Economics*, 106, 614-634.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110, 712-729.

Jiang, L., Levine, R., & Lin, C. (2016). Competition and bank opacity. *Review of Financial Studies*, 29, 1911-1942.

Jones, J. S., Lee, W. Y., & Yeager, T. J. (2013). Valuation and systemic risk consequences of bank opacity. *Journal of Banking and Finance*, 37, 693-706.

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

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

Kim, H., Howland, P., & Park, H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6, 37-53.

Korangi, K., Mues, C., & Bravo, C. (2023). A transformer-based model for default prediction in mid-cap corporate markets. *European Journal of Operational Research*, 308, 306-320.

Koskela, E., & Stenbacka, R. (2000). Is there a tradeoff between bank competition and financial fragility? *Journal of Banking and Finance*, 24, 1853-1873.

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, 302, 309-323.

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

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34, 366-374.

Leledakis, G. N., & Pyrgiotakis, E. G. (2022). U.S. bank M&As in the post-Dodd–Frank Act era: Do they create value? *Journal of Banking and Finance*, 135, Article 105576.

Leledakis, G. N., Mamatzakis, E. C., Pyrgiotakis, E. G., & Travlos, N. G. (2021). Does it pay to acquire private firms? Evidence from the U.S. banking industry. *European Journal of Finance*, 27, 1029-1051.

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

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

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69, 1643-1671.

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

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

Manne, H. G. (1965). Mergers and the market for corporate control. *Journal of Political Economy*, 73, 110-120.

Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. The MIT Press.

Manthoulis, G., Doumpos, M., Zopounidis, C., & Galariotis, E. (2020). An ordinal classification framework for bank failure prediction: Methodology and empirical evidence for US banks. *European Journal of Operational Research*, 282, 786-801.

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011, May). Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5528-5531). IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 3111-3119.

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

Morgan, D. P. (2002). Rating banks: Risk and uncertainty in an opaque industry. *American Economic Review*, 92, 874-888.

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

Neophytou, E., & Mar Molinero, C. (2004). Predicting corporate failure in the UK: A multidimensional scaling approach. *Journal of Business Finance and Accounting*, 31, 677-710.

Nguyen, B. H., & Huynh, V. N. (2022). Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach. *Journal of the Operational Research Society*, 73, 102-121.

O'hara, M., & Shaw, W. (1990). Deposit insurance and wealth effects: the value of being "too big to fail". *Journal of Finance*, 45, 1587-1600.

Pai, P. F., & 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.

Pasiouras, F., Gaganis, S., & 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., & 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.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (pp. 1532-1543).

Powell, R. G. (2001). Takeover prediction and portfolio performance: A note. *Journal of Business Finance and Accounting*, 28, 993-1011.

Prasad, R. M., & Melnyk, Z. L. (1991). Positioning banks for acquisitions: A research note. *Economics Letters*, 35, 51-56.

Ramaswamy, K. (1997). The performance impact of strategic similarity in horizontal mergers: Evidence from the US banking industry. *Academy of Management Journal*, 40, 697-715.

Rhoades, S. A. (1993). Efficiency effects of horizontal (in-market) bank mergers. *Journal of Banking and Finance*, 17, 411-422.

Rhoades, S. A. (1998). The efficiency effects of bank mergers: An overview of case studies of nine mergers. *Journal of Banking and Finance*, 22, 273-291.

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

Rogers, J. L., Van Buskirk, A., & Zechman, S. L. C. (2011). Disclosure tone and shareholder litigation. *Accounting Review*, 86, 2155-2183.

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

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

Slowinski, R., Zopounidis, C., & Dimitras, A. I. (1997). Prediction of company acquisition in Greece by means of the rough set approach. *European Journal of Operational Research*, 100, 1-15.

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, 295, 758-771.

Thompson, S. (1997). Takeover activity among financial mutuals: An analysis of target characteristics. *Journal of Banking and Finance*, 21, 37-53.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, (pp. 6000-6010).

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

Very, P., Metais, E., Lo, S., & Hourquet, P. G. (2012). Can we predict M&A activity. In S. Finkelstein, & C. L. Cooper (Eds.), *Advances in mergers and acquisitions* (Vol. 11, pp. 1-32). Emerald Group Publishing.

Wang, P., Qian, Y., Soong, F. K., He, L., & Zhao, H. (2015). A unified tagging solution: Bidirectional LSTM recurrent neural network with word embedding. arXiv preprint arXiv:1511.00215.

Wheelock, D. C., & Wilson, P. W. (2000). Why do banks disappear? The determinants of US bank failures and acquisitions. *Review of Economics and Statistics*, 82, 127-138.

Wu, C. H., Tzeng, G. H., Goo, Y. J., & 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.

Yıldırım, M., Okay, F. Y., & Özdemir, S. (2021). Big data analytics for default prediction using graph theory. *Expert Systems with Applications*, 176, Article 114840.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big bird: Transformers for longer sequences. arXiv preprint arXiv:2007.14062.

Zheng, Y. (2020). Does bank opacity affect lending? *Journal of Banking and Finance*, 119, Article 105900.