Using textual analysis to identify merger participants: Evidence from the U.S. banking industry

by

Apostolos G. Katsafados
Department of Accounting and Finance
School of Business
Athens University of Economics and Business
Greece

Ion Androutsopoulos
Department of Informatics
Athens University of Economics and Business
Greece

Ilias Chalkidis
Department of Informatics
Athens University of Economics and Business
Greece

Emmanouel Fergadiotis
Department of Informatics
Athens University of Economics and Business
Greece

George N. Leledakis*
Department of Accounting and Finance
School of Business
Athens University of Economics and Business
Greece

and

Emmanouil G. Pyrgiotakis
Essex Business School
University of Essex
U.K.

This version: April, 2020

*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), ion@aueb.gr (I. Androutsopoulos), ihalk@aueb.gr (I. Chalkidis), fergadiotis@aueb.gr (E. Fergadiotis), gleledak@aueb.gr (G. Leledakis), e.pyrgiotakis@essex.ac.uk (E. Pyrgiotakis). We would like to thank Jonathan Batten (the Chief Editor), an anonymous referee, Manthos Delis, Vassilis Efthymiou, Athanasios Episcopos, Emmanuel Mamatzakis, Nickolaos Travlos and the participants at the 2018 National Conference of the Financial Engineering and Banking Society (FEBS) for their valuable comments and suggestions. Apostolos Katsafados acknowledges financial support from the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY) of Greece. 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.

Using textual analysis to identify merger participants: Evidence from the U.S. banking industry

Abstract

In this paper, we use the sentiment of annual reports to gauge the likelihood of a bank to participate in a merger transaction. We conduct our analysis on a sample of annual reports of listed U.S. banks over the period 1997 to 2015, using the Loughran and McDonald's lists of positive and negative words for our textual analysis. We find that a higher frequency of positive (negative) words in a bank's annual report relates to a higher probability of becoming a bidder (target). Our results remain robust to the inclusion of bank-specific control variables in our logistic regressions.

JEL classification: G14, G21, G34, G40

Keywords: *Textual analysis; text sentiment; bank mergers and acquisitions; acquisition likelihood*

1. Introduction

In the previous decades, the U.S. banking industry has experienced intense consolidation through mergers and acquisitions (M&As). In the literature, there is a general agreement on the broad forces that affect bank merger activity (DeYoung et al., 2009). However, to date, there is elusive evidence on the factors that influence the probability of bank to participate in a merger. Furthermore, the majority of the relevant studies examine this issue only from the target banks' perspective (Prasad and Melnyk, 1991; Pasiouras et al., 2007).

In this paper, we attempt to fill this gap in the literature by studying the underlying characteristics of banks that become either bidders or targets. We differentiate from the existing empirical work, since we use the sentiment of annual reports (i.e., Form 10-K) to gauge the banks' acquisition likelihood. Hence, our study adds to the growing literature that relates textual analysis to the banking industry. For instance, Gandhi et al. (2019) use textual data as a proxy for banks' financial distress. In a similar fashion, Del Gaudio et al. (2019) investigate the relationship between bank stability and the tone of the annual reports. To the best of our knowledge, our study is the first to utilize textual information in the context of bank M&As.

Prior literature suggests that potential bidders differ systematically from potential targets in their characteristics. In their early studies, Hannan and Rhoades (1987), Thompson (1997), and Hadlock et al. (1999) find that target banks tend to be in a worse financial condition compared to bidding banks. In detail, larger, well-capitalized and more profitable banks are anticipated as likely bidders, whereas smaller and less profitable banks as potential targets (Becher, 2009). In this regard, the sentiment of annual reports may have an adverse effect on a bank's likelihood to become bidder or target. For this reason, we expect banks with a higher fraction of positive (negative) words in their annual reports to be likely bidders (targets). To test our prediction, we perform several logistic regressions, where we use both textual data

and bank-specific financial variables. Our goal is to quantify whether and to what extent the use of textual information can enhance the ability of our logistic regressions to determine banks' acquisition likelihood.

Our results provide novel evidence that text sentiment constitutes a key element in determining the likelihood of bank acquisitions. Consistent with our expectations, more positive (negative) language in the bank's annual report is associated with a higher probability of becoming a bidder (target) in the subsequent year after the filing. It is also noteworthy that our findings are not only statistically, but also highly economically significant. Finally, this documented positive relationship between textual information and bank acquisition likelihood is not influenced by the inclusion of bank-specific financial variables.

One potential concern of our analysis is the quality of information included in the banks annual reports, due to the managers' incentives to conceal financial distress. However, there are substantial reputational costs to managers who don't report accurate information to their shareholders (Skinner, 1994). It has also been documented that less pessimistic language in annual reports leads to higher litigation risk (Rogers et al., 2011). In fact, the majority of lawsuits filed by a firm's investors involve cases of information misstatement (Kim and Skinner, 2012). Considering these facts, it is less likely that bank managers would be tempted to disclose inaccurate information in their annual reports.

Our findings are important to investors, but more importantly to regulators. In fact, any model that improves the ability of regulators to identify potential bidders and targets is beneficial, since it enables policy makers to a priori evaluate any merger-related anticompetitive effects and the degree of competition in the banking industry (Pasiouras et al., 2010). Furthermore, our proposed methodology might also be of use to bank managers who are interested in identifying potential acquirers or targets (Pasiouras et al., 2007).

The rest of the paper is organized as follows. Section 2 describes our sample collection process and methodology. Section 3 reports our empirical findings, and Section 4 concludes.

2. Data and Methodology

2.1. Sample selection

We obtain data from three different sources. First, we collect bank annual reports (10-Ks, 10-K405s, 10-KSBs, and 10-KSB40s), excluding amended documents, from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database. To be included in the sample, a bank's fiscal year end should be during the calendar years 1997 to 2015. Further, we require at least 2,000 words to appear in the SEC filings. Using this filter, we omit 31 observations. In addition, 2 banks had more than one filing in the same fiscal year, and as a result, we exclude them from the sample, following Loughran and McDonald (2011). Therefore, our selection process results in an initial sample of 16,012 bank-year observations.

Second, we collect bank characteristics from the Federal Reserve Bank of Chicago (FRBC).[1] To do so, we use the Federal Reserve Bank of New York's CRSP-FRB link, which provides the RSSD IDs for all publicly-traded U.S. banks. Then, in order to match the RSSD IDs of the banks with their Central Index Keys (CIK), we merge the FRBC data with our initial sample from EDGAR, using the bank names and locations (state and city if possible). To maximize the number of usable observations, we also use the National Information Centre (NIC) database, where we manually match the bank RSSD IDs with their CIKs. Our final sample includes 8,068 bank-year observations.

Third, we collect bank M&As data from the Thomson ONE database for deals announced between April, 1997 and March, 2017.[2] Similar to Leledakis and Pyrgiotakis (2020), we use the following criteria to filter our bank M&As sample: (1) both bidders and targets are

[1] We obtain financial information of bank holding companies (BHCs) from the FR Y-9C reports and of commercial banks and savings institutions from Call Reports.

[2] To be included in the merger sample, a bank should be a bidder or target in the subsequent year after the filing date. The earliest filing date of our sample was in the end of March, 1997 and the latest in the end of March, 2016.

commercial banks, savings institutions, or bank holding companies, (2) the bidder is public, (3) the target is a public firm, a private firm, or an unlisted subsidiary of a public firm, (4) all public firms are listed on NYSE, AMEX, or Nasdaq, and (5) the bidder acquired an interest of above 50% in a target, as long as this interest was initially less than 50%. Our bank merger sample consists of 1,078 observations (751 bidders, and 327 targets).

2.2. Textual analysis and methodology

The retrieved bank annual reports are encoded in hypertext markup language (HTML). Hence, we remove HTML formatting and any other non-textual information, such as embedded images or spreadsheets that may be present in the text (Bodnaruk et al., 2015). We also remove all identified HTML tables, if their numeric character content exceeds 15%. Further, we eliminate punctuation and generic stop words from the text. Finally, the processed texts are encoded as Bag-of-Words scalars.

To measure the sentiment of the bank annual reports, we use two common term weighting schemes: (1) the term frequency (TF), and (2) the term frequency-inverse document frequency (TF-IDF). For this purpose, we use the Loughran and McDonald's (2011) lists of positive and negative words. TF scores are calculated as the proportion of positive (or negative) words relative to the total number of words in each report. TF-IDF approach downweights the TF scores on the basis of how frequently a word appears in the sampled bank reports. We report results based on the TF-IDF weighting scheme. In untabulated results, we also use the TF weighting, and we find similar results.

Finally, in order to estimate the probability of a bank being a bidder or target, we use several logistic regressions (Palepu, 1986; Barnes, 1998; Powel, 2001; Routledge et al., 2017). More precisely, we estimate the following logistic regressions:

$$Bidder\ dummy_{i,t+1} = a + \beta_1 Positive\ TF-IDF_{i,t} + \beta_2 X_{i,t} + \varepsilon_{i,t} \quad (1)$$

$$Target\ dummy_{i,t+1} = a + \beta_1 Negative\ TF-IDF_{i,t} + \beta_2 X_{i,t} + \varepsilon_{i,t} \quad (2)$$

where *Bidder dummy*$_{i,t+1}$ is a dummy variable that equals 1 if a bank announced an acquisition in the subsequent year after the filing date, and 0 otherwise, *Target dummy*$_{i,t+1}$ is a dummy variable that equals 1 if a bank was identified as a target in the subsequent year after the filing date, and 0 otherwise, and $X_{i,t}$ denotes a vector of financial variables, frequently-used in the banking literature (Pasiouras et al., 2010).[3] A detailed description of our financial variables is included in Table 1. Finally, summary statistics for all variables are presented in Table 2.

Insert **Tables 1 & 2** here

3. Empirical results

Table 3 presents the results of the logistic regressions, where we examine the likelihood of a bank to become a bidder. In the first three columns, the dependent variable equals 1 if a bank became a bidder in the subsequent year after the filing date, and 0 if it became a target in the same year, or if it was not involved in a merger. In the last three columns, the dependent variable equals 1 if a bank became a bidder in the subsequent year after the filing date, and 0 if it was not involved in a merger.

Column 1 of Table 3 uses only *positive TF-IDF* as the predictor. The positive and statistically significant coefficient of this variable indicates that the positive sentiment of the 10-K filing is associated with higher probability of a bank becoming a bidder. Next, we repeat our analysis adding financial variables (columns 2 and 3). We run two separate logistic regressions, due to the fact that *Cost efficiency* and *ROA* are highly correlated. The findings suggest that larger, better-capitalized banks, with higher loan activity and lower loan loss provisions, are more likely to become bidders. In addition, higher efficiency and profitability translate to higher bidder's likelihood. The inclusion of the financial variables improves the Pseudo R^2 in both cases. The important thing in our analysis however, is that the coefficient of *positive TF-IDF* remains positive and highly statistically significant, even when we control

[3] All continuous variables are winsorized by year, at 1% and 99% level.

for the financial characteristics of the sampled banks. Finally, the results of the last three columns of Table 3 are qualitatively similar.

At this point, it is important to note that our textual variable is not only statistically, but also economically significant. In fact, the marginal effect of *Positive TF-IDF* is 1.971 (as estimated in column 2), and its standard deviation is 0.570. In addition, the mean of the bidding banks' dummy equals to 0.093. Therefore, a one-standard deviation increase in the percentage of positive words in a bank's annual report is associated with a 12.08% (1.971*0.570/0.093) higher probability that the bank will become a bidder in the subsequent year after the filing date.

Insert **Table 3** here

Table 4 reports the results of the logistic regressions, where we examine the likelihood of a bank to become target. We conduct two sets of regressions, in a similar spirit as in Table 3. Column 1 of Table 4 uses only *negative TF-IDF* as the predictor. Interestingly, we find that this variable enters the regression with a positive and statistically significant coefficient at the 5% level. This finding implies a positive relation between the negative sentiment of the 10-K filings and the banks' probability to be acquired. The Pseudo R^2 of this regression equals 2%. After including the financial variables (columns 2 and 3), we find that banks with lower non-interest income and loan loss provisions are also more likely to be acquired. In contrast with the bidding banks, lower efficiency and profitability are associated with higher likelihood of becoming a target. Notably, the inclusion of financial variables does not substantially improve the Pseudo R^2 (2.7% and 2.8% in columns 2 and 3, respectively). This result reflects the difficulty of accurately predicting a target firm, and it is consistent with the non-financial literature (Betton et al., 2008). However, *negative TF-IDF* is positive and statistically significant in all regressions, highlighting the importance of textual information in determining banks' acquisition probability. Columns 4 to 6 report similar results.

Strikingly, the economic significance of our textual variable is also high in the case of target banks. In particular, the marginal effect of *Negative TF-IDF* is 0.239 (as estimated in column 2) and its standard deviation equals to 2.277. Furthermore, the mean of the target banks' dummy equals to 0.041. Hence, a one-standard deviation increase in the fraction of negative words in the bank's annual report translates to a 13.27% (0.239*2.277/0.041) higher probability that the bank will become a target in the subsequent year after the filing date.

Insert **Table 4** here

4. Conclusion

In this paper, we perform textual analysis to identify potential merger participants in the U.S. banking industry. Our findings indicate that the sentiment of annual reports sheds light on a bank's likelihood to be bidder or target. In fact, we find that banks with a higher fraction of positive (negative) words in their annual report have a higher probability of becoming bidders (targets) in the subsequent year after the filing. Notably, this positive relationship holds even when we include several bank-specific financial variables in our logistic regressions. This finding highlights the importance of including textual variables in models that access the banks' acquisition likelihood.

As a concluding remark, we would say that there is still much to explore in this issue. For example, it would be interesting to explore text representations obtained from neural encoders (Goldberg, 2017), instead of TF-IDF features based on lists of positive and negative words. However, we hope that our study will provide fertile ground for a more in-depth examination on the role of textual analysis in determining the likelihood of bank acquisitions.

References

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.

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

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

Del Gaudio, B.L., Megaravalli, A.V., Sampagnaro, G., Verdoliva, V. 2019. Mandatory disclosure tone and bank risk-taking: Evidence from Europe. Economics Letters, forthcoming.

DeYoung, R., Evanoff, D.D., Molyneux, P., 2009. Mergers and acquisitions of financial institutions: A review of the post-2000 literature. Journal of Financial Services Research 36, 87-110.

Gandhi, P., Loughran, T., McDonald, B., 2019. Using annual report sentiment as a proxy for financial distress in U.S. banks. Journal of Behavioral Finance, forthcoming.

Goldberg, Y., 2017. Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers.

Hadlock, C., Houston, J., Ryngaert, M., 1999. The role of managerial incentives in bank acquisitions. Journal of Banking and Finance 23, 221-249.

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.

Kim, I., Skinner, D.J., 2012. Measuring securities litigation risk. Journal of Accounting and Economics 53, 290-310.

Leledakis, G.N., Pyrgiotakis, E.G., 2020. U.S. bank M&As in the post-Dodd–Frank Act era: do they create value? Journal of Banking and Finance, Forthcoming.

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

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.

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.

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.

Skinner, D.J., 1994. Why firms voluntarily disclose bad news. Journal of Accounting Research 32, 38-60.

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

Table 1

Financial variables definition

Variables	Description	Commercial Banks (Call Reports)	Bank Holding Companies (FR Y-9C)
LnSize	Logarithm of Total Assets	ln(RCFD2170)	ln(BHCK2170)
Capital Strength	Equity to Total Assets	RCFD3210/RCFD2170	BHCK3210/BHCK2170
Loan activity	Loans to Total Assets	RCFD2122/RCFD2170	BHCK2122/BHCK2170
Non-interest Income	Non-Interest Income to Total Income	RIAD4079/ (RIAD4074+RIAD4079)	BHCK4079/(BHCK4074+BHCK4079)
Loan loss provisions	Loan Loss Provisions to Loans	RIAD4230/RCFD2122	BHCK4230/BHCK2122
Cost efficiency	Non-Interest Expense to Total Income	RIAD4093/ (RIAD4074+RIAD4079)	BHCK4093/(BHCK4074+BHCK4079)
ROA	Net Income to Total Assets	RIAD4340/RCFD2170	BHCK4340/BHCK2170

Table 2

Summary statistics

This table reports summary statistics for the 8,068 bank-year observations of our sample

Variables	N	Mean	Median	Std. Dev.	Min	Max
Positive TF-IDF %	8,068	1.72	1.72	0.57	0.01	4.66
Negative TF-IDF %	8,068	6.23	6.25	2.28	0.01	17.09
LnSize	8,068	14.35	14.00	1.57	9.90	21.67
Capita strength %	8,068	9.42	9.08	2.90	−8.27	69.13
Loan activity %	8,068	66.90	68.03	11.85	4.65	96.21
Non-interest income %	8,068	22.50	20.59	13.61	−255.01	89.95
Loan loss provisions %	8,068	0.58	0.29	0.99	−2.36	15.73
Cost efficiency %	8,068	67.90	65.19	21.33	−35.61	754.03
ROA %	8,068	0.73	0.90	1.05	−16.19	7.73

Table 3

Logistic regressions of bidder dummy on textual sentiments and financial variables

This table illustrates the estimations from logit model by using textual and financial variables. In each column, bidder dummy is used as dependent variable which is a dummy variable that equals 1 if a bank announced an acquisition in the subsequent year after the filing date, and 0 otherwise. The first three columns report logistic regressions where the non-bidder sample includes both targets and non-merged banks, whereas the remaining three columns include only non-merged banks. Heteroskedasticity-robust *z*-statistics are reported in parentheses. Calendar year dummies and a constant are included without being presented. The symbols *, **, and *** denote statistical significance at the 0.10, 0.05 and 0.01 levels, respectively, using a 2-tail test.

	Both targets and non-merged			Non-merged only		
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Positive TF-IDF	0.179**	0.319***	0.324***	0.188***	0.325***	0.329***
	(2.53)	(4.45)	(4.50)	(2.61)	(4.47)	(4.50)
LnSize		0.415***	0.427***		0.416***	0.427***
		(14.38)	(15.75)		(14.21)	(15.55)
Capital strength		0.059***	0.052***		0.061***	0.054***
		(3.91)	(3.33)		(3.98)	(3.42)
Loan activity		0.016***	0.015***		0.016***	0.016***
		(4.56)	(4.48)		(4.60)	(4.53)
Non-interest income		−0.003	−0.009**		−0.003	−0.009**
		(−0.81)	(−2.49)		(−0.94)	(−2.51)
Loan loss provisions		−0.607***	−0.529***		−0.630***	−0.558***
		(−4.56)	(−4.01)		(−4.59)	(−4.10)
Cost efficiency		−0.011***			−0.010***	
		(−2.95)			(−2.78)	
ROA			0.340***			0.321***
			(3.76)			(3.48)
N	8,068	8,068	8,068	7,741	7,741	7,741
Pseudo R^2	0.035	0.102	0.103	0.036	0.103	0.104

Table 4

Logistic regressions of target dummy on textual sentiments and financial variables

This table illustrates the estimations from logit model by using textual and financial variables. In each column, target dummy is used as dependent variable which is a dummy variable that equals 1 if a bank was identified as a target in the subsequent year after the filing date, and 0 otherwise. The first three columns report logistic regressions where the non-target sample includes both bidders and non-merged banks, whereas the remaining three columns include only non-merged banks. Heteroskedasticity-robust *z*-statistics are reported in parentheses. Calendar year dummies and a constant are included without being presented. The symbols *, **, and *** denote statistical significance at the 0.10, 0.05 and 0.01 levels, respectively, using a 2-tail test.

	Both bidders and non-merged			Non-merged only		
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Negative TF-IDF	0.073**	0.070**	0.069**	0.070**	0.067**	0.066**
	(2.38)	(2.21)	(2.19)	(2.24)	(2.11)	(2.07)
LnSize		0.039	0.030		0.069	0.060
		(0.82)	(0.63)		(1.39)	(1.23)
Capital strength		−0.014	−0.008		−0.008	−0.002
		(−0.56)	(−0.33)		(−0.31)	(−0.09)
Loan activity		0.007	0.007		0.008	0.007
		(1.43)	(1.38)		(1.51)	(1.46)
Non-interest income		−0.017***	−0.012**		−0.018***	−0.013**
		(−2.87)	(−2.15)		(−3.00)	(−2.31)
Loan loss provisions		−0.204*	−0.329**		−0.241*	−0.365**
		(−1.66)	(−2.32)		(−1.89)	(−2.52)
Cost efficiency		0.009***			0.009***	
		(3.00)			(2.91)	
ROA			−0.270***			−0.272***
			(−3.19)			(−3.19)
N	8,068	8,068	8,068	7,317	7,317	7,317
Pseudo R^2	0.020	0.027	0.028	0.022	0.029	0.030