Finance Research Letters xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

An alternative approach to predicting bank credit risk in Europe with Google data

Marcos González-Fernández, Carmen González-Velasco*

Departamento de Dirección y Economía de la Empresa, Facultad de Ciencias Económicas y Empresariales, Universidad de León, Campus de Vegazana, s/n. 24071 León, España

ARTICLE INFO

Keywords:
Sentiment index
Google data
Credit risk
Credit default swaps

JEL Classification:
G10
G17
G40

ABSTRACT

The aim of this paper is to construct an alternative approach based on a sentiment index to measure bank credit risk in European countries using an alternative approach instead of traditional measures. Specifically, we use Google data for a set of keywords related to bank credit risk to capture investor sentiment. The resulting index shows a great similarity to traditional indexes based on bank CDS. The out-of-sample analysis demonstrates that our sentiment index is helpful for predicting bank credit risk during periods of financial distress, since it enhances the accuracy of the estimations.

1. Introduction and related literature

The financial crisis that started in 2008 showed the importance of the financial sector in the development of the economy. Citizens, companies and even sovereign states suffered the consequences of the financial turmoil. However, one positive lesson the crisis provided was that it served as a turning point in the management of risk in the banking sector. The evaluation of credit risk has become a crucial issue for financial markets. Since then, risks are more carefully evaluated and better proxied in multiple ways.

In this sense, there are two areas of the literature associated with bank credit risk assessment. The first area is related to financial and economic variables. Regarding this area, the LIBOR-overnight index swap spread (LIBOR-OIS spread) and the CDS spreads (Smales, 2016) are usually employed as measures of bank credit risk. The former is calculated as the LIBOR rate minus the overnight index swap (OIS) to remove markets expectations about central banks policies (Smales, 2016). Thus, the resultant LIBOR-OIS spread is a better proxy for bank credit risk that is usually employed in the literature during periods of financial turmoil (Aït-Sahalia et al., 2012; Black et al., 2016). On the other hand, the CDS spreads represents the periodic rate a protection seller receives from a buyer for transferring the risk of a credit event (Guesmi et al., 2018) and are commonly used in the literature as indicators of bank credit risk (Avino and Cotter, 2014). Along these traditional indicators, other authors suggest that credit risk can be modeled by macroeconomic factors (Yurdakul, 2014), since the weakening of the banks financial balance-sheet is also determined by macroeconomic factors in addition to bank specific variables (Cucinelli et al., 2018).

An alternative area is related to investor sentiment literature. In this sense, Smales (2016) determines that the CDS spread is significantly related to news sentiment. To this end, Smales constructs a sentiment variable that takes into account news related to banks, sorting the news into positive, negative or neutral. The results indicate that there is a significant relationship between

* Corresponding author.
E-mail address: carmen.gvelasco@unileon.es (C. González-Velasco).

https://doi.org/10.1016/j.frl.2019.08.029
Received 30 May 2019; Received in revised form 18 July 2019; Accepted 28 August 2019

sentiment and CDS spread, although this relationship is asymmetric. Hence, the surges in the CDS spread are higher for negative news than the reductions for positive news. Similarly, Beetsma et al. (2013) and Apergis (2015) analyze the effect of news on sovereign CDS for peripheral countries. Both studies determine that bad news contributes to increased sovereign risk. These results are in line with other researchers (Da et al., 2015; Tetlock, 2007) who state that negative events are more helpful than positive events in identifying sentiments and generating a larger investor attention.

In this paper, we follow this latter area of research by deepening in the analysis of market investor sentiment and its relationship to bank credit risk. Specifically, we use internet investor sentiment through Google data to create a sentiment index that allows us to measure bank credit risk. Google data have frequently used in the literature to proxy investor sentiment in stock markets (Ben-Rephael et al., 2017; Da et al., 2011; Da et al., 2015) and commodities markets (Han et al., 2017a, 2017b; Peri et al., 2014; Vozlyublennaia, 2014) as well as for other macroeconomic variables, such as unemployment (D'Amuri and Marcucci, 2017; Fondeur and Karamé, 2013; González-Fernández and González-Velasco, 2018). However, there is a lack of literature that addresses the relationships among investor sentiment, Google data, and bank credit risk. In this sense, this paper adds to the literature by filling this gap. The remainder of the paper unfolds as follows. Section 2 summarizes the data and the methodology. Section 3 shows the empirical analysis and the main results. Finally, the last section concludes the paper and offers some implications for the results obtained.

2. Data and empirical methodology

The aim of this paper is to construct a Google-based sentiment index to measure bank credit risk. Therefore, we first need to ensure that this index is functional by relating it to a well-established measure of credit risk. Following the literature (Alemany et al., 2015; Drago et al., 2017; Thornton and Tommaso, 2018), we proxy bank credit risk using CDS data. Namely, since the aim of our paper is to construct a Google-based sentiment index, we use a CDS index for European banks. Specifically, we use the Europe banks sector 5Y CDS Index. This index, constructed by Thomson Reuters Datastream, summarizes European bank CDS data representing an equally weighted portfolio of the evolution of the more liquid bank CDS, i.e., 5-year. Therefore, we assume that this index sensibly reflects bank credit risk.[1]

To approximate investor sentiment, we use the Google trends tool. Specifically, we use what is called Google search volume index (GSVI hereinafter) which is commonly used in the literature, since it is the web browser with a large number of users. The GSVI is constructed by dividing the number of searches s for a given keyword j ($V^s_{j,t}$) into a random sample of all searches ($V^s_{all,t}$) in the same period t. Hence, it is worth noting that the GSVI does not represent the total number of searches, but it is rescaled into an index that ranges between 0 and 100, with the latter representing the maximum level of attention for a given keyword Dergiades et al. (2015) as shown in Eq. (1):

$$GSVI_t = \frac{V^s_{j,t}}{V^s_{all,t}} \times \frac{100}{r^*} \tag{1}$$

Therefore, we obtain the GSVI for a set of keywords related to bank credit risk. First, we compile four primitive keywords which reflect bank credit risk. This first selection is arbitrary as in previous studies (Dzielinski, 2012; Smith, 2012; Vozlyublennaia, 2014, inter alia) but it intuitively reveals concerns about bank credit risk. These primitive keywords are *bank credit risk, bank crisis, banking crisis* and *European bank crisis*. Following Da et al. (2015), the next step is to refine this primitive list of keywords by focusing on Google top related searches, i.e., keywords that users who have searched for the primitive keywords are also interested in. The aim of this step is to expand the number of keywords and cover all the possible terms related to bank credit risk. This procedure provides a list of 74 keywords. Then, we retain only the keywords related to the goal of our paper[2] and we remove those which are duplicated. This step gives us a final list of 40 keywords, which are summarized on Panel A of Table 1.

Subsequently, we download the weekly data for the final keywords[3] from January 2008 to December 2017 (520 weeks) and winsorize the series at 5% (2.5 in each tail). The next step is to identify the specific keywords that better reflect investor sentiment to bank credit risk. For this aim, we run backward rolling regressions of the GSVI for our 40 keywords on the Europe banks sector 5Y CDS Index. We use a fixed rolling window of two years (104 weeks) for the regressions. Thus, the first regression includes the period January 2008 to December 2009, and then the window moves one week ahead in both extremes. This procedure results in 417 rolling regressions that allow us to determine the historical link (Da et al., 2015) between our keywords and bank credit risk. We select the keywords with positive slopes, i.e., those that lead to an increase in bank credit risk, with a significant t-statistic (Table 1, Panel B). The results indicate that there are six keywords that meet these requirements: *crisis loan, Deutsche bank crisis, European bank crisis, European banking crisis, European debt crisis* and *European sovereign debt crisis.* We construct the sentiment index by aggregating the GSVI for these keywords in a similar fashion as in Da et al. (2015) or Han et al. (2017b). Moreover, we also construct an index with

[1] Smales (2016) determines that his news sentiment index is related to CDS spreads as a measure of bank credit risk. Therefore, we assume that our Google-based sentiment index is also related to CDS data.

[2] For instance, within the top related searches for *bank crisis* there are several related searches about the nullification crisis that took place in the U.S. in the XIX century.

[3] It is worth noting that Google trends data can slightly differ depending on the day that the search is conducted, since Google uses a random sample in the denominator from Eq. (1) to increase response speed. To address this bias, we have collected the data on three different dates (March 14th, 15th and 16th, 2019) and averaged the data.

Table 1
Keywords sample.[a]

Panel A. List of 40 keywords in alphabetical order

2008 crisis, 2008 banking crisis, bank credit crisis, bank crisis, Bank of America crisis, bank rate, bank risk management, banking crisis, banking crisis definition, banking crisis UK, banking crisis US, banking regulation, central bank, counterparty credit risk, credit crisis, credit risk management, crisis loan, Cyprus banking, Cyprus banking crisis, Deutsche bank credit risk, economic crisis, European bank crisis, European banking crisis, European Central Bank, European debt crisis, European sovereign debt crisis, financial crisis, financial crisis 2008, global crisis, global banking crisis, global financial crisis, Irish banking crisis, interest rate risk, liquidity risk, risk management, the banking crisis, US bank crisis, what is credit risk and *World Bank crisis.*

Panel B. Keywords with a significant historical relationship to Europe banks sector 5Y CDS

Keywords in alphabetical order	Coefficient (Standard errors)	T-statistic
Crisis loan	2.499 (0.920)	2.71***
Deutsche bank credit risk	3.560 (1.420)	2.51**
European bank crisis	1.123 (0.492)	2.28**
European banking crisis	1.197 (0.496)	2.41**
European debt crisis	3.060 (0.897)	3.41***
European sovereign debt crisis	1.356 (0.474)	2.86***

The table shows the list of 40 keywords related to bank credit risk in Panel A, and the list of those keywords that show a larger average coefficient with a significant t-statistic in Panel B. Average robust standard errors are shown in parentheses.

** significant at the 0.05 level.

*** significant at the 0.01 level.

[a] All the searches have been performed using lowercase letters since we assume that people use nonstandard spelling when performing the searches and the most frequently bias involves the omission of capital letters (Lyddy, Farina, Hanney, Farrell, and Kelly O'Neill, 2014). There are no relevant differences in Google data depending on whether we use only lowercase letters or the right spelling. For grammatical purposes we use the right spelling along the paper.

the same keywords but without winsorizing the data and two other indexes using a principal components analysis (PCA) and retaining the first component (with and without previously winsorizing the data).

Table 2 shows the correlations between the sentiment indexes constructed and bank credit risk measured through the Europe banks 5Y CDS Index. As it can be observed, the correlations between banks CDS and the sentiment index range between 0.63 and 0.69, and all the correlations are highly significant. From the four sentiment indexes, the first index, which uses aggregated winsorized data, shows the higher correlation to bank CDS. This sentiment index is represented in Fig. 1 with the dots indicating the evolution without winsorizing the data. As shown, the evolution exhibits a similar pattern, especially during the maximum financial distress period, i.e., between 2008 and 2013. All the indexes show a very similar pattern among themselves with almost no differences, since their correlations are over 90% in all cases.

3. Empirical results

We run regression models to test whether the sentiment indexes constructed can be helpful for short and long term predictions for banks credit risk. We proxy banks credit risk through Europe banks sector 5Y CDS as in the previous sections. First, we perform an AR (1) model by regressing the bank sector index on itself lagged one period, which represents the benchmark model with which to compare the results. The coefficient for this benchmark is close to one, suggesting that it is a random walk process (Choi and Varian, 2012). Hence, ex-ante expectations are that the benchmark is the best forecast for bank credit risk. In subsequent regressions, we include the indexes constructed in Section 2.

Models 1 and 2 include the sentiment index constructed by aggregating the relevant keywords, and models 3 and 4 those constructed through PCA. The coefficients for the sentiment indexes are positive as assumed and significant in all the models, except for model 3. The R^2 and *RMSE* values slightly improve with the inclusion of the sentiment indexes. In light of the results from Table 3, the constructed Google-based sentiment indexes seem to have some ability to predict bank credit risk. To confirm this observation, we perform an out-of-sample evaluation analysis. For this purpose, we first estimate each model until the last week of December 2009, and then a one-step-ahead forecast is created for the next week. We repeat this procedure until the end of the sample. Thus, our out-

Table 2
Correlations between Europe banks sector 5Y CDS and Google-based sentiment indexes.

	(1)	(2)	(3)	(4)	(5)
1- Europe banks sector 5Y CDS	1.00				
2- Sentiment Index	0.69***	1.00			
3- Sentiment Index (no winsorization)	0.67***	0.99***	1.00		
4- Sentiment Index PCA	0.64***	0.95***	0.94***	1.00	
5- Sentiment Index PCA (no winsorization)	0.63***	0.93***	0.91***	0.99***	1.00

The table displays the pairwise correlations between the Europe banks sector 5Y CDS Index form Thomson Reuters Datastream and the Sentiment indexes constructed based on Google data.

***significant at the 0.01 level.

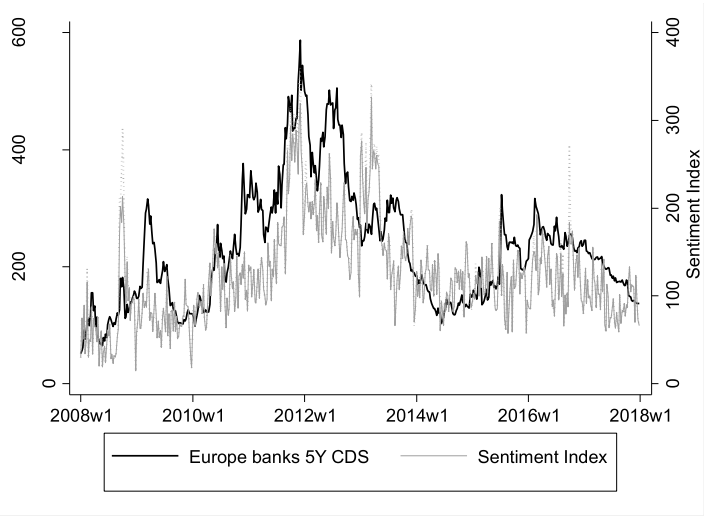

Fig. 1. Evolution of the sentiment index and Europe Banks CDS Index 5Y.

Table 3
Regression analysis.

	Benchmark	Model 1	Model 2	Model 3	Model 4
Europe banks sector 5Y CDS Index $_{t-1}$	0.980***	0.961***	0.961***	0.969***	0.966***
	(0.011)	(0.013)	(0.012)	(0.012)	(0.012)
Sentiment index$_t$		0.051**			
		(0.021)			
Sentiment index (no winsorization)$_t$			0.049**		
			(0.019)		
Sentiment index PCA$_t$				1.135	
				(0.692)	
Sentiment index PCA (no winsorization)$_t$					1.427*
					(0.729)
Constant	4.683**	2.666	2.778	7.223***	7.884***
	(2.321)	(2.503)	(2.472)	(2.672)	(2.677)
R^2	0.9652	0.9656	0.9657	0.9654	0.9655
RMSE	19.182	0.995	0.994	0.998	0.996

The table shows the regressions for the benchmark and the rest of the models. Models 1 and 2 represent the sentiment index constructed by aggregation of the relevant keywords. Models 3 and 4 are the sentiment index constructed by PCA. For the benchmark model, the actual RMSE is exhibited while, for the rest of the models, we report the ratio of the competitor's model and the benchmark. Therefore, whether the RMSE is below one, it indicates that the prediction improves the benchmark and vice versa.
*significant at the 0.10 level; **significant at the 0.05 level; ***significant at the 0.01 level.

of-sample data covers the period 2010 to 2017. Within this time span, we consider different subperiods to forecast two shorts samples and a long sample. Specifically, we consider the period of maximum financial distress from January 2010 to December to 2012 and a stabilization phase from January 2013 to December 2017 as short samples, and the entire time horizon (January 2010 to December 2017) as a long sample. Ex-ante expectations are for a better prediction of the sentiment index on bank credit risk in the short sample of high distress (January 2010 to December 2012), since previous research highlights that Google data better reflect investor sentiment in negative events.

Table 4 reports the results for the out-of-sample estimations. Regarding the entire time horizon, we can observe that the forecasts for the benchmark are slightly better than those including the sentiment indexes. However, the Diebold and Mariano (1995) test indicates that there are no significant differences between the forecasts. Hence, the inclusion of the sentiment indexes does not worsen the predictions. Similar results occur during the stabilization phase (January 2013 to December 2017) in which the benchmark outperforms the models with the sentiment indexes but with small or no significance. On the other hand, during the period of maximum financial distress, the models that include sentiment indexes outperform the benchmark[4] as expected. Moreover, these differences are significant for all the models indicating that the accuracy of the prediction significantly enhances, including our Google-based sentiment index.

[4] The results are robust for shorter time horizons within the period January 2010 to December 2012 in which bank credit risk has been more volatile.

Table 4
Out-of-sample forecast evaluation .

	Benchmark	Model 1	Model 2	Model 3	Model 4
RMSE (entire time horizons)	19.919	1.002	0.999	1.004	1.005
RMSE (January 2010-December 2012)	26.963	0.983***	0.983**	0.979***	0.979***
RMSE (January 2013-December 2017)	14.092	1.043	1.033	1.059*	1.059

The table shows the RMSE values for the out-of-sample estimates. We first run a regression until December 2009, and then, we generate a one-step-ahead forecast until the end of the sample for each model. Model 1 includes the sentiment index; model 2 includes the sentiment index (no winsorization); model 3 includes the sentiment index PCA and model 4 includes the sentiment index PCA (no winsorization). For the benchmark model, the RMSE is exhibited, while for the rest of the models, we report the ratio of the competitor's model and the benchmark. Therefore, whether the RMSE is below one, it indicates that the model's prediction improves the benchmark and vice versa.
***, ** and * indicate rejection at the 1%, 5% and 10% respectively of the Diebold and Mariano (1995) test under the null hypothesis that the forecast adequacy is equal between the benchmark and each of the models. Therefore, a rejection indicates that the forecast is significantly improved or worsened.

In summary, the presented results are in line with previous literature that underlines the ability of Google data to measure investor sentiment, especially during negative events and in the presence of bad news (Apergis, 2015; Beetsma et al., 2013; Smales, 2016). Here, we prove this behavior regarding bank credit risk. Thus, we confirm that our Google-based sentiment indexes improve the forecast in periods in which bank credit risk is facing complications, and their inclusion does not significantly worsen the results in other long and short samples.

4. Conclusions and implications

This paper studies the ability of an alternative approach based on a Google-based sentiment index to predict bank credit risk in Europe measured using European banks CDS data. In this sense, the contribution of this paper is twofold. First, we construct a sentiment index to measure investor sentiment towards bank credit risk using Google data. For this purpose, we select those Google searches that better reflect investor sentiment and use them to construct the index. Second, we test the ability of the constructed Google-based sentiment index to predict bank credit risk through regression analysis and out-of-sample procedures. The empirical results show that the sentiment index exhibits a high correlation to bank CDS, and its inclusion in the regression analysis slightly enhances the estimations. Moreover, the out-of-sample process indicates that, overall, the presence of the sentiment index does not significantly worsen the benchmark results, and on the other hand, it significantly enhances the forecast accuracy in periods of financial instability.

These empirical findings may have important implications for banks to assess credit risk, since Google data is easily available and has the advantage that it is more transparent than other alternatives for measuring investor sentiment (Da et al., 2015). Thus, these data can be used by banks and financial authorities to construct their own credit risk models. The inclusion of investor sentiment measures will allow banks and financial authorities to anticipate changes in bank credit risk, which should be especially noteworthy during episodes of turmoil.

Acknowledgments

We would like to thank the editor and the referees for providing us with constructive comments to improve the initial version of this paper. This work was supported by the Ministerio de Economía, Industria y Competitividad, Gobierno de España[research project number ECO2017-89715-P, entitled “El Análisis del Riesgo en los Mercados Financieros”].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.frl.2019.08.029.

References

Aït-Sahalia, Y., Andritzky, J., Jobst, A., Nowak, S., Tamirisa, N., 2012. Market response to policy initiatives during the global financial crisis. J. Int. Econ. 87 (1), 162–177.
Alemany, A., Ballester, L., González-Urteaga, A., 2015. Volatility spillovers in the European bank CDS market. Finance Res. Lett. 13, 137–147.
Apergis, N., 2015. Forecasting credit default swaps (CDSs) spreads with newswire messages: evidence from European countries under financial distress. Econ Lett 136, 92–94.
Avino, D., Cotter, J., 2014. Sovereign and bank CDS spreads: two sides of the same coin. J. Int. Financ. Mark. Institut. Money 32, 72–85.
Beetsma, R., Giuliodori, M., de Jong, F., Widijanto, D., 2013. Spread the news: the impact of news on the European sovereign bond markets during the crisis. J Int Money Finance 34, 83–101.
Ben-Rephael, A., Da, Z., Israelsen, R.D., 2017. It depends on where you search: institutional investor attention and underreaction to news. Rev. Financ. Stud. 30 (9), 3009–3047.
Black, L., Correa, R., Huang, X., Zhou, H., 2016. The systemic risk of European banks during the financial and sovereign debt crises. J. Bank Financ. 63, 107–125.
Choi, H., Varian, H., 2012. Predicting the present with Google trends. Econ. Record 88 (s1), 2–9.
Cucinelli, D., Battista, M.L.D., Marchese, M., Nieri, L., 2018. Credit risk in European banks: the bright side of the internal ratings based approach. J. Bank. Financ. 93,

213–229.
D'Amuri, F., Marcucci, J., 2017. The predictive power of Google searches in forecasting US unemployment. Int. J. Forecast. 33 (4), 801–816.
Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. J. Finance 66 (5), 1461–1499.
Da, Z., Engelberg, J., Gao, P., 2015. The sum of all fears investor sentiment and asset prices. Rev. Financ. Stud. 28 (1), 1–32.
Dergiades, T., Milas, C., Panagiotidis, T., 2015. Tweets, Google trends, and sovereign spreads in the GIIPS. Oxf. Econ. Pap. 67 (2), 406–432.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Statist. Jul 13 (3), 253–263.
Drago, D., Tommaso, C.Di, Thornton, J., 2017. What determines bank CDS spreads? Evidence from European and US banks. Finance Res. Lett. 22, 140–145.
Dzielinski, M., 2012. Measuring economic uncertainty and its impact on the stock market. Finance Res. Lett. 9 (3), 167–175.
Fondeur, Y., Karamé, F., 2013. Can Google data help predict French youth unemployment. Econ. Model. 30, 117–125.
González-Fernández, M., González-Velasco, C., 2018. Can Google econometrics predict unemployment? Evidence from Spain. Econ. Lett. 170, 42–45.
Guesmi, K., Dhaoui, A., Goutte, S., Abid, I., 2018. On the determinants of industry-CDS index spreads: evidence from a nonlinear setting. J. Int. Financ. Mark. Institut. Money 56, 233–254.
Han, L., Li, Z., Yin, L., 2017a. The effects of investor attention on commodity futures markets. J. Future. Mark. 37 (10), 1031–1049.
Han, L., Lv, Q., Yin, L., 2017b. Can investor attention predict oil prices? Energy Econ. 66, 547–558.
Lyddy, F., Farina, F., Hanney, J., Farrell, L., Kelly O'Neill, N., 2014. An analysis of language in university students' text messages. J. Comput.-Med. Commun. 19 (3), 546–561.
Peri, M., Vandone, D., Baldi, L., 2014. Internet, noise trading and commodity futures prices. Int. Rev. Econ. Finance 33, 82–89.
Smales, L.A., 2016. News sentiment and bank credit risk. J. Emp. Finance 38, 37–61.
Smith, G.P., 2012. Google internet search activity and volatility prediction in the market for foreign currency. Finance Res. Lett. 9 (2), 103–110.
Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. J. Finance 62 (3), 1139–1168.
Thornton, J., Tommaso, C.di, 2018. Credit default swaps and regulatory capital relief: evidence from European banks. Finance Res. Lett. 26, 255–260.
Vozlyublennaia, N., 2014. Investor attention, index performance, and return predictability. J Bank Financ 41, 17–35.
Yurdakul, F., 2014. Macroeconomic modelling of credit risk for banks. Procedia - Soc. Behav. Sci. 109, 784–793.