MONITORING MORTGAGE CREDIT RISK IN BANKING THROUGH SOCIAL MEDIA DATASET

Harish Kamath[1], Noor Firdoos Jahan[2]

[1]Research Scholar, ISBR Business School Bangalore
[2]Professor, R V Institute of Management, Bangalore

Email: [1]bharishkamath@gmail.com, [2]noor.firdoos@gmail.com

Received: 14 April 2020 Revised and Accepted: 8 August 2020

ABSTRACT: Credit risk is one of the critical risks, all the banks, and financial institutions. The challenge is the capability to predict the risk and the timing of such a prediction. The impact of any risk is higher when the probability of occurrence is not known ahead of time. Financial institutions can make informed decisions if they know of any possible delinquencies in that month or next. Often the classification of default occurs only after a payment is missed by the borrower. Continuous monitoring can help set the alert mechanisms, to indicate the possible defaults. Banks and financial institutes use the static dataset, which was produced at the loan disbursement time. Such data is stale and there have been so many events occurred, impacting the disposable income, in the life of a borrower. This study examines the possibility of using the Facebook posts, Likes and shares and classifies them as spend related or income-related and use that classification for finding out if this data can be used for predicting the possible default in the near term, say a month or two. Loan default is a function of willingness and capability. One's willingness and the capability to repay plays a larger role in the default and this study focus on identifying such indicators from Facebook. However, this study discounts the fake accounts and detection, analysis of such accounts.

KEYWORDS: Credit risk, Social media Machine learning, Classification, Sentiment analysis.

I. INTRODUCTION

Credit risk is a bigger risk for any financial institution, in the mortgage industry. Despite several set guidelines based on experience and mathematical models, the loan might have been lent, but there is an element of uncertainty over repayment or possibility of missing payment from the borrower, during the long loan servicing lifecycle (about 20-30 years). Two factors play a great role in repayment. "Willingness" and "Capability". Willingness starts coming down when the worth of the subject asset value comes down and there is no incentive for the borrower to payback. There might be other factors also which can influence the willingness. If a non-payment leads to no action or less severe action, indirectly people are incentivized to not to pay. Capability to repay may change over time just like willingness. The borrower might have lost the job or source of income. There is no way to detect such dynamic changes, in place in any credit risk models, implemented in financial institutes. Social media can capture the sentiments and possibly predict such risks. The current credit risk management systems hardly capture dynamic data. Often the default risk is raised to various levels only after they occur. In other words, it's not truly monitoring but report post the incident occurs. This paper checks the possibility of using dynamic data to assess and predicts the possibility of a borrower being defaulter. There is a huge cache of data that is available on social media. This paper looks at the elements that can give the clue about a possible default, in other words, the possibility of missing a payment in the coming month. This can be possibly achieved by understanding underlying sentiments in each exhibited social media activity and relate them to Capability and Willingness. The prediction can later be used with the right weights to build an improved machine learning model. Social media contains the dynamic and most current data points which can be used to analyse the lifestyle, status of the job, increased expenses, etc.

II. LITERATURE REVIEW

There has been a lot of work done in the area of credit risk management using the Social media data set, though not in this specific area. (Smales 2016) seeks to consider the relationship between the sentiment of newswire messages for a set of major international banks and changes in two important credit measures, namely the LIBOR-OIS spread and the CDS spread. It is found that there is a negative relationship between news sentiments and change in CDS spreads. (Vaidyanathan 2013) made a study using real-life scenarios. His study explains the assessment of Credit Risk with concepts like Business Evaluation, Financial evaluation, Credit Appraisal system, Credit Rating and Model with BASEL reference (The Basel committee 2006), Examines Classification of Non-performing assets (a.k.a. NPA). NPA is the impact part of the Credit Risk. It's an outcome of the risk occurring. The study emphasizes the need for Monitoring. The Monitoring Mechanisms are not elaborated in the Book. This is the gap, we want to address. The preventive actions described in the book are primarily on the Loan origination part. One of the best researches so far in this field is by Ntwiga (Ntwiga 2016), is a systematic analysis of Social network for credit risk management. This paper look extends the concept and attempts to understand if social media data can be used for monitoring the credit risk actively.

Analysis

Our method aims at predicting the possible default/delinquency during the loan servicing cycle of the loan lifecycle. At the origination lifecycle, the risk is already mitigated by verifying the standard set of documents verifying capability and willingness to repay the loan. During the servicing cycle, which is the largest timeline, varying up to 30 years. During this period, there could have been a lot of changes in the individual's life during the loan tenure, which may adversely impact the loan payment. It is important to understand them ahead of time, ideally before it happens. The method contains the following steps, First, the Loan information is collected from multiple financial institutes, with name, Loan status (delinquent or regular). Second, mine the Facebook posts of all the borrowers for whom, we have collected the loan details. Classification of sentiments is based on lexical analysis of the content. The context of the "likes" and "share" on Facebook are also captured and subjected to sentiment analysis.

The actual defaulted loans and the regular loan data was collected. The data contained the key data points, which could be used to link the Facebook account. This data was collected from multiple banks to ensure that the different geography, gender, range of loan amount are being covered.

For each of the identified borrowers, be its regular or default, the Facebook data on "like", "Share" and new "post" were grabbed using Fb data grabber tools. In the case of defaulters, the months before the actual non-re-payment of the loan were carefully studied. In case of re-payment resumed after months, even such data was carefully captured.

Data from Facebook had to be cleansed before its being used for machine learning following are the few techniques deployed in our case. Removing the Escaping HTML characters: At times, the HTML escaping characters are present in the dataset. Like "&", """ etc., these have no bearing on the outcome and must be standardized. Keeping the Encoding (like UTF 8) format of the data to standard encoding, that's useful for the research.

Removal of Stop-words using NLTK When data analysis needs to be data have driven at the word level, the commonly occurring words (stop-words) like space, tab or any such whitespaces, should be removed.

Apostrophe Lookup To avoid any word sense disambiguation in text, it is advised to maintain the right structure in it and to go by the rules of context-free grammar. When apostrophes are used, chances of disambiguation increase. Such changes don't have bearing on the outcome if cleansed

Removal of URLs Hyperlinks or URLs could be in a dataset like comments, reviews and posts should be removed for Lexical analysis, as they don't have any bearing on the analysis.

Removal of Expressions Textual data especially transcripts may contain human expressions like [laughing], [Crying], [chuckling], etc. These expressions are usually nonrelevant to the content of the speech for lexical analytics. So, these need to be removed to make a useful dataset.

Slangs lookup (removal or replacement) Social media comprises of most slang words. Such words need to be altered into the standard words list to make slang free text. The words like "Luv u sssooo much" need to be converted to "love you so much", "Helo baby" to "hello baby"

Split Attached Words data set originated from textual content by typing or copying from sources can result in conjoined words. As an example, I had good food today. In this sentence "Ihad" and "goodfood" are conjoined and such instances must be split properly.

At times the words are not in proper format or phrase. They could be cryptic compared to standard dictionary words, but the humans can make sense and derive sentiment. For example: "congs bro" congratulations brother. "Thanks, Sssooo very very very much" "Thanks So very much" etc. Any data which is not in the context could

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

just skew the research. A word by word analysis could lead to skewed research and possibly strengthen the biases. It's important to understand the posts or likes or shares in the right context.

Sometimes when we do the word by word analysis few words may contradict each other and can bring in a skew in personality profiling. Adding the context to the dataset brings a different perspective. It can help understand the behavioural traits and responses to various situations a person goes thru. Such analysis will have more affinity to the real world.

Sentiment Analysis is the most prevalent text classification algorithm. It analyses the underlying message and tells whether the sentiment expressed is positive, negative or neutral. With the R tool, the Facebook data is extracted based on the allowed permission sets. Sentiment analysis is a form of the classification algorithm. Using the R tool, we can do the sentiment Analysis on Facebook data (Bhargava and Rao 2018),(Kaushik and Mishra 2014). there are many case studies of this. Sentiment Analysis will help classify the data into negative or positive sentiment or neutral. A bag of words will be used for this classification. A list of keywords is entered into a text file as a Negative word set or positive word set. These two text files will be picked up for a set of words and any words matching from the files will be marked and accounted as positive or negative, based on the file where the word is matching

Classification (or supervised learning) methods are capable of mapping input vectors into one of the various preferred output classes through learning by examples. A classifier can be learned by computing the rough distance between input-output instances and correctly labelling outputs out of the training set. This procedure is named as the model generation stage. After generating the model, the resulting classifier can classify an unidentified example based on the learned classes in the training set. Various classification techniques are available. We have used the supervised learning approach as we already have the default status for each loan. For this research Support Vector Machines (SVM), KNN, LDA, Linear, Logistic regression, Decision tree, LDA and Naive Bayes algorithms are used for the classification. Following diagram (diagram 1) explains the approach used.

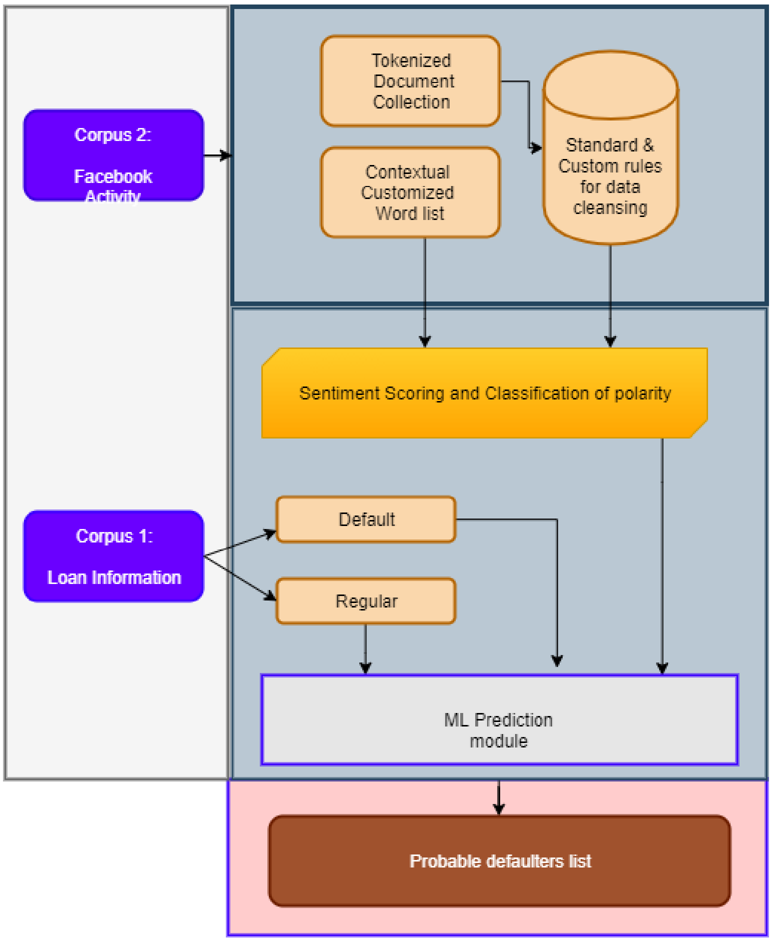

Diagram 1: Approach

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

III. DISCUSSION

Loan data for 1500 loan info was gathered. The following metrics summarized the loan data information. The genal trend seen is, about 2.87% of the loans are delinquent in the mortgage loan (TransUnion 2018) as of Q3 2018. As the target was only the Home loan, Only the 628 regular and 453 default loans were considered. Out of all the eligible loans, Facebook data had access restrictions, hence the numbers eligible loan numbers reduced further. Outliers were separated post-Facebook data gathering and few more were removed from the dataset.

Facebook mining tools, Natural Language Tool Kit (NLTK), sentiment analysis with lexicon (sample is shown below) was done using python code. For each “likes” the context of the “like” was mined and the classification of sentiment was done. Similarly, the post (original or shared) was mined sentiment analysis was performed. Following is the sample lexicon (partial). The classification is based on the impact on disposable income. If the disposable income is impacted by spending, it’s a negative sentiment. An increase in savings, that is classified as the positive sentiment. Following (refer to table 1) are the sample list of keywords /events used for sentiment polarization

Itemized indicators (Positive and Negative both are listed)
Family outing - very often
Addition of family members
Divorce
Marriage
Family functions
Newly found Love interest
Betting
Frequent to Gaming sites
The elevation is one’s Job (promotion)
House price is lesser than the purchase price
Salary Increase
Real estate market is not doing well
probable Natural calamity in the borrower’s locality
An increased expense due to Natural calamity
Transfer to other location within the country
Transfer to other location outside the country
Moving to a location which has a cheaper cost of living
probable Natural calamity in the borrower’s locality
An increased expense due to Natural calamity

Table 1: sample list of events considered for sentiment analysis

Following are the data points used as a part of the dataset

Loan Id: masked with a dummy number to mask the identity and the Personally identifiable data. This Id is created to refer to the original data, which can’t be shared publicly.

Spend Sentiments (Negative): is the ratio of total negative sentiments against the total (Positive + Neutral + Spend) sentiments accounted for Spend

Income Sentiments (Positive): is the ratio of Positive sentiments among the total sentiments detected.

Neutral Sentiments are those “posts”, “Likes”, “Share” which have no impact on disposable income.

Results from Sentiment Analysis: The following image shows the sample data from the social media dataset, post sentiment analysis tabulated and randomly picked from the loan dataset for illustration purpose. The following table (table 2) contains the sample data for illustration.

Id	s2iRatio	Spend	Income	Neutral	Posts	Likes	class
168	10.24	4.78	2.14	93.08	5	11	Default
17	9.2	3.4	2.71	93.89	3	1	Default
34	7.85	2.92	2.69	94.39	2	13	Default

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

45	2.01	1	2.02	96.98	2	2	Regular
51	1.94	0.96	2.01	97.02	3	6	Regular
53	1.82	1.8	1.01	97.19	1	7	Regular
55	1.72	1.72	1	97.28	1	3	Regular

Table 2: Sample list of tabulated data

The diagram here below(Diagram 2) Sentiment Analysis chart contrasting the Income and spend ratio segregated after the sentiment analysis.

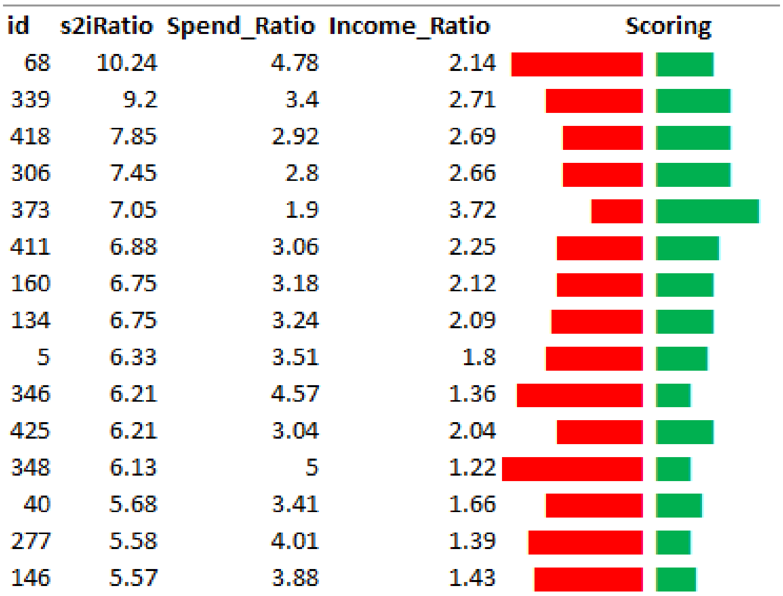

Diagram 2: Sentiment analysis score chart

Following diagram (diagram 3) explains the correlation between spending, income, Posts, Likes and shares with the loan status (defaulted and regular state). From the diagram, it is very clear by colours and numbers as to how each of the variables is correlated with other variables and what is the correlation found from the data set. From the image it is evident that there is a strong correlation between Income indicators to Loan status, Spend Indicators to Loan Status has a significant negative correlation. The polarization of sentiments has a greater role in the Loan default status.

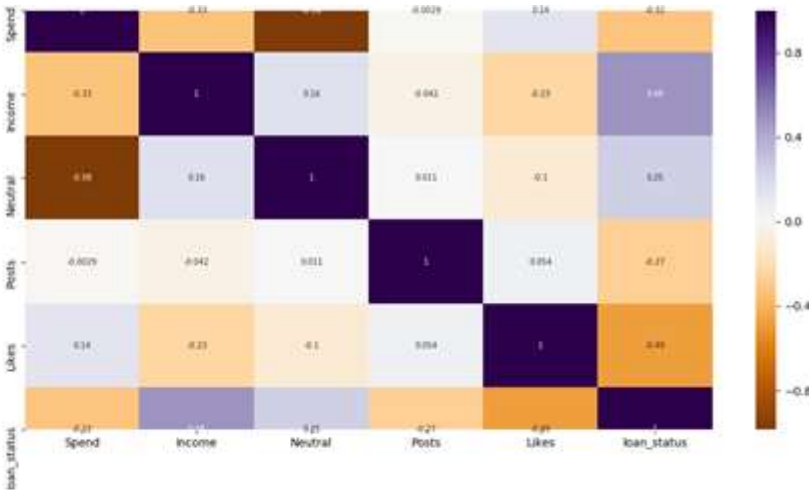

Diagram 3: Correlation chart

Various algorithms of Machine learning for classifiers were used for testing the prediction of the possible default. Using the supervised learning we classified each loan record as regular or defaulter. Multiple algorithms can be used in the machine learning and this research has tried the best-known algorithm to check if we can get a classification accurately. From the above table, it can be inferred that all the algorithms (Linear, Sigmoid, Polynomial Logistic Regression, KNN, LDA, Naïve Bayes) gives 100% accuracy but for the Decision Tree and the SVM. So, there is more than one machine learning algorithm, which can produce 100% accuracy for the given dataset. in other words, we have more than one algorithm, which can classify the defaults and regulars based on the polarity of the sentiments exhibited on Facebook.

Parameters	Linear	Sigmoid	Polynomial	Logistic Regression	Decision Tree	KNN	LDA	Naïve Bayes	SVM
Accuracy	100.00%	100.00%	100.00%	100.00%	99.00%	100.00%	100.00%	100.00%	77.94%
Precision	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	77.94%
F1 score	100.00%	100.00%	100.00%	100.00%	99.00%	100.00%	100.00%	100.00%	88.00%

Table 3: Accuracy matrix of prediction

Following diagram (diagram 3) explain the classification for positive sentiments into defaulter vs regulars.

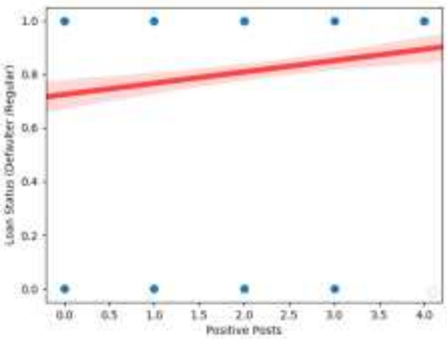

Diagram 3: classification of Positive posts into defaulter and regular

Accuracy of the Sentiment analysis plays a greater role in the success of this model. Hence the larger set of keywords for classifying the sentiments of the exhibited social behaviour into income-related or spend related. It also needs to be observed that income-related will be quite low often than the spend related. Hence The ratio may look like skewed but during the machine leaning this will be automatically taken care in the training part of the machine learning model. SO far the multiple sets of data have been tested and almost all have yielded similar results. Hence a larger dataset was attempted in the research.

IV. CONCLUSION

Based on the table 3, there is a clear indication that we can gather Facebook data for the borrowers and perform a sentiment analysis of the data in the context of spend sentiment or income-related sentiment or neutral sentiment. This dataset can then be subjected to any or many machine learning algorithms and predict the probability of a borrower becoming a defaulter.

This paper doesn't focus on video analysis or image analysis, but the text analysis. That is a possible subject for future research. Someone may make a fake post about self about high spending or huge income. Though there are algorithms available to detect such fakery, this mechanism has not been adopted into the paper. This is a possible application in the commercial adoption of this paper.

This has the potential to be used as a commercial application. The model shall be retrained after every cycle of prediction to improve accuracy. It acts as a feedback loop and improves the accuracy of the prediction.

V. REFERENCES

[1] Pekar, Viktor, and Jane Binner. 2018. "Forecasting Consumer Spending from Purchase Intentions Expressed on Social Media" In, 92–101, DOI: 10.18653/v1/W17-5212.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

[2] Schoen, Harald et al. 2013. “The Power of Prediction with Social Media.” Internet Research 23(5): 528–43.
[3] Tumasjan, Andranik and Sprenger, Timm O. and Sandner, Philipp and Welpe, Isabell M., Election Forecasts with Twitter - How 140 Characters Reflect the Political Landscape (January 1, 2011). Social Science Computer Review, Vol. 29, pp. 402-418.
[4] Culotta, Aron. 2010. “Towards Detecting Influenza Epidemics by Analyzing Twitter Messages.” In SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics,
[5] Arman Khadjeh Nassirtoussi, Saeed Reza Aghabozorgi, Ying Wah Teh, David Chek Ling Ngo: Text mining for market prediction: A systematic review. Expert Syst. Appl. 41(16): 7653-7670 (2014)
[6] Tan, Tianhui and Phan, Tuan Q., Social Media-Driven Credit Scoring: The Predictive Value of Social Structures (July 30, 2018). Available at SSRN: https://ssrn.com/abstract=3217885 or http://dx.doi.org/10.2139/ssrn.3217885
[7] Smales, Lee A. 2016. “News Sentiment and Bank Credit Risk.” Journal of Empirical Finance 38: 37–61
[8] Vaidyanathan. 2013. “Credit Risk Management for Indian Banks.” South Asia Economic Journal.
[9] The Basel Committe. 2006. “Principles for the Management of Credit Risk.” IFAS Extension.
[10] Ntwiga, Davis Bundi. 2016. “Social Network Analysis for Credit Risk Modeling.”
[11] Akaichi, Jalel, Zeineb Dhouioui, and Maria Jose Lopez-Huertas Perez. 2013. “Text Mining Facebook Status Updates for Sentiment Classification.” In 2013 17th International Conference on System Theory, Control and Computing, ICSTCC 2013; Joint Conference of SINTES 2013, SACCS 2013, SIMSIS 2013 - Proceedings,
[12] Zhao, Yanchang. 2013. R and Data Mining R and Data Mining.
[13] Pozzi, Federico Alberto, Elisabetta Fersini, Enza Messina, and Bing Liu. 2016. “Challenges of Sentiment Analysis in Social Networks: An Overview.” In Sentiment Analysis in Social Networks
[14] Kulkarni A., Shivananda A. (2019) Exploring and Processing Text Data. In: Natural Language Processing Recipes. Apress, Berkeley, CA
[15] Cortes, Alvaro, Christopher E. Herbert, Erin Wilson, and Elizabeth Clay. 2007. “Factors Affecting Hispanic Homeownership: A Review of the Literature.” Cityscape.
[16] Cortés, Kristle, Ran Duchin, and Denis Sosyura. 2016. “Clouded Judgment: The Role of Sentiment in Credit Origination.” Journal of Financial Economics 121(2): 392–413.
[17] Zheludev, Ilya, Robert Smith, and Tomaso Aste. 2014. “When Can Social Media Lead Financial Markets?” Scientific Reports.
[18] Wharton. 2014. “The Surprising Ways That Social Media Can Be Used for Credit Scoring.” Knowledge@Wharton.
[19] Ghailan, Omar, Hoda M.O., and Osman Hegazy. 2016. “Improving Credit Scorecard Modeling Through Applying Text Analysis.” International Journal of Advanced Computer Science and Applications.
[20] Saia, Roberto, and Salvatore Carta. 2016. 268 Lecture Notes in Business Information Processing An Entropy Based Algorithm for Credit Scoring.
[21] Education, Marwadi et al. 2014. “An Improved Sentiment Classification Using Lexicon into SVM.” Journal of Chemical Information and Modeling 53(1): 37–42.
[22] Wei, Liwei, Jianping Li, and Zhenyu Chen. 2007. “Credit Risk Evaluation Using Support Vector Machine with Mixture of Kernel.” In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),.
[23] Chen, Qing, Hui Feng Xue, and Li Yan. 2011. “Credit Risk Assessment Based on Potential Support Vector Machine.” In Proceedings - 2011 7th International Conference on Natural Computation, ICNC 2011,
[24] Kruppa, Jochen, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. 2013. “Consumer Credit Risk: Individual Probability Estimates Using Machine Learning.” Expert Systems with Applications.
[25] Dahiya, Shashi, S.S. Handa, and N.P. Singh. 2015. “Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set.” Industrija.
[26] Papouskova, Monika & Hájek, Petr. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. Decision Support Systems. 118. 10.1016/j.dss.2019.01.002.
[27] Bao, Wang, Ning Lianju, and Kong Yue. 2019. “Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment.” Expert Systems with Applications 128: 301–15. https://www.sciencedirect.com/science/article/pii/S0957417419301472 (September 15, 2019).
[28] Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. “Machine Learning Models and Bankruptcy Prediction.” Expert Systems with Applications.
[29] Bhargava, Mandava Geetha, and Duvvada Rajeswara Rao. 2018. “Sentimental Analysis on Social Media Data Using R Programming.” International Journal of Engineering and Technology(UAE).
[30] Kaushik, Chetan, and Atul Mishra. 2014. “A Scalable, Lexicon Based Technique for Sentiment Analysis.” International Journal in Foundations of Computer Science & Technology.