

Social Media Analytic-based Corporate Credit Rating Forecasting

Yuh-Jen Chen*, Professor.

Ling-Han Hong, Graduate Associate.

Yu-Chen Chen, Graduate Associate.

*Department of Accounting and Information Systems
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, ROC*

Email: yjchen@nkust.edu.tw

Tel: +886-7-3814526 ext. 16631

Fax: +886-7-6151302

Abstract—This study develops an approach to forecasting corporate credit ratings by analyzing public opinion toward corporations on social media to assist financial institutions in effectively evaluating and controlling corporate risk. This objective is achieved through the following steps: (i) designing a corporate credit rating forecasting process based on big data from social media, (ii) developing techniques for corporate credit rating forecasting, and (iii) implementing and evaluating the corporate credit rating forecasting mechanism.

The experimental results of this research show that the accuracy of corporate credit rating prediction based on social media big data is higher than that of traditional financial report, corporate governance and macroeconomic indicators. Moreover, the adopted forecasting model, K-Nearest Neighbor (KNN), is superior to the other machine learning models in terms of accuracy.

Keywords: Credit rating, Credit rating forecasting, Social media, Big data

I. INTRODUCTION

After the financial crisis of 2008, frequent business failure and default have resulted in considerable losses for investors and a large number of dead accounts for financial institutions. In response, financial institutions have prioritized assessing credit ratings for corporate loans. Previously, financial institutions determined corporate credit ratings on the basis of corporate financial and governance indicators. With advancements in the Internet and the popularity of social media, the appeal of enterprises on social media has become a relevant research topic. Social media represents an alternative method for financial institutions to determine corporate credit ratings. Therefore, the large amounts of data from social media can be used to effectively analyze and forecast corporate credit ratings in risk management departments of financial institutions in the field of financial technology (FinTech).

In view of the related literature review in Section II, financial institutions mostly used financial reports, corporate governance, and macroeconomics as feature indicators for forecasting corporate credit risk ratings. Moreover, there are

also a few studies that have tried to add social platform-related indicators, but the accuracy rate has not reached 80%. The main reason was that the collection of information from the community is not easy, resulting in insufficient data samples. Therefore, if the data extraction technique can be improved to obtain a sufficient amount of social platform data, and innovative text analysis methods can be proposed to extract features, then a more accurate corporate credit risk prediction model should be constructed.

Moreover, policy makers need to pay attention to the astounding number of errors in credit reports that are the result of misaligned economic and legal incentives. A considerable proportion of companies have a “potentially material error” in their credit file that makes them look riskier than they are. Lenders respond to this incorrect data by offering companies higher interest rates, less favorable terms, or denying credit if the error makes companies look too risky. Credit bureaus have little economic incentive to conduct proper disputes or improve their investigations.

Therefore, this study develops an approach to forecasting corporate credit ratings by analyzing public opinion toward corporations on social media to assist financial institutions in effectively evaluating and controlling corporate risk. This objective is achieved through the following steps: (i) designing a corporate credit rating forecasting process based on big data from social media, (ii) developing techniques for corporate credit rating forecasting, and (iii) implementing and evaluating the corporate credit rating forecasting mechanism.

The remainder of this paper is organized as follows. Section II presents the literature review related to the corporate credit rating and risk evaluation. Section III presents the corporate credit rating forecasting process based on big data from social media. Section IV describes the techniques involved in the corporate credit rating forecasting process. Section V presents an evaluation of the corporate credit rating forecasting mechanism. Section VI presents the conclusions and recommendations for subsequent studies.

II. RELATED WORK

Various studies on corporate credit rating and risk evaluation have been conducted. For example, Karminsky and Khromova (2016) constructed a reliable model based on public information for the practical usage of interested agents, regulators and banks themselves. In the study, a table of representative variables that have potential influence on ratings was constructed. The experimental results found that macroeconomic indicators increased the explanatory power of the model. The forecasting accuracy for rating agencies such as Fitch, Standard & Poor, and Moody reached 62%, and the deviation of more than 95% did not exceed one grade in the actual rating. The experimental data revealed that the predictive ability of the model was useful for banks and supervisory institutions. After the 2007-2008 crisis, corporate credit scoring is becoming a key role in credit risk management. Luo et al. (2017) investigated the performance of credit scoring models using credit default swap (CDS) dataset and compared the classification performance of deep learning algorithms (such as deep belief networks with restricted Boltzmann machines) with that of commonly used credit scoring models (including logistic regression, multi-layer perceptron, and support vector machine). The performance was assessed using the classification accuracy and the area under the receiver operating characteristic curve. The results found that DBN yields the best performance. Corporate credit ratings are widely used in financial services for risk management, investment, and financing decisions. The application of statistics and machine learning techniques to establish corporate credit ratings has been extensively studied. Hsu et al. (2017) proposed a bio-inspired computation-based classification model using the credit rating dataset of Compustat and adopted an artificial bee colony (ABC) approach and support vector machine (SVM) to enhance the forecasting of credit rating and credit rating changes. The experimental outcomes indicated that the proposed model provided improved prediction accuracy than other traditional statistical or soft-computing approaches, and it can be used for potential credit rating or change predicting. Pérez-Martín et al. (2018) believed that the volume of databases that financial companies manage is so great that it has become necessary to address this problem, and the solution to this can be found in big data techniques applied to massive financial datasets for segmenting risk groups. In the study, the presence of large datasets was approached through the development of some Monte Carlo experiments using known techniques and algorithms. Additionally, a linear mixed model (LMM) has been implemented as a new incremental contribution to calculate the credit risk of financial companies. These computational experiments were developed with several combinations of dataset sizes and forms to cover a wide variety of cases. Results revealed that large datasets need big data techniques and algorithms that yield faster and unbiased estimators.

Moreover, some studies that emphasized the comparison of various forecasting models in corporate credit rating and risk evaluation are also conducted. For example, Doumpos and Figueira (2019) used multicriteria outranking and ELECTRE TRI-NC to develop internal credit-rating models by using an

expert-based evaluation framework. The models were developed with multicriteria classification settings, and the results were analyzed on the basis of the internal attributes and deviations of risk rating categories as defined by rating agencies. Caridad et al. (2019) indicated that forecasting companies' long-term financial health was provided by credit rating agencies such as S&P, Moody's, Fitch group and others. Estimates of rates were based on publicly available data, and on the so-called 'qualitative information'. Nowadays, it is possible to produce quite precise forecasts for these ratings using economic and financial information that is available in financial databases, utilizing statistical models or, alternatively, artificial intelligence techniques. Forecasting credit ratings might be a suitable application of big data analytics. As machine learning is one of the foundations of intelligent big data analytics, Wallis et al. (2019) presented a comparative analysis of traditional statistical models and popular machine learning models for the prediction of Moody's long-term corporate debt ratings. Machine learning techniques such as artificial neural networks, support vector machines, and random forests generally outperformed their traditional counterparts in terms of both overall accuracy and the Kappa statistic. Moscatelli et al. (2020) analyzed the performance of a set of machine learning models in predicting default risk, using standard statistical models, such as the logistic regression, as a benchmark. When only a limited information set was available, for example in the case of an external assessment of credit risk, the study found that machine learning models provided substantial gains in discriminatory power and precision, relative to statistical models. This advantage diminished when confidential information, such as credit behavioral indicators, was also available, and it became negligible when the dataset was small. The study used a large dataset containing the financial ratios and credit behavior indicators of approximately 300,000 non-financial Italian companies from 2011 to 2017. The results revealed that machine learning models provided predictions with high discrimination and precision.

III. DESIGN OF A CORPORATE CREDIT RATING FORECASTING PROCESS

Figure 1 displays the corporate credit rating forecasting mechanism, which consists of two stages: the development and training stage and the implementation and adjustment stage. The development and training stage involve the surveying and preprocessing of public opinions, which are then coded using the name of the enterprise (e.g., Quanta) as well as a corresponding number (e.g., 2382) and an alias (e.g., Quanta Computer) for the enterprise. In the implementation and adjustment stage, dimensions are extracted from public opinion toward a target enterprise, and polarity is analyzed to create a repository for the dimensions and a repository for the public opinion content with sentence polarity. These two repositories and the TCRI from the TEJ database constitute the model. The results of the dimension extraction and polarity analysis for each target enterprise are used to forecast credit rating. Loan reviewers can increase the accuracy of the model by including an enterprise's repayment and default records.

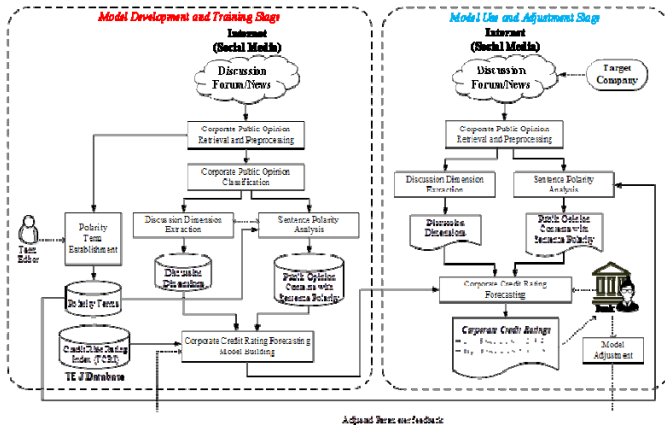


Figure 1. Corporate Credit Rating Forecasting Process based on Big Data from Social Media

IV. DEVELOPMENT OF TECHNIQUES FOR FORECASTING CORPORATE CREDIT RATINGS

Based on the forecasting mechanism designed in Section III, the relevant techniques involved in the mechanism are developed. They are corporate public opinion retrieval and preprocessing, corporate public opinion classification, polarity term establishment, discussion dimension extraction, sentence polarity analysis, and corporate credit rating forecasting model building and use, all of which are discussed in the following subsections.

A. Corporate Public Opinion Retrieval and Preprocessing

The web crawlers of well-known social media platforms in Taiwan, such as stock forums Mobile01 and PTT, and financial news outlets Udn Finance and ETtoday automatically collect information related to public opinion toward an enterprise from social media platforms based on the layout of their web pages. Figure 2 presents a web crawler in Pseudo Code. Public opinions are subjected to sentence and word segmentation, part-of-speech (POS) tagging, and stop word filtering in the Chinese Knowledge and Information Processing (CKIP) Chinese Parser (Ma & Chen, 2003) to acquire a meaningful POS tagged vocabulary set (Figure 3).

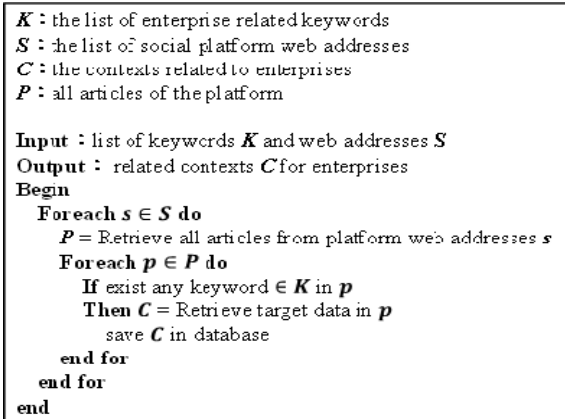


Figure 2. Algorithm for Retrieving Corporate Public Opinions from Online Social Media (Pseudo Code)

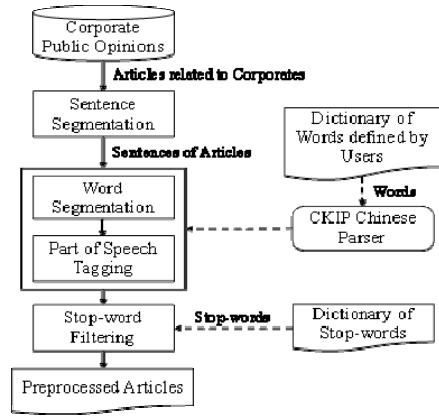


Figure 3. The Process for Preprocessing Corporate Public Opinions

B. Corporate Public Opinion Classification

To determine companies' credit risk rating for loans, the public opinions detailed in Section A are categorized in accordance with the framework for listed companies in Taiwan. The keywords used for classification are an enterprise's name, code, and industry. Figure 4 presents an example of the public opinion classification framework for the electronics industry, which consists of enterprise related to semiconductors, computers and peripheral equipment, optoelectronics, communication and internet, electronic parts/components, electronic product distribution, information service, other electronics, and semiconductor.

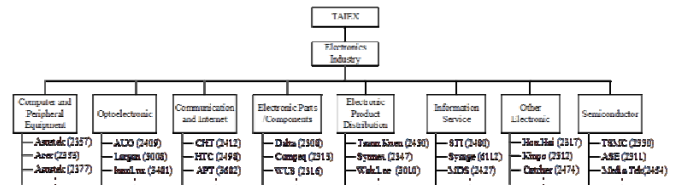


Figure 4. Corporate Public Opinion Classification Framework for the Electronics Industry (An Illustrative Example)

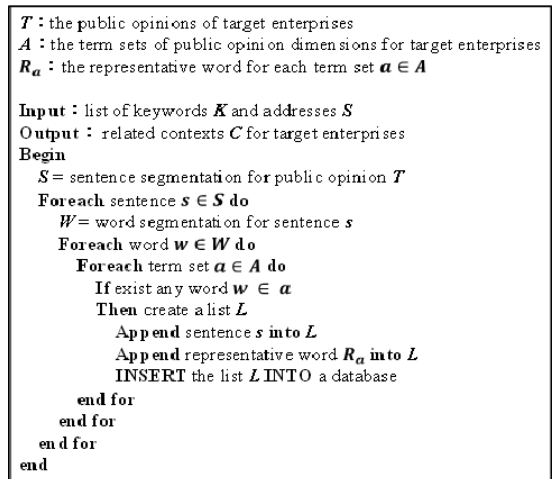


Figure 5. Algorithm for Classifying Corporate Public Opinions (Pseudo Code)

C. Polarity Term Establishment

The polarity of a sentence (i.e., positive or negative) conveying public opinion can be evaluated in terms of its nouns (N), verbs (V), and adjectives (ADJ), which are used to create a library of polarity terms. Nouns, verbs, and adjectives are selected from the sentences, and an editor manually identifies the polarity terms, which are then stored in the library (Figure 6). For example, “產能(N+)利用創新高(V+) (product capacity(N+) utilization hits record high(V+))” .

```

D : all the processed public opinions of target enterprises
P : the POS set ∈ { noun, verb, adjective }

Input : the public opinion D
Output : the polarity of term Tp and p ∈ P
Begin
  Foreach POS p ∈ P do
    Tp = build an empty list
    Sp = extract the list term of POS p from D
    Foreach term s ∈ Sp do
      If determine the s available is true
      THEN
        o = Determine the polarity of s by a term editor
        Append the tuple (s, o) into Tp
      end for
    end for
  end

```

Figure 6. Algorithm for Establishing Polarity Terms (Pseudo Code)

D. Discussion Dimension Extraction

This study uses the LDA (Latent Dirichlet Allocation) Gibbs sampling algorithm proposed by Darling (2011) to extract discussion dimensions from public opinions, such as investment technology, international situation, business management, stock market performance, and individual stock performance. Figure 7 presents the algorithm.

```

A : the public opinion articles of target enterprises
Na : the number of words in an article a ∈ A
K : the number of topics in articles
wi,a : the word w, where i ∈ Na in an article a ∈ A
na,k : the number of words assigned to a topic k ∈ K in a document
nk,w : the number of times words is assigned to a topic k ∈ K
nk : the total number of times any word is assigned to a topic k ∈ K

Input : all the words wi,a in an article a ∈ A
Output : topic assignments z and counts na,k, nk,w and nk
Begin
  randomly initialize z
  grouping all the words wi,a by an article a ∈ A
  Foreach the words wa in an article a ∈ A do
    For i = 0 to Na - 1 do
      word, topic = wi,a, zi
      na,topic, ntopic,word, ntopic += -1, -1, -1
    For k = 0 to K - 1 do
       $p(z = k | \cdot) = (n_{a,k} + \alpha_k) \frac{n_{k,w} + \beta_w}{n_k + \beta \times N_a}$ 
    end for
    topic = sample from p(z | ·)
    zi = topic
    na,topic, ntopic,word, ntopic += 1, 1, 1
  end for
end for
return z, na,k, nk,w and nk
end

```

Figure 7. Algorithm for LDA Gibbs Sampling (Pseudo Code)

E. Sentence Polarity Analysis

After public opinion content preprocessing (described in Section A) and the identification of polarity terms (described in Section C), the POS combination-based polarity analysis method and POS-combination rules (7514 pieces; Table 1 lists the partial rules) proposed by Chen et al. (2018) are used for sentence polarity analysis (i.e., positive or negative evaluation) to facilitate the establishment and use of the model. Figure 8 presents the process of sentence polarity evaluation using the POS combination. A preprocessed sentence is analyzed for terms in the polarity term library. If a term matches a term in the library, then the POS combination for the sentence is paired with the POS-combination rules. If a POS-combination rule for sentence polarity analysis is observed in a sentence, the polarity of the POS-combination rule is considered the polarity of the sentence; otherwise, the polarity term in the sentence is considered the polarity of the sentence.

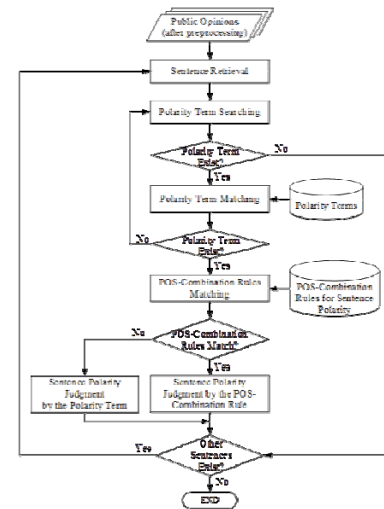


Figure 8. Sentence Polarity Judgment Process

Table 1. Partial POS-Combination Rules for Sentence Polarity Judgment

POS-Combination Rules	Sentence Polarity Judgement
(N)(ADV1)(Vi)(nn)	Positive
(ADV1)(Vi)(N)(nn)	Positive
(A)(ADV)(Vi)(nn)	Negative
(ADV)(Vi)(Vi)(nn)	Negative

F. Corporate Credit Rating Forecasting Model Establishment and Use

According to the dimension and polarity of sentences acquired from the above-mentioned public opinion analysis, five public opinion indicators are identified for model training: the percentages of positive sentences regarding investment technology, business management, stock market performance, individual stock performance, and the percentage of discussion volume regarding an enterprise. Equations (1) and (2) present the calculation steps. Corporate credit ratings retrieved from the TEJ database are used as the labels for the public opinion indicators. These indicators and the labeled

data sample are divided into training and testing datasets on the basis of proportion. The k-nearest neighbor (KNN) algorithm (Guo et al., 2003; Harrou et al., 2020; Moldagulova & Sulaiman, 2017; Xing & Bei, 2019; Zhang & Zhou, 2007) is used as the forecasting model. KNN is a non-parametric and lazy learning algorithm that does not require learning sample data; instead, training data are used as a reference for decision boundaries. Thus, the model uses only a training dataset for testing and adjustment. Figure 9 presents the KNN algorithm, in which the Euclidean distance is employed in the distance calculation formula, as shown in Eq. (3).

$$\begin{aligned} & \text{Percentage of Positive Reviews for the Dimension} \\ &= \frac{\text{Number of Positive Sentences for the Dimension}}{\text{Total Number of Sentences for the Dimension}} \end{aligned} \quad (1)$$

$$\begin{aligned} & \text{Percentage of Discussion Volume for the Corporate} \\ &= \frac{\text{Total Number of Sentences for the Corporate}}{\text{Number of Sentences for all Corporates in the Industry}} \end{aligned} \quad (2)$$

$$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

where $D(x, y)$ denotes the distance between data x and y ;
 x denotes the data vector including $x_i, \forall i = 0, \dots, n$;
 y denotes the neighbor data vector including $y_i, \forall i = 0, \dots, n$;

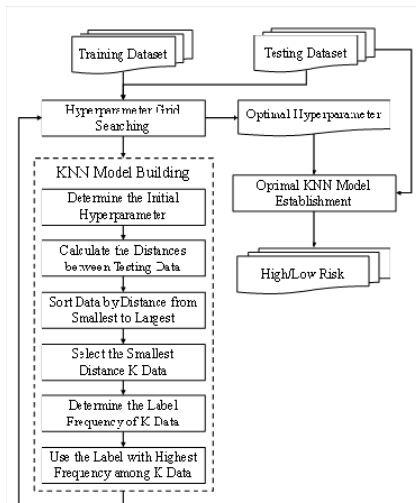


Figure 9. Algorithm for K-Nearest Neighbor (KNN)

V. IMPLEMENTATION AND EVALUATION OF A CORPORATE CREDIT RATING FORECASTING MECHANISM

Programming languages are used to implement the techniques for corporate credit rating forecasting and to determine the feasibility of the proposed approach. Forecasting accuracy is evaluated by comparing the effectiveness of the models used in this study with that of other forecasting models.

The system implementation environment comprises the Windows Server 2019 operating system, the MySQL database system (version 8.0.23), the Visual Studio Code compiler

(version 1.57.1), and the Python programming language (version 3.8.5). The hardware comprises an Intel Xeon Silver 4210 CPU with 2.20 GHz and 128 GB of RAM. The third-party service comprises Google Search and the Scikit-Learn forecasting model suite.

A total of 8,093,440 data points for public opinions (January 1, 2016~December 21, 2020) are compiled from well-known online forums, discussion boards, and news outlets in Taiwan, and the credit ratings of 540 companies listed in Taiwan are selected from the TCRI for the model. The 11,427 data points used in this study are divided into 8,790 training data points and 2,637 testing data points, with a 3:1 ratio, to train and test the KNN. The KNN model adopted in this study is used to evaluate and compare the effectiveness of forecasting corporate credit ratings with financial report/corporate governance/macro-economics indicators and social media data. Moreover, the corporate credit rating in this study focuses on corporate operational risks (i.e., high or low risk), corporate credit ratings are thus divided into two levels of high and low, so that financial institutions can understand the dynamic changes of the company in the market, including growth, stability or decline and factors such as business cycle, which are used as the reference basis for determining the degree of operational risk faced by the company. Table 2 presents the experimental results in a confusion matrix. The results indicate that the accuracy of corporate credit rating forecasting based on big data from social media is higher than the accuracy of forecasting made on the basis of financial report/corporate governance/macro-economics indicators and social media data discussed in the literature review. In practice, financial institutions will not actually approve any loans from companies with low credit ratings. Conversely, any company with a non-low credit rating (such as medium credit rating) may be safe and can obtain a financial loan.

KNN			
Corporate Credit Rating Forecasting		Predictive Values	
KNN		High	Low
Actual Values	High	1268	52
	Low	315	1002
Accuracy (%)		86.08%	

Table 2. Confusion Matrix for Testing the Corporate Credit Rating Forecasting Model

The sample is used to compare the accuracies of eight classification models, including Naive Bayes, Random Forest, XGBoost, SVM, GA-SVM, QGA-SVM, DNN, and KNN. Their performances are presented in Table 3.

Forecasting Models	Performance Measures			
	Accuracy	Precision	Recall	F1-score
Naive Bayes	48.24%	19.09%	40.00%	25.84%
Random Forest	77.02%	74.32%	82.18%	78.24%
XGBoost	48.24%	19.09%	40.00%	25.84%
SVM	75.54%	80.25%	74.28%	77.15%
GA-based SVM	77.28%	81.75%	74.52%	77.97%
QGA-based SVM	78.88%	77.79%	80.91%	79.32%
DNN	70.46%	72.01%	71.94%	72.06%
KNN	80.17%	86.56%	71.24%	78.12%

Table 4. Performance Measures for the Forecasting Models

VI. CONCLUSION AND FUTURE RESEARCH

In the past, most of the corporate credit rating forecasting mechanisms provide only the analysis of structured financial performance information, and none tackles the issues of the textual analysis of unstructured non-financial performance information, such as corporate public opinion. In this study, we proposed novel methods for forecasting corporate credit ratings using corporate public opinions on social media. Thus, this study first designed the process for forecasting corporate credit ratings based on big data from social media, and then developed the related techniques involved in the process. Finally, the mechanism for forecasting corporate credit ratings based on big data from social media was implemented based on the techniques.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Science and Technology R.O.C, Taiwan, for financially supporting this research under Contract No. MOST109-2410-H-992-019. Additionally, we deeply appreciate the editor, associate editor and reviewers for their constructive comments and suggestions on the work.

REFERENCES

- [1] Caridad, D., Hančlová, J., Bousselmi, H. W., & López del Río, L. C. (2019). Corporate rating forecasting using artificial intelligence statistical techniques. *Investment Management & Financial Innovations*, 16(2), 295-312.
- [2] Chen, Y. J., Chen, Y. M., Kao, S. C., & Wu, J. H. (2018). Development of a technology for part of speech combination supported Chinese eWOM analysis. *International Journal of Computer & Software Engineering*, 3(2).
- [3] Doumpos, M., & Figueira, J. R. (2019). A multicriteria outranking approach for modeling corporate credit ratings: an application of the ELECTRE TRI-NC method. *Omega*, 82, 166-180.
- [4] Guo, G., Wang, H., Bell, D. A., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, 986-996.
- [5] Harrou, F., Zeroual, A., & Sun, Y. (2020). Traffic congestion monitoring using an improved kNN strategy. *Measurement*, 156, 107534.
- [6] Hsu, F. J., Chen, M. Y., & Chen, Y. C. (2017). The human-like intelligence with bio-inspired computing approach for credit ratings prediction. *Neurocomputing*, 279, 11-18.
- [7] Karminsky, A. M., & Khromova, E. (2016). Extended modeling of banks' credit ratings. *Procedia Computer Science*, 91, 201-210.
- [8] Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65, 465-470.
- [9] Moldagulova, A., & Sulaiman, R. B. (2017). Using KNN algorithm for classification of textual documents. In *2017 IEEE eighth international conference on information technology (ICIT)* (pp. 665-671). IEEE.
- [10] Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161, 113567.
- [11] Pérez-Martín, A., Pérez-Torregrosa, A., & Vaca, M. (2018). Big data techniques to measure credit banking risk in home equity loans. *Journal of Business Research*, 89, 448-454.
- [12] Wallis, M., Kumar, K., & Gepp, A. (2019). Credit rating forecasting using machine learning techniques. *Managerial perspectives on intelligent big data analytics*, 180-198.
- [13] Xing, W., & Bei, Y. (2019). Medical health big data classification based on KNN classification algorithm. *IEEE Access*, 8, 28808-28819.
- [14] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7), 2038-2048.