Credit Scoring Model Construction Based On LinkedIn Social Media Data

Dian Puteri Ramadhani, Putri Mentari Wijaya and Andry Alamsyah
School of Economic and Business
Telkom University
Bandung, Indonesia
dianpramadhani@telkomuniversity.ac.id, putrimntrw@student.telkomuniversity.ac.id, andrya@telkomuniversity.ac.id

Abstract

In the credit acceptance process, the financial institutions analyze the borrowers' creditworthiness through their demographic data based on the 5C principle; character, capacity, conditions, capital, and collateral. However, the legacy credit scoring methods have drawbacks, including not having an excellent credit reputation as it is limited to the structural nature of demographic data. We construct a credit scoring model by combining the demographic element and adding two social media elements; content and network. The content considers creditworthiness by assessing borrowers' posts, which consist of opinions and conversations on social media. In comparison, the network considers borrowers' connectivity to their social community. The paper proposes a new credit scoring model better to represent the quality of borrowers' characteristics and behavior. The data is collected from LinkedIn, which is suitable to represent the professional network. The proposed model has been verified through expert judgment, including the credit providers, and has been simulated through a machine learning approach to automate credit acceptance decisions.

Keywords
Credit Scoring, Expert Judgment. Prediction Analysis, Creditworthiness

1. Introduction
Financial institutions previously measured creditworthiness based on financial history, socioeconomic data, and demographics (Niu et al. 2019). Credit scoring is generally carried out with 5C analysis, consisting of character, capacity, capital, condition, and collateral. Financial institutions use the 5C analysis principle to determine the borrower's ability and intention to return the credit received and avoid non-performing loans when the borrower's condition does not fulfil the requirements (Basori and Wahyuningsih 2018). Due to financing problems, the legacy credit scoring method is considered not optimal enough (Eprianti 2019). Other drawbacks in credit scoring using legacy methods include customers' poor credit history, outdated data, and data availability (Guo et al. 2016).

On the other hand, nowadays, we can employ social information derived from social media; as explained by Zhang et al. (2016), social information is suitable for describing the borrower's behavior characteristics. One of the suitable social media is LinkedIn. Its proved by Garg et al. (2016), we can reveal the character and behavior of LinkedIn users by analyzing profile information data and LinkedIn users' friendship networks. Moreover, LinkedIn also provides user conversation data. Thus, LinkedIn social media data is suitable for describing the user characteristics regarding demographics, conversations, and relationships. Finally, the social media data generated by users might be an alternative data source for conducting creditworthiness evaluations (Guo et al. 2016).

The eligibility of prospective borrowers to obtain credit from financial institutions can be analyzed using social media data. This study analyzes three 5C credit scoring principles; character, capacity, and conditions. The other two analytical instruments, capital and collateral, are considered private and not publicly available. The three principles used to consist of character, capacity, and conditions. Each is represented in the user profile account features, including the demographic, conversation, and relationship (Guo et al. 2016). Social media data is heterogeneous and contain much irrelevant information. For this reason, we need a specific way to handle and transform the data to evaluate the creditworthiness based on character, capacity, and condition principles (Guo et al. 2016).

The challenges posed by using social media data as a new data source for credit scoring might be overcome by using expert judgment methods and appropriate data processing, such as data mining. The expert judgment method is used when the financial institutions do not have high-quality data or do not have data that records sufficient bad debts for statistical development. Even with high-quality data, no financial institutions should rely solely on statistical models in lending decisions because some loans cannot rely solely on modelling procedures. So providing bank loans requires further analysis and expert judgment is vital (Saardchom 2012). Meanwhile, data processing uses data mining methods can find regular patterns in the data in an automatic or semi-automatic way from a data set that allows for making predictions (Ledhem 2021). The use of data mining in this study uses a prediction analysis model in the form of a classification method is feasible or not and utilizes a decision tree algorithm to get the final result so that this technique produces variables related to certain target variables (Syed Nor et al. 2019). The decision tree algorithm is considered a competitive alternative in credit scoring, and decision trees can provide an elegant way to rank the most significant attributes in credit scoring (Bastos 2008).

Previous research revealed that credit scoring analysis conducted by financial institutions to measure borrowers' creditworthiness using the 5C principle and primary data still has many drawbacks. There are conditions where the borrower fails to fulfil the requirements, such as bad credit history. We propose using social information to analyze the character and behaviour of borrowers through social media. Acknowledged one study revealed that several principles of 5C, such as character, capacity, and conditions, could be represented through social information. Social information can be extracted from professional platforms through demographic, conversation, and relationship data. However, the data are heterogeneous and do not explicitly explain the creditworthiness of borrowers. Therefore, we use the expert judgment method and the decision tree algorithm in the data processing. Ultimately, the credit scoring model was built based on demographic, conversation, and relationship data using data sources from LinkedIn social media. The model could be adapted to the financial institutions' specific needs, where each prospective borrower is analyzed in depth using the proposed method.

1.1 Objectives

The main goal of this study is to propose a credit scoring model to determine a person's eligibility to receive credit by using social information from social media LinkedIn. The findings in this study may be helpful for financial institutions to produce proper credit scoring models. Thus, they can also serve borrowers with bad credit history by identifying the character and behavior of borrowers on social media.

2. Literature Review

Credit is the ability to carry out a purchase or a loan with the assurance that the borrowers will defer the payment for an agreed period (Syaifuddin 2019). In offering credit, financial institutions conduct credit scoring to increase approval rates and reduce credit losses at a meagre cost (Guo et al. 2016). Then credit scoring is a statistical analysis conducted by a financial institution or credit bureau to evaluate a borrower's creditworthiness and is based on credit history, credit behavior, and demographic data (Niu et al. 2019). Credit scoring usually uses the 5C principle, which consists of character, capacity, condition, capital, and collateral (Mardhotillah 2020). There is a numerical score in the information used in credit scoring to assist financial institutions in lending decisions to borrowers (Goel and Rastogi 2021).

Information from social media can be managed and extracted for business. He et al. (2017) pointed out that social media is an entity consisting of technologies, practices, or online communities used to generate content or share insight, opinions, experiences, and perspectives. Social media users can share demographic data and conversations and build relationships through social media. Demography is a group of demographic variables such as age, gender, number of family members, family life cycle, income, occupation, education, religion, race, and nationality (Kottler and Keller 2012). User-generated content comes from people who voluntarily contribute data, information, or media that appear helpful or entertaining and has two types: precisely textual data and online photo data. Moreover, there are two types of relationship data that users can decide in their social media, namely public or semi-public profile, that provides service for individuals to share and connect to the system (Perangin-angin and Zainal 2018).

Several studies that have been conducted using many methods and data have been tested. Yu et al. (2020) researched data cleaning collected from social media. Research about credit scoring conducted by Saardchom (2012) creates alternative scoring models, where a professional opinion is required for model validation. Therefore, expert judgment is a technique in which judgment is provided based upon a specific set of criteria and expertise acquired in a specific knowledge area, application area, product area, a particular discipline, or industry. The expert judgment could be any group or person who may provide such expertise with specialized education, knowledge, skill, experience, or training (Sotille 2018). The data mining method is famous for discovering the hidden, buried pattern under large-scale data and providing new meaningful information (Tyas et al. 2021). Bastos (2008) has used data mining using a decision tree algorithm. The decision tree algorithm is an inverted tree structure where each branch shows the test result (Zhang et al. 2016). Decision trees have advantages in representing alternative graphical decisions efficiently that can be expressed quickly and clearly (Teles et al. 2020). From the research by several researchers, we conclude that social media data can be employed to become something valuable with proper data processing.

Based on research developments through literature review, currently, financial institutions can rely on various kinds of data and do not rely on primary data only in conducting credit scoring. The innovations made by several researchers, especially in credit, open up new alternatives for borrowers with limited access to credit. Currently, data on social media can be processed and produce outputs that are useful for financial institutions in determining the borrower's creditworthiness. Therefore, it is crucial in the selection and processing of relevant social information in the process of developing a credit scoring model.

3. Methods

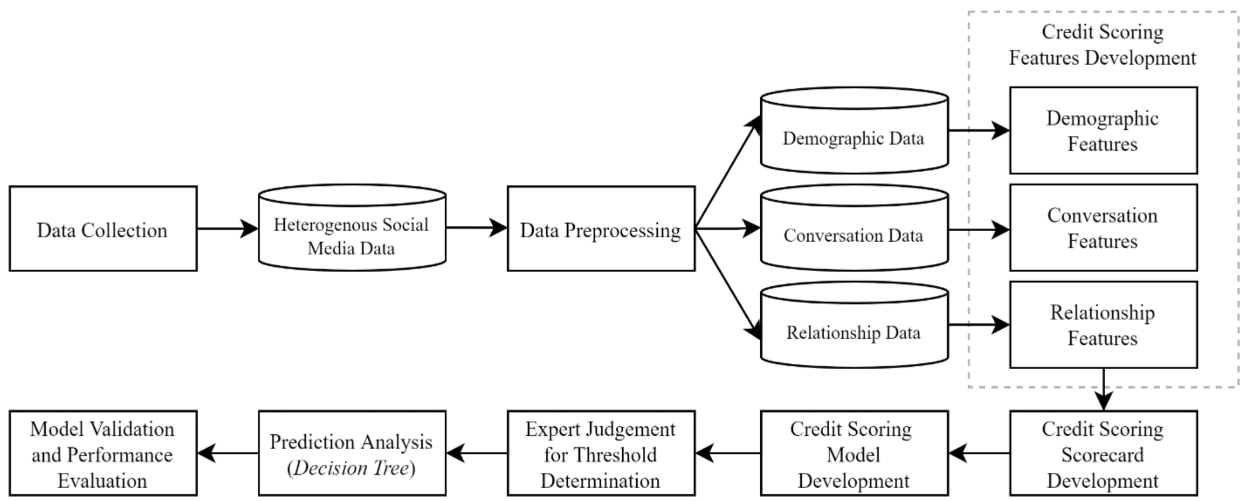

Figure 1. Research Methods

Figure 1 shows the study workflow in several steps. The steps are explained below.

3.1 Data Collection

We collect the LinkedIn social media data through the scrapping method using Phantombuster web scrapping service. We utilize three Phantombuster web scrapping features. First, we use LinkedIn Profile Scrapper to collect user profile data for the demographic feature. Next, we use LinkedIn Followers Insight to collect friendship connection data from LinkedIn users for the relationship feature. The last one is LinkedIn Activities Extractor to collect users' activity data for the conversation feature. After finishing the data collection stage, we have the heterogenous raw social media data.

3.2 Data Pre-processing

We need to preprocess the heterogeneous social media data. The data is preprocessed in several stages. Generally, we cleaned the irrelevant data and handled the empty data. The conversation data is still in text form. We analyze the textual data using LIWC software to obtain the analytical thinking, clout, authenticity, and emotional tone values. Attributes in conversation data have several measured values, which are described in Table 1. We also measure the value contained in conversation data, including the ratio of like to post, the ratio of comment to post, and the number of posts to describe a person's mental health (Sherman et al. 2016) and affect one's loneliness (Deters and Mehl 2013). The number of posts also is mentioned to describe a person. Good credit users tend to grant for content rather than consume content on social media (Guo et al. 2016).

Table 1. LIWC Summary Measures

LIWC Measurement	Definition	Post Example
Analytical Thinking	Measure a person's formal and logical thinking based on the writing.	"Alhamdulillah Telkom University in 2021 successfully entered the QS World University Ranking by Subject Ranking: *Computer Science and Information System* for the first time. Ranked 551-600 in the world and ranked 6th nationally. Thank you for the contribution of all academicians, alumni, and stakeholders Tel-U 🙏🏻 🙏🏻 🙏🏻 "
Clout	Represent social status, self-confidence, or leadership based on the writing.	"It is an honor to be a partner of PT Mirae Asset Sekuritas Indonesia in implementing Salesforce #salesforce #enjoytransformation."
Authenticity	Measure someone expressing themselves genuinely or honestly based on the writing.	"I'm trying to explain my understanding of the quicksort algorithm in this medium article: https://lnkd.in/g_TW2Vv"
Emotional Tone	Measure negative or positive emotional tendencies based on the writing.	"Passion is the umpteenth key, the important thing is strategy and implementation."

The relationship data attributes are also in each account's friendship name list. We transform the friendship name list into a mutual connection pattern using Gephi software to produce the proposed social networking metrics. Then the mutual connection pattern is processed into an ego-network model to produce several proposed social networking metrics, degree centrality, and density. After that, an account is selected with a complete demographic, conversation, and relationship data based on the data attributes of the company, job title, school, conversation in uploaded content, and connections. Relationship features are developed to identify the value formed from the social relationship of an account with other accounts based on the degree centrality and density attributes. The degree of centrality can measure the popularity of actors in the relationship, while density can measure the interactions between actors.

3.3 Credit Scoring Features Development

Data features are established to get quality data and produce the best relevant input data. The features attribute determines a person's creditworthiness through the credit scoring model, as in Table 2.

Table 2. Data Features

Feature	Notation	Feature Attribute	Notation
Demographic Feature	F_1	Gender	G
		Age	A
		Job (Salary)	J
		Education	E
Conversation Feature	F_2	Analytical Thinking	AT
		Clout	C
		Authenticity	Ac

		Emotional Tone	ET
		Like to Post Ratio	LR
		Comment to Post Ratio	CR
		Number of Posts	NP
Relationship Feature	F_3	Degree Centrality	DC
		Density	D

3.4 Establishment of Credit Scoring Scorecard

The credit scoring scorecard is formed by referring to the statement from the Indonesian Bankers Association (2014), which stated that the formation of a scorecard model to construct a creditworthiness model is better adapted to the criteria and needs of each financial institution. The details of assigning value to each attribute are based on the following sources:

a) The grouping of gender attributes is based on the April 2020 Monthly Socio-Economic Data report by the Central Statistics Agency.
b) The age attribute grouping is based on the report on the Development of Fintech Lending for the March 2020 Period by the Financial Services Authority.
c) The grouping of job attributes is based on the April 2020 Socio-Economic Data Monthly report by the Central Statistics Agency (with adjustment).
d) The grouping of education attributes is based on the April 2020 Socio-Economic Data Monthly report by the Central Statistics Agency.
e) The grouping of conversation attributes is based on a literature review of research conducted by Vaughn (2019) by selecting parameters using the values contained in the Linguistic Inquiry Word Count (LIWC), including analytical thinking, clout, authenticity, and emotional tone.
f) The grouping of relationship attributes, including degree centrality and density, is based on a literature review of the research entitled "Employee Communication Network Mapping Analysis Using Social Network Analysis" (Zusrony et al. 2019).

Based on the sources above, the exact value of each attribute is shown in Table 3 below:

Table 3. Credit Scoring Scorecard

Feature	**Notation**	**No Attribute**	**Feature Attribute**	**Notation**	**Score**	**Criteria**
Demographic Feature	F_1	1	Gender	G	2	Woman
					2	Man
		2	Age	A	4	Less than 19 years old
					8	19 to 34 years old
					12	35 to 54 years old
					16	More than 54 years old
		3	Job (Salary)	J	4	IDR 1,000,000 to IDR 5,000.000
					8	IDR 5,000,001 to IDR 10,000.000
					12	IDR 10,000,001 to IDR 20,000.000
					16	IDR 20,000,001 to IDR 40,000.000
					20	More than IDR 40,000.000
		4	Education	E	2	Primary School
					4	Junior High School
					6	Senior High School
					8	Diploma
					12	Bachelor

Conversation	F_2	5	Analytical Thinking	AT	1	0 to 35.0
					2	35.1 to 70.0
					3	70.1 to 100
		6	Clout	C	1	0 to 35.0
					2	35.1 to 70.0
					3	70.1 to 100
		7	Authenticity	Ac	1	0 to 35.0
					2	35.1 to 70.0
					3	70.1 to 100
		8	Emotional Tone	ET	1	0 to 35.0
					2	35.1 to 70.0
					3	70.1 to 100
		9	Like to Post Ratio	LR	1	0 to 10.0
					2	10.1 to 20.0
					3	More than 20.0
		10	Comment to Post Ratio	CR	1	0 to 2.0
					2	2.1 to 5.0
					3	More than 5.0
		11	Number of Posts	NP	1	Less than 10
					2	More than 10
Relationship	F_3	12	Degree Centrality	DC	2	0 to 50
					5	More than 50
		13	Density	D	2	0 to 0.050
					5	More than 0.050

3.5Credit Scoring Model Development
After developing the credit scoring scorecard, we develop the credit scoring model in two steps by way of expert judgment, and prediction analysis explained below.

1) Expert Judgment
   At this stage, the credit scoring model combines demographic data, conversation data, and relationship data. Experts agree that the credit scoring model is made by considering the priority in the weighting to produce one number to show the priority for each attribute so that the formation of the model can impact the attributes that are given a value.
2) Prediction Analysis
   Data processing was carried out using RapidMiner software by forming a predictive process diagram by selecting the required features in the software used. The model's design uses a decision tree algorithm and a data set with preprocessing data and data features that have been formed, then separating the data into training data and testing data. After the prediction process is mapped, the RapidMiner software will process the inputted data and provide final results and interpretations in accuracy values based on the processed data.

3.6Validation and Evaluation of Prediction Model Performance
This study uses RapidMiner software for validation and evaluation steps using cross-validation and confusion matrix methods to analyze the proposed models' prediction results. According to Salmu, S., and Solichin (2017), the performance measurement of the classification method is carried out to test the results by machines. In contrast, cross-validation and confusion matrices are applied to measure accuracy, precision, and recall values to provide classification performance information.

4. Result and Discussion

1) Expert Judgment

In constructing a credit scoring model, we divided it into three parts, including demographic features, conversation features, and relationship features.

Demographic Feature

The first feature is demographic features consisting of attributes of gender, age, job (salary), and education. The job attribute has the most significant weight with a weight of twice the value because it is assumed that a high score on this attribute indicates the borrower's ability to repay the credit. On the other hand, a low job attribute score indicates the risk of default or potential bad credit. Rabuh (2020) also stated that demographic data are usually included in financial data modelling and are usually used by lenders. In comparison, other attributes have the same weight as one attribute value. So demographic feature formulation is as follows:

$$\boldsymbol{F_1 = G + A + (2 \times J) + E} \quad \textbf{(1)}$$

Where $\boldsymbol{F_1}$ is the demographic feature formulation, $\boldsymbol{G}$ is the gender of the borrower, $\boldsymbol{A}$ is the borrower's age, $\boldsymbol{J}$ is the amount of salary, and $\boldsymbol{E}$ is the last education.

Conversation Feature

The second feature is the conversation feature. The conversation feature consists of several attributes: analytical thinking, clout, authenticity, emotional tone, like to post ratio, comment to post ratio, and number of posts. Each attributes weights one attribute value. The conversation feature formulation is as follows:

$$\boldsymbol{F_2 = AT + C + Ac + ET + LR + CR + NP} \quad \textbf{(2)}$$

Where $\boldsymbol{F_2}$ is the conversation feature formulation, $\boldsymbol{AT}$ is the analytical thinking score derived from the LIWC model, $\boldsymbol{C}$ is the clout score derived from the LIWC model, $\boldsymbol{Ac}$ is the authenticity score derived from the LIWC model, and $\boldsymbol{ET}$ is the emotional tone score derived from LIWC model, $\boldsymbol{LR}$ is the user's account like to post ratio, $\boldsymbol{CR}$ is the user's account comment to post ratio, and $\boldsymbol{NP}$ is the user's account number of post.

Relationship Feature

The third feature is the relationship feature. The relationship features consist of degree centrality and density. Each attributes weights one attribute value. The relationship feature formulation is as follows:

$$\boldsymbol{F_3 = DC + D} \quad \textbf{(3)}$$

Where $\boldsymbol{F_3}$ is the relationship feature formulation, $\boldsymbol{DC}$ is the user's degree centrality value within the network, and $\boldsymbol{D}$ is the density of the user's network.

Credit Scoring Formula

After formulating each attribute, we construct a credit scoring model by combining demographic features, conversation features, and relationship features as follows:

$$\boldsymbol{CS = \textstyle\sum_{n=1}^{3} \mathrm{Fn}} \quad \textbf{(4)}$$

$\boldsymbol{CS}$ is the credit scoring formulation, and $\boldsymbol{F_n}$ is the feature formulation of the n^{th} feature formulation. $\boldsymbol{CS}$ is equal to the sum of 1st feature ($\boldsymbol{F_1}$) – demographic feature formulation, 2nd feature ($\boldsymbol{F_2}$) – conversation feature formulation, and 3rd feature ($\boldsymbol{F_3}$) – relationship feature formulation.

In the proposed credit scoring model, the credit score can be calculated by adding up each attribute's value and weight. After the proposed model is made, the expert determines the threshold to determine the minimum score for someone who can be declared eligible for credit. Based on the literature review and considering priority in the weighting, the expert determined that the threshold used was 64, so a person was declared creditworthy. After the threshold is acquired, score weighting is implemented on all data from each feature according to the value of their respective attributes. Data can be declared eligible for a credit if the score exceeds the specified threshold. Table 4 shows the sample of data attribute with creditworthiness class; creditworthy and not creditworthy. The data is labelled eligible for a credit if the total score exceeds the specified threshold. The data is labelled not eligible for a credit if the total score is below the specified threshold.

Table 4. Creditworthiness Decision by Expert Judgment

ID	**Attribute**														
	G	**A**	**J**	**E**	**AT**	**C**	**Ac**	**ET**	**LR**	**CR**	**NP**	**DC**	**D**	**TS**	**Label**
1	Woman	26	13.924.383	Bachelor	97.0	81.8	4.3	99.0	81.6	3.6	5	235	0.073	72	Eligible
2	Man	58	3.369.963	Bachelor	80.45	68.9	0.59	50.0	1	0	4	150	0.108	59	Not Eligible
3	Woman	22	6.014.650	Bachelor	82.2	79.7	9.3	83.7	293	15	2	83	0.173	65	Eligible
4	Man	22	15.255.680	Diploma	54.1	39.0	23.9	98.0	8.2	1.2	5	28	0.037	57	Not Eligible
5	Man	32	102.166.280	Bachelor	80.6	62.7	17.2	99.0	80.67	29	6	3	0.076	85	Eligible

2) Prediction Analysis

After developing a credit scoring model with the complete determined creditworthiness label, we use the data to measure the model accuracy. We perform prediction analysis using the classification methods with the decision tree algorithm. In the classification model based on the decision tree algorithm, the class label for data is predicted from the root of the three to the last branch. The decision tree algorithm determines the best split to optimize the accuracy of the prediction results. We can follow the branch corresponding to a certain data value in the analysis and proceed to the next node. The decision tree result for the credit scoring model is visualized in Figure 2.

Based on the decision tree visualization, a person declared eligible for credit if having a salary between IDR 10,000,000 to IDR 20,000,000, or IDR 20,000,000 to IDR 40,000,000, or more than IDR 40,000,000. The algorithm will consider the comment to post ratio if the salary is between IDR 5,000,000 to IDR 10,000,000. If the comment to post ratio is more than 5.0 and the borrower is female, she is declared eligible for credit. However, if the borrower is a male, he is declared not eligible for credit. A person is declared not eligible if having a salary between IDR 1,000,000 to IDR 5,000,000. The algorithm will consider the comment to post ratio if the salary is between IDR 5,000,000 to 10,000,000. If the comment to post ratio is less than 5.0, the borrower is declared not eligible for credit. From the explanation and visualization of the decision tree, the job becomes the root attribute, among other attributes.

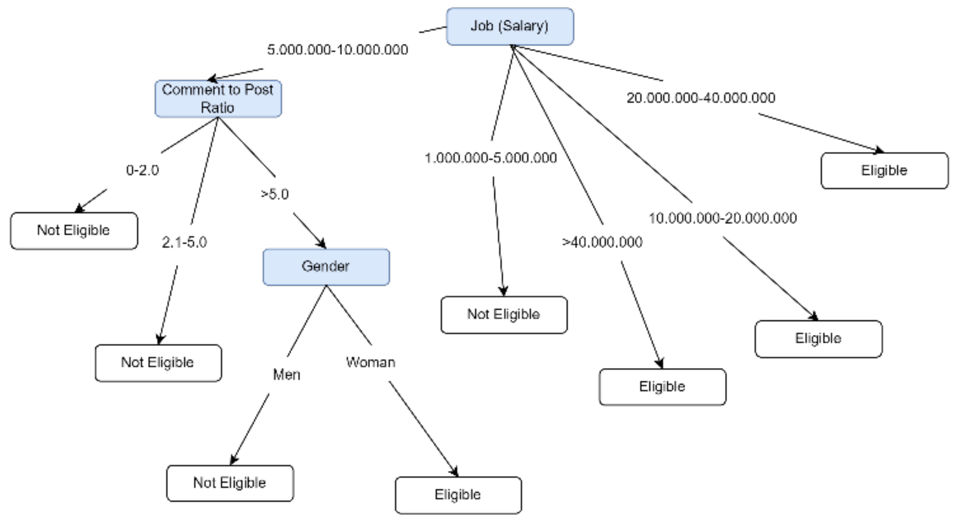

Figure 2. Visualization of Decision Tree

We test the labelled data for prediction accuracy. The predicted sample data is shown in Table 5. Each data is given a creditworthiness label based on the specified threshold and then compared with the predicted creditworthiness label based on the decision tree algorithm. The data prediction comparison is then recorded in a confusion matrix for the evaluation stage using RapidMiner. The confusion matrix is shown in table 6.

Table 5. Prediction Analysis using Decision Tree Algorithm

ID	Attribute														
	G	A	J	E	AT	C	Ac	ET	LR	CR	NP	DC	D	Label	Prediction
1	Man	19-34	5.000.001-10.000.000	Bachelor	70.1-100	35.1-70.0	70.1-100	0-35.0	0-10.0	0-2.0	>10	0-50	>0.050	Not Eligible	Not Eligible
2	Man	35-54	>40.000.000	Bachelor	70.1-100	70.1-100	0-35.0	70.1-100	>20.0	0-2.0	>10	>50	>0.050	Eligible	Eligible
3	Woman	19-34	1.000.000-5.000.000	Bachelor	35.1-70.0	70.1-100	0-35.0	35.1-70.0	>20.0	>5.0	<10	0-50	>0.050	Not Eligible	Not Eligible
4	Woman	19-34	10.000.001-20.000.000	Bachelor	70.1-100	70.1-100	0-35.0	70.1-100	0-10.0	0-2.0	>10	0-50	0-0.050	Eligible	Eligible
5	Woman	19-34	5.000.001-10.000.000	Bachelor	70.1-100	70.1-100	0-35.0	70.1-100	10.0-20.0	0-2.0	<10	>50	>0.050	Not Eligible	Eligible
6	Man	19-34	5.000.001-10.000.000	Bachelor	70.1-100	70.1-100	70.1-100	70.1-100	>20.0	>5.0	>10	>50	>0.050	Eligible	Not Eligible
7	Woman	19-34	5.000.001-10.000.000	Bachelor	35.1-70.0	70.1-100	70.1-100	70.1-100	>20.0	>5.0	>10	>50	>0.050	Eligible	Not Eligible
8	Woman	>54	10.000.001-20.000.000	Bachelor	35.1-70.0	35.1-70.0	0-35.0	35.1-70.0	>20.0	>5.0	<10	0-50	>0.050	Eligible	Eligible
9	Man	>54	1.000.000-5.000.000	Senior High School	70.1-100	35.1-70.0	0-35.0	35.1-70.0	0-10.0	0-2.0	<10	>50	>0.050	Not Eligible	Not Eligible

Table 6. Decision Tree Performance Metrics Result

<table>
<tr><td colspan="2" rowspan="2"></td><th colspan="2">Predicted Class</th></tr>
<tr><th>Eligible</th><th>Not Eligible</th></tr>
<tr><th rowspan="2">Actual Class</th><th>Eligible</th><td>True Positive (TP)
The sum of borrowers predicted as eligible for credit and eligible for credit.</td><td>False Negative (FN)
The sum of borrowers was predicted as not eligible for credit and eligible for credit.</td></tr>
<tr><th>Not Eligible</th><td>False Positive (FP)
The sum of borrowers predicted as eligible for credit or not.</td><td>True Negative (TN)
The sum of borrowers was predicted as not eligible for credit and not eligible for credit.</td></tr>
</table>

The TP, FN, FP, and TN value is then supplied for the model evaluation calculation. We calculate the model accuracy by comparing true positive and true negative with the whole data, precision by comparing true positive with true positive and false positive, recall by comparing true positive and true negative with true positive and false negative, and f-measure by comparing twice of precision and recall with precision and recall. The higher the evaluation value indicates a better model. The evaluation result is shown in table 7.

Table 7. Decision Tree Performance Metrics Result

Accuracy	Precision	Recall	F-Measure
85,95%	91,50%	73,33%	81,41%

Based on the decision tree classification algorithm, the proposed credit scoring model accuracy is 85.95%, the precision value is 91.50%, the recall value is 73.33%, and the f-measure value is 81.41%. We consider the credit scoring model fairly good since the high evaluation rate.

5. Conclusion

Financial institutions have the freedom to implement various types of credit scoring models. Financial institutions can use data tools to build their model if they have abundant data. The model we show here could be generalized into generic; financial institutions can make modifications according to their personalized, domain-based needs. The model built is an initial model for credit scoring based on LinkedIn social media data. For us, the most important thing is how to combine various types of data to evaluate creditworthiness. We recommend using other possible features, data sources, and predictive models to get an advanced model.

References

Basori, O. R., & Wahyuningsih, S. D., Analisis Penilaian Prinsip 5C dalam Pemberian Kredit terhadap Non Performing Loan guna Menilai Tingkat Kesehatan Bank pada PT BPR Harta Swadiri Pandaan. *Penelitian Manajemen Terapan (PENATARAN)*, vol. *3, no.1,* pp. 54–63, 2018.

Bastos, J., *Munich Personal RePEc Archive Credit scoring with boosted decision trees Credit scoring with boosted decision trees*. *8156*, 2008.

Deters, F. große, & Mehl, M. R., Does Posting Facebook Status Updates Increase or Decrease Loneliness? An Online Social Networking Experiment. *Social Psychological and Personality Science*, vol. *4, no.5,*pp. 579–586, 2013. https://doi.org/10.1177/1948550612469233

Eprianti, N., Penerapan Prinsip 5C Terhadap Tingkat Non Performing Financing (Npf). *Amwaluna: Jurnal Ekonomi Dan Keuangan Syariah*, *3*(2), 2019. . https://doi.org/10.29313/amwaluna.v3i2.4645

Garg, P., Rani, R., & Miglani, S., Emerging Research in Computing, Information, Communication and Applications. *Emerging Research in Computing, Information, Communication and Applications*, 1–9., 2016. https://doi.org/10.1007/978-81-322-2553-9

Goel, A., & Rastogi, S. (2021). Credit scoring of small and medium enterprises: a behavioural approach. *Journal of Entrepreneurship in Emerging Economies*, *ahead-of-p*(ahead-of-print). https://doi.org/10.1108/jeee-03-2021-0093

Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From footprint to evidence: An exploratory study of mining social data for credit scoring. *ACM Transactions on the Web*, *10*(4). https://doi.org/10.1145/2996465

He, W., Wang, F. K., & Akula, V. , Managing extracted knowledge from big social media data for business decision making. *Journal of Knowledge Management*, vol. *21, no.2,* pp. 275–294, 2017. https://doi.org/10.1108/JKM-07-2015-0296

Ledhem, M. A. (2021). Data mining techniques for predicting the financial performance of Islamic banking in Indonesia. *Journal of Modelling in Management*. https://doi.org/10.1108/JM2-10-2020-0286

Mardhotillah, I. C. (2020). 5c Factor Analysis Aspects Of The Designation Of Lending Of Working Capital In Pd. Bpr Bank Regional Lamongan. *Jurnal Studi Manajemen Dan Bisnis*, vol. *6, no.1,* pp.1–7, 2020. https://doi.org/10.21107/jsmb.v6i1.6672

Niu, B., Ren, J., & Li, X., Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. *Information (Switzerland)*, vol. *10, no.12,* 2019. . https://doi.org/10.3390/INFO10120397

Perangin-angin, L. L. K., & Zainal, M., Partisipasi Politik Pemilih Pemula Dalam Bingkai Jejaring Sosial Di Media Sosial. *Jurnal ASPIKOM*, vol. *3, no.4,* pp. 737, 2018. . https://doi.org/10.24329/aspikom.v3i4.210

Rabuh, A., *Developing a Credit Scoring Model Using Social Network Analysis*. *November*, 1–194, 2020.

Saardchom, N., Expert Judgment Based Scoring Model. *Journal of Business and Economics*, vol. *3, no.3,*pp. 164–175, 2012 . https://doi.org/10.15341/jbe(2155-7950)/03.03.2012/002

Salmu, S., D., & Solichin, A. (2017). Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta Prediction of Timeliness Graduation of Students Using Naïve Bayes : A Case Study at Islamic State University Syarif Hidayatullah Jakarta. *Prosiding Seminar Nasional Multidisiplin Ilmu*, *April*, 701–709.

Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M., & Dapretto, M., The Power of the Like in Adolescence: Effects of Peer Influence on Neural and Behavioral Responses to Social Media. *Psychological Science*, vol. *27, no.7,*pp. 1027–1035, 2016. https://doi.org/10.1177/0956797616645673

Syaifuddin, D. T. (2019). *Buku 1. Manajemen Perbankan* (Issue December).

Syed Nor, S. H., Ismail, S., & Yap, B. W. (2019). Personal bankruptcy prediction using decision tree model. *Journal of Economics, Finance and Administrative Science*, vol. *24, no.47,* pp. 157–170, 2019. https://doi.org/10.1108/JEFAS-08-2018-0076

Teles, G., Rodrigues, J. J. P. C., Saleem, K., Kozlov, S., & Rabêlo, R. A. L. , Machine learning and decision support system on credit scoring. *Neural Computing and Applications*, vol. *32, no.* pp . 9809–9826, 2020. https://doi.org/10.1007/s00521-019-04537-7

Tyas, S. J. S., Febianah, M., Solikhah, F., Kamil, A. L., & Arifin, W. A., Analisis Perbandingan Algoritma Naive Bayes Dan C.45 Dalam Klasifikasi Data Mining Untuk Memprediksi Kelulusan. *Jurnal Teknologi Informasi Dan Komunikasi*, vol. *8, no.1,* pp. 86–99, 2021.

Yu, X., Yang, Q., Wang, R., Fang, R., & Deng, M. (2020). Data cleaning for personal credit scoring by utilizing social media data: An empirical study. *IEEE Intelligent Systems*, vol. *35, no.2,* pp. 7–15, 2020. https://doi.org/10.1109/MIS.2020.2972214

Zhang, Y., Jia, H., Diao, Y., Hai, M., & Li, H., Research on Credit Scoring by Fusing Social Media Information in Online Peer-to-Peer Lending. *Procedia Computer Science*, *91*(Itqm), 168–174, 2016. https://doi.org/10.1016/j.procs.2016.07.055

Biographies

Dian Puteri Ramadhani received her bachelor's (2018) and master's (2019) degrees from Telkom University, Bandung, majoring in management with a specialty in big data analytics and complex network. Since 2017, Dian has become a researcher at the Social Computing and Big Data Laboratory, Telkom University, who is actively conducting research and publications related to big data analytics in business and management. In 2020, Dian has been a full-time faculty member in the Department of Management, Telkom University, teaching courses including big data and data analytics, econometrics, and managerial economics. Dian has published an Introduction to Social Network Analysis: Konsep dan Praktis.

Putri Mentari Wijaya is a last-semester student at Telkom University, majoring in Management Business Telecommunication and Informatics, focusing on Operations and ICT Management. Putri has experience as a laboratory assistant for Big Data and Data Analytics courses and is interested in working with data.

Andry Alamsyah received the B.S in Mathematics from ITB ITB Bandung in 1996, the M.S. degree in Information System from Universite Picardie de Jules Verne, France, in 2004, and the Ph.D. degree in Electrical Engineering and Informatics from ITB Bandung in 2017. Since 2011, Andry has been a researcher at the School of Economics and Business, Telkom University. He is the founder and Chief of the Lab. Social Computing and Big Data. He was Director of the Digital Business Ecosystem Research Center from 2018-to 2020. He is also the founder and president of Asosiasi Ilmuwan Data Indonesia (AIDI), the only formal association for Data Scientists in Indonesia. By 2021, Andry has appointed as an Honorary member of Asosiasi Blockchain Indonesia. Research Interest: Social Computing, Social Network, Complex Network / Network Science, Big Data, Blockchain Technology & Opportunities, Technopreneurship, Business Model.