A Sentiment Analysis Approach in Analyzing Risk for SME Business Financing

Irfan Wahyudin, Taufik Djatna, Wisnu Ananta Kusuma
Computer Science, Mathematic and Natural Science, Bogor Agricultural University
Dramaga IPB, Bogor, 16680
Tel: (0251) 86228448, Fax: (0251) 8622986
e-mail: irfanwahyudin@apps.ipb.ac.id, taufikdjatna@ipb.ac.id, ananta@ipb.ac.id

Abstract

In risk analysis, the implementation of qualitative model is relatively more difficult than the quantitative model, since most of the banks chanelling loan for Small Medium Enterprise financing have no specific criteria to approve the proposals. In this research, we proposed a sentiment analysis approach to construct a proper model to be used by banks in enhancing their risk analysis against the submitted SME financing proposals. Thus, we focused on three objectives to overcome the problems that oftenly occur in qualitative model implementation. First, we modelled risk clusters using K-Means clustering, optimized by Pillar Algorithm to get the optimum number of clusters. Secondly, we performed risk measurement by calculating term-importance scores using TF-IDF combined with term-sentiment scores based on SentiWordNet 3.0 for Bahasa Indonesia. Eventually, we summarized the result by correlating the featured terms in each cluster with the 5Cs Credit Criteria. The result shows that the model is effective to group and measure the level of the risk and can be used as a basis for the decision makers in approving the loan proposal.

Keywords: *risk analysis, SME business, centroid optimazion, K-Means, opinion mining, sentiment analysis*

1. Introduction

Currently, the goverment of Indonesia requires all national banks in Indonesia to support SME business by providing working capital loan. On the other hand, the central bank also insists the national banks to have risk assessment before granting a loan to minimize the risks that may occur, such as loan default. To measure the performance of the banks on channeling loan, one of main indicators that the central bank uses is Non Performing Loan (NPL) ratio, which would be issued by the bank through its annual report.

There are two models that are widely used to implement risk assessment in financing, namely a quantitative model and a qualitative model [1]. Risk assesment for SME business in national banks in Indonesia is commonly dominated by the implementation of credit scoring system (quantitative model). Unfortunately, not all of the banks have succesfully implemented these models. This condition is as shown in a national private bank, where the NPL ratio has an inclining trend in the last three years. This condition drove the top management of the bank to encourage the risk management division to refine the implementation of risk management.

From our observation, they already succesfully used the quantitative model, implemented in their Loan Origniating System. A problem was found in the qualitative model where the implementation was conducted in delivering the risk opinion and risk mitigation through a risk analysis document. The problem is that the qualitative model was not significantly used by the authorities in making decision since there were no decision criteria as a baseline in making the decision.

The objective of this research is to perform clustering tasks to group the risk opinion document based on its concept, and once the cluster has been built, the risk level in each cluster is measured using term-importance and sentiment weighting. An evaluation is also performed for both clustering and sentiment measurement to reveal the implication and the criteria in assessing

Received February 23, 2014; Revised May 29, 2014; Accepted June 12, 2014

the loan risk. By conducting this research, the expectation is that the risk cluster and the risk measurement can be used as a baseline in making the decision to help the bank evaluate and refine current risk analysis implementation.

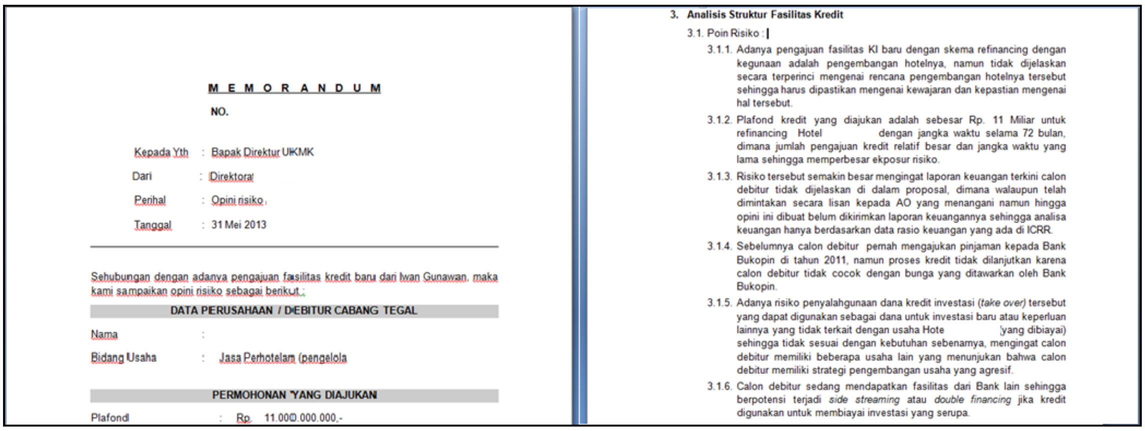
Regarding to the techniques used in a sentiment analysis, there are two major techniques commonly used; those are machine learning based and lexicon based [2]. Supervised machine learning techniques that are commonly used are such as Support Vector Machine [3], Neural Networks [4], and Naive Bayes [5]. In addition, for unsupervised machine learning, there are several clustering techniques, let say K-Means [5] and hierarchical clustering [6].

For the lexicon based, there are various lexicon resources that can be utilized as a dictionary to determine the polarity of terms, such as SentiWordNet [7], which is derived from a well known corpus, that is, WordNet, an English dictionary for word synonyms and antonyms. The next one is SentiStrength [8], which is a lexicon based technique, distributed as a desktop application tool that is already combined with several popular supervised and unsupervised classifier algorithms: SVM, J48 classification tree, and Naive Bayes. The other is emoticon based sentiment analysis, which is considered the simplest one [9]. Unfortunately, most of the lexicon dictionaries and corpus resources are designated for English. Some efforts have been done to overcome this shortfall, for instance, by translating either the observed corpus object [10] or the lexicon dictionary and the labeled corpus [11]. Moreover, the purpose of the sentiment analysis is mostly dominated by how businesses determine the opinion and judgement of their customers upon their products from open resources, such as social media instead of performing sentiment analysis using closed resources, such as risk analyst opinion upon bank customers' businesses.

Support Vector Decomposition (SVD) for dimension reduction and concept extraction is performed as an initialization followed by a clustering task using K-Means, optimized by centroid initialization namely Pillar Algorithm [12]. A method to measure the term importance from the risk opinion corpus is performed using the widely use TF-IDF, combined with positive-negative polarity measurement using the SentiWordNet 3.0 library [7]. Unlike in English, as of today there is only one international research publication utilizing SWN 3.0 in Bahasa Indonesia that aims to detect sarcasm in social media [11]. The translation problems are overcomed by utilizing tools and techniques such as Google Translate, Kateglo (*Kamus Besar Bahasa Indonesia* based dictionary), and by asking banking experts for specific banking and finance terms.

2. Research Method

As a case study, we conducted it in one of national private banks in Indonesia where the SME financing is one of their core businesses. We collected about 519 risk analysis documents from the Risk Management division. All of the documents are in Microsoft Words (*.doc an *.docx format), consisting of narrative opinions in Bahasa Indonesia. There are seven opinion points delivered in the documents; those are 1) Credit Scoring 2) Financial Performance 3) Proposed Loan Facility 4) Business Performance 5) Repayment Ability and Cash Flow 6) Legal Analysis, and 7) Foreign Exchange (optional). Here in Figure 1, we depict the sample of one risk analysis document.



M E M O R A N D U M

NO.

Kepada Yth : Bapak Direktur UKMK

Dari : Direktorat

Perihal : Opini risiko

Tanggal : 31 Mei 2013

Sehubungan dengan adanya pengajuan fasilitas kredit baru dari Iwan Gunawan, maka kami sampaikan opini risiko sebagai berikut :

DATA PERUSAHAAN / DEBITUR CABANG TEGAL

Nama :

Bidang Usaha : Jasa Perhotelam (pengelola

PERMOHONAN YANG DIAJUKAN

Plafond : Rp. 11.000.000.000,-

3. Analisis Struktur Fasilitas Kredit

3.1. Poin Risiko :

3.1.1. Adanya pengajuan fasilitas KI baru dengan skema refinancing dengan kegunaan adalah pengembangan hotelnya, namun tidak dijelaskan secara terperinci mengenai rencana pengembangan hotelnya tersebut sehingga harus dipastikan mengenai kewajaran dan kepastian mengenai hal tersebut.

3.1.2. Plafond kredit yang diajukan adalah sebesar Rp. 11 Miliar untuk refinancing Hotel dengan jangka waktu selama 72 bulan, dimana jumlah pengajuan kredit relatif besar dan jangka waktu yang lama sehingga memperbesar ekposur risiko.

3.1.3. Risiko tersebut semakin besar mengingat laporan keuangan terkini calon debitur tidak dijelaskan di dalam proposal, dimana walaupun telah dimintakan secara lisan kepada AO yang menangani namun hingga opini ini dibuat belum dikirimkan laporan keuangannya sehingga analisa keuangan hanya berdasarkan data rasio keuangan yang ada di ICRR.

3.1.4. Sebelumnya calon debitur pernah mengajukan pinjaman kepada Bank Bukopin di tahun 2011, namun proses kredit tidak dilanjutkan karena calon debitur tidak cocok dengan bunga yang ditawarkan oleh Bank Bukopin.

3.1.5. Adanya risiko penyalahgunaan dana kredit investasi (*take over*) tersebut yang dapat digunakan sebagai dana untuk investasi baru atau keperluan lainnya yang tidak terkait dengan usaha Hote (yang dibiayai) sehingga tidak sesuai dengan kebutuhan sebenarnya, mengingat calon debitur memiliki beberapa usaha lain yang menunjukan bahwa calon debitur memiliki strategi pengembangan usaha yang agresif.

3.1.6. Calon debitur sedang mendapatkan fasilitas dari Bank lain sehingga berpotensi terjadi *side streaming* atau *double financing* jika kredit digunakan untuk membiayai investasi yang serupa.

Figure 1. Risk analysis documents snapshot

All of the parts were analyzed based on 5Cs Credit Criteria (Character, Capacity, Capital, Condition, and Collateral).

As seen in Figure 2, the research framework was divided into 4 parts; those are 1) Preprocessing 2) Risk Clustering 3) Risk Measurement, and 4) Evaluation. We will discuss the details and results in the following section.

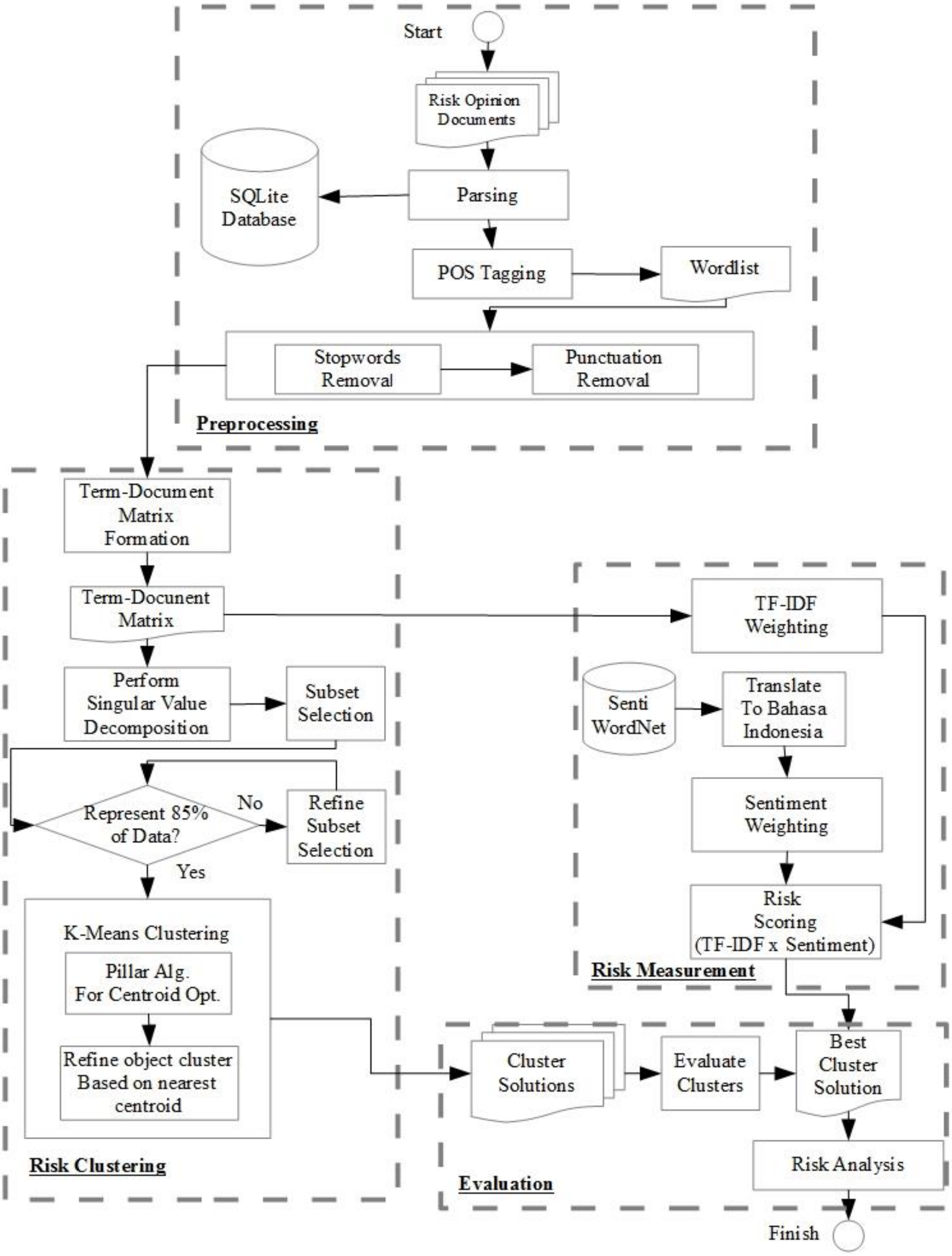


Figure 2. Research Framework

3. Results and Analysis

3.1. Preprocessing

As commonly done in text mining, preprocessing is the prerequisite task to remove stop words, punctuation, and unimportant terms, and to formalize the terms used as feature vector for the next task [13]. After scanning the seven parts of the corpus, there were 3289 terms found. From the term collection, there were several things to do such as fixing typos and formalizing the terms. To do so, a mini dictionary was created to be used as a reference for the program.

Furthermore, the formalization was done for some terms like “stratejik”→”strategi” (strategy), “spekulas”→”spekulasi” (speculation), “melamah”→”melemah” (declining). The formalization was also important since some terms could not be found in the SWN 3.0 lexicon, although those are not typos. For instance, terms like “komperehensif” was converted to “komprehensif”; “identity” was converted to “identitas”; “volatility” was converted to “volatilitas”.

3.2. Singular Value Decomposition

The term-document matrix was considered a high dimensional matrix, so it could take big amount of time to have computation. Thus, the term-document matrix dimension was reduced by using Singular Value Decomposition. The best K-vectors to represent the whole dataset [14] were selected based on formulation (1)

$$q \leftarrow \left(\sum_{i=1}^{k} v_i / \sum_{i=1}^{n} v_i\right) \qquad (1)$$

where *q* is the threshold that represents the whole dataset and *v* is the vector column-row that represents a term-document relationship. By initializing *q = 0.80*, and after some measurements from SVD formulation (2) were done, 300 subsets were taken from matrix *V*, that is, the matrix of document representation, because *k* = 300 represented 80% of whole dataset and it was used as feature vector for clustering.

$$A_{300} = U_{300}\, \Sigma_{300}\, V_{300} \qquad (2)$$

3.3. K-Means Clustering and Centroid Optimization

K-Means clustering, a widely used centroid based partitioning algorithm, was used in order to find how many risk clusters exist in the corpus. The original proposal of K-Means clustering used random centroid selection in the beginning of iteration. This is not an issue when it computes a small size of datasets. However, dealing with a large size of data could take a lot of time to produce the best centroid selection.

The algorithm was inspired by the function of pillars of a building or a construction. It is common that a pillar in a building is deployed at each edge or at each corner in a building, so that the mass of the building is concentrated in each pillar. The same idea was adopted for the clustering task that the best initial centroids were presumed to exist in the edge of the dataset, or in other words, those *k*-furthest objects in the dataset were selected as initial centroids, where *k* is the number of clusters to be observed. Hence, to find the best cluster solution, we iterated some possible numbers of clusters from K=2 to K=10, and each iteration was preceded by centroid optimization using Pillar Algorithm as seen in Figure 3.

```
alpha = 0.4
beta = 0.6
while (K ≥ 2 and K ≤ 10)
    while (alpha ≥ 0.4 and alpha ≤ 1.0)
        while (beta ≥ 0.6 and beta≤ 1.0)
            Centroids = Pillar_Algorithm(P, K, alpha, beta)
            Solution = K_Means(K, Centroids)
```

Figure 3. Pseudocode to find the best cluster solution

Complete steps of original Pillar Algorithm paper are described in Figure 4, where the mean calculation was done for each *t* variable term, that is, available terms are in the term-document matrix list *P*, *n* is the number of document, and ***m*** is the mean vector of the term. After getting the mean of all terms as a starting point of the iteration, the algorithm selected the *k* furthest distance objects from *m*, defined as Ж or initial centroids, and checked whether Ж already existed in *SX* list; if not, it would be stored to *SX*.

```
Pillar_Algorithm(P, K, alpha, beta)
    m = GetMeanFromEachVariable(P)
    Distances = []
    n = length(P)
    MaximumIteration = n
    nmin = (alpha * n) / K//The minimum number of neighbor objects in radius of dmax
    for i = 1 to n
        m[i] = Sum(P[i]) / NumberOfVariables
        d = EuclidianDistance(P[i] - m[i])
        Distances←d //Stores the distance to mean calculation
                    //into matrix Distances

    D = SortDescendingly(Distances) //Get the furthest object from the mean by sort
                                    //descendingly
    DM = D

    while (NumberOfCentroid < K and iteration < MaximumIteration)
        dmax = DM[0]
        nbdis = beta * dmax //The furthest distance that has to be fulfilled
        for x = 0 to n-1 //Iterates the objects within the matrix Distances
            if DM[x] not in SX //If current object not in SX list
                Ж = P[DM[x]]
                SX← Ж //Stores to SX list

        for x = 0 to n:
            d = EuclidianDistance(P[i] - Ж) //Calculate distance between
                                            //object with Ж
            DTemp1←d //Stores into DTemp1

        D = SortDescendingly(DTemp1) //Sort descendingly

        no = 0
        for x = 0 to n: //Calculate the number of neigbors that in radius of nbdis
            if D[0] <= nbdis //Check if the distance < nbdis
                no++    //Add value of no variable if fulfill max distance
                        //criteria

        if no >= nmin //Check if the number of neigbors >= nmin
            NumberOfCentroid++ //Add the number of centroid
            C.append(zhe)

            for x = 0 to n
                d = EuclidianDistance(P[x] - Ж)
                DTemp2←d
            DM = SortDescendingly(DTemp2)
        else //Otherwise, continue exploration through other objects
            iteration++
        if iteration == MaximumIteration //If it has reach max iteration allowed
            return                       //then exit
```

Figure 4. Pillar Algorithm modified from Barakbah 2009

The selection method was simply by sorting the distance matrix dataset containing each term vector distance to the mean. The distance formula we used here is the basic Euclidian distance measurement.

3.4. TF-IDF Weighting

In general, TF-IDF [13] was used to identify how important each available term in the corpus is, where *tf$_{t,d}$*, the frequency of term *t* in document *d*, is defined as formula (3).

$$tf_{t,d} = \frac{f_{t,d}}{argmax(tf_d)} \tag{3}$$

For *idf$_i$*, inverse document frequency for term *t* in the corpus *D* is defined as formula (4).

$$idf_t = log_2 \frac{N}{n_t} \tag{4}$$

where *N* is the number of document available in corpus, n_t is occurrence number of term *t* in all documents in the corpus. There was a little modification in implementing the formula above. The term in the basic TF-IDF was selected distinctly based only on how the term was spelled, and disregarded the term preposition in sentence. Since the SWN 3.0 was also based on the term preposition, the term position in the term list needed to be added, as obtained in the POS Tagging task[1].

3.5. Sentiment Weighting

Before performing sentiment weighting by using Google Translate API, SWN 3.0 lexicon needed to be translated into Bahasa Indonesia [11]. By using Levensthein distance measure, not all terms in the corpus were perfectly matched, so the polarity values were manually set for special terms in banking and finance e.g., “collectability" or “coll”, “bowheer” (project employer) and “jaminan” (collateral) to get the precise positive(pos) or negative(neg) polarity value. If the weight was not manually defined, there would be misinterpretation in the sentiment weighting task, since those terms were not found in the lexicon.

There were 197 terms categorized as “FIX”, considered as typos, and needed to be formalized. Of all the terms, 316 terms consist of person name, place name, and special terms like “retaksasi” (reassessed collateral), “pinjaman rekening koran" (checking account based loan) categorized as “BNK” that are considered special terms in banking and finance. Moreover, about 520 terms were categorized as “KAT” or terms that could not be found in SWN 3.0 lexicon and their proper synonyms needed to be searched out in Kateglo[2] database. As the SWN 3.0 lexicon provided both positive and negative scores, the term polarity was defined by comparing the positive score and negative score. If the positive score is greater than negative score, the sentiment weight is 1, otherwise it is equal to -1 [7].

3.6. Evaluation

3.6.1. Silhouette Function

After performing the clustering task, the cluster evaluation was done by using silhouette function [15]. By using silhouette function, it would be easy to understand how good an object placed in a cluster is; therefore, the quality of a clustering task was ensured for the risk documents. The purpose of silhouette function is to replace the usage of variance analysis in the original paper of Pillar Algorithm since the variance analysis cannot describe the quality level of cluster result just like silhouette has, that is, $s \in [-1.00,1.00]$.

From the experiments that have been conducted, the values of *α* and *β* play a significant role in silhouette score. We noticed that the lowest values of *α* and *β* are 0.4 and 0.6, feasible to $2 \leq K \leq 10$. Any combination value lower than those combinations are only feasible to *K=2*. There are 715 cluster solutions, thus, it is hard to observe all the cluster solutions, so we decided to pick up the cluster solutions with the highest silhouette score as listed in Table 2.

Table 2. Best cluster solution for each *k*, based on silhouette score

K	*A*	*B*	*Silhouette Score*	*Number of Empty Cluster*
2	0.85	0.9	0.494237	0
3	0.75	0.85	0.660766	0
4	0.95	0.85	0.660766	1
5	0.65	0.75	0.642436	2
6	0.75	0.75	0.642436	3
7	0.7	0.8	0.701234	1
8	0.55	0.75	0.70496	1
9	0.95	0.75	0.574014	4
10	0.7	0.75	0.624935	3

From Table 2, it seems that K=8 is the best cluster solution to model the risk document based on the term relationship. Nevertheless, when we take a closer look, it has a cluster without any member in it, so we can conclude that the clustering task did not place the document properly. Then, the exploration continued with additional conditions to select the best cluster solution. Additional conditions are defined as follows: the best cluster solution is only selected (1) if it has no empty cluster and, (2) if it has no negative average silhouette score as figured in Figure 5.

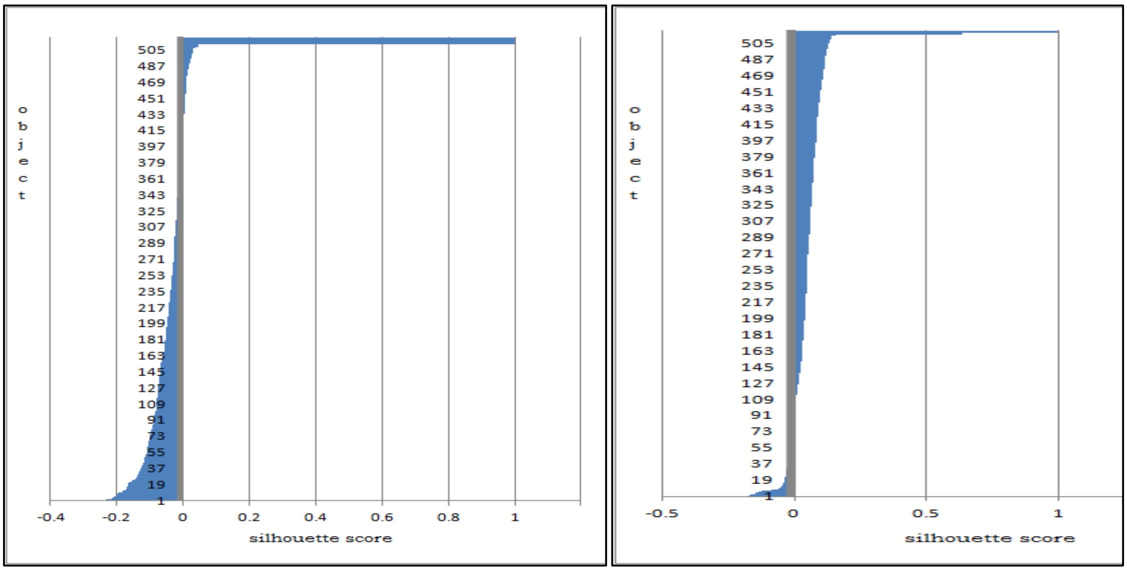


Figure 5. Comparison between bad cluster solution and negative average silhouette score (left), and good cluster solution without average negative silhouette score (right)

The second additional condition has been added because high silhouette score does not always represent a solution with good quality for each cluster within. For example, for *k* = 7, *α* = 0.65, and *β* = 0.8, from the silhouette score, they may be considered one of best cluster solutions, but when it is observed deeper, most of its objects have negative silhouette score. Our exploration found that the best cluster solution is *k* = 6, *α* = 0.65, and *β* = 0.8, because it has the highest silhouette score, *s* = 0.30205, and fulfills both additional criteria.

3.6.2. Sum Squared of Error

Table 3. List of the SSE of first top 5 cluster solutions for *k* = 6, and the best cluster solution

α	*β*	*Number of Cluster with Negative Avg Silhouette*	*Number of Empty Cluster*	*SSE*
0.75	0.75	1	3	7786.063
0.7	0.75	0	3	7802.287
0.6	0.8	2	0	7358.057
0.55	0.8	2	0	7358.057
0.5	0.8	2	0	7358.057
0.65	**0.8**	**0**	**0**	**3974.671**

This evaluation also helps to understand the nature of the cluster solution. It has been noticed that the greater the number of clusters in a cluster solution is, the lower the SSE is resulted. For instance, for k = 6, the top 5 cluster solutions are listed in Table 3 above. Table 3 empowers our reason to add the additional conditions since the cluster solutions that have not fulfilled both additional conditions tend to have higher SSE. Thus, those are not recommended as the best solutions, despite having high silhouette score.

3.6.3. Sentiment Analysis Performance

The sortation was limited up to 200 mostly presented terms which represent the character of the cluster, and Table 4 shows the most weighted terms (by selecting the terms with negative polarity then accumulating those TF-IDF and sentiment weight) in each cluster. To get proportional measurement, the Risk Score was gained by dividing the total term score with the total number of documents in the cluster. For instance, in cluster 1, the total term weight is -6262.070, and the total number of documents is 221; thus, the Risk Score is -28.335.

Table 4. Risk cluster analysis and its corresponding 5Cs criteria

Cluster (Rank)	Risk Analysis	Number of Documents	Risk Score	Corresponding 5Cs Criteria
1(2)	Related to collateral and asset (fix asset and current asset), e.g. “asset”, “piutang” (claim), “tanah”(land), “jaminan”(collateral)	221	-28.335	Collateral
2(3)	Related to income, e.g. “net profit”, “copat” (cash operating profit after tax), “pendapatan”(profit), “leverage” (gain and loss ratio)	213	-27.502	Capacity
3(4)	Related to production capacity, e.g. “persediaan”(stock), “kapasitas”(capacity), “penjualan”(sales)	47	-27.447	Capacity
4(5)	Related to both income and financial measurements, e.g. “net profit”, “copat” (cash operating profit after tax), “pendapatan” (profit), “leverage”(gain and loss ratio), “equity”, and “roe”(return of equity)	34	-26.102	Capacity
5(1)	Related to business condition, e.g. “persaingan” (business competition), “wilayah”(territory), “demonstrasi warga” (protest from local residents)	2	-35.0645	Condition
6(6)	Related to business financial measurement, e.g. “perputaran” (business cycle), “quick ratio”, “equity”, “return of equity”, “roe”(return of equity) and “profit”	2	-24.5485	Capacity

[1] POS Tagging is done by utilizing Pebahasa-Bahasa Indonesia POS Tagger Python API based on Hidden Markov Model. https://github.com/pebbie/pebahasa

[2] Kateglo, Kamus, Tesaurus, and Glossarium for Bahasa Indonesia, an online resource which provide an API for Bahasa Indonesia natural language processing http://kateglo.com

To indicate how big the risk is, the sum of TF-IDF score and negative sentiment score was accumulated in each cluster. While each cluster has unique characteristics, on the contrary, each of them has variation sector of business. This indicates that specific business does not always has a specific risk, so, the bank may be more aware and more thoroughly in analyzing every loan proposal that comes in, not treating it in the same way as analyzing previous proposals with the same business sector. The result also shows that the type of risk found in the cluster solution is related to four of 5Cs (Character, Capacity, Capital, Condition, and Collateral) Credit criteria [16] that are commonly used to make lending decision. The mostly found criterion in the corpus is Capacity, while Character and Capital are not considered too significant, as the top ranked terms do not reflect these criteria.

This result can be used by the bank to resharp the risk analysis since only three of the 5Cs criteria are exposed significantly in at least one cluster. The analysts may have difficulities in analyzing Character since it needs more in-depth investigation in the field. However, they must also improve the Character analysis since it is the most important criterion. Capital is the criterion that the analysts may rely on the scoring system, so that they will not be too concerned with delivering the opinions.

4. Conclusion

Through conducting this research, a model for risk analysis has been developed to help the bank, particularly the risk analysts and decision makers in understanding the risk exposure in their past and recent SME business financing customers. The results also show that cluster solution consisting of 6 clusters, is relevant to the risk evaluation, that is, 5C's Credit criteria. For future works, our model can be integrated with existing Loan Originating System, and expected to empower the decision making by benchmarking the risk opinion against the risk cluster.

Sentiment analysis is now dominated by industry to gain information from the market reception upon their products. We found that there are plenty of resources that are still left behind, have potential to be observed, and wait for researchers to use the as their research objects. The closed resources such as risk opinion document can be utilized to measure the level of risk within the SME business. Moreover, language is now no longer considered a barrier since there are many resources, research, and tools available to overcome this problem.

References

References

[1] Soares J, Pina J, Ribeiro M, Catalao-Lopes M. Quantitative vs. Qualitative Criteria for Credit Risk Assessment. Frontiers in Finance and Economics. 2011. 8(1):68-87.

[2] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. 2014. pp. 1093-1113.

[3] Xu K, Shaoyi S, Li J, Yuxia S. Mining comparative opinions from customer reviews for Competitive Intelligence. Decision Support Systems. 2011. 50:743-754.

[4] Ghiassi M, Skinner J, Zimbira D. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications. 2013. 40(16):6266-6282.

[5] Li N, Wu DD. Using text mining and sentiment analysis for online forums hotspot detection. Decision Support Systems. 2010. 48:354-368.

[6] Xu H, Zhai Z, Liu B, Jia P. Clustering Product Features for Opinion Mining. Proceedings of the fourth ACM international conference on Web search and data mining. ACM. . 2011. doi: 10.1145/1935826.1935884. 347:354.

[7] Esuli, Sebastiani. Sentwordnet: A publicly. International Conference on Language Resources and Evaluation (LREC). 2006. 1:417-422.

[8] Thelwall M. Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and Technology, Volume Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. 2010. 1:2544–2558.

[9] Gonçalves P, Benevenuto F, Araujo M, Cha M. Comparing and Combining Sentiment Analysis Methods. 2013. Conference on Online Social Networks (COSN). 1:27-38.

[10] Lunando E, Purwarianti A. Indonesian Social Media Sentiment Analysis with Sarcasm Detection. Advanced Computer Science and Information Systems (ICACSIS). 2013. 1:195 - 198.

[11] Denecke K. Using SentiWordNet for Multilingual Sentiment Analysis. International Council for Open and Distance Education Conference. 2008. doi: 10.1109/ICDEW.2008.4498370

[12] Barakbah AR, Kiyoki Y. A Pillar Algorithm for K-Means Optimization by Distance Maximization for Initial Centroid Designation. IEEE Symposium on Computational Intelligence and Data Mining (CIDM). 2009. 1:61-68. doi:10.1109/CIDM.2009.4938630.

[13]Manning CD, Prabakhar R, Schütze H. 2009. An Introduction of Information Retrieval. Cambridge University Press. Cambridge:118-119.

[14] Osinski SL, Stefanowski J, Weiss D. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. Master Thesis. Poznan University of Technology. Poznan; 2003.

[15]Rousseeuw PJ. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. 1986. 20:53-65.

[16] Gumparthi S. Risk Assessment Model for Assessing NBFCs' (Asset Financing) Customers. International Journal of Trade, Economics and Finance. 2010. 1(1):121-130.