ization-Sentiments-of-Analysis-from-Tweets-in-myXLCare-using.pdf

by

Submission date: 15-Dec-2020 03:52PM (UTC+0700)
Submission ID: 1475638439
File name: ization-Sentiments-of-Analysis-from-Tweets-in-myXLCare-using.pdf (796.31K)
Word count: 3245
Character count: 18097

Optimization Sentiments of Analysis from Tweets in myXLCare using Naïve Bayes Algorithm and Synthetic Minority Over Sampling Technique Method

IOP Conf. Series: Journal of Physics: Conf. Series **1471** (2020) 012014 doi:10.1088/1742-6596/1471/1/012014

Optimization Sentiments of Analysis from Tweets in myXLCare using Naïve Bayes Algorithm and Synthetic Minority Over Sampling Technique Method

Dedi Dwi Saputra[1], Windu Gata[2], Nia Kusuma Wardhani[3], Ketut Sakho Parthama[4], Hendra Setiawan[5], Sularso Budilaksono[6], Dimas Yogatama[7], Agus Hadiyatna[8], Endah Putri Purnamasari[9], Bryan Pratama[10], Deny Novianti[11]

[1]Master of Computer Science-Postgraduate Program, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, Indonesia
[2]Graduate School Master's Degree Computer Science, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, Indonesia
[3]Master Faculty of Computer Science-Postgraduate Program, University of Mercubuana, Jakarta, Indonesia
[4]Graduate School Master's Degree Computer Science, University Paramita Indonesia, Jakarta, Indonesia
[5]Graduate School Master's Degree Computer Science, STMIK Bani Saleh, Jakarta, Indonesia
[6]Graduate School Master's Degree Computer Science, University of Persada Indonesia YAI, Jakarta, Indonesia
[7]Graduate School Master's Degree Computer Science, University of Gadjah Mada, Jakarta, Indonesia
[8]Master of Computer Science-Postgraduate Program, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, Indonesia
[9]Master of Computer Science-Postgraduate Program, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, Indonesia
[10]Master of Computer Science-Postgraduate Program, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, Indonesia
[11]Master of Computer Science-Postgraduate Program, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, Indonesia

Abstract. Text mining can be used to classify opinions about complaints or not complaints experienced by XL customers. This study aims to find and compare classifications in the sentiments of analysis from the view of XL customers. This dataset was derived from tweets of XL customers written on myXLCare Twitter account. In text mining techniques, "transform case", "tokenize", "token filters by length", "n-gram", "stemming" were used to build classification and sentiments of analysis. Gataframework tools were used to help during pre-processing and cleansing processes. RapidMiner is used to help create the sentiment of analysis to search and compare two different classifications methods between datasets using the Naïve Bayes algorithm only and Naïve Bayes algorithm with Synthetic Minority Over-sampling Technique (SMOTE). The results of the two methods in this study found that the highest results were using the Naïve Bayes algorithm with Synthetic Minority Over-sampling Technique (SMOTE) with an accuracy of 86.33%, precision 82.85%, and recall ratio 92.38%. **Keywords**—Text Mining, Naïve Bayes Algorithm, SMOTE Method, Sentiments of Analysis, Twitter

IOP Conf. Series: Journal of Physics: Conf. Series **1471** (2020) 012014 doi:10.1088/1742-6596/1471/1/012014

1. Introduction

As a customer-oriented telecommunications service provider, currently PT XL Axiata, Tbk. uses social media as a form to collect concerns and complaints from its customers. Twitter is one of the most frequently used social media in expressing opinions or complaints experienced by XL customers. Twitter is a social media that allows users to send messages in real-time [1]. Opinions from customers in the form of Twitter data can be used to find out whether opinions are customers complaints or not. The purpose of the classification is to categorize a target class into the selected category. On the other hand, text mining is one of the techniques that can be used for classification. Text mining is a variation of data mining which tried to find interesting patterns from the collection of large amounts textual data [2]. Therefore, it is necessary to perform classification of sentiment analysis to determine whether an opinion is a complaint from XL customers on social media Twitter or not. Sentiment analysis or opinions mining is an understanding process of extraction and processing textual data to obtain information contained in opinion sentences. It is a process to get attractive designs and relationships and can applicable in large volumes of data [2]. Therefore, it is possible for the text mining method to be used for sentiment analysis purposes. RapidMiner was used to help create sentiment analysis to find and compare two different classification methods between datasets that only use the Naïve Bayes algorithm and the Naïve Bayes algorithm with the Minority Over-Sampling Technique (SMOTE). The objective and purpose of this study are to classify tweets from XL customers on Twitter. The classification process was performed by separating the tweets as complaints or not complaints through sentiment analysis approach using different classification methods between datasets using only Naïve Bayes Algorithms and datasets using Naïve Bayes algorithms with Minority Over-Sampling Technique (SMOTE).

2. Literature Study

Research on sentiment analysis of tweets on Twitter uses Arabic to analyze and predict the correct sentiment [3], [4]. Research [5] has proposed analytical sentiments from English tweets using RapidMiner. The RapidMiner approach is also carried out for analytical sentiment in business [6]. Other research also uses Twitter to carry out linguistic analysis and then build highly efficient classifiers [7], [8]. Consequently, the use of Twitter social media has been widely used for the analytical sentiment. The sentiment of analysis from Twitter must focus on classification problems [9], [10], [11]. Classification is the process of finding a model or function that can explain and distinguish a concept or class of data. Various classification studies using Naïve Bayes such as research [4], [12], [13], have been applied, and those who find the best results are research [4]. The example of sentiment classification is from Twitter media [4]; they used the classification of sentiment analysis using Naïve Bayes by stemming method. From the results of the study, an accuracy value of 83.17%, precision 79. 07 %, recall ratio 90.37% were obtained. In addition to RapidMiner, this research also uses web based Gataframework software. The Gataframework provides text preprocessing that can help in stemming the Indonesian language that can be accessed [14]. The features of the worksheet software used for this research are annotation removal, and transformation to remove URL, tokenization regular expression, Indonesia stemming, and Indonesian “stop” word removal. In the research [12], the Naïve Bayes classification problem with an unbalanced dataset could lead to additional/extra time to search a parameter value that is not optimal.

3. Methodology

3.1. Object

This research has an object about twitter post which related to the MyXLCare on Twitter.

3.2. Research Method

Research method that will be used is collecting tweet data. Tweet data is taken by the Crawling method from Twitter. The data taken is only tweets in Indonesian, which is 2000 tweets with the words MyXLCare. Data are taken randomly from either ordinary user accounts or online media accounts on Twitter.

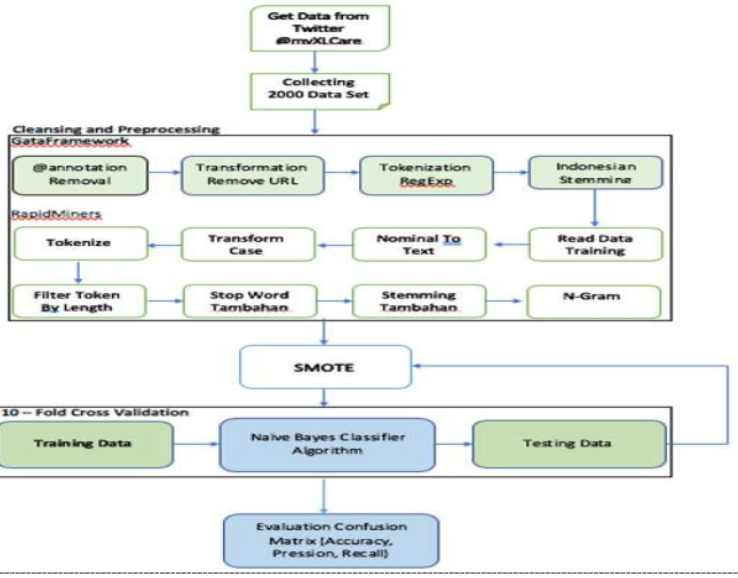

Figure 1. Research Method for Classification Sentiments Analysis.

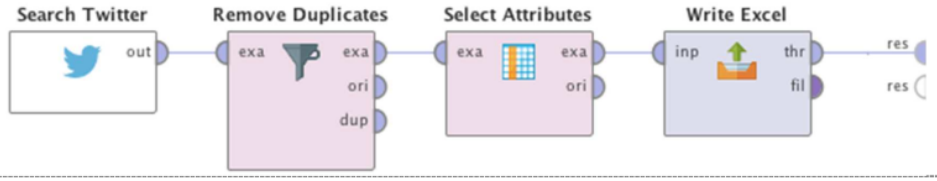

Figure 2. Get Data Retriveal from Twitter.

4. Result and Discussion

4.1. Collecting and Labeling Data

The first stage of the sentiment analysis process is collecting data. Data taken from Twitter with the search for **"@myXLCare"** gets 2000 datasets using RapidMiner applications. Figure 2 shows the retrieval data from social media Twitter using the operator **"Search Twitter,"** and then the operator **"Remove Duplicates"** is used to delete the same tweets, to make them a unique tweet. The **"Select Attribute"** operator is used to retrieve data that is text-only and save the data into an excel file using the **"Write Excel"** operator. The next step is labeling. Labeling function is to divide the data into several sentiment classes that will be used. The number of sentiment classes used is two classes, namely "Complaint" or "Not Complaint." The purpose of this labeling process is to divide the dataset into two parts, namely training data and testing data. Training data is data used to train the system to recognize the pattern being sought, while testing data is data that is used to test the results of the training that has performed.

4.2. Data Preprocessing

At this stage, which is preparing data for pre-processing, is the stage where data is prepared for the analysis. This stage also uses two pre-processing applications. The first application uses the Gataframework accessible via [14]. Pre-processing when using Gataframework, namely "@annotation removal", "Transformation URL", "Tokenize RegExp", "Normalization Emoticon", "Indonesian Stemming", "Indonesian Stop Word Removal".

4.3. Annotation Removal

Text is parsed based on white space. In this process, all the annotations contained in the tweet will be removed and change the entire capital letter to lowercase:

Table 1 - Example of Annotation Removal

Before	After
@myXLCare Simcard xl baru. Kok tidak bisa dipakai buat menyimpan nomor baru ya? Mohon solusi.	myXLCare Simcard xl baru. Kok tidak bisa dipakai buat menyimpan nomor baru ya? Mohon solusi.

4.4. Transformation Remove URL

It is performed after the @annotation removal process is carried out. In this process, the link or URL contained in the tweet is omitted:

Table 2. Example Transformation Remove URL

Before	After
@Ini hp internet ga aktif lg udah 1 thn belakang. tp kenapa pulsa berkurang terus yaa. padahal mobile datanya off. @myXLCare @myxl https://t.co/cngp3wAYvS.	*Ini hp internet ga aktif lg udah 1 thn belakang. tp kenapa pulsa berkurang terus yaa. padahal mobile datanya off.*

4.5. Regular Expression Tokenization

This step is completed after the transformation removes URL process, followed by the regular expression tokenization process. Where in this process, all the words in each document are collected to remove punctuation such as periods (.) and commas (,). This process also removes symbols, special characters, and anything that is shaped with letters:

Table 3. Example of Tokenization Regular Expression

Before	After
myXLCare Simcard xl baru. Kok tidak bisa dipakai buat menyimpan nomor baru ya? Mohon solusi.	*myxlcare simcard xl baru kok tidak bisa dipakai buat menyimpan nomor baru ya mohon solusi*

4.6. Indonesian Stemming

In the next stage, the results of regular expression tokens will be stemming. In the stemming stage, the word with a conjunction is changed to become the basic word for tweets in Indonesian:

Table 4. Example of Indonesian Stemming

Before	After
*mXLCare Simcard xl baru. Kok tidak bisa **dipakai** buat **menyimpan** nomor baru ya? Mohon solusi.*	*myxlcare simcard xl baru kok tidak_bisa **pakai** buat **simpan** nomor baru ya Mohon solusi*

4.7. Indonesian Stopword Removal

Next stage is the removal of Indonesian stop word. The process involves deletion of the words which are not relevant, as the word **"tetapi"**, **"ke"**, **"dengan"**. These words are without meaning if separated by other words and are not related to adjectives related to sentiment:

Table 5. Example of Indonesian Stopword Removal

Before	After
*admin **mau tanya saya pengguna** xl priority **hari ini baru aja** reset **tp kenapa** langsung sisa **k** remaining balance**nya ya***	*admin xl priority reset langsung sisa remaining balace.*

4.8. Transform Cases

At this stage, the Indonesian stop word results are continued by the Transform Cases process from RapidMiner. It is used to convert all words to uppercase letters:

Table 6. Example of Transform Cases

Before	After
myXLCare Simcard xl baru. Kok tidak bisa dipakai buat menyimpan nomor baru ya? Mohon solusi.	*myxlcare simcard xl baru kok tidak bisa dipakai buat menyimpan nomor baru ya mohon solusi*

4.9. Tokenization

At this stage, the Indonesian stop word results are continued by the Tokenization process from RapidMiner. It is used so that all the words in each document are collected and omitted punctuation, and omitted if there are symbols or special characters that are not letters and break the sentence into words:

Table 7. Example of Tokenization

Before	After
myXLCare Simcard xl baru. Kok tidak bisa dipakai buat menyimpan nomor baru ya? Mohon solusi.	*myxlcare simcard xl baru kok tidak bisa dipakai buat menyimpan nomor baru ya Mohon solusi*

4.10. N-gram Generation

The results of the tokenization continue in the last process, i.e., Generate N-gram. Generate N-gram is an operator having the function to collect the words given in a paragraph and calculate n-grams by moving one word forward. In this stage, it uses bi-gram:

Table 8. Example of Generate N-Gram

Before	After
myxlcare simcard baru kok tidak bisa pakai buat simpan nomor baru mohon solusi.	*myxlcare myxlcare_simcard simcard simcard_baru baru baru_kok kok kok_tidak tidak tidak_bisa bisa bisa_pakai pakai pakai_buat buat buat_simpan simpan simpan_nomor nomor nomor_baru baru baru_mohon mohon mohon_solusi solusi*

4.11. Weighting Word

Weighting feature is a process of giving value to each feature based on the relevance and its influence on the results of the classification. This value can later be used as a basis for feature selection based on the minimum weight calculated from each feature. Weighting is done using the TF-IDF method (Frequency's Term Frequency- Inverse Document). TF-IDF algorithm is one algorithm in the feature weighting method in text mining. TF-IDF has the following formula:

$$TF - IDF = TF\ X\ IDF$$

4.12. Synthetic Minority Over-Sampling Technique (SMOTE)

In research [13] SMOTE balances data sets by synthesizing minority data synthetically in the input room based on their environmental information. Training datasets consist of minority data points (Smin) and majority data points (Smin). For each (Xi, Yi) ∈ Smin, the most data points set (Smaj).

For each (Xi, Yi) ∈ Smin, SMOTE generates a new minority data point along the line segment that joins (Xi) and one of the closest neighbors chosen randomly. SMOTE has the following formula:

$$\chi_{new} = \chi_i + (\chi_j - \chi_i) \times \delta$$

4.13. Naïve Bayes Classifier

Classification is a learning function that classifies an element of data into one of several defined classes. One classification method that can be used is the Naïve Bayes method, which is also often called Naïve

Bayesian Classification. Naïve Bayes is a learning algorithm based on Bayes theory using strong (Naïve) assumptions.
The essence of Naïve Bayes is to find the highest probability of data. The Bayes formula is as follows:

$$Pcd = \frac{P(c)X\ Pdc}{P(d)}$$

Remarks Formula:
Ped is the probability of class c after entering class c.
P (c) is the previous class c probability.
Pdc is d probability in class c.
Pd is d probability.

4.14. Results and Performance Analysis

This section explains the results of experiments and analyzes their performance results:

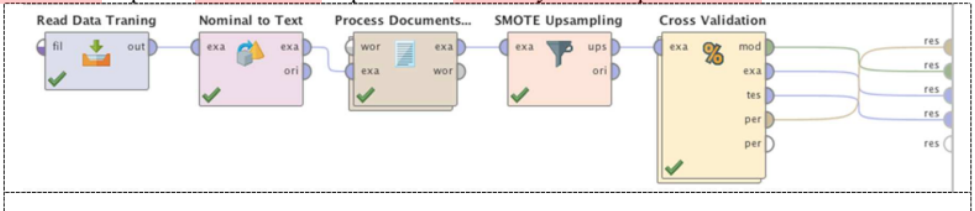

Figure 3. Primary Process on RapidMiner Naïve Bayes Algorithm with Synthetic Minority Over-sampling Technique (SMOTE) Method.

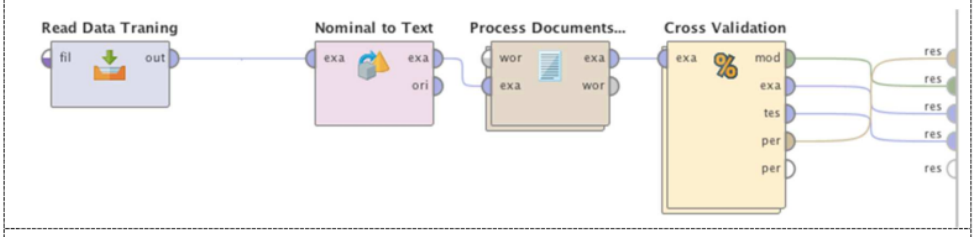

Figure 4. Primary Process on RapidMiner Naïve Bayes Algorithm without Synthetic Minority Over-sampling Technique (SMOTE) Method.

Figure 3 shows the main processes in the RapidMiner application using SMOTE. The "Read Data Training" operator is used to read data in an excel file. The operator "Nominal to Text" is used for the nominal attribute type selected for text, this operator also maps all attribute values to the appropriate string values. Operator "Document Process" for pre-processing data. Operator "Up sampling SMOTE" is used so that the data becomes balance. And for the "Cross Validation" operator, it is used for classification and evaluation of sentiment analysis with ten-fold cross validation experiments. Figure 4 shows the main processes in the application RapidMiner without SMOTE method. The "Read Data Training" operator is used to read data in an excel file. The operator "Nominal to Text" is used for the nominal attribute type selected for text, this operator also maps all attribute values to the appropriate string values. Document Processing" operator to pre-process data. And for the "Cross Validation" operator, it is used for classification and evaluation of sentiment analysis with ten-fold cross-validation experiments.

Table 9. Confusion Matrix Naïve Bayes with SMOTE Method

Method	TP	FP	TN	FN
Naïve Bayes Algorithm + SMOTE	267	22	232	57
Naïve Bayes Algorithm	156	47	232	57

Based on the research data that has been done with the Confusion Matrix in Table 9 The values of average accuracy, precision, and recall are shown in Table 10 as follows:

Table 10. Value of Accuracy, Precision, and Recall

Method	Accuracy	Precision	Recall
Naïve Bayes Algorithm + SMOTE	86.33%	82.58%	92.38%
Naïve Bayes Algorithm	78.88%	74.43%	76.90%

5. Conclusion

From the results of this study indicate that the Naïve Bayes Algorithm if it is optimized by using feature Synthetic Minority Over-sampling Technique (SMOTE) the method gets Accuracy value of 86.33%, Precision 82.58% and Recall 92.38%. And based on the results of the study also shows that the calculation of the Naïve Bayes algorithm without being optimized by the Synthetic Minority Oversampling Technique method gets the Accuracy value of 78.88%, Precision 74.43% and Recall 76.90. Sobased on the results of this study, it can be concluded that the optimization of the Naïve Bayes Classifier Algorithm with feature Synthetic Minority Over-sampling Technique (SMOTE) is better classifier. That can be used, using a tweet dataset from XL customers via @myXLCare social Twitter media. Compared, using Algorithm Classifier Naïve Bayes only. And it can be inferred from these results Naïve Bayes algorithm with Synthetic Minority Over-sampling Technique (SMOTE) method can classify tweets "Complaint" or "Not Complaint "from the tweet of XL customers.

6. References

[1] R. Passonneau, “Sentiment Analysis of Twitter Data,” no. June, pp. 30–38, 2011.
[2] M. Ronen Feldman, Bar-Ilan University, Israel , James Sanger, ABS Ventures, Boston, *The Text Mining Handbook*. Cambridge University Press, 2006.
[3] S. Alotaibi, “Sentiment Analysis Challenges of Informal Arabic Language,” vol. 8, no. 2, pp. 278–284, 2017.
[4] M. A. Ibrahim and N. Salim, “Sentiment Analysis of Arabic Tweets : With Special Reference Restaurant Tweets,” vol. 4, no. 3, pp. 173–179, 2016.
[5] P. Tripathi, S. K. Vishwakarma, and A. Lala, “Sentiment Analysis of English Tweets Using RapidMiner,” 2015.
[6] J. Ahmed, “Sentiment Analysis and Classification of Tweets Using Data Mining,” pp. 4–7, 2017.
[7] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” no. December, 2013.
[8] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” vol. 2, no. 1, 2008.
[9] M. Singh, “Sentiment Analysis and Similarity Evaluation for Heterogeneous-Domain Product Reviews,” vol. 144, no. 2, pp. 16–19, 2016.
[10] A. Bifet and E. Frank, “Sentiment Knowledge Discovery in Twitter Streaming Data.”
[11] J. Isabella, “Analysis and evaluation of Feature selectors in opinion mining,” vol. 3, no. 6, pp. 757–762, 2013.
[12] D. Liliya and K. Irina, “Improving the Classification Quality of the SVM Classifier for the Imbalanced Datasets on the Base of Ideas the SMOTE Algorithm,” vol. 02002, pp. 8–11, 2017.
[13] J. Mathew, M. Luo, C. K. Pang, and H. L. Chan, “Kernel-Based SMOTE for SVM Classification of Imbalanced Datasets,” pp. 1127–1132, 2015.
[14] W. Gata, “Gataframework,” *gataframework*, 2017. [Online]. Available: http://www.gataframework.com/textmining/.

.

ization-Sentiments-of-Analysis-from-Tweets-in-myXLCare-using.pdf

ORIGINALITY REPORT

19%	0%	19%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

11%

★ Achmad Bayhaqy, Sfenrianto Sfenrianto, Kaman Nainggolan, Emil R. Kaburuan. "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes", 2018 International Conference on Orange Technologies (ICOT), 2018

Publication

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		