Building automatic customer complaints filtering application based on Twitter in Bahasa Indonesia

D Gunawan*, R P Siregar¹, R F Rahmat¹ and A Amalia²

¹Department of Information Technology, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155
²Department of Computer Science, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

*Email: danigunawan@usu.ac.id

Abstract. Twitter has become a media to provide communication between a company with its customers. The average number of Twitter active users monthly is 330 million. A lot of companies realize the potential of Twitter to establish good relationship with their customers. Therefore, they usually have one official Twitter account to act as customer care division. In Indonesia, one of the company that utilizes the potential of Twitter to reach their customers is PT Telkom. PT Telkom has an official customer service account (called @TelkomCare) to receive customers’ problem. However, because of this account is open for public, Twitter users might post all kind of messages (not only complaints) to Telkom Care account. This leads to a problem that the Telkom Care account contains not only the customer complaints but also compliment and ordinary message. Furthermore, the complaints should be distributed to relevant division such as “Indihome”, “Telkomsel”, “UseeTV”, and “Telepon” based on the content of the message. This research built the application that automatically filter twitter post messages into several pre-defined categories (based on existing divisions) using Naïve Bayes algorithm. This research is done by collecting Twitter message, data cleaning, data preprocessing, training and testing data, and evaluate the classification result. This research yields 97% accuracy to classify Twitter message into the categories mentioned earlier.

1. Introduction

Customer care holds an important role in a company. Nowadays, along with the development of information technology, there are alternative ways to communicate between the company and customer. For example, Twitter has become the alternative communication to the customer care division as the number of its active users reach 330 million monthly (as per third quarter of 2017) [1]. Usually, a company has one official Twitter account to communicate between the customer and the company. In addition to the convenience offered by using Twitter, it has a drawback. For a well-known company, the twitter posts can reach thousands a day. The problem is the twitter posts to the customer care division are not only customer complaints, but also other twitter posts that are not related to customer care. This requires more manpower to manually select the related twitter posts manually. Therefore, the company requires an application that can automatically filter twitter posts to related division in customer care.

This research uses a twitter account called @TelkomCare which is owned by a well-known telecommunication company in Indonesia, PT Telkom. It has four divisions to handle the customer complaints such as “Indihome”, “Telkomsel”, “UseeTV”, and “Telepon”. This research will filter the
Naïve Bayes algorithm has been used in many applications or researches. For example, previous research [2] used Naïve Bayes to implement focused crawler. This algorithm was used to classify the web contents in order to obtain a specific topic. Later, the focused crawler was improved with distributed system to obtain more contents [3]. Both researches used bag of words method to classify the contents. Meanwhile, as in our research the number of the same keywords does not influence the meaning of the twitter post, then we do not calculate the number of keywords appearance to classify the contents. However, we do not consider the usage of multiword expression [4] to extract the keywords as used in content clustering research [5].

Previous study classified twitter users based on their posts into three classes, such as political, ethnical or business-oriented [6]. They used Gradient Boosted Decision Trees (GBDT) to classify the twitter posts. This research yield 76% accuracy.

Other study used Naïve Bayes classifier to filter twitter timeline (front page of each twitter account) to five categories such as news, opinions, deals, events and private messages [7]. This research used eight features selection to eliminate twitter posts which are not written in English, less than three words and only contain URL address without other information. The accuracy of the system with eight features selection reached 90%.

Recently, a study also used Naïve Bayes algorithm to classify emotion in twitter posts based on the contents [8]. There are six categories of emotion such as happiness, fear, sadness, anger, surprise, and disgust. The evaluation result of this research is 83.10%.

This paper is organized as follow: the first section is the introduction about the background of the conducted research. Then, section two discusses the research methodology to obtain the result. Section three discusses the result of this research and the last section is conclusion.

2. Research Methodology
This research was done by conducting several steps. The first step is collecting and analyzing the data sources. In this step we examined the pattern of twitter post by the users. After collecting and analyzing the data, the next step is extracting keywords to determine each customer care division. Next, we implement text pre-processing to twitter post in order to match the required format for data training. After implementing text pre-processing, we clean the unwanted data such as the twitter post which has only two or three words. Then the data will be divided into two types, data training and data testing. The last step, we built the application to show the demonstration of the twitter post classification. The details of the research methodology are as follow:

2.1 Data Source
We obtained the data from Telkom Care twitter that contain the posts from Telkom subscribers through Twitter. The data is divided into two parts, such as training data and testing data. Training data is used to train the system in order to classify new tweets which are not defined in certain class yet. Testing data is used to test the system in order to confirm the classification method. To define the classification label, we use several keywords that are determined by manual observation.

This research uses 120 tweets (user posts in Twitter) for training data. The training data are obtained from the official account of customer care PT Telkom Indonesia, @TelkomCare. The data are obtained by using Streaming Twitter API method. We were collecting tweet for 60 minutes.

After collecting tweets from the official account of PT Telkom, the next step is class definition. We obtained 250 tweets in 60 minutes. All this tweets are mixed between the potential data training or not. We selected 120 unique tweets manually which have potential content to represent the classification.

After manually selected the potential data training, we divided them into several categories according to the current condition of PT Telkom division. There are four classes such as Indihome, UseeTV, Telkomsel, Telepon. The number of training data for each class is shown in table 1. We did not use the
data training for uncategorized tweets. The tweet that is not categorized in one of the four classes above will automatically defined as uncategorized. The example of data training is shown in table 2.

| Class       | Number of Tweets |
|-------------|------------------|
| Indi Home   | 39               |
| UseeTV      | 27               |
| Telkomsel   | 26               |
| Telepon     | 27               |

Table 2. Tweet pre-defined classification

| Tweet                                                                 | Category |
|-----------------------------------------------------------------------|----------|
| @TelkomCare cara berlangganan indihome gimana?                         | 39       |
| @TelkomCare harga untuk paket telkomsel flash reguler 1 gb berapa ya?  | 27       |
| @TelkomCare min ini iflix gabisa dibuka yg di tv saya. Disuruh update. Gmana update nya klo di tv? | 26       |
| @TelkomCare mohon dibantu saat ini saluran telpoon di tempat saya tidak ada tone sehingga speedy di tempat saya mati. | 27       |

2.2 Extracting Keywords

Keywords is the parameter used to train the system in determining the suitable class for each tweet. This research had through several steps to determine the keywords. The first step was discussion with the twitter administrator about the words which would be the reference to determine subscriber’s complaint. As the result, there are four main keywords to represent each product, such as Indihome, UseeTV, Telkomsel, and Telepon.

The first implementation of the classification using these keywords, the result is still messy. Generally, every tweet will include one of those keywords. However, there are a lot of cases, that the system cannot classify it because of lack of the keywords inside the tweet. And also, there are tweets classified as uncategorized, although if we observe the tweet semantically, it is included in one of the four classes. Therefore, other keywords are required to handle the issue.

The next step was conducting a new research by pulling the tweets as much as possible from the users who mention Telkom account. After that, the next step is extracting frequently used words which is potentially used to be supporting main keywords to determine tweet classification. From this process we obtain 202 keywords. By adding these keywords, we obtain much better result than the first implementation. However, for certain cases there are tweets which contain strong keywords. To solve the problem, the system requires a learning machine. In this research we implement Naïve Bayes classifier.

2.3 Text Pre-processing

Preprocessing step is required before we conduct training dataset and testing dataset. There are some steps to minimize noise by making the structure uniform and reducing widen dimension. After the dataset pass through some pre-process steps, then the dataset will be represented in matrix. The whole pre-processing steps are filtering, case folding, tokenizing, stemming, and normalization. Those are explained below:
A. Filtering
Twitter is an expressive social media. It means Twitter allows users to insert symbols such as emoticon and special characters in a post to express their feelings. These expressive symbols tend to cause the wrong interpretation in classification system. Thus, the special characters should be eliminated before processing the classification. For example, the characters such as # and @ are common characters in Twitter.

B. Case folding
Each twitter posting may contain uppercase or lowercase. This condition will influence text pre-processing because between “A” and “a” will be considered as different character. To overcome this condition, all the characters will be converted to lowercase.

C. Tokenizing
Tokenizing is the process to extract the tweet to separated words. This process is done as a preliminary process to obtain the keywords of each tweet. We use space as delimiter to split the tweet to words.

D. Stemming
Each token in tweet may contain the same root word in different format. For example, the root word may have suffix. These words may be recognized as different word because of different word form. To overcome this situation, we have to convert all words in to its root word. We use existing stemming algorithm for Bahasa Indonesia called Enhanced Confix Stripping Stemmer [9]. For example, the words “berlangganan” and “ditagihkan” become “langgan” and “tagih”.

E. Spelling normalization
After normalization process is completed, the next step is the keyword search from the extracted tokens. Tokens will be checked with the previously defined keywords. If the token matches the keywords, then it will be served for the next process. The token that does not match the existing keywords will be eliminated. We found that there is a problem in matching the token with the available keywords. Some tokens are typo or abbreviated so that they are not perfectly match with the keywords. To overcome this problem, we did the spelling normalization by implementing levenshtein algorithm. This algorithm will test the similarity between the token with the keywords.

| No. | Keywords | Label | Category (Probability) |
|-----|----------|-------|------------------------|
| 1   | saldo    | x1    | Telkomsel              |
| 2   | indihome | x2    | Indihome               |
| 3   | maintenance | x3 | Indihome, Telepon, Telkomsel, UseetV |
| 4   | telepon  | x4    | Telepon                |
| 5   | jaringan | x5    | Indihome, Telepon, Telkomsel, UseetV |
| 6   | useetv   | x6    | UseetV                 |
| ... | ...      | ...   | ...                    |
| 201 | speed    | x201  | Indihome, Telkomsel, UseetV |
| 202 | grapari  | x202  | Telkomsel              |
2.4 Data Modeling
All the training and testing data are required to be modeled into matrix in order to form structured data. The structured data will avoid the mistake in calculation. The example of data modeling for training data is shown in table 3. The data modeling for testing data is similar with the example in table 3, unless the value in the class column is not determined yet.

2.5 Classifier
The existing problem is the customer complaint posts are divided manually by the human. This process consumes a lot of time because there are also many posts that are not related to the complaints. This will cause the complaints are not handled as soon as possible by the customer care staffs. There are many existing classifiers that can classify the posts. This research uses Naïve Bayes Classifier because of the its speed in classification process.

2.6 Evaluation
To evaluate the result of classification method, we use accuracy, precision, recall and F-score. As shown in formula (1), the accuracy is obtained by dividing the total number of correct classification data with the total number of data testing.

\[
Accuracy = \frac{\text{total number of correct classification}}{\text{total number of data testing}} \times 100\% \tag{1}
\]

Precision and recall can be mapped in matrix as shown in table 4. According to the matrix, recall and precision are formulated as in (2) and (3).

| Found | Relevance | Irrelevance | Total |
|-------|-----------|-------------|-------|
| Found | a (hits)  | b (noise)   | a+b   |
| Not found | c (misses) | d (rejected) | c+d   |
| Total | a+c       | b+d         | a+b+c+d |

\[
Recall = \frac{a}{a + c} \times 100\% \tag{2}
\]

\[
Precision = \frac{a}{a + b} \times 100\% \tag{3}
\]

3. Result and Discussion
3.1 Classification Result
This research uses 660 twitter posts which mention @telkomcare as the data testing. These data testing are analyzed manually by human to obtain the actual classification. We found that Indihome category has 198 posts, UseeTV category has 122 posts, Telkomsel category has 110 posts, Telepon category has 110 post and there are 120 uncategorized posts. Next, we implemented text-preprocessing to all the data testing. Finally, we applied Naïve Bayes classifier to obtain the automatic classification of twitter posts. The result is shown in table 5.
According to table 5, the system identifies 180 posts in Indihome category, 120 posts in UseeTV category, 120 posts in Telkomsel category, 120 posts in Telepon category and 120 posts in Uncategorized. However, from the manual analysis, we found that 178 posts are related to Indihome category, 118 posts are related to UseeTV category, 110 posts are related to Telkomsel, 112 posts are related to Telepon category and 120 posts are uncategorized. The example of irrelevant posts classification is shown in table 6.

Table 5. Classification result

| Category     | Relevant | Irrelevant | Number of classified posts | Actual number of the data |
|--------------|----------|------------|----------------------------|---------------------------|
| Indihome     | 178      | 2          | 180                        | 198                       |
| UseeTV       | 118      | 2          | 120                        | 122                       |
| Telkomsel    | 110      | 10         | 120                        | 110                       |
| Telepon      | 112      | 8          | 120                        | 110                       |
| Uncategorized | 120      | 0          | 120                        | 120                       |

After analyzing the classification result, we found 20 twitter posts are identified as “false positive”. This result is influenced by the keywords and the number of data training. For example, the classification result in the table 6 shows that the twitter post in the third row is classified as Telkomsel category. However, the actual classification is Indihome. This is influenced by the number of keywords between the Telkomsel and Indihome, such as “internet”, “paket”, “indihome”, and “telkomsel”. Those keywords will produce almost the same probability between Telkomsel and Indihome category. The number of data training will influence Naïve Bayes algorithm to decide the classification result.

3.2 Evaluation
The system performance can be evaluated by using accuracy, precision, recall and F-score. According to the classification result, we calculate the accuracy according to the formula (1) as follow:

$$Accuracy = \frac{636}{660} \times 100\% = 96.3\%$$

As mentioned earlier, to determine the actual classification we use manual classification approach. This manual classification is obtained by discussion and interview with the person in charge. The result of evaluation result in table 7 shows that the average of precision value is 97%, recall is 97.6% and F-
score is 97.2%. According to the theory by Pao [10], this system can be determined as an ideal system because of the ratio of recall and precision have nearly the same value.

| Category   | Relevant (a) | Irrelevant (b) | Total (a+b) | Not found (c) | Total (a+c) | Recall [a/(a+c)] x100% | Precision [a/(a+b)] x100% | F-score |
|------------|--------------|----------------|-------------|---------------|-------------|------------------------|---------------------------|---------|
| Indihome   | 178          | 2              | 180         | 18            | 192         | 91%                    | 99%                       | 95%     |
| UseeTV     | 118          | 2              | 120         | 2             | 122         | 97%                    | 98.3%                     | 98%     |
| Telkomsel  | 110          | 10             | 120         | 0             | 110         | 100%                   | 92%                       | 96%     |
| Telepon    | 112          | 8              | 120         | 0             | 112         | 100%                   | 93.4%                     | 97%     |
| Uncategorized| 120          | 0              | 120         | 0             | 120         | 100%                   | 100%                      | 100%    |
| **Average**|              |                |             |               |             | **97.6%**              | **97%**                   | **97.2%**|

4. Conclusion

According to the result, this application is able to classify twitter posts into four main classes, namely “Indihome”, “UseeTV”, “Telkomsel” and “Telepon” with the accuracy of 97%. This research yields 97% and 97.6% of the precision and recall average respectively. According to the precision and recall, the F-score calculation yields 97.2%. This means this system can be determined as an ideal system because of the precision and recall almost have the same value.

References

[1] Statista 2017 Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2017 (in millions) Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[2] Amalia A, Gunawan D, Najwan A, and Meirina F 2016 Focused crawler for the acquisition of health articles 2016 Int. Conf. on Data and Software Engineering (ICoDSE) pp 1–6

[3] Gunawan D, Amalia A, and Najwan A 2017 Improving Data Collection on Article Clustering by Using Distributed Focused Crawler J. Comput. Appl. Informatics I pp 39–50

[4] Gunawan D, Amalia A, and Charisma I 2016 Automatic extraction of multiword expression candidates for Indonesian language 6th IEEE Int. Conf. on Control System, Computing and Engineering (ICCSCE) pp 304–309

[5] Gunawan D, Amalia A, and Charisma I 2017 Clustering Articles in Bahasa Indonesia Using Self-Organizing Map 2017 Int. Conf. on Electrical Engineering and Informatics

[6] Pennacchiotti M and Popescu A 2011 A Machine Learning Approach to Twitter User Classification Proc. of the Fifth Int. AAAI Conf. on Weblogs and Social Media pp 281–288

[7] Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, and Demirbas M 2010 Short Text Classification in Twitter to Improve Information Filtering SIGIR ’10 Proc. of the 33rd int. ACM SIGIR conf. on Research and development in information retrieval pp 841–842

[8] Wikarsa L and Thahir S N 2015 A text mining application of emotion classifications of Twitter’s users using Naive Bayes method 2015 1st Int. Conf. on Wireless and Telematics (ICWT) pp 1–6

[9] Adriani M, Asian J, Nazief B, and Tahaghoghi S M M 2007 Stemming Indonesian: A confix-stripping approach ACM Trans. Asian Lang. Inf. Process. 6 4 pp 1–33

[10] Pao M L 1989 Concepts of Information Retrieval Libraries Unltd Incorporated