# ANDROID SHORT MESSAGES FILTERING FOR BAHASA USING MULTINOMIAL NAIVE BAYES

Shaufiah, Imanudin and Ibnu Asror
Data Mining Center Laboratory, School of Computing, Telkom University, Bandung, Indonesia
E-Mail: shaufiah@telkomuniversity.ac.id

## ABSTRACT

The presence of Short Message Service (SMS) that indicate fraud acts is rising and very disturbing for SMS users which is known as spam SMS. Therefore, it is very important to automatically detect or filter spam SMS. This research developed a system that could classify SMS between SMS spam with not spam (ham) in Bahasa (Indonesian Language). This system conducted with Multinomial Naïve Bayes classification with the feature weighting Term Frequency - Inverse Document Frequency (TF-IDF). Before the classification, data had been preprocessed using tokenization, slang handling, stopword, and stemming. The evaluation is done by using cross validation and were conducted by comparing several test scenarios based on the selected preprocessing technique. From the experiment the best results were obtained 94.44% in accuracy with preprocessing slang handling and stemming. This best result were implemented on the mobile Android with adding rule if the sender of SMS is not in the contact list, then the incoming SMS would be processed to test whether it is spam or ham. From the experiment on Android mobile application accuracy raised until 94.74%.

**Keywords:** SMS Spam, classification, multinomial naïve bayes, term frequency - inverse document frequency, android.

## 1. INTRODUCTION

Today, SMS is one commonly used to communicate. Along with the increased intensity of its use, SMS oftenly misused by irresponsible people to commit crimes such as fraud via SMS. The rise of these scams resulted in insecurity and inconvenience for the SMS recipient. Those kinds of SMS is known as spam SMS.

Spam SMS is indicated by some criterias such as its containing a demand of mobile phone credit to a specific number on behalf of mother or father and other schemes. SMS scams that are circulating widely in the community tend to have a certain pattern. The problem is that people usually have lack of knowledge about it and got fooled by the SMS.

Therefore, this research built an automatic SMS spam filter to avoided undesired SMS. This automatic SMS spam filter is conducted by identifying patterns of SMS whether it is spam or ham so that people can be more cautious to follow up the SMS received and crimes committed via SMS could be avoided.

Data mining could be addressed sms spam filtering problem using classification task. There are so many previous studies on the classification to filter SMS spam with various techniques such as SVM, Naïve Bayes, ID3, and C45. But the effectiveness of each technique varies and mostly used for SMS in English, Spanish and other Languages. Therefore there is chance to examined SMS spam filtering specially in Bahasa (Indonesian Language).

Another challenge in the process of identification and classification of SMS spam is its unstructured, open and irregular form due to abbreviation and slangs. So that the data need to be preprocessed beforehand. The preprocessing held in order to clean or simplify the data without changing the information its contains, so that the computation time in the classification process [1].

Based on study[2]Multinomial Naïve Bayes algorithm has the highest level of effectiveness that reached 98.2% compared with other methods on classification. From these facts Multinomial Naïve Bayes is chosen to be applied in SMS spam filter classification which combined with some preprocessing techniques.

## 2. THEORETICAL BASE

Spam SMS is part of spam that involves the use of text messages sent to cell phones via short message service. Spam SMS usually contains unexpected promotions, prize draws, or fraud by the receiver from random target[3].

### 2.1 Classification

Classification is the process of finding a set of models or functions that describe and distinguish classes of data with the aim of predicting the class of an unknown object class (supervised learning). The classification process is divided into two phases, which are learning and testing. In the learning phase, data's that has been known class of data (training set) are used to build the model. Later in the test phase, the formed model is tested with most other data to determine the accuracy of the model. If the accuracy is sufficient enough, then the model can be used to predict the class of unknown data[4], [5].

### 2.2 Feature weighting

TF-IDF algorithm was first proposed by Salton and Buckley in 1988 and used for information retrieval, which later participated as one of the algorithms used in the method of feature weighting in text mining. TF-IDF has the following formula[6]:

$$TFIDF(word) = \log(tf_{ij} + 1)\ x\ idf_i \tag{1}$$

The formula can be translated into term frequency of feature i in document j multiplied by the IDF of the feature i, where the IDF stands for Inverse Document Frequency. IDF itself can be calculated by [6]:

$$IDF_i = \log(\frac{D}{df}) \tag{2}$$

*D*is total amount of document while df is number of documents contains feature i.

### 2.3 Multinomial Naïve Bayes

Multinomial Naïve Bayes algorithm is an algorithm developed from naïve Bayes classifier theorem. This algorithm uses multinomial distribution on conditional probabilities function. Although using multinomial distribution, this algorithm can be applied for case of text mining by changing the text data into a form that can be calculated with the nominal value of the integer.

Generally, the Multinomial Naïve Bayes algorithm for text classification cases where document d can be categorized in class c can be calculated by:

$$P(c|d) \propto P(c) \prod_{i=1}^{n_d} P(w_i|c). \tag{3}$$

P$(c)$ is the prior probability of a class can be calculated by P$(c)$ = $\frac{s_i}{s}$ , where $s_i$ is the number of training samples of the class and s is the total number of training samples. And $(w_i|c)$ is the conditional probability of $w_i$ feature appears in a document d in the class c. This equation calculates the contribution of each $w_i$ in document d to the class c. Because the ultimate goal is to find a class c is most likely to a document d based on feature that appears on the document, then:

$$c = argmax_{c\in\{+,-\}} P(c) \prod_{i=1}^{n_d} P(w_i|c) \tag{4}$$

Where argmax function count and take classes with a maximum value of each class which are calculated. With the use of the feature weighting in Multinomial Naïve Bayes, TF-IDF can be used to substitute the conditional probability function so that the equation of Multinomial Naïve Bayes with TF-IDF feature weighting to classify a document d in the class c can further be represented to[6], [7]:

$$P(c_j|d) = P(c_j) + \sum_{i=1}^{n_d} TF_i - IDF(X_i)\ldots \tag{5}$$

Where $TF_i$ is a feature of terms frequency from feature i of the entire document with the class c, and $ID(X_i)$ is the IDF value of the feature $X_i$ in the entire document.

## 3. SYSTEM DESIGN

This system is divided into two lines, which are offline and online. An offline workflow is classification process stage that is conducted outside of the Android mobile devices as a learning process while the online workflow was done in the Android mobile device which is a classification stage. The general design of the system can be seen in Figure-1.

### 3.1 Dataset

Classes used in the SMS classification SMS consists of two classes, Spam and not Spam (Ham). Stages in the classification consist of a training data processing and testing data processing. To get the training data and testing data at the test this thesis, the data distribution used is the distribution by k - fold cross - validation which divides the obtained data as mush as k sampling data. Then k times experiment was performed, where each experiment used the k$^{th}$ partition data as a testing data and utilizes the remaining partitions as training data. In this research, data distribution with k - fold cross - validation used value of k = 5 and k = 10 which is a common value that is often used.

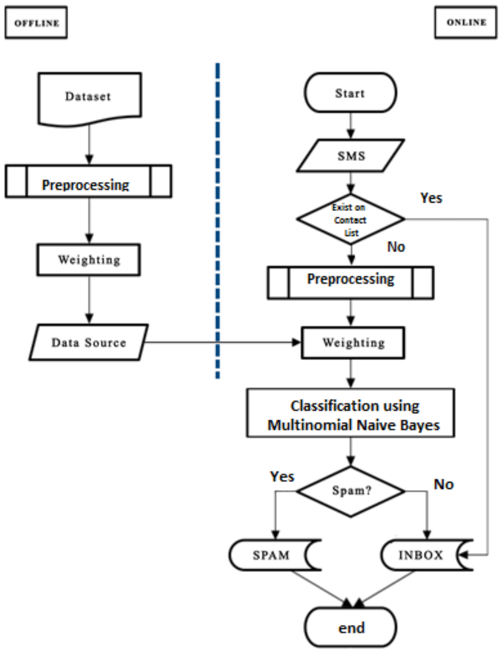


**Figure-1.** General design.

### 3.2 Preprocessing

Preprocessing was started with dataset collection into a csv file. The file was entered into the SQL. The training datasets have doc_content and doc_type fields. The testing datasets have fields while division of fields in testing data is doc_contentorigin_class. Doc_content is the content of SMS. While doc_type, origin_class is the category of it which spam or non-spam (ham) were carried out the following process.

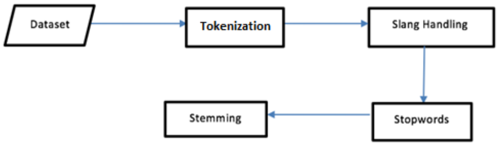


**Figure-2.** Preprocessing.

Figure-2 describes about step by step preprocessing dataset. the data was divided into two kind of data. The data training and the data testing. In the data training have doc_content and doc_type, the data testing have doc_content and origin_class. The doc_contentrepresentsSMS content, doc_type and origin_class represent SMS category classes, namely spam or not spam(ham) that will be processed for the next step :

**A. Tokenizing**

Tokenizing process was applied in every word in documents. In this step every punctuation mark or spacing (example: ' - ) ( \ / = . , : ; ! ?.) will be transform into delimeter '#' that separate every word as token. And every token will be transform into lowercase. Some examples can be viewed in Figure-3.

| SMS | Tokenizing |
|---|---|
| Tlng beliin Papa pulsa 50rb dino ini 081355814488 PENTING, kbtln Papa ada diKantor POLISI, nanti Papa ganti uangnya. | #tlng#beliin#papa#pulsa#dino#ini#{PHONExyz}#penting#kbtln#papa#ada#dikantor#polisi#nanti#papa#ganti#uangnya# |

**Figure-3.** Example result tokenizing.

**B. Slang handling**

Slang handling process was applied in every words in documents which will be checked with the slang word. There are 282 slang words in database. The slang words or dictionary slang words as reference replace the SMS content. Some examples can be viewed in Figure-4.

| SMS | Slang Handling |
|---|---|
| Tlng beliin Papa pulsa 50rb dino ini 081355814488 PENTING, kbtln Papa ada diKantor POLISI, nanti Papa ganti uangnya. | #tolong#beliin#papa#pulsa#dinomor#ini#{PHONExyz}#penting#kebetulan#papa#ada#dikantor#polisi#nanti#papa#ganti#uangnya# |

**Figure-4.** Example slang handling.

**C. Stopword**

The stopword process will remove some word in the content SMS which stopword dictionary is used for stopword as reference. The objective in this process is remove the words that have not value in the corpus. There are two stopword dictionary with different words. The first dictionary has 327 words and the second dictionary has 758 words. The differences of stopword dictionary will be used to find the quality of stopword process. Some examples can be viewed in Figure-5.

| SMS | Stopword |
|---|---|
| Tlng beliin Papa pulsa 50rb dino ini 081355814488 PENTING, kbtln Papa ada diKantor POLISI, nanti Papa ganti uangnya. | #tolong#beliin#papa#pulsa#dinomor#{PHONExyz}#kebetulan#papa#dikantor#polisi#papa#ganti#uangnya# |

**Figure-5.** Example stopwords.

**D. Stemming**

The Algorithm that will be used in this stemming process is NaziefAndriani Algorithm. In this process will check every word in documents (content SMS) then find the basic word/root word. The other things this process will remove every prefix or postfix in every words. There are 28506 dictionary stemming words that will be used. The objective in this process is decrease words that have same meaning. Like word "menghubungi", "hubung", "hubungi". After this process the words become "hubung", "hubung", "hubung".The first example has three different words and the result is only one word. Some examples can be viewed in Figure-6.

| SMS | Stemming |
|---|---|
| Tlng beliin Papa pulsa 50rb dino ini 081355814488 PENTING, kbtln Papa ada diKantor POLISI, nanti Papa ganti uangnya. | #tolong#beliin#papa#pulsa#nomor#081355814488#betul#papa#kantor#polisi#papa#ganti#uang# |

**Figure-6.** Example stemming.

**E. Weighting**

After preprocessing, every word in the documents will be weighted using TF-IDF. And then we got weight result for every word in documents.

**F. Data source**

After preprocessing, every document weighting will be gathered into one data source. This data source will be used in the next step which to classify the content SMS in mobile application (Android).

**3.3 Mobile Application System Design (Online)**

This Mobile application system design is step-by-step process that will be used. Every content SMS will be filtered based on user contacts. If the sender is not in user contacts then will be processed below:

### A. Preprocessing

This process will be used same method in offline method. Where the steps are tokenizing, slang handling, stopword and stemming. The difference process is in Android system.

### B. Weighting

This process will be used same method in offline method. The differenceprocess is in Android system.

### C. Classification mining (Multinomial Naïve Bayes)

This process will be used feature weight from weighting process. Thatfeature weight will be used for classification using Multinomial Naïve Bayes. The classification needs to find the occurrence of words in the documents. The occurrence of words will be used to find the probability of the content SMS. From that probability will be used for classifying the SMS content. The SMS content will be divided into two classes; one is SPAM and the othersisHAM (Not Spam).

## 4. TESTING

### 4.1 The dataset

The dataset will be used for the data training and the data testing whichdivided based on k-fold cross validation algorithm. The dataset will be divided into k sample of the dataset. The k value in this research is 5 and 10, becausethat values are usually used in the algorithm. The datasets have 180 SMS, 100 SMS as spam, and 80 SMS as not spam (Ham).

### 4.2 The testing scenario process

The testing scenario will beselectedpreprocessing scenario. After preprocessing will be process with multinomial Naïve Bayes. This is scenario will be used as below:

- Sometime slang handling can be use or not in the scenarios
- Stopword comparing dictionary long stopword and dictionary short stopword in the scenarios
- Stemming will be used in allscenarios.

The testing scenarios show in Table-1.

**Table-1.** Testing scenarios.

| Scenarios | Preprocessing | Information |
|---|---|---|
| 1 | 1,2,4 | 1 = *SlangHandling*<br>2 = *Stopword* (*long*)<br>3 = *Stopword*(*short*)<br>4 = *Stemming* |
| 2 | 1,3,4 | |
| 3 | 1,4 | |
| 4 | 2,4 | |
| 5 | 3,4 | |
| 6 | 4 | |

Data partitions with 5-fold and 10-fold for all scenarios show in Table-2 and Table-3.

**Table-2.** Data partition 5-Fold**.**

| Data training | Data testing | Scenarios | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| BCDE | A | 1,2,4 | 1,3,4 | 1,4 | 2,4 | 3,4 | 4 |
| ACDE | B | | | | | | |
| ABDE | C | | | | | | |
| ABCE | D | | | | | | |
| ABCD | E | | | | | | |

**Table-3.** Data Partition 10-Fold.

| Data Training | Data Testing | Scenarios | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| BCDEFGHIJ | A | 1,2,4 | 1,3,4 | 1,4 | 2,4 | 3,4 | 4 |
| ACDEFGHIJ | B | | | | | | |
| ABDEFGHIJ | C | | | | | | |
| ABCEFGHIJ | D | | | | | | |
| ABCDFGHIJ | E | | | | | | |
| ABCDEGHIJ | F | | | | | | |
| ABCDEFHIJ | G | | | | | | |
| ABCDEFGIJ | H | | | | | | |
| ABCDEFGHJ | I | | | | | | |
| ABCDEFGHI | J | | | | | | |

Finally, all scenarios will find value of precision, recall, and accuracy.

### 4.3 Result

This paper result want to find the best scenarios in this SMS spam filtering. In Figure-7 shows the comparison for all scenarios with 5-fold. In figure 8 shows comparison for all scenarios with 10-fold.

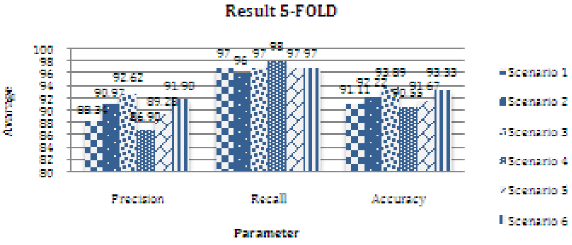


**Figure-7.** Average result all scenario with 5-fold.

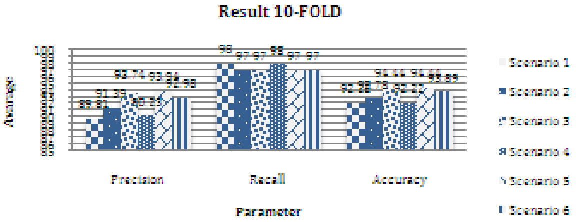
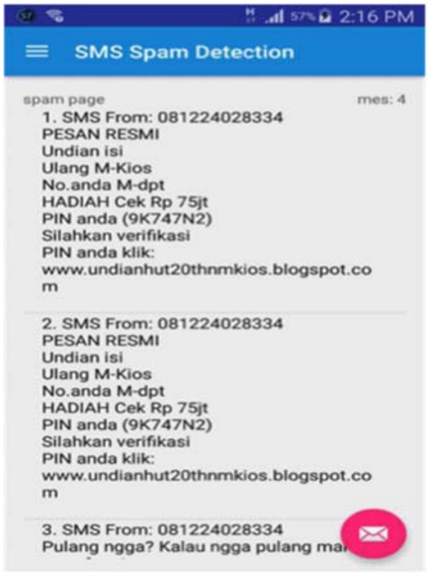


**Figure-8.** Average result all scenario with 10-fold.

In Figure-7 and Figure-8 have the comparison of average result for precision, recall and accuracy? The lowest value of precision in scenario 4 and 1used preprocessing 2. The preprocessing 2 used stopword elimination and stopword list dictionary long. From above preprocessing can be concluded that using stopword list dictionary long is not good for SMS spam filtering. The recall value is more than 96% for all scenarios, which can be concluded this system can filter SMS spam is good. The accuracy value is more than 94%, which can be concluded this system can determine SMS spam and SMS ham is good. For all result can be concluded that Multinomial naïve Bayesis effective used in SMS spam filtering. Example can be viewed in Figure-9 and Figure-10.

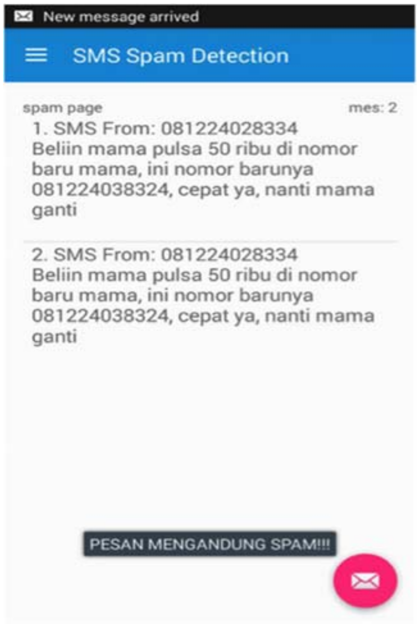


1. SMS From: 081224028334
PESAN RESMI
Undian isi
Ulang M-Kios
No.anda M-dpt
HADIAH Cek Rp 75jt
PIN anda (9K747N2)
Silahkan verifikasi
PIN anda klik:
www.undianhut20thnmkios.blogspot.com

2. SMS From: 081224028334
PESAN RESMI
Undian isi
Ulang M-Kios
No.anda M-dpt
HADIAH Cek Rp 75jt
PIN anda (9K747N2)
Silahkan verifikasi
PIN anda klik:
www.undianhut20thnmkios.blogspot.com

3. SMS From: 081224028334
Pulang ngga? Kalau ngga pulang ma

**Figure-9.** SMS SPAM filter in android.



1. SMS From: 081224028334
Beliin mama pulsa 50 ribu di nomor baru mama, ini nomor barunya 081224038324, cepat ya, nanti mama ganti

2. SMS From: 081224028334
Beliin mama pulsa 50 ribu di nomor baru mama, ini nomor barunya 081224038324, cepat ya, nanti mama ganti

PESAN MENGANDUNG SPAM!!!

**Figure-10.** SMS SPAM filter incoming SMS.

## 5. CONCLUSIONS

It can be conclude after doing some secenario on Android short messages filtering for Bahasa using multinomial naïve bayes algorithm, namely; preprocessing with stemming handling or using stopword (short) or stemming alone are the best preprocessing. The accuracy reached 93.33% - 94.44%. The application is divided into two main processes, namely learning algorithm and classification. The performance system is precession 93.74%, recall 97% and accuracy 94.44%.

## REFERENCES

[1] S. A. Abdurrasyid and B. Indonesia. 2009.Implementasidanoptimasialgoritmanaziefdanadrianiuntuk stemming dokumenbahasaindonesia. pp. 1-8.

[2] K. Mathew and B. Issac. 2011. Intelligent spam classification for mobile text message.in Proceedings of 2011 International Conference on Computer Science and Network Technology. 1: 101-105.

[3] G. V. Cormack. 2008. Email Spam Filtering: A Systematic Review. Found. Trends® Inf. Retr. 1(4): 335-455.

[4] J. Han, M. Kamber and J. 2012. (Computer scientist) Pei, Data mining: concepts and techniques. Elsevier/Morgan Kaufmann.

[5] I. Asror, D. Saepudin, and Shaufiah. Combination Rough Set Theory and Genetics Algorithm for Intrusion Detection System.

[6] Kibriya Ashraf M. 2004. Multinomial Naïve Bayes for Text Categorization Revisited.

[7] J. Jerin Jose and K. Kedhareswari. 2015. Information Refinement over Multi-Media Question Answering Applying Ranking and Naïve Bayes Classification. 10(18).