Analysis and implementation of cross lingual short message service spam filtering using graph-based k-nearest neighbor

Dyah Ayu Cyntya Dewi, Shaufiah, Ibnu Asror
Data Mining Center Laboratory, School Of Computing, Telkom University, Bandung 40257, Indonesia
E-mail: cintyadewi.ring@gmail.com, shaufiah@telkomuniversity.ac.id, iasror@telkomuniversity.ac.id

Abstract. SMS (Short Message Service) is one of the communication services that still be the main choice, although now the phone grow with various applications. Along with the development of various other communication media, some countries lowered SMS rates to keep the interest of mobile users. It resulted in increased spam SMS that used by several parties, one of them for advertisement. Given the kind of multi-lingual documents in a message SMS, the Web, and others, necessary for effective multilingual or cross-lingual processing techniques is becoming increasingly important. The steps that performed in this research is data / messages first preprocessing then represented into a graph model. Then calculated using GKNN method. From this research we get the maximum accuracy is 98.86 with training data in Indonesian language and testing data in Indonesian language with K 10 and threshold 0.001.

1. Introduction
SMS (Short Message Service) is one of the communication services that are still the main choice in various parties. Along with the development of other communication media, the country lowered the SMS rate to remain the top choice for mobile users [1]. This results in increased spamming of spam that some irresponsible people use, such as advertising and fraud. So with the number of SMS fraud or advertising action is causing discomfort for most SMS recipients. SMS spam is an SMS that contains information unwanted by the recipient of the message. Spam SMS is sent from one sender to many numbers obtained randomly (randomly).

Given the multi-lingual nature of documents in a message, web, etc., the need for effective multilingual or cross-lingual processing techniques is becoming increasingly important. Therefore, the hope of this research system can recognize a message with multi-language exactly that is Indonesian or English language and classify it into class of spam or ham (not spam)[2][3].

In this research, a message will be classified in the category of spam or not spam (ham). The dataset in this research using English language and Indonesian language. For the classification process, the data already obtained needs to be preprocessed in advance to clear the data without changing any contained information [4].

There are classification methods that have been used in previous research, such as Nave bayes [5][6][7], KNN, GKNN with accuracy maximum 95.5. From the research mentioned that Graph-based K-nearest Neighbor can perform SMS classification with accuracy level reaches 98.9 [8].
Based on this research, GKNN method was chosen as algorithm used to perform the process of classification on SMS filtering in this research. From this research the system will classify data in spam or ham class (not spam).

2. Related Works
Palmieri et all says that The mobile networks are indeed very popular in our lives but raises serious security concerns their increasing vogue and widespread coverage [9]. It means that smartphones may now represent an most ideal target for malicious authors. Kim et all says SMS is not only used for texting among people but it is also used as security like authentication method, token (mobile banking, double authentication, one-time password delivery, etc)[8].

When we embarked on this line of research, we did not nd any publications addressing the area of Cross-lingual SMS spam filtering. On the other hand, there is a rich literature addressing the related problem of Cross-Lingual Information Retrieval (CLIR) or Cross-Lingual Text Categorization (CLTC), olsson et al says that Both CLIR and CLTC are based on some computation of the similarity between texts, comparing documents with queries or class profiles[10]. The main contrasting between CLIR and CLIC is that CLIR is queries based, consisting of a few words only, whereas in CLTC have class and each class is describe by an major profile (may be seen as a adequate documents collection)[10]. In this research we used CLTC to solve our problem.

3. Proposed Schema
System development will be divided in four phase, these are: documents translation as needed, preprocessing, build graph, classification using graph-based K-Nearest Neighbor. In figure 3-1, represent system overview.

**Figure 1. System Overview.**

3.1. Dataset
A dataset is an object that represents data. The dataset that used are contain of unstructured text which later data will be transformed into a form of data that is structured to facilitate the process of storing, calling, and classification on the system. The Indonesian dataset is the result of manual collection from previous researchers[6][5], while the English dataset is obtained from SMS Corpus: SMS Spam Collection v.1.

3.2. Document Translation
The data used will be adjusted to the data partition scenario. For a cross-lingual approach, data will be translated into other languages with the help of machine translator i.e. Google Translate.
Documents translation is only done for test scenario 3, 4 and 5. Before preprocessing, data will be translated. For example, English data will be translated into Indonesian language, figure 2 illustrated this example.

![Data Translation Diagram](image)

**Figure 2.** Data Translation.

3.3. Preprocessing Data

At this phase, training data and testing data will be processed to produce better data to be processed in the next phase. In the preprocessing phase, there are 5 step of the process that must be done. Here is an explanation of each step done on preprocessing:

3.3.1. Case Folding

Case Folding and remove punctuation: is the process of converting capital letters contained in the dataset into lowercase for all datasets and also remove characters and punctuation.

3.3.2. Tokenizing

Tokenizing is the process to separate the sentence contained in the review into a snippet of words, this piece of words that will be input of the next step after preprocessing.

3.3.3. Slang Handling

In dataset, there are many informal words, it called slang word. To solve these cases, so we created a dictionary containing the words slang and equipped with the meaning of the word. Glossary for the English slang dictionary from “Slang Dictionary-Text Slang Internet Slang Words” on the website http://www.noslang.com/dictionary/, while for Bahasa Indonesia obtained from http://en.wikipedia.org/wiki/Indonesian slang.

3.3.4. Stopword Removal

Eliminate Common words that have lesser significance. To remove the words included in the stopword list, then use a word matching technique with a dictionary that lists the word stopwords. For the list of word stopwords, in this study used dictionary word stopword Indonesian sourced from the site https://github.com/masdevid/ID-Stopwords, and dictionary word stopword English sourced from the site http://www.ranks.nl/Stopwords.

3.3.5. Stemming

In the dataset there are many words that have affixes, so it takes stemming that aims to change back words that have additional affixes or changes in the form of words. The stemmer for Indonesian language is Sastrawi Indonesian stemmer is sourced from https://pypi.python.org/pypi/Sastrawi/1.0.1. And for English stemmer use Porter Stemmer sourced from NLTK library: https://pypi.python.org/pypi/stemming/1.0.
3.4. Build Graph
After preprocessing, it will generate data that ready to be processed in the next step. Data will be represented in graph. Formation of graph in data training is done by dividing data into 5 documents / messages into 1 (one) graph. So suppose, there are 100 documents/messages, then split into 20 graphs. While for data testing representation, each message/document will be 1 (one) graph.

Each node that forms in a graph shows a token selected at the preprocessing step. Edge is formed based on the order of occurrence between 2 words. The weights of the edge are shown with the Feature Weight Matrix. Graph formed consists of 2 graphs, namely training graph and testing graph. Training graph will group the data by category, spam and ham. While testing graph is a graph formed from data testing.

3.5. Classification Using G-KNN
Graph-based K-Nearest Neighbors (GKNN) is a evolution of K-Nearest Neighbors, which in GKNN data will be represented in the form of graph models. From that model will be calculated the similarity between documents classified by the large number of documents [8]. In a graph consists of nodes, edges, and weights of edges. To measure the similarity between 2 graphs (spam graph and ham graph), then the classification is measure the similarity. There is a Feature Weight (FW) which defines the similarity between 2 graphs based on the weight of the nodes and edges on the graph. The Node Fit Percent (NFP) shows how many nodes in their sample graph with their weights bigger than zero also appear in the test graph NFP can be formulated as follows:

\[ Nfp = \frac{\left| \{ \text{node} \mid \text{node} \in tg \cap \text{node} \in sg \} \right|}{\text{Number of Feature terms}} \]  

Calculate the value of NFP from testing graph and training graph by calculating the frequency of each feature on the graph that is the value of W (i, i). If the Nfp value is greater than the threshold, it will calculate the FW value of 2 pieces of graph. Otherwise if the Nfp value is smaller than the threshold, then the 2 calculated graphs are not in one category, so there is no need to calculate the FW value. For the calculation of NFP and Feature Weight can be seen in Figure 3 and Figure 4.

![Figure 3. Calculation of NFP.](image)

The process of comparing the values of NFP and FW is done until all training and testing graph have been compared. In the end, the initially empty RL list will be filled with the FW value and the category of the training graph. The most categories that appear on the list will be a new category of testing graphs tested that represent category of SMS testing.

4. Analysis and Experiment
For the classification in the Graph K-Nearest Neighbor method is used K value that will determine the length of list / array. The K value used in this test is 10, as well as by varying the threshold value to see the effect of the threshold value in find the most optimal threshold value in this research.
4.1. First Scenario Testing Analysis
The first test was conducted using Indonesian training data, and data testing using Indonesian Language. The threshold values used in this test are 0.001, 0.025 and 0.05. Based on the first test (Figure 5), with 10-fold by changing the threshold value obtained the highest accuracy of the highest average accuracy with the threshold of 0.001 is 84.44 with the highest accuracy value of 100 available in the sample data C. In the Indonesian language dataset especially the spam class has a pattern Almost similar, it creates a weighted value on each edge that is getting higher, so when searching for FW value, then having a higher weight value will have a higher chance of being in that category.

4.2. Second Scenario Testing Analysis
The second test result for 10-fold cross validation that consist of training data using Indonesian, and data testing using English (without translated) with a combination of threshold value 0.0005 and 0.0001.

Based on the results of the second test (Figure 6), by changing the threshold value obtained average accuracy of 86.4 and the average value of F1 of 2.84. It can be seen that the accuracy values have the same and low result of each sample of data testing. This low F1-measure result can be influenced by the language of different training data and data testing, so in the
4.3. Thrid Scenario Testing Analysis

The third test result for 10-fold validation that consist of data training using Indonesian and English (without translation), and data testing using English (translated to Indonesian) with the combination of threshold value 0.0001, 0.005 and 0.05.

Based on the third test (Figure 7), by changing the threshold value obtained the highest accuracy with the highest threshold 0.0001 is 90.82, with the highest accuracy of 92.4 in sample data A and C. It can be seen also that the value of F1-measure is low, it is also caused by the use of different language in training data and data testing.
4.4. Fourth Scenario Testing Analysis

The fourth test was performed using 10-fold validation using Indonesian and English training data (translated into Indonesian), and testing data using Indonesian and English (translated to Indonesian). The threshold values used in this test are 0.001, 0.0075 and 0.05.

Based on the fourth test (Figure 8), by changing the threshold value obtained the highest accuracy on the threshold 0.001 of 96.27, with the highest accuracy of 97.68 in a sample data G. From the results of this test is considered the most optimal for cross-lingual handling, because in this scenario has a good combination of data sharing and data translation. The best results of the overall experimental results are from this fourth scenario.

Figure 8. Fourth Graph of Testing Accuracy.

Figure 9. Fifth Graph of Testing Accuracy.
4.5. Fifth Scenario Testing Analysis
The fifth test was performed using 10-fold validation using Indonesian (translated to English) and English training data and testing data using Indonesian (translated to English) and English. The threshold values used in this test are 0.001, 0.075 and 0.05.

Based on the fifth test (Figure 9), by changing the threshold value obtained the highest accuracy on the threshold 0.001 of 93, with the highest accuracy of 94.78 in a sample data F. From the results of this test is considered the most optimal for cross-lingual handling, because in this scenario has a good combination of data sharing and data translation, but this scenario is not the best result.

5. Conclusion
Based on experiment result we found that the fourth scenario is the optimum one. with the accuracy 97.86. with based model in indonesian language(Bahasa), data training translated into bahasa and data testing translated into bahasa. That can be conclude with this dataset is the best using Indonesia language/Bahasa as primary language for analysis.

References
[1] A. Garz, Cross-Lingual Web Spam Classification, pp. 11491156, 2013
[2] M. Xiao and Y. Guo, Semi-Supervised Representation Learning for Cross-Lingual Text Classification, no. October, pp. 14651475, 2013.
[3] M. Xiao and Y. Guo, A Novel Two-Step Method for Cross Language Representation Learning, Adv. Neural Inf., pp. 19, 2013.
[4] Sunardi, M., Listiyono, H. (2009). Aplikasi SMS Gateway. Jurnal Teknologi Informasi DINAMIK, XIV(1).
[5] T. Akhir, Sistem SMS Spam Filtering Berbahasa Indonesia Pada Smartphone Android Menggunakan Algoritma Multinomial Nave Bayes SMS Spam Filtering System For Indonesian SMS On Android Smartphone Using Multinomial Naive Bayes Algorithm Program Studi Sarjana Teknik Inf, 2016.
[6] , Shaufiah, Imanudin and Asror, Ibnu, ANDROID SHORT MESSAGES FILTERING FOR BAHASA USING MULTINOMIAL NAIVE BAYES, 2016.
[7] Gomez Hidalgo, J.M., Bringas, G.C., Sanz, E.P., Garca, F.C.: Content based SMS spam filtering; Proceedings of the 2006 ACM symposium on Document engineering, pp. 107-114, 2006.
[8] S. Kim, Graph-based KNN Algorithm for Spam SMS Detection, vol. 19, no. 16, pp. 24042419, 2013.
[9] Palmieri, Michela and Impey, Soren and Kang, Hyojin and di Ronza, Alberto and Pelz, Carl and Sardiello, Marco and Ballabio, Andrea, "Human molecular genetics", vol 2011, number 19, pages 3852-3866, year 2011, publisher Oxford University Press
[10] J. S. Olsson, D. W. Oard, and J. Haji, Cross-Language Text Classification, pp. 645646.