



Stop Words Removal on Textual Data Classification

Taye Oladele Aro¹, Funmi Dada², Abdullateef Oluwagbemiga Balogun³, Samuel Ayodeji Oluwasogo⁴

¹Department of Mathematical and Computing Sciences, Kola Diasi University, Ibadan, Oyo, Nigeria

^{2,3}Department of Computer Science, University of Ilorin, Ilorin, Nigeria

⁴Management Information System Unit, Federal Polytechnic Offa, Kwara State, Nigeria

Abstract

Text data is highly voluminous and performing mining tasks on it can be daunting due to large memory usage, thus researchers have considered different techniques to reduce the data while still maintaining or increasing the level of accuracy. Stop word removal is one of the pre-processing techniques used in text data mining. This paper investigates the effect of stop words removal on the text data mining performance. The machine learning algorithms used are: C4.5 Decision Tree and Multinomial Naïve Bayes (MNB) on two text datasets; Sentiment Analysis and SMS Spam dataset. Results revealed that the removal of stop words had no influence on the classification accuracy of text mining model, but actually reduced the level of confidence of prediction.

Keywords: *Textual data, Stop words, C4.5 Decision Tree, Multinomial Naïve Bayes, SMS Spam, Sentiment Analysis*

1. Introduction

Text mining is a process of obtaining intriguing and nontrivial patterns from large amount of text documents (Talib, Hanif, Ayesha, & Fatima, 2016). The technique of text mining involves the selection of relevant features in unstructured text with the purpose of getting information that is of high-quality from text (Singh, 2016; Usai, Pironti, Mital, & Mejri, 2018). Undefined information is obtained through the application of text mining process of diverse techniques (Kaushik & Naithani, 2016). The text mining can be used in many domains like information retrieval, analysis of text, clustering, visualization, information extraction, electronic medical records, classification and learning algorithms (Dang & Ahmad, 2014; Yanhong, Anmeng, & Jingling, 2017; (Sun et al., 2018). Text Classification (TC) is a process of classifying a document in a predefined class using the text content (Iskonomakis & Kotsiantis, 2005), such as categorizing whether a message is a spam or ham message (Almeida, Gomez Hildago, & Yamakami, 2011; Nagwani & Sharaff, 2017).

Text Classification plays a significant role in text mining (Sailaja, Padamsree, & Mangathauaru, 2016) as it represents knowledge engineering by which expert obtain knowledge on how to classify automated document such that similar documents can be placed in their respective categories the way the expert would have categorized (Chin, Zappone, & Zhao, 2016). It has been identified by several studies over the years that SMS spam filtering and sentiment analysis had shown different experimental outputs with respect to the significance of stop words removal from datasets during text mining process.

Studies have established that the existence of stop words in datasets affects classification processes negatively (Al-Talib & Hassan 2013; Bakliwal et al. 2012; Nagwani & Sharaff 2017) and thus proposed the removal of stop words. Alsudais, Leroy, & Corso (2014); (Kanavos et al., 2017; Kotzias et al.2015) in their respective research, did not remove stop words during pre-processing. Almeida, Hidalgo & Yamakami (2011) claimed that non-removal of stop words actually make classification better.

This paper considered the influence of stop word removal on text mining classification. Two text datasets : Spam SMS and Sentiment Analysis dataset were used to evaluate the proposed text mining model which were of short sentences so as to determine the advantages and relative disadvantages of stop words removal before performing classification.

2. Related Works

Davcheva (2018) presented text mining mental health forums learning from user experiences The study revealed how natural text processing of large datasets can automate and accelerate data collection, processing, and insight generation from millions of posts by tens of thousands of individuals. Being able to back up many of the findings of this study with previous findings in psychological literature and in a few cases also practice, demonstrated how and why text mining of large public datasets is a viable and dependable tool that can bring about a veritable change in how research mental health, and approach of the question for improving and administering therapy and treatments.

A two-level knowledge discovery approach for SMS thread identification presented by (Nagwani & Sharaff 2017). At the first level, state of the art classification algorithms were used to classify the texts into spam and ham messages. At the second level, non-negative matrix factorization (NMF) text clustering was applied on the ham SMS messages to create a cluster of those ham messages. During experimental analysis, the study removed stop words using a stop word list.

Chin, Zappone and Zhao (2016) analyzed the sentiments on the presidential elections in the United States in 2016 to assess the relevant associations of emotions towards the candidates and a part of their conclusion was that removal of stop words did not have a significant impact on classification accuracy as they only slightly improved classification effectiveness.

Kotzias et al. (2015) considered the challenge of learning algorithms to perform predictions at the instance level rather than at the group level on reviews to determine sentiments. A new objective function was proposed which encouraged the better inferred instance-level labels based on instance-level comparison, while the still showing respect to group-level label constraints. The system was used in predicting labels for sentences given labels for review using a convolutional Neural Network to infer sentence similarity. The method was evaluated on review datasets from IMDB, Yelp and Amazon. The proposed system was accurate and scalable compared with the existing systems.

Alsudais et al. (2014) developed a classifier to predict the place a user is tweeting from, they applied the content of the text, the meta data and the geographical area the user is tweeting from to produce features. The study employed six classes for their location which include active life, eating out, hotels, night life, shopping and shows. The six locations were identified using random forest. Stop words were not removed during pre-processing of the data.

Al-talib & Hassan (2013) presented a study analyzing the classification of SMS using TF-IDF weighting. The study applied method that improved classification significantly. Stop words were removed from data used as the study claimed that not removing it will affect the classification accuracy negatively.

A system for identifying sentiments in tweets using a scoring function (Bakliwal, Arora & Madhappan 2012). The study used prior information to classify and weight various sentiment bearing words/phrases in tweets. It was reported during experimental analysis that stop words contributed negatively in sentiment classification and that it also added ambiguity in the model formation.

3. METHODOLOGY

The text data mining model aims at determining a better classification process for SMS filtering and sentiment analysis by considering how the removal of stop words affects general performance of text mining. WEKA (Waikato Environment for Knowledge Analysis) data mining software tool was used in carrying out experimental analysis. The stratified 10-fold cross validation was used. Two selected classification algorithms: Multinomial Naïve Bayes and the C4.5 decision tree were applied on the SpamSMS and Sentiment labelled instances datasets from the UCI (University of California, Irvine) repository. The datasets were pre-processed by separating the sentences into words, converting all words to lower case, digits and special characters were converted to strings. The stop words non -removal baseline method was used in this study. So while in set of analysis stop words were not removed, also the stop words were removed in the other set of analysis and the results were compared.

3.1 Description of Datasets

The data for the study was obtained from the Irvine repository in University of California. The two datasets contained short sentences in each unit. The data are described as follows:

(i) SMS Spam Collection

The SpamSMS collection is a public set of SMS labelled messages that have been gathered for mobile phone spam research. The dataset contains SMS messages that are grouped into either spam or ham. The dataset contains 4827 ham messages and 747 spam messages all together making 5574 sentences

(ii) Sentiment Labelled Instances

The dataset consists of 3000 instances of texts that have been labelled as positive and negative sentiment, the texts were extracted from movies, products and restaurants reviews. The sentences were obtained from three websites which are imdb.com, amazon.com and yelp.com, each of the websites contributes 500 positive sentences and 500 negative sentences respectively. In dataset collation, it was confirmed that the statements were noticeably positive or negative in order not to produce neutral statements.

3.2 Dataset Pre-processing

Datasets were converted to standard format of arff"; this represents one of the generally acceptable format in WEKA software to operate on. In tokenizing the documents the following characters were removed `\r\t,;:\'\"()?!@#$%^&*()_+ |\?><-=[]{}[\^?~\`"-` and so were not part of the letters making up the words considered by the algorithms in making classification decisions. All documents were normalized and tokenized with words converted to lower cases.

3.3 Algorithm for the Proposed Text Mining Model

The text mining model for this study is shown in Figure 1

-
- Step 1: Collect Standard dataset for text classification.
 - Step 2: Normalize documents
 - Step 3: Tokenize text.
 - (i) Separate sentences into words.
 - (ii) Convert all characters into lowercase.
 - (iii) Remove special characters and digits then convert them into word string.
 - Step 4: Remove stop words using stop words list
 - Step 5: Apply classification algorithms: MNB and C4.5
 - Step 6: Evaluate and compare results
-

Figure 1: Algorithm of Developed System

3.4 Loading of Dataset

The graphical interface for the loading of datasets is shown in Figure 2.

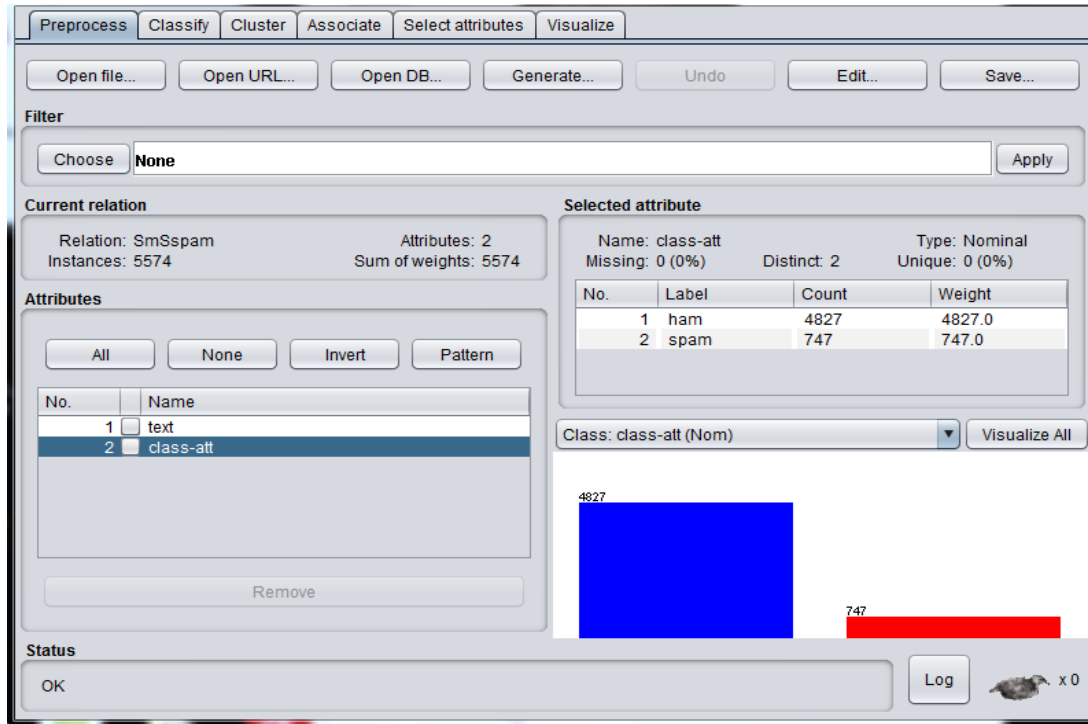


Figure 2: Interface for Loading Dataset

3.5 Pre-processing Phase

The interface of the pre-processing phase for the two datasets is shown in Figure 3.

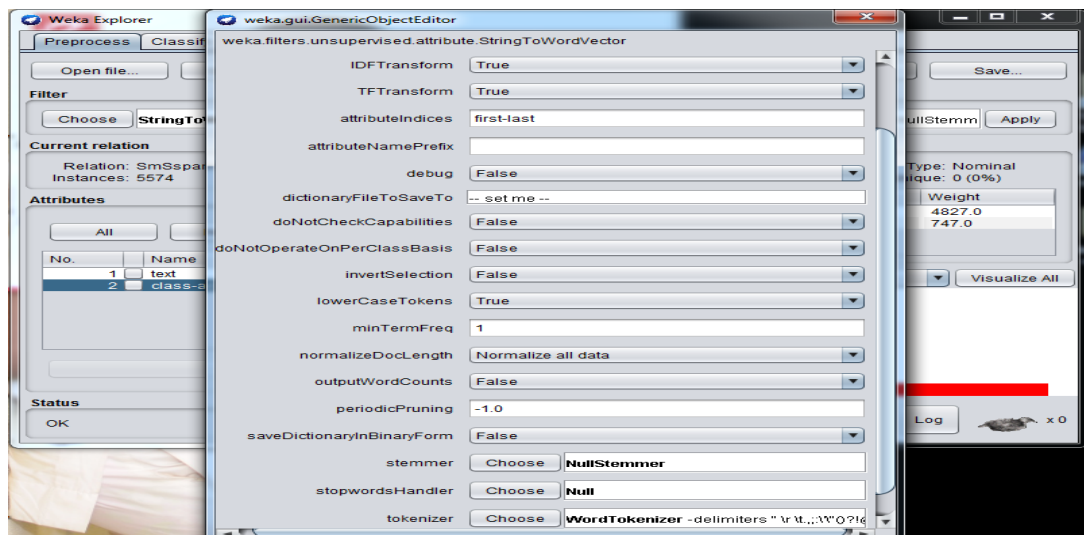


Figure 3: Pre-processing using TF-IDF

3.6 Performance Evaluation Metrics

The performance evaluation parameters used in this paper are discussed as follows:

- (i) **Accuracy:** This is correctly and incorrectly classified instances, which shows the percentage of test instances that were correctly and incorrectly classified while the unclassified instances show the percentage of test instances incorrectly classified. The percentage of correctly classified instances is often called accuracy. The percentage of incorrectly classified instances is obtained by subtracting the correctly classified instances from 100 as shown in Equation (1)

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (1)$$

Where TP = True positive, FP = False positive, TN = True negative and FN = False negative

- (ii) **Recall:** It is also known as sensitivity which is the true positive rate; it measures the proportion of actual positives that are correctly classified as given in Equation (2).

$$\text{Recall} = \frac{TP}{TP+FP} \quad (2)$$

- (iii) **Precision:** Precision represents the positive predictive value; it measures the proportion of actual negatives that are correctly classified as shown in Equation (3).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

- (iv) **Kappa Statistic:** It is a measure of how closely the instances classified by the machine learning classifier matched the data labelled as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy.

4. RESULTS AND DISCUSSION

Different results were obtained from procedures used in performing the investigation of influence of stop removal in the text mining model. This is discussed in the following subsections:

4.1 Evaluation Metrics of the Proposed Text Mining Model

The sample results of WEKA data mining tool software interface for general performance for the text classification using C4.5 Decision Tree and Multinomial Naïve Bayes is shown in Table 1.

Table 1: Sample of Evaluation Metrics

Evaluation Metrics	MNB	C4.5
Accuracy	82.1%	65.6%
Precision	0.822	0.696
Recall	0.821	0.656
Kappa Statistic	0.642	0.312

4.1.1 Results of Accuracy for the Proposed Text Mining Model

The results of classification accuracy of text data for the two datasets: SMS Spam dataset and Sentiment Analysis using C4.5 Decision Tree and Multinomial Naïve Bayes are shown in graphical representations of Figure 4 and Figure 5

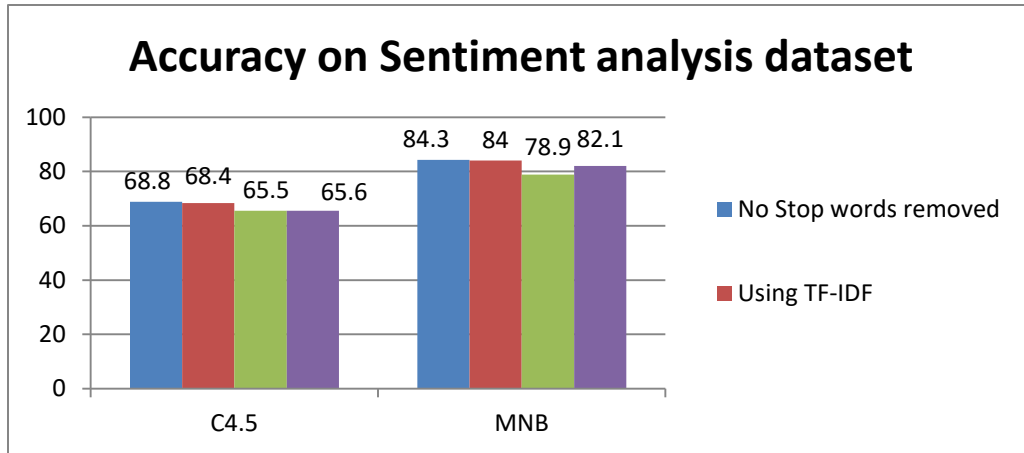


Figure 4: Accuracy on Sentiment Analysis Dataset

From Figure 4 , on the sentiment analysis data, C4.5 without the removal of stop words gave the highest classification accuracy while the others which involved the removal of stop words gave lower accuracies, though TF-IDF reduction in accuracy was just a slight drop to the value of 0.4, others accuracies were lower. Removal of stop words produced similar effect on MNB with little reduction effect showed when TF-IDF was applied, while others had greater reduction in accuracy.

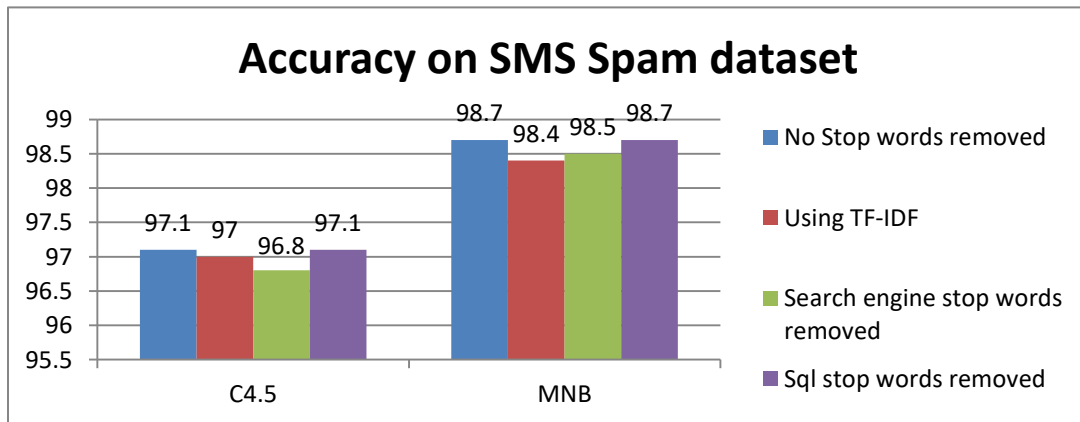


Figure 5: Accuracy on SMS Spam Dataset

In Figure 5, the removal of stop words did not increase the accuracy of the classification of the SMS spam dataset but rather reduced the accuracy. It was revealed that the results from the two datasets, using sql stop words removal did not produce any effect on the accuracy.

4.1.2 Results of Precision for the Proposed Text Mining Model

The results of Precision for two datasets: SMS Spam dataset and Sentiment Analysis using C4.5 Decision Tree and Multinomial Naïve Bayes are shown in graphical representations of Figure 6 and Figure 7.

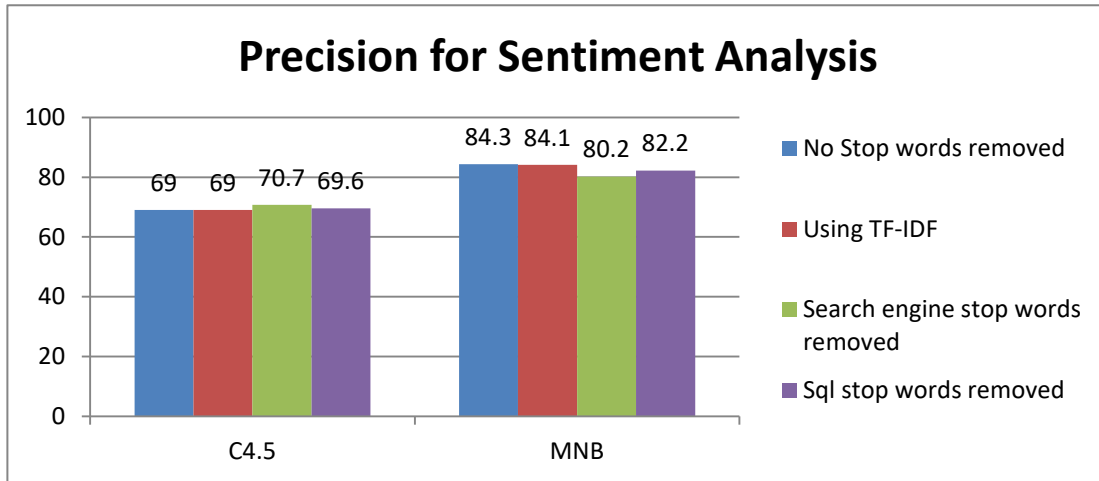


Figure 6: Precision for Sentiment Analysis Dataset

In Figure 6, MNB gave better precision than C4.5 Decision Tree as it had above 80% in all situations while the highest for C4.5 was 70%. The influence of stop word removal on C4.5 was little as it resulted in a slight increase in precision while for MNB, removal of stop words reduced the precision's capability.

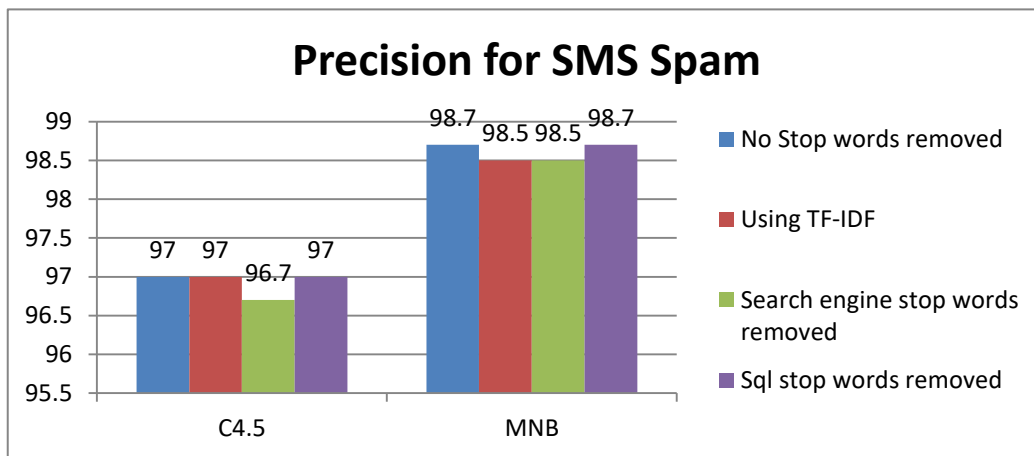


Figure 7: Precision for SMS Spam dataset

In Figure 7, stop words removal with different methods did not show any significant impact on precision values for the two learning algorithms: C4.5 Decision Tree and MNB using SMS spam dataset. The highest precision values of 97% for C4.5 and 98.7% for MNB were recorded when there was no stop words removal

4.1.3 Result of Kappa Statistics for the Proposed Text Mining Model

The results of Kappa Statistics for two datasets: SMS Spam dataset and Sentiment Analysis using C4.5 Decision Tree and Multinomial Naïve Bayes are shown in graphical representations of Figure 8 and Figure 9.

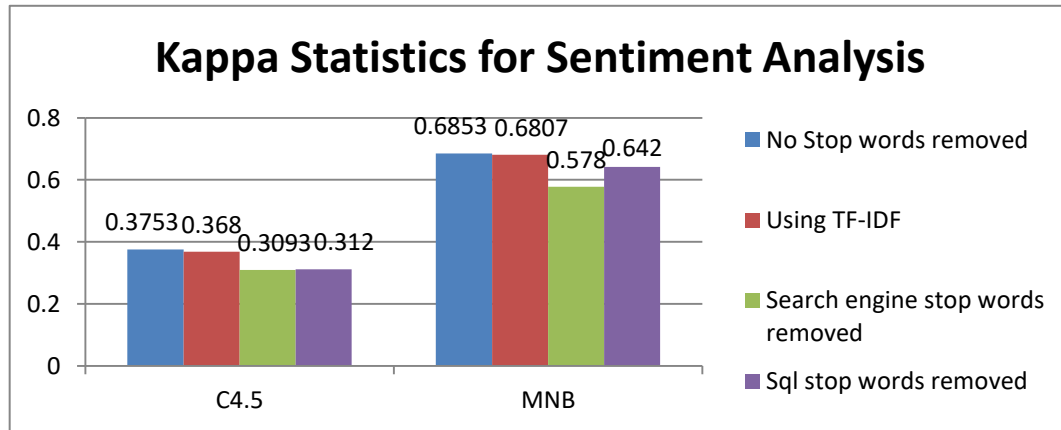


Figure 8: Kappa Statistics for Sentiment Analysis Dataset

In Figure 8, all C4.5 implementations performed better than chance all of MNB. This shows that MNB classified with much more confidence than C4.5. In both situations, the confidence of prediction of the algorithms before stop words were removed was better than the confidence of the algorithms when stop words have been removed.

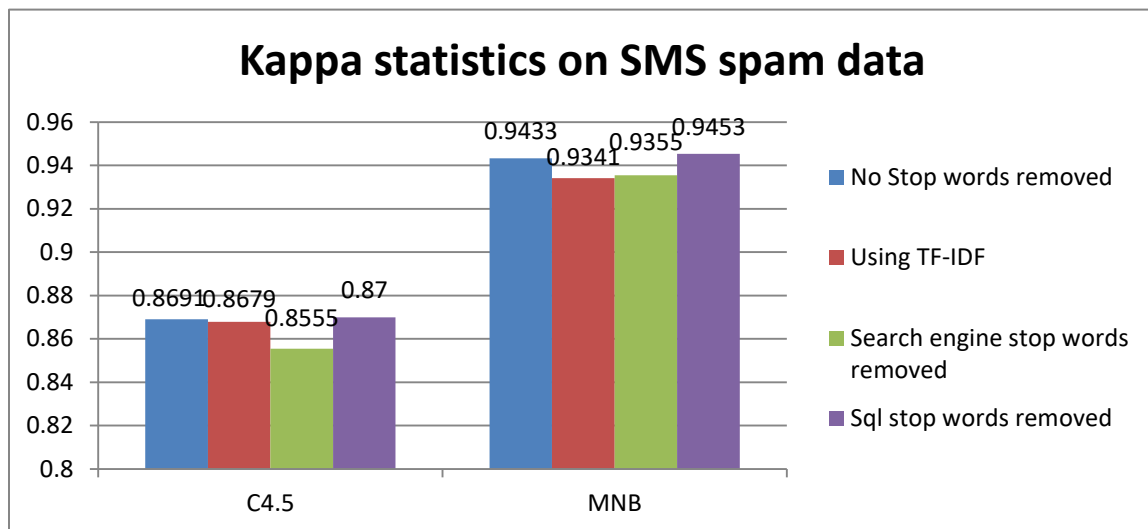


Figure 9: Kappa statistics for SMS spam dataset

algorithm has a greater confidence than those of when the stop words have been removed, except in the case of MySQL stop words removal in which the confidence was increased a little. This exception shows the possibility that stop words can also increase confidence of classification.

5. CONCLUSION

The paper studied the influence of stop words removal in text data mining performance. Three stop words removal techniques: TF-IDF, search engine stop words and sql stop words removal were employed to extract stop words from text datasets. The results showed that the removal of stop words from text data did not increase the

classification accuracy or confidence of classification but in fact reduces it with the exception of the removal of MySQL stop words on the SMS spam dataset which informs of the possibility that removing stop words can increase the confidence of prediction albeit slightly. Multinomial Naïve Bayes algorithm (MNB) performs much better in the classification of text data than C4.5.

REFERENCES

- Alsudais, A., Leroy, G., & Corso, A. (2014). We know where you are tweeting from: Assigning a type of place to tweets using natural language processing and random forests. In *Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014* (pp. 594–600). <http://doi.org/10.1109/BigData.Congress.2014.91>
- Al-Talib, G., & Hassan, H. (2013). A Study on Analysis of SMS Classification Using TF-IDF Weighting. *International Journal of Computer Networks and Communications Security*, 1(5), 189–194.
- Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining Sentiments from Tweets. *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, (July), 11–18.
- Chin, D., Zappone, A., & Zhao, J. (2016). Analyzing Twitter Sentiment of the 2016 Presidential Candidates. *International Journal of Intelligent Systems and Applications(IJISA)*.
- Dang, S., & Ahmad, P. H. (2014). Text Mining: Techniques and its Application. *International Journal of Engineering & Technology Innovations, ISSN (Online)*, 1(4), 866–2348.
- Davcheva, P. (2018). Text Mining Mental Health Forums-Learning from User Experiences. In *Twenty-Sixth European Conference on Information Systems* (pp. 1–11).
- Iskonomakis, M, S. Kotsiantis, and V. T. (2005). Text Classification Using Machine Learning Techniques.” WSEAS Transactions on Computer. In *WSEAS Transactions on Computers* (pp. 966–974).
- Kanavos, A., Nodarakis, N., Sioutas, S., Tsakalidis, A., Tsolis, D., & Tzimas, G. (2017). Large scale implementations for twitter sentiment classification. *Algorithms*, 10(1), 1–21. <http://doi.org/10.3390/a10010033>
- Kaushik, A., & Naithani, S. (2016). A Comprehensive Study of Text Mining Approach. *International Journal of Computer Science and Network Security*, 16(2), 69–76.
- Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015). From Group to Individual Labels Using Deep Features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15* (pp. 597–606). <http://doi.org/10.1145/2783258.2783380>
- Nagwani, N. K., & Sharaff, A. (2017). SMS spam filtering and thread identification using bi-level text classification and clustering techniques. *Journal of Information Science*, 43(1), 75–87. <http://doi.org/10.1177/0165551515616310>
- Sailaja, N V., Padamsree, L. & Mangathauaru, N. (2016). Survey of Text Mining Techniques , Challenges and their Applications. *International Journal of Computer Applications*, 146(11), 30–35.
- Singh, T. (2016). A Comprehensive Review of Text Mining. *Interntional Journal of Computer Science & Information Technology*, 7(1), 167–169.
- Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Review Article Data Processing and Text Mining Technologies on Electronic Medical Records : A Review. *Journal of Healthcare Engineering*, 1–10.
- Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *IJACSA) International Journal of Advanced Computer Science and Applications*, 7(11), 414–418. Retrieved from www.ijacsa.thesai.org
- Usai, A., Pironti, M., Mital, M., & Mejri, C. A. (2018). Knowledge discovery out of text data : a systematic review via text mining. *Journal of Knowledge Management*, 1–19. <http://doi.org/10.1108/JKM-11-2017-0517>
- Yanhong, L., Anmeng, S., & Jingling, W. (2017). A Survey of Current Work in Medical Text Mining---Data Source Perspective. *International Journal of Biomedical Science and Engineering*, 5(3), 29–34. <http://doi.org/10.11648/j.ijbse.20170503.13>