Vol. 8(1), pp. 1-9, July-December 2019
DOI: 10.5897/IJCER2019. 0281
Article Number: C9AC37C61384
ISSN 2141-6494
Copyright © 2019
Author(s) retain the copyright of this article
http://www.academicjournals.org/IJCER

ACADEMIC JOURNALS
expand your knowledge

International Journal of Computer Engineering Research

Full Length Research Paper

Analysis and result of classification algorithm on email classification

Elifenesh Yitagesu Desta* and Tekalign Tujo Gurmessa

Department of Computer Science, College of Computing, Madda Walabu University,P. O. Box 247, Bale Robe, Ethiopia.

Received 21 January, 2019; Accepted 13 May, 2019

In this time, one of the most and fastest forms of communication is electronic mail or what we call e-mail. However, the increase of e-mail users has resulted in the dramatic increase of spam emails in the past few years. Spam is the use of electronic messaging systems to send bulk data. In this paper, e-mail data were classified as ham email and spam email using supervised learning algorithms. Three different classifiers such as Naïve Bayesian (NB) classifier, K-nearest neighbor (KNN) classifier and Support Vector Machine (SVM) classifier were used. The experiment was performed by applying filtering on the classifiers. The result shows the difference between the classifier before and after applying filtering algorithm. To examine the performance of the selected classification methods or algorithms, namely Naïve Bayes, SVM and KNN, true positive, false positive, precision, recall and F-measure were validated. There was a time difference using those classification algorithms. KNN and SMO algorithms are almost the best classifiers among the three before applying filtering algorithm. Sequential minimal optimization (SMO) is an algorithm used to solve quadratic programming (QP) problem that arises during the training of support vector machines (SVM) and after applying filtering algorithm. SMO algorithm is the best classifier algorithm. For this experiment, the data mining tool called WEKA was used.

Key words: WEKA, classifier, K-nearest neighbor (KNN), support vector machines (SVM), Naïve Bayesian (NB), boosting.

INTRODUCTION

Email is the short form of electronic mail and it is defined as the exchange of information through communication channel. Typically, emails come from different email addresses rather than being entered from the key board or electronic files stored on the disk. Most mainframes, minicomputers, and the emailing system are applied on the computer network. The term electronic mail can also be written as Email or e-mail. Email address is required to send and receive email messages. The majority of internet service providers provide a free email account to

*Corresponding author. E-mail: elifeneshyitagesu@gmail.com.

customers. Email has been tested to be one of the Internet's preferred services; it is used for international communications. But, it is criticized for its insecurity, spam, as well as viruses and malware being unfold through email attachments. E-mail offers a way for web users to simply transfer information globally. E-mail presents a super way to send millions of commercials free of charge for the sender, but the bad thing is that these days' emails are appreciably exploited. Generally, receiving an e-mail from an unknown supply comprises contents that are of no importance to the user. As a result, due to these e -mails, many people are getting cluttered with all unsolicited bulk e-mails also referred to as "spam" or "unsolicited mails" (Vinod et al., 2013). Spam often causes unwanted information or bulk information to get transmitted to email accounts. Spam mail could be a set of electronic spam involving nearly identical messages sent to numerous recipients. Spam emails conjointly embrace malware as scripts or alternative executable file attachments. Spam is waste of time, storage space and communication bandwidth. If spam continues to increase, it will be unmanageable in the near future to handle such huge spam.

Automatic e-mail filtering looks like the foremost effective methodology to counter spam at the moment and has a good competition between spammers and spam-filtering ways. In the past, most of the spams were treated by the interference of e-mails coming from sure addresses or filtering out of messages with sure subject lines. Spammers began to use many difficult ways to beat the filtering ways like victimization of random senders' addresses and/or appending of random characters to the start or the tip of the message subject line. Spam emails vicinity unit is used to spread a virus or malicious code for fraud in banking, publishing, advertising and much more (https://pdfs.semanticscholar.org/c2ea/4bf0282b9b39a6b a773581332bb0587ec4ab.pdf). (Nilam et al.,2017) So to avoid this kind of bulk email it is essential to use spam filtering technique which is a machine learning algorithm. In this study, we cover the performance of three widely used supervised machine learning method for data classification and identify the best classifier algorithm. Those supervised machine learning algorithms are K-nearest neighbor (KNN), Support vector machines (SVM), and Naive Bayesian (NB). Supervised learning is one of the methods associated with machine learning which involves allocating labeled data so that a certain pattern or function can be deduced from that data. It is worth noting that supervised learning involves allocating an input object, a vector, while at the same time anticipating the most desired output value, which is mostly referred to as the supervisory signal. The bottom line property of supervised learning is that the input data are known and labeled appropriately. In a study by Binh et al. (2018), four Bayesian machine learning algorithms (NB, NBT, BN and DTNB) were selected and compared with one of the benchmark landslide models of the SVM for landslide susceptibility assessment at Pauri Garhwal district, Uttarakhand State, India. Results show that the SVM model was highly reliable followed by the NBT, DTNB, BN and NB. This is in accordance with the results of statistical index based methods and the ROC curve.

In this work, supervised machine learning is used rather than unsupervised machine learning because it is worth noting that both methods of machine learning require data to analyze to produce certain functions or data groups. However, the input data used in supervised learning are well known and labeled. This means that the machine is only tasked with the role of determining the hidden patterns from already labeled data. However, the data used in unsupervised learning are not known nor labeled. It is the work of the machine to categorize and label the raw data before determining the hidden patterns and functions of the input data. It does not only input data accurately but also gives accurate and reliable results. To review the performance outcome of the three machine learning strategies six terms were used: true positive, false positive, recall, precision, f measure and accuracy. These are called imagining the algorithms. The entire machine learning algorithms give different results on the same dataset. This paper focuses on effective and efficient email classification techniques based on data filtering method used for the training model and accuracy of the algorithm before and after filtering the classification method. It also compares the accuracy of algorithms before and after boosting.

METHODOLOGY

K-nearest neighbor (KNN)

For the KNN classifier, three nearest neighbors and Euclidean distance function were used. In the classification phase, KNN searches the training example and calculates the distance between every sample set of the class label and observation. After that, it recognizes the nearest neighbors and then predicts the final classification output (Atia et al., 2019).

Support vector machine (SVM)

Support vector machines were introduced by Cortes and Vapnik (1995) to simultaneously minimize classification error and maximize the margin between two classes (Vahid et al., 2018). Support vector machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates a set of objects having different class memberships (Kishore et al., 2012).

Naïve Bayes classification (NB)

Naïve Bayes model is a probabilistic classifier based on the Bayes theorem with the simplistic (naïve) assumption that features are independent. Although this assumption is usually violated in practice, NB classification still performs well (Vahid et al., 2018). In 1998 the Naïve Bayes classifier was proposed for spam recognition. Bayesian classifier works on dependent events and the probability of an event occurring in the future based on previous occurrence of the same event (Almeida et al., 2011). This technique can be used to classify spam e-mails. Words probabilities play the main rule here (Awad and EL Seuofi, 2011).

RESULTS AND DISCUSSION

Performance evaluation

Spam filtering researchers use the best analysis methods for performance evaluation. So here we used the following:

(1) Spam Precision (SP)
(2) Spam Recall (SR)
(3) Accuracy (A)

Spam precision (SP) represents the range of relevant documents identified as,

$$SP= \frac{\#\ of\ spam\ correctly\ classified}{total\ \#\ of\ message\ classifies\ as\ spam} = \frac{N_{spam\_spam}}{N_{spam\_spam}+N_{ham\_spam}}$$

Percentage of all documents known: This shows the noise that filter presents to the user, that is many of the messages classified as spam will actually be spam emails.

Spam recall (SR): This performance evaluation is the percentage of all spam emails correctly classified as spam emails.

$$SR = \frac{\#\ of\ spam\ correctly\ classified}{total\ \#\ of\ email} = \frac{N_{spam\_spam}}{N_{spam\_spam}+N_{spam\_ham}}$$

Accuracy (A) is the compulsory performance evaluation method that shows the percentage of all emails that are correctly categorized or classified.

$$A = \frac{\#\ of\ emails\ correctly\ classified}{total\ \#\ of\ emails} = \frac{N_{ham\_ham}+N_{spam\_spam}}{N_{ham}+N_{spam}}$$

where N_{ham_ham+} and N_{spam_spam} are the number of emails correctly classified to the legitimate email and spam email, respectively; $N_{ham_spam}$and $N_{spam_ham}$are the number of legitimate or ham email and spam email that have been misclassified; that means that spam as ham and ham as spam message; N_{ham} and N_{spam} are the whole number of legitimate or ham and spam messages to be classified.

Performance comparison

To review the performance outcome of the three machine learning strategies, six terms were used: true positive, false positive, recall, precision, f measure and accuracy. The entire machine learning algorithms give different results on the same dataset. The performance criteria for evaluating the classifiers are: classification accuracy, kappa statistic, mean absolute error, root mean squared error, virtual absolute error, and time.

(1) Classification accuracy: This is one of the performance criteria and it includes both correct and incorrect classified accuracy. And also it is the ability to predict categorical class labels. It calculates the proportions of correctly classified instance as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

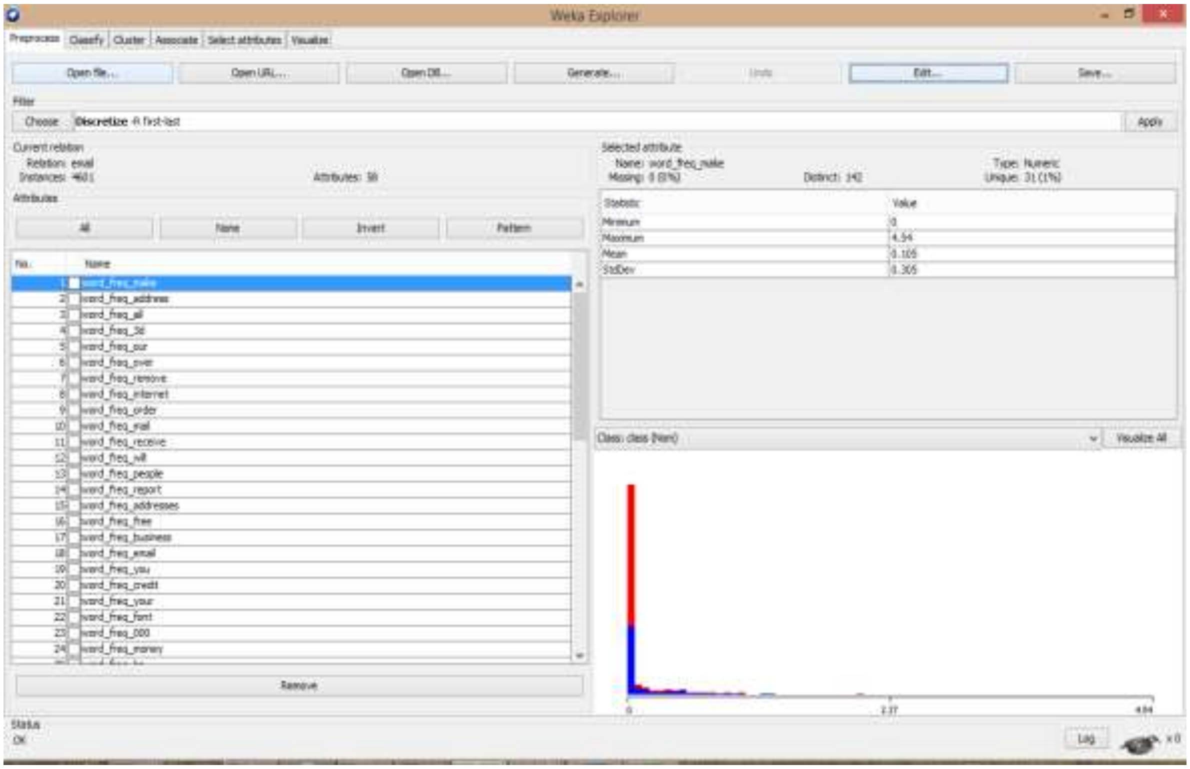
(2) True positive (TP): Is the amount of examples classified as spam among all examples which truly have spam.
(3) False positive (FP): is the amount of examples classified as spam but goes to a different class among all examples which are not spam previously.
(4) Precision: Is the amount of examples which actually have spam amongst all those which were classified as spam earlier.
(5) F- Measure is simply calculated as $\frac{2*Precision*Recall}{Precision+Recall}$, a collective measure for both precision and recall imagining methods (Kishore et al., 2012).

WEKA result

The procedure of Email classification through WEKA works is shown in Figure 1 with results. We used the dataset on the link http://archive.ics.uci.edu/ml/dataset/Spambase as seen in the design model. It contains both spam and non-spam email data sets (Figure 1).

The dataset contains a total of 4061 instances; from this 1813 are spam (39.4%), 2788 are ham (61%) and 58 (57%) are continuous, with 1 nominal class label. Attributes of the spam email dataset are loaded into Weka data mining tool. The loaded dataset is in .arff data file format because weka accepts the data in .arff file format alone. An Attribute-Relation File Format (ARFF) file is an ASCII text file that explains a list of instance distribution from a set of attributes (Figures 2 and 3) methods or algorithms namely Naïve Bayes, SVM and KNN, true positive, false positive, precision, recall and F-measure were validated.

From the three classifier algorithm, SMO is the best



Filter

Discretize -R first-last

Current relation

Relation: email

Instances: 4601

Attributes: 58

Selected attribute

Name: word_freq_make

Missing: 0 (0%)

Distinct: 142

Type: Numeric

Unique: 31 (1%)

Statistic	Value
Minimum	0
Maximum	4.54
Mean	0.105
StdDev	0.305

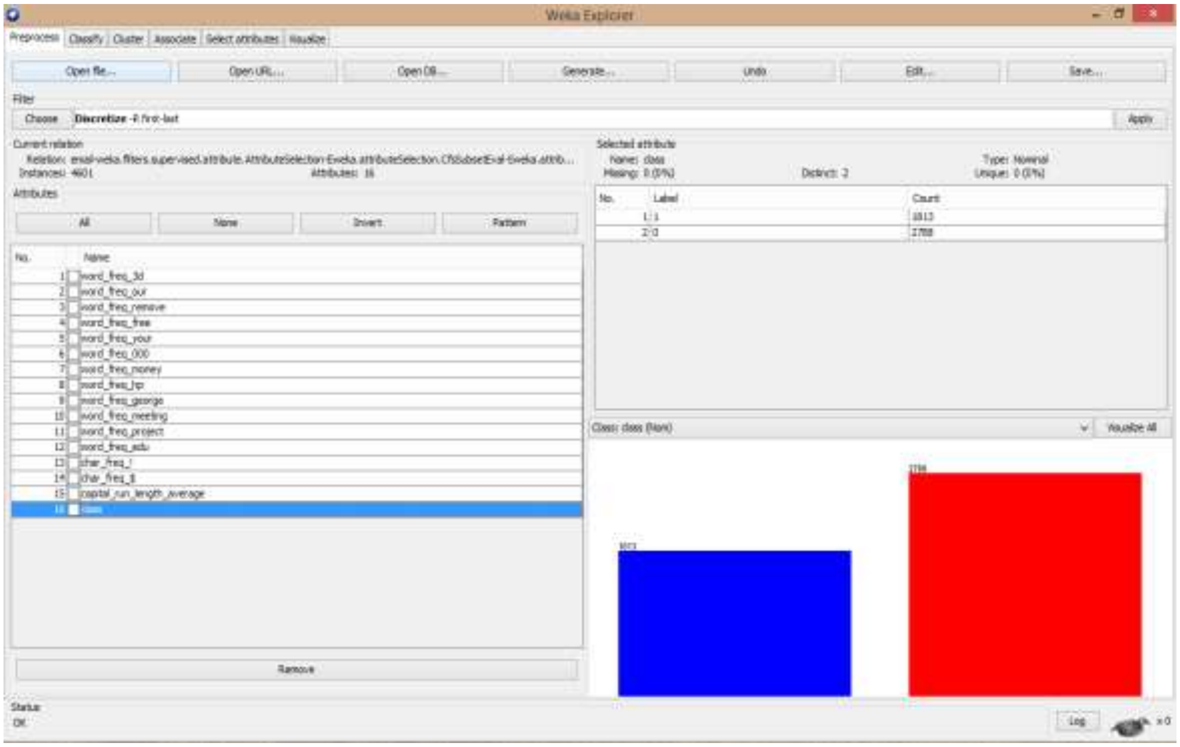
Attributes

No.		Name
1	□	word_freq_make
2	□	word_freq_address
3	□	word_freq_all
4	□	word_freq_3d
5	□	word_freq_our
6	□	word_freq_over
7	□	word_freq_remove
8	□	word_freq_internet
9	□	word_freq_order
10	□	word_freq_mail
11	□	word_freq_receive
12	□	word_freq_will
13	□	word_freq_people
14	□	word_freq_report
15	□	word_freq_addresses
16	□	word_freq_free
17	□	word_freq_business
18	□	word_freq_email
19	□	word_freq_you
20	□	word_freq_credit
21	□	word_freq_your
22	□	word_freq_font
23	□	word_freq_000
24	□	word_freq_money

Status

OK

Figure 1. Loading of the spam dataset into weka tool.



Filter

Discretize -R first-last

Current relation

Relation: email-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval-Eweka.attrib...

Instances: 4601

Attributes: 16

Selected attribute

Name: class

Missing: 0 (0%)

Distinct: 2

Type: Nominal

Unique: 0 (0%)

No.	Label	Count
1	1	1813
2	0	2788

Attributes

No.		Name
1	□	word_freq_3d
2	□	word_freq_our
3	□	word_freq_remove
4	□	word_freq_free
5	□	word_freq_your
6	□	word_freq_000
7	□	word_freq_money
8	□	word_freq_hp
9	□	word_freq_george
10	□	word_freq_meeting
11	□	word_freq_project
12	□	word_freq_edu
13	□	char_freq_!
14	□	char_freq_$
15	□	capital_run_length_average
16	□	class

Status

OK

Figure 2. Weka explorer showing the data set after applying filtering.

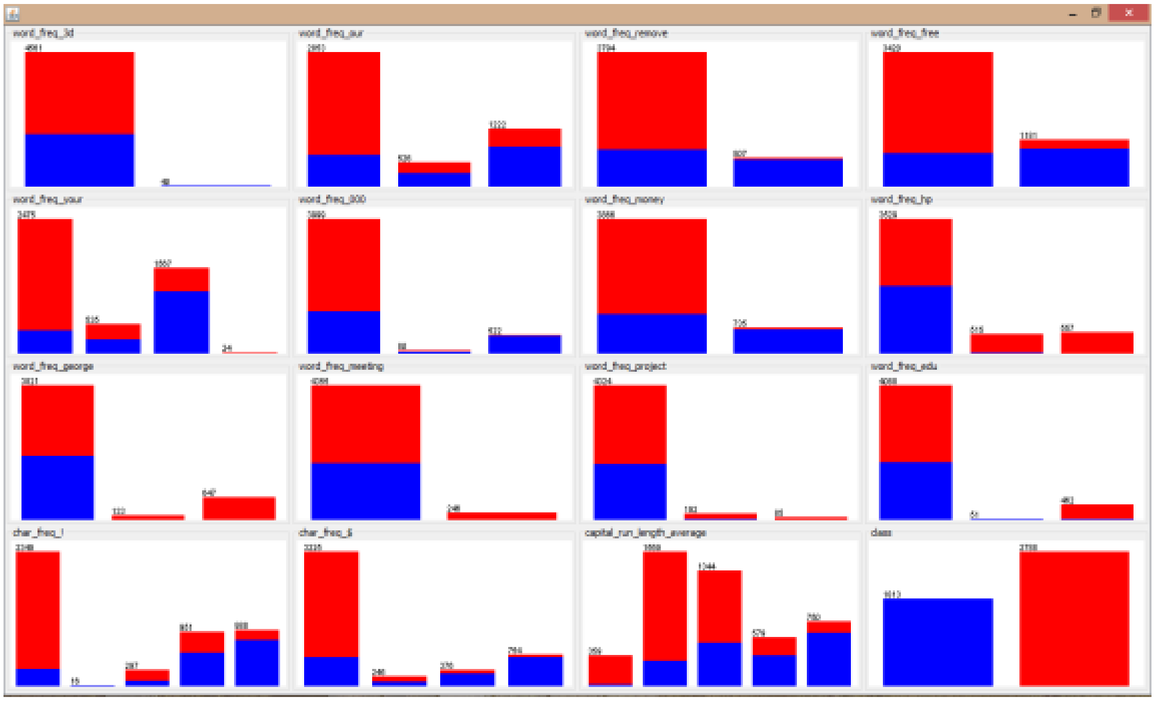


Figure 3. Visualizing the entire attribute after applying filtering.

Table 1. Result before filtering of classification methods.

Algorithm	TP	FP	Precision	Recall	F-measure
KNN	0.908	0.103	0.908	0.908	0.908
NB	0.793	0.152	0.842	0.793	0.794
SMO	0.904	0.121	0.905	0.904	0.903

Table 2. Result after filtering of classification methods.

Algorithm	TP	FP	Precision	Recall	F-measure
KNN	0.93	0.081	0.929	0.93	0.91
NB	0.927	0.093	0.927	0.927	0.926
SMO	0.939	0.075	0.939	0.939	0.938

spam filtering methods in both ways, before and after filtering mechanisms.

The accuracy of SMO in both experiments, that is, before and after applying filtering method, had the best accuracy than the KNN and NB algorithms. In the study by Binh et al. (2018), SVM classifier had the best accuracy than Bayesian algorithms like NB, NBT, BN and DTNB (Jin et al., 2003). (Tables 1 to 4 and Figures 4 to 8).

Boosting

Boosting is a machine learning ensemble Meta algorithm used primarily to decrease bias and also variance in supervised learning; it is a family of machine learning algorithm which converts weak learners to strong learners by boosting those algorithms. It incrementally structures an ensemble by training each new classical instance to highlight the training instance that previous

Table 3. The accuracy of the algorithms before filtering.

Algorithm	Correctly classified instance (%)	Incorrectly classified instance (%)	Time
KNN	90.7629	9.7629	0
NB	79.2871	20.7129	0.03
SMO	90.4369	9.5631	0.43

Table 4. The accuracy of the algorithms after filtering.

Algorithm	Correctly classified instance (%)	Incorrectly classified instance (%)	Time
KNN	92.9581	7.0419	0
NB	92.6538	7.3462	0
SMO	93.8709	6.1291	1.62

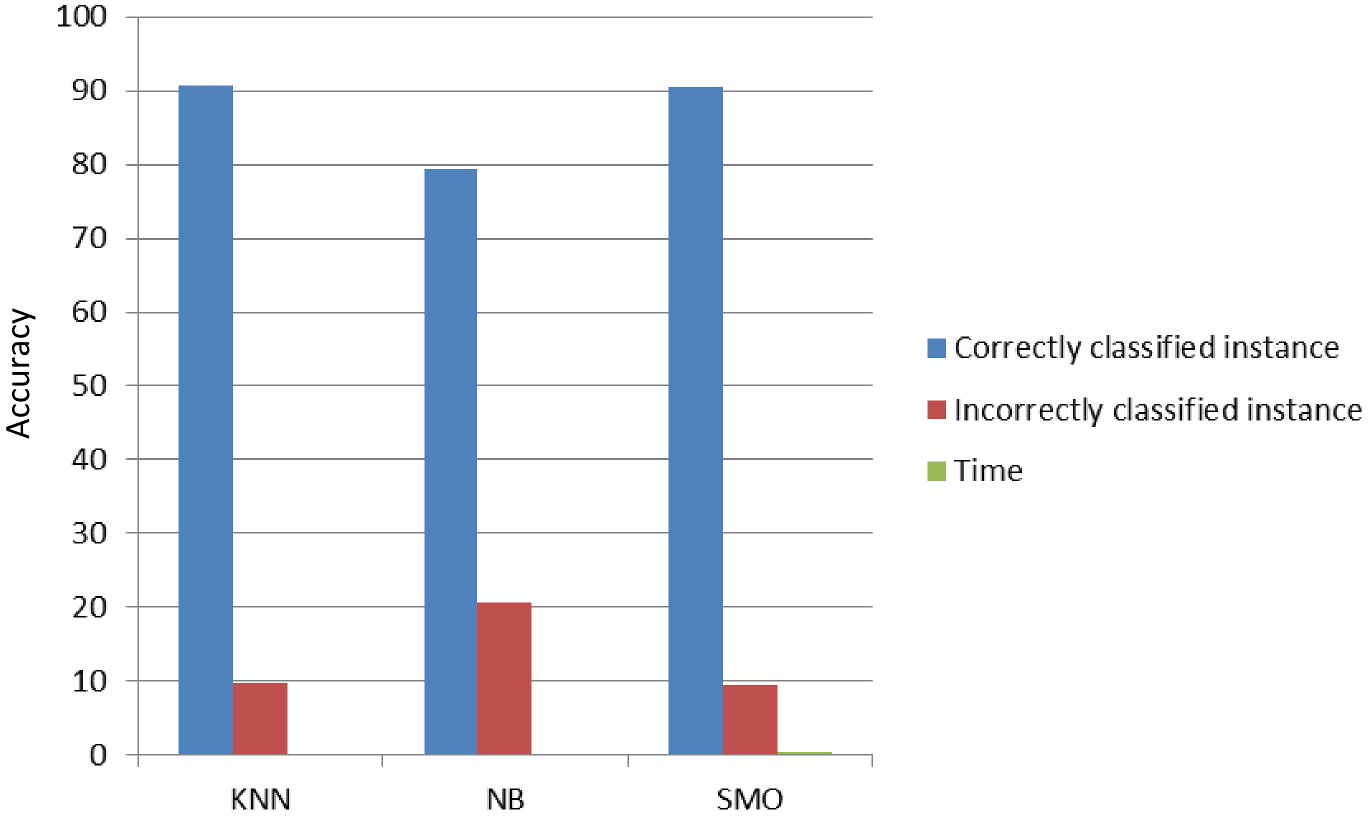


Figure 4. The accuracy of the algorithms before filtering.

models miss-classified. Here, to modify or improve the classifier accuracy, we just used the attribute selection method on the preprocessing section; this means that for attribute selection all the attributes are not used on the classifier before using selection filtering method to improve the accuracy of the classifier.

Conclusion

Effective filtering techniques should be used to avoid spam or irrelevant mail. In this work, different supervised classifier techniques such as instance based 1, Naïve Bayes, and Support Vector Machine were used.

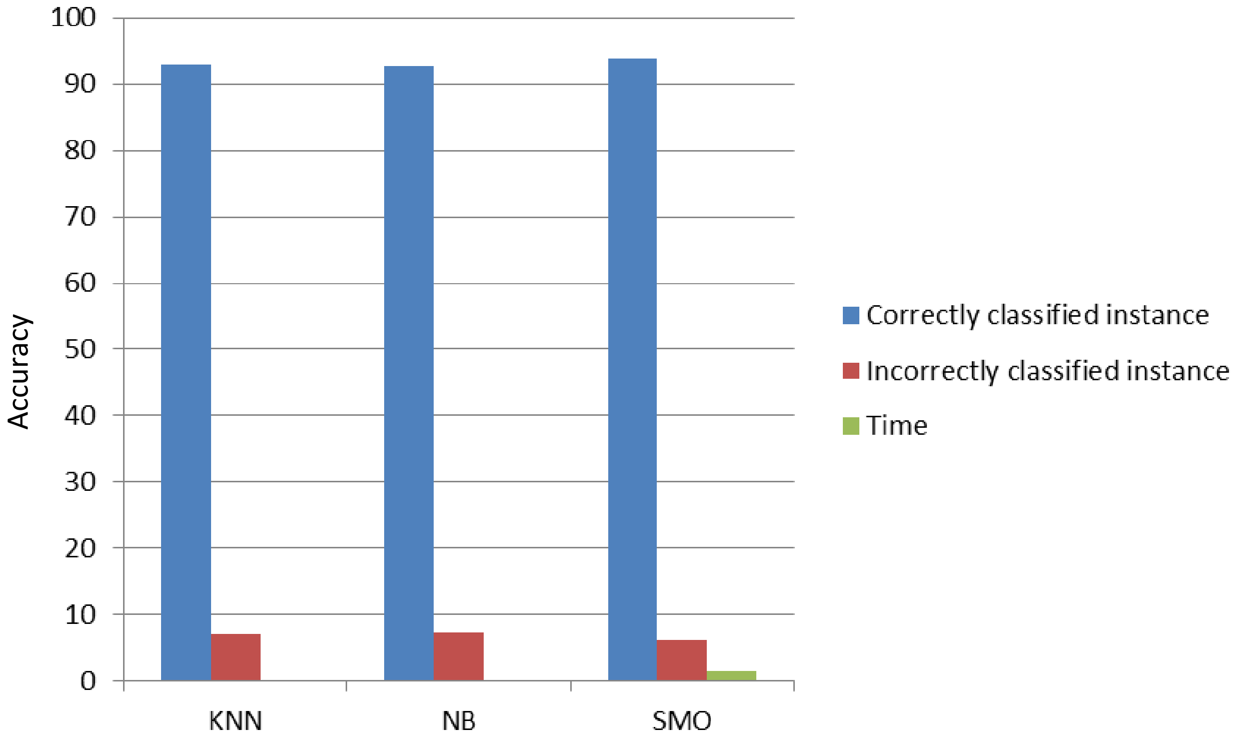


Figure 5. The accuracy of the algorithms after filtering.

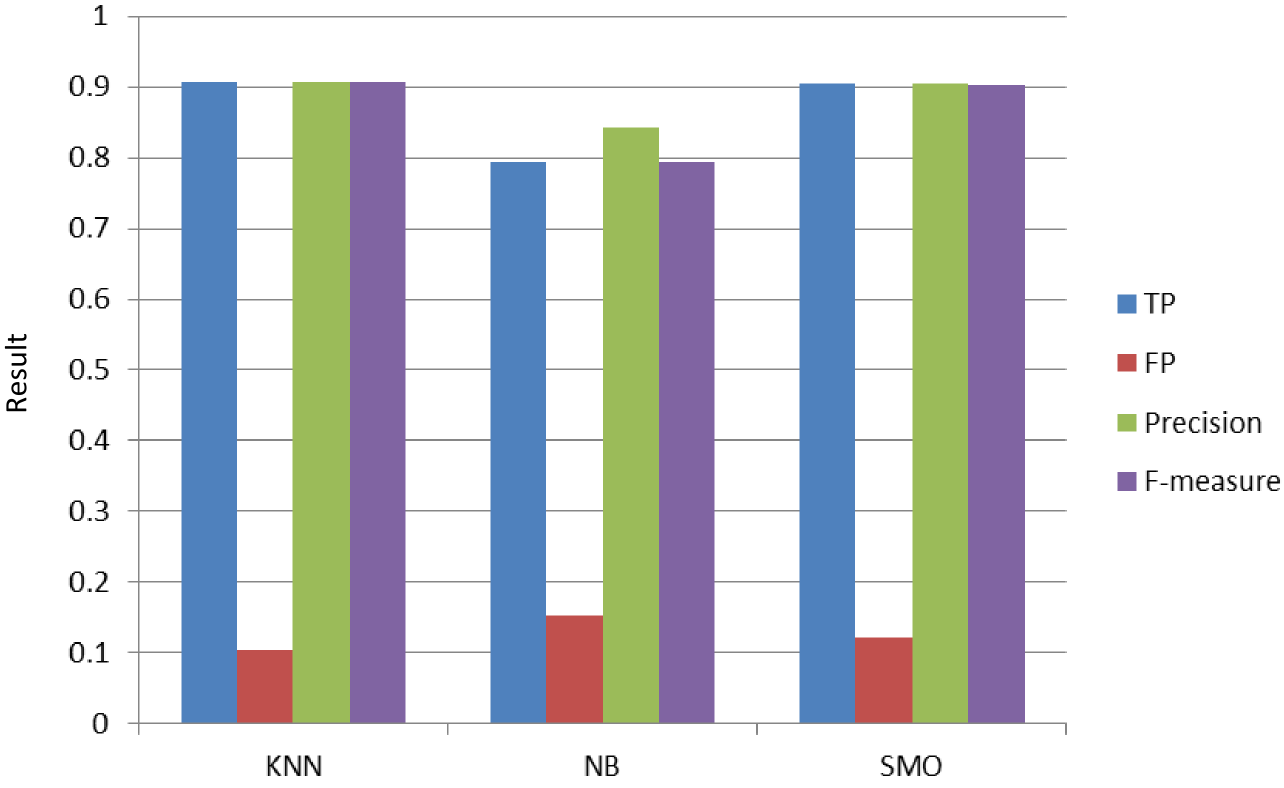


Figure 6. Result before filtering of classification methods.

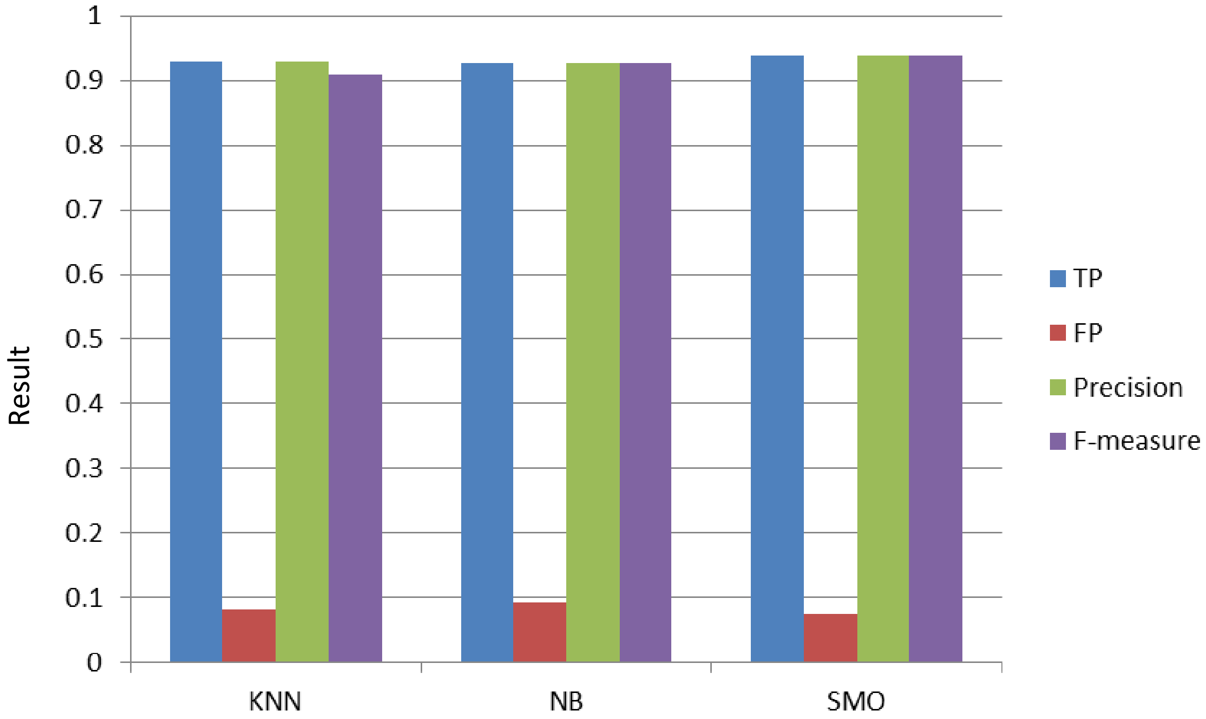


Figure 7. Result after filtering of classification methods.

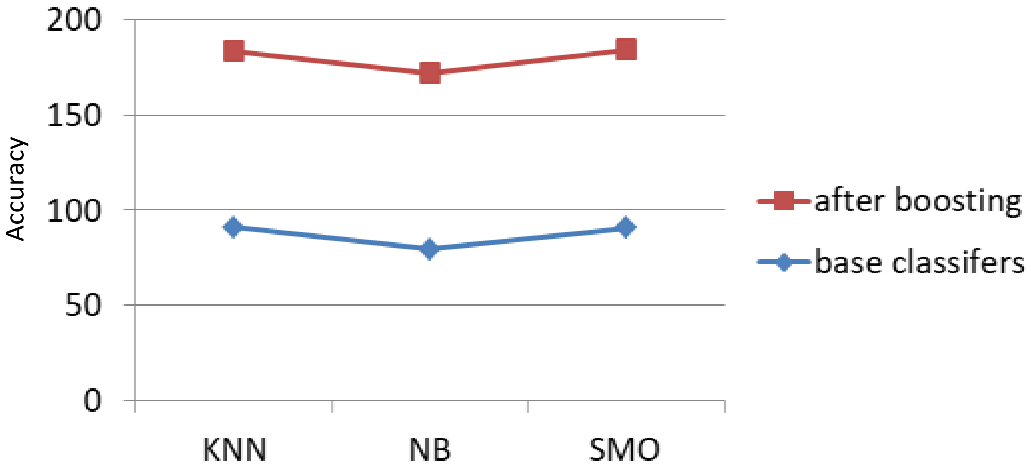


Figure 8. Comparison of accuracy with and without boosting.

Classifiers used for the datasets without using filtering techniques had low accuracy value. But by using the filtering technique called attribute selection method on the data set, we got the best accuracy than the previous method. The accuracy of KNN and SVM in the first experiment (before applying filtering technique) was 90.76 and 90.43%, respectively. But in the second experiment (after applying filtering technique), SVM is a classifier with the best accuracy (93.87%) than the other two.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

ACKNOWLEDGEMENTS

The author is grateful to her Computer Science Department for supporting her work and Computer Science non-teaching staff for helping with the hardware and software issue.

REFERENCES

Almeida T, Almeida J, Yamakami A (2011).Spam filtering:how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. Journal of internet services and applications 1(3):183-200.

Atia J, Nadeem J, Zahid W ,Tanzila S, Osama S, Muhammad Q S, Mohammad E A (2019). Machine Learning Algorithms and Fault Detection for Improved Belief Function Based Decision Fusion in Wireless Sensor Networks. sensors 19(6).

Binh T P, Indra P, Khabat K, Kamran C, Phan TT, Trinh QN, Seyed V H, Dieu T B (2018) . A comparison of Support Vector Machines and Bayesian Algorithms for Landslide Susceptibility Modeling. Geocarto International pp. 1-23.

Nilam B, Namrata C, Ronit C, Shraddha M (2017). Spam E-mail detection using classifiers and Adaboost. International Journal of Computer Engineering and Application *XI*(VIII). https://pdfs.semanticscholar.org/c2ea/4bf0282b9b39a6ba773581332bb0587ec4ab.pdf

Kishore RK, Poonkuzhali G, Sudhakar P, Member LAENG (2012). Comparative Study on Email Spam Classifier using Data Mining Techniqes. Proceedings of the international Multi Conference of Engineering and Computer Scientists 1:14-16

Vahid N, Sepideh N, Stavros A, Julie C (2018). Classification of thermally treated wood using machine learning techniques. Journal of the international academy of wood science 53(1):275-288.

Vinod P, Divakar S, Anju S (2013). A Novel Technique of Email Classification for Spam Detection. International Journal of Applied Information Systems 5(10).

Awad WA, ELseuofi SM (2011). Machine learning methods for spam e-mail classification. International Journal of Computer Science and Information Technology 3(1):173-184.