International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X;pISSN:2321-2241 Volume: 13; Issue: 7; July- 2024

Advanced Techniques for Mobile Phone Spam Detection

Mridula Shukla
Assistant Professor
Department of Master of Computer Applications
The Oxford College of Engineering
mridulatewari005@gmail.com

Yogasathish S
PG Student
Department of Master of computer Applications
The Oxford College of Engineering
yogasathishmca2024@gmail.com

Abstract:

Despite the proliferation of messaging programs on mobile phones these days, SMS remains the preferred method of communication. But these days, when SMS tariffs are reduced, here is too a rise in SMS spam, which some people employ as a substitute for fraud and advertising. Because it can annoy and hurt consumers, it becomes a serious problem.. Accuracy is one of the maximum difficult aspects of SMS spam filtering. In this study, we suggested merging two data mining tasks—association and classification—to improve SMS spam filtering performance. The Naive Bayes Classifier is used to determine whether an SMS is spam or ham, and FP-growth in association is used to mine frequent patterns on SMS. Previous research's SMS spam collecting was used to create the training data. When Naive Bayes and FP-Growth work together, the dataset SMS Spam Collection v.1 achieves the greatest average accuracy of 98, 506%, and 0,025% better than it would have without FP-Growth. It also improves the precision score, meaning the classification result is more accurate.

Keywords- SMS spam; Naive Bayes; FP-Growth; text classification

I. INTRODUCTION

SMS is a text-based messaging app that lets workers send brief texts with others using mobile phones (often, these texts are much longer than 160 characters in 7-bit). The fact that it is the record generally used and popular form of communication does not change the detail that many people utilize it for illegal activities like fraud and media advertising. Since the SMS tariff in China is significantly less than $0.001, single of the reasons for the rise in SMS spam is the decreased SMS rate. Furthermore, this number is higher than email spam, according to Korea Information Security (KISA). For example, each week, 1.1 billion SMS spam are sent to US mobile users, while 8,29 billion SMS spam are sent to Chinese mobile users.

The above-mentioned issues can be resolved with a pardoned solution. In this circumstance, SMS are filtered giving to the text

classification. Popular text categorization methods include neural networks, nearest neighbours, decision trees, Naive Bayes, rule induction, and Support. The Vector Machine. However, the organization of SMS differs from that of ordinary document texts or emails because SMS texts are typically casual, have a maximum text length of 160 7-bit characters, and are frequently truncated. The question "Is the feature good enough to distinguish between SMS spam and non-spam?" arises when an SMS is extremely brief. Furthermore, there are now a plethora of different sorts of SMS, thus another technique is required to include structures that can discern between SMS spam and non-spam. But all forms of SMS that are currently in use follow a similar trend, especially SMS spam. That situation can serve as the foundation for using the approach where words appear one after the other concurrently.

Two approaches work together in this experiment: the FP-Growth Algorithm frequent itemset and the Naive Bayes classifier. In machine learning for data recovery, naive bayes is considered to be one of the most significant and successful learning algorithms. Furthermore, giving to the mentioned paper, as compared to the Naive Bayes implementation alone, the accuracy can be improved by implementing a user-specified minimum support scheme. Every frequently occurring word is regarded as solitary, independent, and mutually exclusive in addition to being mutually independent as the minimum support yields the frequent itemset as an extra characteristic. It also has the skill to increase opportunity scores and produce a classification system that is more accurate. The frequent itemset is obtained using the Apriori Algorithm in the cited paper.

II. DESIGN OF SYSTEM

The built system typically contains of two stages: the general system design seen in figure 1 and the exercise and challenging procedure.

A.Analysis

To create the classification model, the data training procedure is used. Aside from that, the testing procedure involves evaluating the classification results derived from the obtained model.

Preprocessing the facts is the initial step. To make the next steps easier, the testing and preprocessing of the drill data are done independently. Before the training (using training data) and testing (using testing data) processes, preprocessing is done at an early stage. The preprocessing steps being conducted are as follows:

1. Example Erasing characters by folding them

In order to homogenize the data and eliminate all characters other than letters,

numbers, and punctuation, all text is transformed to lowercase.

2. The use of tokens

Token processing is done first in order to break the string into a token or a single word, which might help with the token search process, before the subsequent steps (processes 3 through 6) are taken.

3. Managing Slang Terms

There are many colloquial terms in the dataset that are used as slang terms. To contract with those terms, a dictionary that includes the slang words together with their true meanings is made. The list of words in the Slang dictionary comes from the.

4. Eradication of Stopwords

The word-dictionary matching technique is used with stopword lists that are downloaded from the website in instruction to eliminate the terms that are share of the stopwords.

5. Using stems

Many of the standings in the dataset have prefixes; consequently, stemming is required to return those words to their base forms. Its goal is to reduce word variants that occur when two words with the same sense but dissimilar affix forms are used. The Porter Stemmer algorithm, which is implemented by the Snowball Tartarus library, is approved out by this method.

6. Manage the Number

All that this process does is manage the phone number's numeric characters. It is done because, according to observations, a big amount of phone numbers—many of which fall into the category of spam—appear in the SMS dataset. As a result, the phone numbers could provide a distinctive characteristic for the classification of SMS texts. In order to accommodate a numeric character in a token, a provision measuring 2'. 7 (the length of a minimum standard telephone number) has been created. To further homogenize all the phone number data from the numeric characters set into a single word, a character token made up of those digits is transformed into a "phone number" string. Should the digits come together in a token

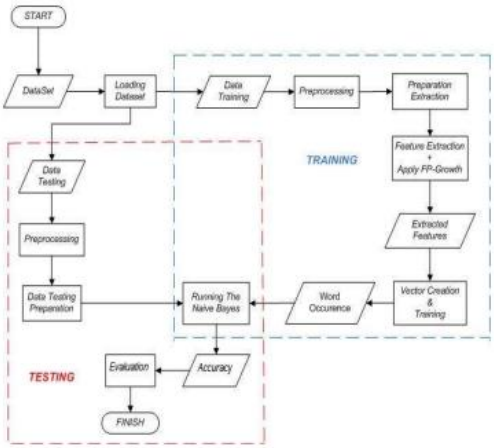

Fig. J. The general description of the system

B. The Extraction of Features

The FP-Growth method is used in the drill data feature extraction process to obtain the frequently occurring itemset, as depicted in the

figure.2.

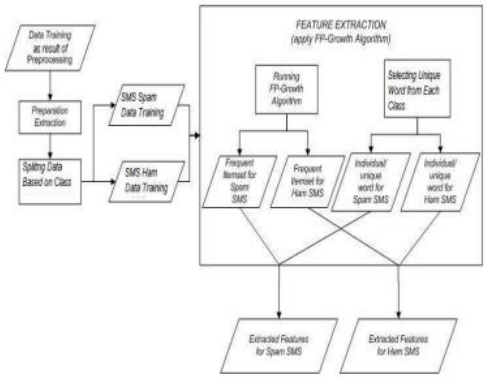

Fig. 2. The procedure of feature removal

The following are the explanations for Figure 2.

1. The extraction preparation step transforms the word-formatted SMS data into a numeric format before it is analysed. The data is then divided into each class, resulting in the acquisition of two input files for additional processing.

2. The minimal support required for FP-Growth operating is determined by the tests,

3. Frequent itemset become the new features for each class in the classification process as a result of the FP-Growth process.

4. The New features are paired with specific elements of

C. Vector Creation and Training

In this process, the calculation for each of the name that has been haul out in each class is performed. To simplify the calculations, the vector table is created and converted to a form of word occurrence table.

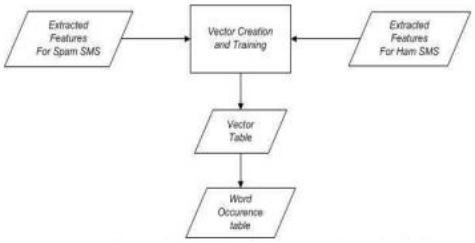

Fig. 3. The procedure of vector creation and training

A vector table displays the features or words that appear in each SMS sentence. The word occurrence table, on the other hand, is a table that lists every word that appears in every class.

D. Utilizing the Naive Bayes Framework

The Naive Bayes technique is being used to classify data at this point. This involves counting the words on the word occurrence table, calculating the total, and determining the prior chance for each class (spam and ham). Subsequently, data testing is entered to complete the preparation process and do the categorization. To prevent the O probability score during the arrangement stage, the Laplace estimator or Laplace smoothing computation is used.

Fig. 4. is an overview of the classification with Naive Bayes.

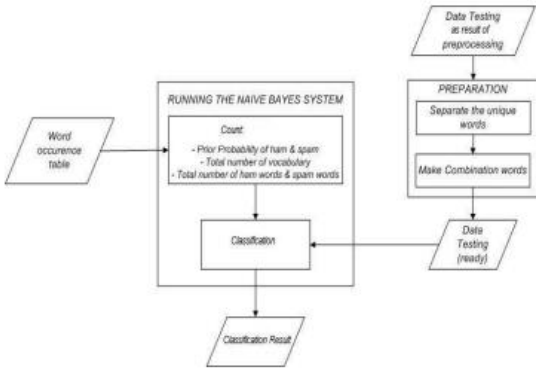

Fig. 4. running naive bayes

E. Evaluation (Testing)

By computing the accuracy score, it is possible to determine through the evaluation process whether or not the produced model is feasible to apply. It implies that the outcomes of each class's instruction constitute the model. procedures in the procedure of scores for variables like previous probability, the quantity of words learned in each lesson, and the sum of acquired vocabularies. It is likely to use the model in the categorization process of a new SMS if the accuracy result achieves a high score. Measuring recall (r) and precision (p) is an excellent approach to assess how well a text classification to a word performed. The precision is the degree of agreement between the user's requested info and the system's responses. Meanwhile, the recall measures how well the system rediscovers information.

$$p = \frac{tp}{tp+fp} \qquad r = \frac{tp}{tp+fn}$$

$$F\ measure = \frac{2 \times p \times r}{p+r}$$

False positives are valid messages (ham) that are mistakenly viewed as spam, false negatives are spam that is mistakenly interpreted as ham, and true positives can be understood as messages that are considered spam.

The accuracy can be computed once the meaning of each prec1s10n and recall has been established. The degree of similarity between the expected and actual scores is known as accuracy, and it can be expressed as follows:

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

The next model, which is founded on the calculation of word occurrence, can be employed in the course of a new SMS predictive categorization after achieving good accuracy through training and testing. It is then categorized to discriminate among SMS spam and ham.

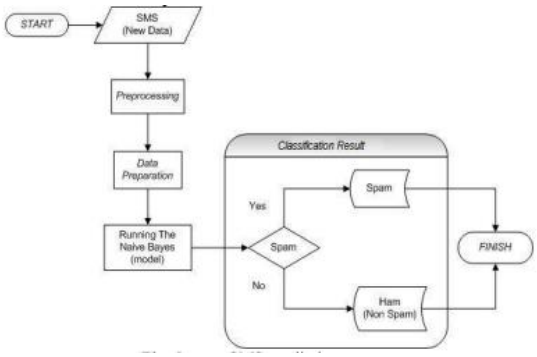

Fig. 5. new SMS prediction process

III. TESTING

A. Dataset

The dataset from SMS Spam Collection v.l with SMS amounted to 5574 SMS consisting of 4,827 ham SMS and 747 spam SMS, as fine as from SMS Spam Corpus v.0.1 Big with SMS amounted to 1324 SMS consisting of 1002 ham SMS and 322 spam SMS, are used to determine the routine of the SMS filtering system built in this research.

B. The Testing Process

Six testing scenarios are run through this test in the subsequent order:

1. FP-growth-free Naive Bayes testing with the SMS Spam Corpus v.0.1 large dataset
2. FP-growth Naive Bayes Testing with SMS Spam Corpus v.0.1 Large-scale Dataset
3. Using the SMS Spam Collection dataset v.1, Naive Bayes Testing without FP-growth is performed.
4. Using the SMS Spam Collection dataset v. l., Naive Testing Bayes with FP-growth
5. Using both datasets, naive Bayes testing without FP-growth
6. Using both datasets, naive bayes testing with FP-growth
7. Naive Bayes testing with FP-growth conferring to the dataset properties

RESULT:

The study of the application of the optimal minimal support and a contrast of the use of Naive Bayes only with the custom of the Naive Bayes and FP-Growth cooperation are carried out founded on the l sat until the sixth testing procedure.

The study of the data characteristics appropriate for implementation using the FP-Growth approach is then carried out in the seventh testing procedure.

1) The FP-Growth minimum support analysis

The outcomes of applying the minimum support score vary reliant on on the dataset. The best minimum support findings for every dataset are examined founded on the test results that were achieved. Fig. 6 is included to make the analysis procedure easier.

TABLE FOR TEST CASES:

Test Case ID	Description	Input Data	Expected Result	Actual Result	Status
TC01	Detect spam with common spam keywords	"Win a free prize now"	Classified as spam	Classified as spam	Pass
TC02	Detect ham with common non-spam keywords	"Meeting at 5 PM tomorrow"	Classified as ham	Classified as ham	Pass
TC03	Detect spam with misspelled spam keywords	"Congratz, you won a pr1ze"	Classified as spam	Classified as spam	Pass
TC04	Detect ham with business-related content	"Invoice for last month's services"	Classified as ham	Classified as ham	Pass
TC05	Detect spam with links	"Visit http://scam.com for a free gift"	Classified as spam	Classified as spam	Pass

With a minimum support of 3%, the precision current score of Fig. 6 in the SMS Corpus v.0.1 Big yields the maximum score. It is implemented as a result of numerous new features created especially for this type of spam; as a result, the system is more accurate

at responding to requests for information. However, the outcome is in direct opposition to the recall results, which yield the lowest score. Figure 6 illustrates the 3% minimum support for SMS data Corpus v.0.1 Big, demonstrating how the functionality that generates thousands of spam SMS is achieved. The word opportunity formula states that the amount of times a word seems in a class is inversely related to the. Conferring to the word opportunity formula, a word's frequency in a class is inversely correlated with its performance in that class; the more times a word appears in a class, the less likely it is that the word opportunity will occur there. Furthermore, the previous possibility score of the smaller amount of spam data is obviously lower because its composition is around one-third that of the ham data. It outcomes in a large volume of misclassified spam SMS messages. Meanwhile, the best minimal support is currently reached in SMS Corpus Big v.0 I by 6% to 98.308% accuracy results.

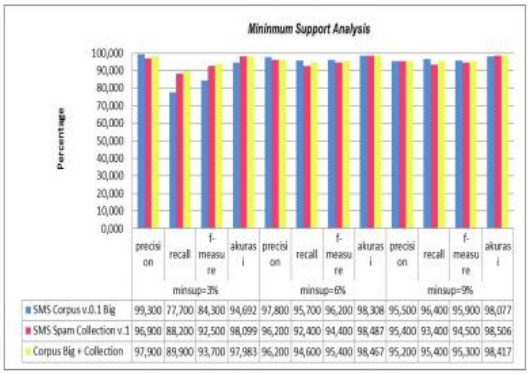

Fig. 6. The minimum support analysis

When linked to other datasets, the SMS Spam Collection dataset consistently yields a higher accuracy score. It demonstrates that using the three minimal support parameters in this dataset makes sense. The maximum accuracy in this Collection Spam dataset is achieved when 98.506% is equal to the 9% minimum support.

However, using both combined datasets yields almost the similar exactness when using the SMS Spam Collection dataset. This could be because any number of used datasets closely has the same quantity; therefore, there are a lot of similarities when applying it to the three minimum support parameters. The maximum accuracy in this dataset is achieved when 98.467% is identical to the 6% minimum support. After testing with three datasets, it was determined that both the amount and superiority of the datasets had an impact on achieving a high score. The volume of minimum support parameters score is inversely correlated with the amount of dataset used. Until then, it is non advised to utilize the extremely tiny minimum support score for small amounts of data, as the SMS Corpus Big v.0.1, as this will result in a significant portion of the worst new feature. However, in order to obtain the best new features for really large amounts of data, such a blend of the two datasets, a smaller parameter score is required.

Since the new features won't be given at all and the accurateness score won't increase if the minimal support is really high.

2) The Comparative analysis of both methods

Using the average evaluation score, the accuracy rate of the assessment findings in both ways (by FP Growth and without FP-Growth) will be compared.

It is demonstrated by the test results in Figure 7 that using the FP-Growth approach in conjunction with Naive Bayes consistently leads in an f-measure score and higher accuracy. It shows that the system is more suited to convey out the classification. Furthermore, FP-Growth can raise the precision score considerably. As a result, the system responds to user requests for information with greater accuracy. The recall score is appropriate despite being conversely lower since the document composition ratio is higher than that of ham. Nevertheless, it has an additional benefit in that the class will typically be categorized as hams if a text has undiscovered qualities that were previously discovered through training,

Thus, conferring to the testing, using each dataset to apply FP-Growth results in superior accuracy; using SMS Corpus v.0.1 Big boosts accuracy to 1.154%, using Spam Collection SMS increases accuracy to 0.025%, and using both datasets increases accuracy to 0.184%. With an accuracy of up to 98.506%, the SMS Spam Collection v.1 dataset yields the highest accuracy.

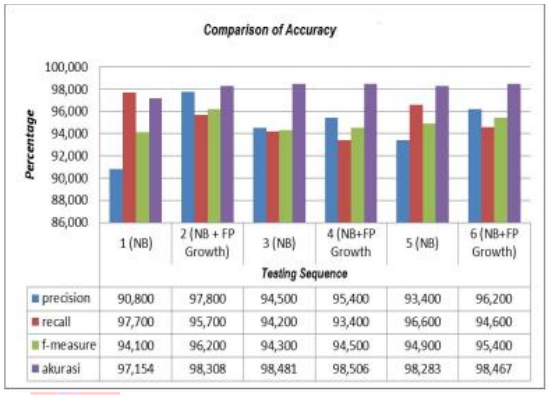

Fig. 7. accuracy comparison

IV. CONCLUSION

The research's tests were analyzed, and the following conclusions can be drawn from the results:

1. The two approaches utilized in the study worked equally well for classifying SMS messages, with an exactness rate on average above 90%. Comparing the average precision for each dataset with the usage of cooperation approaches like Naive Bayes and FP-Growth

is superior. It performs well on the spam Corpus v.0.1 Big SMS, Spam Collection SMS, and combined dataset, excelling by 1.154%, 0.025%, and 0.184%, respectively.

2. The best accuracy is achieved when the SMS Spam Collection v.l dataset with the 9% minimum support is utilized, and the FP-Growth implementation realizes an exactness of up to 98.506%.

3. Because SMS has a character limit, implementing minimum support makes it easier to cope with limited features. As a upshot, new features are created to distinguish between ham and spam SMS.

4. It is acceptable to apply the FP-Growth method to datasets with a change of training data.

5. The quantity of the dataset has an inverse relationship with the minimal support parameter score that is provided. In direction to develop more appropriate new features for the heavier data, a lower minus is utilized (a larger minus prevents the production of new features). Meanwhile, the larger minus is employed for the smaller dataset (if minus is too little, it cannot get the lower and effective new value).

6. It can progress the precision score by using the FP-Growth for feature extraction. As a result, in response to the SMS classification, the system is clever to provide users with information more precisely.

REFERENCES

[1] [l] Shirani-Mehr, Houshmand. "SMS spam identification utilizing AI approach." (2014): 1-6.

[2] Wang, Han Xue, Qian, and Wang Xiaoyu [2]. "Considering of arranging garbage messages dependent on the information mining." Service Science and Management, 2009.

[3]The YouTube Spam Collection, dt.fee.unicamp.br. [On the Internet]. /~tiago/SMS spam collection/ http://www.dt.fee.unicamp.br/ is reachable. [As of March 12, 2015].

[4][On the Internet]. The website http://www.ranks.nl/stopwords is accessible. [Retrieved: April 23, 2015].IEEE International Conference on MASS'09. "Stop words" on Ranks.nl

[5] Jian Pei, Han, Jiawei, and Micheline Kamber. Concepts and methods Version 3.1. Morgan Kaufmann Publishers-