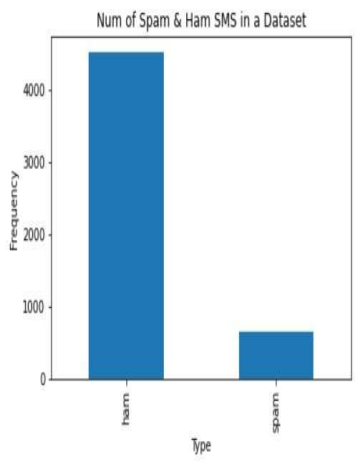
Classification on SMS Spam Collection Dataset

Chandrashekar Panjhazari (cp22@hood.edu)
CS 522 Data Mining, Hood College

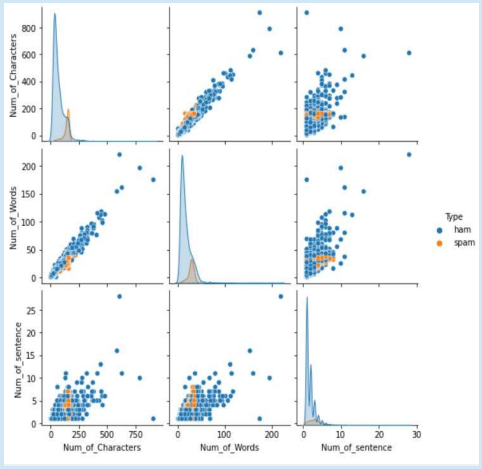
Abstract

Today, we can see that there has been a sharp spike in the number of cell phone users, which has resulted in a substantial rise in SMS spam messages. As compared to other things the cost per SMS usage rate is cheap. which has led numerous individuals and cellphone network operators to simply ignore the problem. We are aware that the current generation of cellphone spam-filtering systems on the market fall short of our expectations, which makes it more challenging to prevent real mobile phone spamming. On the other hand, we can observe that the absence of publicly available datasets for SMS spam, which are essential for testing and comparing different classifiers to demonstrate which one performs the best, represents a substantial disadvantage in academic contexts. Furthermore, since SMS messages are typically brief, content-based spam filters might not work well in this regard. Therefore, in this research, we will assess the performance of numerous popular categories of classification algorithms such as Multinomial Naïve Bayes, Random Forest, K Nearest Neighbor, Support Vector Machine etc.

Keywords: Spam Filtering, Classifiers, SMS Spamming, Datasets etc.



The above figure shows the number of ham & spam SMS in a dataset.



From the above figure you can see the outliers in the ham type.

Introduction

What is a spam? So, we all know that unwanted N number of SMS which are sent to our cell phones is called a spam or spamming. Basically, the risks associated with spam communications for consumers are numerous such as unwanted advertising, the disclosure of personal information, falling prey to fraudulent schemes, getting deceived into malicious and phishing websites, unintentional exposure to offensive content, etc. Spammers try to test the anti-spam infrastructure of the operator by sending different amounts of spam to see if volume obstacles are present. They frequently utilize numerous lines to send messages, thus number blocking as a spam prevention measure is not an option at present. This circumstance calls for some sort of content-based screening that considers both SMS contents and quantity. So, this research paper helps us to find out how to consider a message which is sent to the customers is a spam or a genuine message. For doing that we will use Natural Language Processing techniques and we take the help of Natural Language Tool Kit which is also known as NLTK. It includes libraries which helps in performing NLP (Natural Language Processing) tasks. After performing those tasks with the help of classifiers such as Multinomial Naïve Bayes, Random Forest, K Nearest Neighbor, Support Vector Machine etc. We check for the accuracy and then compare which classifier model produced better accuracy. So, that in future using that classifier model we can develop a specify tools or Spam Filters for reducing SMS Spamming.

Methods

After taking up this research project I started to work on with the dataset which I have downloaded from UCM Machine Learning website. As I have already mentioned spam dataset are rarely available.
It is a Classification problem. So, for this project I have used 4 classifiers namely Naive Bayes, Random Forest, KNearest Neighbor, and Support Vector Machine Classifier.
I have used matplot library for plotting the diagrams in this project.
Word clouds I have used for displaying most used words in the ham and spam dataset.
The Natural Language Toolkit, sometimes known as NLTK, is an open-source collection of programs, libraries, and learning materials for creating Natural Language Programming. By using NLTK I performed some subtasks such as cleaning the Text, Tokenization, Filtering Stop Words, Lemmatizing etc.
Used TFID Vectorizer and then divided the data set into test & train and validated them to get accuracy for all the classifiers and used f1 score because this is an imbalanced dataset and F1score is a good metric for predicting the accuracy for the models.

Results & Discussion

So, we can see that after performing all the NLP tasks with the help of NLTK. We prepared our data for performing classification using different classifiers. As we know that the data was imbalance when we check for their accuracies results were same for all the classifiers. So, to overcome this problem we have used the F1 score which gives us the score by combining the precision and recall scores. So, in the below figure you can clearly see that the accuracy for all the classifiers is same but when you look at their F1scores they are different. By considering that we can say Random Forest classifier has performed well followed by Support Vector Machine, Naïve Bayes and the least performance was given by K Nearest Neighbor classifier.

	Precision	Recall	F1score	Accuracy on Testset	Accuracy on Trainset
NaiveBayes	1.000000	0.577381	0.732075	0.958835	0.997669
RandomForest	0.991870	0.726190	0.838488	0.958835	0.997669
KNeighbours	1.000000	0.208333	0.344828	0.958835	0.997669
SVC	1.000000	0.684524	0.812721	0.958835	0.997669

Conclusion

In my view this research to perform classification on SMS Spam Dataset has worked well by performing the data mining techniques. With the help of NLTK we performed NLP tasks so that we can prepare the data for performing classification. We also used F1score to predict which classifier has performed well because it is a best metric to predict which classifier performance is great. I think this kind of project helps to develop tools based on machine learning algorithms so that they can work accurately and predict the spam messages.

Acknowledgements

I would like to express my deepest gratitude to Professor Dr. Xinlian Liu, for providing me an amazing opportunity to complete this project in his Data Mining class and I am also grateful to my Classmates for sharing their knowledge and experiences in the class throughout the semester.

References

[1] Almeida, T.A., GÃ³mez Hidalgo, J.M., Yamakami, A. Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11), Mountain View, CA, USA, 2011.
[2] GÃ³mez Hidalgo, J.M., Cajigas Bringas, G., Puertas Sanz, E., Carrero GarcÃa, F. Content Based SMS Spam Filtering. Proceedings of the 2006 ACM Symposium on Document Engineering (ACM DOCENG'06), Amsterdam, The Netherlands, 10-13, 2006.
[3] Cormack, G. V., GÃ³mez Hidalgo, J. M., and Puertas SÃ¡nz, E. Feature engineering for mobile (SMS) spam filtering. Proceedings of the 30th Annual international ACM Conference on Research and Development in information Retrieval (ACM SIGIR'07), New York, NY, 871-872, 2007.