On the use of Locality for Improving SVM-Based Spam Filtering

Okesola, J.O.
School of Computing
University of South Africa, South Africa
48948535@mylife.unisa.ac.za

Ojo, F.O.
Department of Computer Science
Federal College of Education
Osiele, Abeokuta, Nigeria.

Longe, O.B.
Department of Computer Science & Mathematics
Adeleke University
Ede, State of Osun, Nigeria
longeolumide@fulbrightmail.org.

ABSTRACT

Recent growths in the use of email for communication and the corresponding growths in the volume of email received have made automatic processing of emails desirable. In tandem is the prevailing problem of Advance Fee fraud E-mails that pervades inboxes globally. These genres of e-mails solicit for financial transactions and funds transfers from unsuspecting users. Most modern mail-reading software packages provide some forms of programmable automatic filtering, typically in the form of sets of rules that file or otherwise dispose mails based on keywords detected in the headers or message body. Unfortunately programming these filters is an arcane and sometimes inefficient process. An adaptive mail system which can learn its users' mail sorting preferences would therefore be more desirable. Premised on the work of Blanzieri & Bryl (2007), we proposes a framework dedicated to the phenomenon of locality in email data analysis of advance fee fraud e-mails which engages Support Vector Machines (SVM) classifier for building local decision rules into the classification process of the spam filter design for this genre of e-mails.

Keywords: Locality, SVM, Spam Filtering, E-mail & Security.

1. INTRODUCTION

There are various definitions of spam (junk mail) and how it differs from legitimate mails (non-spam, genuine mail or ham). The shortest among the popular definitions describes spam as "unsolicited bulk email" (Androutsopoulo et al, 2000, SPAMHAUS, 2005) and in some cases, the word "commercial" is added. The TREC Spam Track rides on this concept to define spam as "unsolicited or unwanted email that is sent indiscriminately, directly or indirectly, by a sender having no current relationship with the user" (Cormack & Lynam, 2005). Therefore one of the most widely accepted definitions of spam was presented by SpamDefined (2001) as "the one or more unsolicited messages, sent or posted as part of a larger collection of messages all having substantially identical content".

Spam is used to advertise different kinds of goods and services, and the percentage of advertisements dedicated to a particular kind of goods or services changes over time (Geoff et al, 2004). However, their changeability as addressed by Delany et al, (2004), is a big challenge, in particular the local nature relating to concept drift in spam. The problem of undesired electronic messages is nowadays a serious issue, as spam constitutes up to 75–80% of total amount of email messages (MAAWG, 2006). Spam causes several problems, resulting to direct financial losses. It causes a misuse of traffic, storage space, and computational power (Mikko & Carl, 2006). Spam waste the processors time leading to loss of work productivity and violation of privacy rights. Spam has been causing several legal problems through pornography advert, pyramid schemes, etc. (Drake et al, 2004). The total worldwide financial losses caused by spam estimated by Ferris Research Analyzer Information Service were over $50 billion (FerrisResearch, 2015).

Virus and Phishing are special cases of spamming activity that are dangerous and difficult to control. While the former is an unwanted program that attacks computing resources (Nicola, 2004), the latter particularly hunts for sensitive information (passwords, credit card numbers, etc.) by imitating requests from trusted authorities such as banks, server administrators or service providers (Christine, et al, 2004). This desirous nature has called for a growing scientific literature to address the characteristics of the spam phenomenon and offer feasible controls.

2. RELATED WORKS

Spamming is a cheap and illegal form of advertisement exploiting the facilities of the electronic mail infrastructure to easily reach thousands of users on the Internet. The implementation of reliable spam filters are imperative as e-mail users have to deal with the growing amount of these uninvited e-mails. Origin or address-based Antispam resident at the recipients' end of the mailing infrastructure typically use network information for Spam classification, while content filters examine the actual contents of email messages.

Several mechanisms are already in use to address spamming but they each have shortcomings that make them less effective. Support Vector Machines (SVMs) and Naïve Bayesian Systems (NBS) have been used to solve problems relating to text classification among others. Data variations or outliers however have very negative impact on the classification efficiencies of these two systems. Longe et al, (2008) developed SPAMAng, a Naïve Bayesian System for outbound e-mail filtering and SVM light a support vector machine open-source implementation was used in the experiment. Findings from the analysis of the performance of the two systems on a set of carefully selected advance fee fraud electronic mail corpus revealed that outliers can introduce some vulnerability into SVMs causing it to be defeated by spammers. This degradation in performance by SVMs is more noticeable in the domain of fraudulent spam mail filtering (419 mails) where the spammers engage in concept drifts using text manipulations, phishing and spoofing to fool spam filters. The comparison of SVMs with SPAMAng showed that SVMs does not always produce the best result in all text classification purposes

SVMs and Supervised Learning
In trying to help people decide which classification technique is best, studies have shown that, for text classification, Support Vector Machines (SVMs) produce the highest classification accuracy when compared to other techniques such as K-NN, Naïve Bayes. 'Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets (Cukier et al., 2006). To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighbouring datapoints of both classes. The larger the margin the better the generalization errors of the classifier from which the samples automatically perform the category assignments. This is a supervised learning problem. Since categories may overlap, each category is treated as a separate binary classification problem. This representation scheme leads to very high-dimensional feature spa (Andrew 2004).

Naïve Bayesian Classifications
A Naive Bayes Classifier (NBS) is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or lack of presence) of a particular feature of a class is unrelated to the presence (or lack of presence) of any other feature. Despite the fact that the far-reaching independence assumptions are often inaccurate, the naive Bayes classifier has several properties that make it surprisingly useful in practice. In particular, the decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution.

Each message is represented by a vector $x = (x_1 , x_2 , x_3 , \ldots., x_n)$, where $x_1 , \ldots, x_n$ are the values of attributes $X_1, \ldots., X_n$. Binary attributes: $X_i = 1$ is used if some characteristic represented by X_i is present in the message; otherwise $X_i = 0$. In spam filtering, attributes correspond to words, i.e. each attribute shows if a particular word (e.g. "adult") is present. To select among all possible attributes, one can compute the mutual information (MI) of each candidate attribute X with the category-denoting variable C . Other methods proposed for addressing the spam filtering task are discussed next.

2.1 E-Mail Transmission Protocol Methods

This is an act of enhancing or completely substituting the existing standards of email transmission by new spam-proof variants. The main drawback of the commonly used Simple Mail Transfer Protocol (SMTP) is that it provides no reliable mechanism of checking the identity of the message source. Sender Policy Framework (SPF) overcomes this challenge by inventing better and secured techniques of packaging the sender's identification. Other variants developed to address the problem include the Designated Mailers Protocol (Gordon, 2003), Trusted E-Mail Open Standard (Vincent et al, 2003), and SenderID mechanism (Sender, 2004). These proposals are fully discussed by Levine and DeKok (2004).

2.2 Learning-Based Methods of Spam Filtering

Filtering is a popular solution to the problem of spam. It can be defined as an automatic classification of messages into spam and legitimate mail. Existing filtering algorithms are quite effective, often showing accuracy of above 90% during the experimental evaluation (Chih-Chin & Ming-Chi, 2004)). It is possible to apply the spam filtering algorithms on different phases of email transmission such as at the routing stage, at the destination mail server, or in the destination mailbox (Agrawal et al, 2005). Although, a filter is known to prevent end-users from wasting their time on junk messages, it does not prevent the misuse of resources because all the messages are delivered nevertheless. Therefore it is always been argued that filtering at the destination only gives a partial and not a total solution to the spam problems.

Figure 1 depicts the various components of an e-mail that can be analysed by a spam filter. In order to classify new messages, a spam filter can analyse these components separately (by checking the presence of certain words in case of keyword filtering) or in groups (by considering that the arrival of a dozen of substantially identical messages in five minutes is more suspicious than the arrival of one message with the same content).

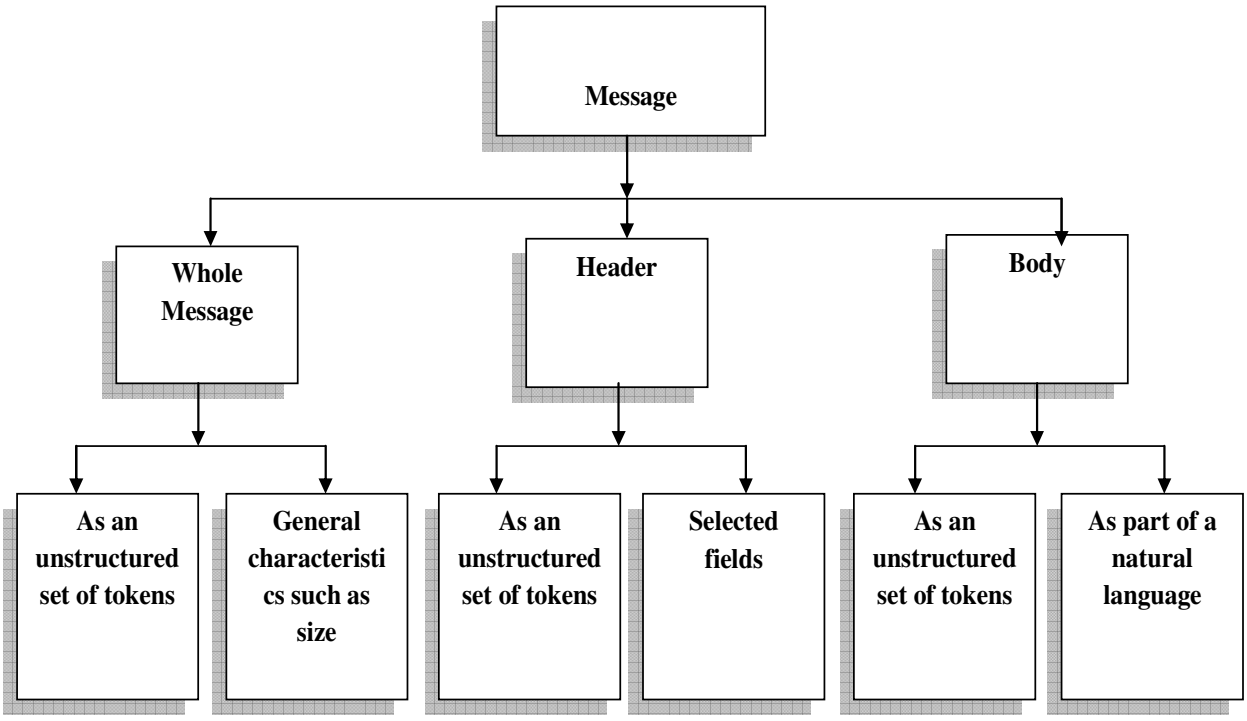

Fig. 1: Components of an e-mail

An e-mail message typically consists of two parts - the body and the header. Message body is usually a text in a natural language, possibly with HTML mark-up and graphical elements while the header is a structured set of fields, each having a name, value, and specific meaning. Some of these fields (such as From, To, or Subject) are standard while others depend on the software involved in message transmission such as spam filters installed on mail servers. Subject field contains what the user sees as the subject of the message and is often treated as a part of the message body. The body is sometimes referred to as the contents of the message. However, non-content features are not limited to the features of the header. For instance, a filter may consider the message size as a feature or a training data (pre-collected messages with reliable judgments), and may be optimised involving users' collaboration to receive multiple user inputs about new messages for analysis.

3. SPAM FILTERING TECHNIQUES

Some notable machine learning techniques currently in use for spam filtering are highlighted in this section.

3.1 Naive Bayes:

The Naïve Bayes classifier is a probability based classification system that when applied to text, can be considered as an improved learning-based variant of keyword filtering. It rests on the naive independence assumption, namely that all the features are statistically independent.

3.2 k-Nearest Neighbour:

The k-Nearest Neighbour (k-NN) classifier was proposed for spam filtering by Androutsopoulos et al (2000). With this classifier the decision to flag a message as spam is simple. K nearest training samples are selected using a predefined similarity function, and then the message x is labelled as belonging to the same class as the majority among this k samples. Christine et al. (2004) showed that this method, due to its local nature, is good in coping with changeability.

3.3 Collaborative Spam Filtering

Researchers have made efforts to achieve better spam filtering through the collaboration of users. This is done by engaging users to share knowledge about spam between peer to peer systems (Lorenzo, 2005; Zhou et al, 2003), or gathering spam reports from the users on a mail server like we have in Gmail and yahoo. Privacy issues become the main challenge in such situations of data exchange between users but this has been addressed by Damiani et al. (2004) who proposed a privacy-preserving approach to P2P spam filtering system.

3.4 Opposing Reactivity

Spammers keep improving on their techniques to outpace filtering methods and make the methods ineffective by disabling them from identifying and categorising threats. Following the systematization proposed by Wittel and Wu (2004) therefore, it becomes easier to categorize attacks on spam filters as follows:

- **Tokenization attacks** - When the spammer intends to prevent correct tokenization of the message by splitting or modifying features such as putting extra spaces in the middle of the words.
- **Obfuscation attacks** - When the content of the message is obscured from the filter (by means of encoding).
- **Statistical attacks -** When the spammer intends to skew the message's statistics. If the data used for a statistical attack is purely random, the attack is called weak. Otherwise it is called strong. An example of strong statistical attack is good word attack as postulated by Daniel and Christopher (2005).

4. CONSIDERING LOCALITY IN SPAM CLASSIFICATION

In machine learning, Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms that analyse data and recognise patterns, commonly used for classification and regression analysis. Given a set of training samples each marked for belonging to one of the two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

An SVM model is a representation of the samples as points in space, mapped in such a way that the samples of the separate categories are divided by a clear very wide gap. New samples are then mapped into that same space and are predicted to belonging to a category depending on the side of the gap they fall into. Additionally, SVMs efficiently perform a non-linear classification using what the "kernel trick" by implicitly mapping their inputs into high-dimensional feature spaces. (Wikipedia, 2015). Spam is not uniform but rather consists of messages on different topics (Hulten et al 2004) and in different genres (Cukier, et al 2006). However, accuracy can be improved upon when classifications are based on the local decision rules. An SVM classifier typically provides a global decision rule independent of the sample which must ordinarily be classified.

The primary research goal of this study is to develop a SVM-based classification algorithm with attendant capability that considers locality in the spam filtering process. This is because correct and accurate classification of spam mails is usually limited by the fact that spam consist of messages on various topics and genres. Local decision rules must therefore be applied in collaborative filtering as opposed to the present application of global rules that sees and classifies spam based on pre-coded information on genre and types. The changeability of data is also likely to have local nature (Delany et al, 2004), and this is applicable to legitimate mails as well. The existence of algorithms which classifies email by topic (Li et al., 2007) provides evidence of both locality in legitimate mail and the possibility to capture it using bag-of-words feature extraction. The simplest spam filtering method which makes use of locality in the data is k-Nearest Neighbour (k-NN), which was shown to be outperformed by SVM on spam versus non-spam classification task (Lai and Tsai, 2004). This suggests that a more elaborate way of building local decision rules is highly required.

4.1 The research framework

The framework of this study shall infuse the existing filtering techniques into the new model in order to develop a learning-based classifier called Higher Probability SVM Nearest Neighbour (HP-SVM-NN). This algorithm is based on the SVM and Nearest Neighbour (SVM-NN) classifier, which is a combination of SVM and k-Nearest Neighbour (k-NN). While SVM-NN requires the locality size as an input parameter, HP-SVM-NN selects the locality size dynamically for each sample to be classified.

HP-SVM-NN may be experimentally evaluated and shown to outperform SVM on the task of spam filtering. A practical spam filtering architecture shall be proposed based on this classifier, and the accuracy and speed of the architecture shall be evaluated. The phenomenon of locality shall be addressed in email data and discussions on the ways in which it influences the problem of spam recognition will be outlined. The study will then show, by experiments with the classification of HP-SVM-NN, that locality is an important issue for the spam filtering task.

5. THE RESEARCH IMPACTS AND EXPECTED CONTRIBUTIONS

This proposal, when implemented, will inculcate the resulting filtering technique into practical anti-spam software solutions. At a more general level, the comparison of local and global classification algorithms will always help to evaluate the importance of the locality phenomenon for learning-based spam filtering and thereby, providing useful information for the design of the future classifiers.

The following interesting research directions are expected to arise from the results of this work when fully implemented. They will be the major contributions to knowledge that will attract researchers' attention for future works:

- A new classification algorithm will be developed
- A practical filtering architecture based on this algorithm will be designed, and
- An exploration of the phenomenon of locality in email classification will be established.

6. CONCLUSION

The combination of these three contributions highlighted in section 5 will yield a new and accurate classification algorithm, called the HP-SVM-NN classifier. Although this classifier may be required to show reasonably high accuracy even when using only message headers and, to a certain extent, be able to cope with image spam without any additional modifications, it would nevertheless be interesting to design a version of the filter able to process images in addition to text, as this would supposedly increase the level of classification accuracy.

REFERENCES

1. Blanzieri, E. & Bryl, A. (2007): Highest Probability SVM Nearest Neighbor Classifier for Spam Filtering. Technical Report March 2007 Technical Report # DIT-07-007. Accessed on January 26, 2014 from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.8142&rep=rep1&type=pdf
2. Androutsopoulos I, Koutsias, J., Konstantinos V. and Constantine, D. (2005): An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '00, pages 160–167, New York, NY,USA, ACM Press. ISBN 1-58113-226-3.
3. Christine, D., Oliver, J. and Koontz, E. (2004): Anatomy of a phishing email. In Proceedings of the First Conference on Email and Anti-Spam, CEAS'2004.
4. Cormack, G., and Lynam, T. (2005): Spam corpus creation for TREC. In Proceedings of Second Conference on Email and Anti-Spam, CEAS'2005. http://ceas.cc/2005/
5. Cukier W., Cody, S. and Nesselroth, E. (2006): Genres of spam: Expectations and deceptions. In Proceedings of the 39th Annual Hawaii International Conference on System Sciences, HICSS '06, Volume 3. Accessed on February 5, 2014 from: http://www.computer.org/csdl/proceedings/hicss/2006/2507/03/250730051a.pdf
6. Damiani, E., Sabrina, D., Stefano, P. and Pierangela, S. (2004): P2P-based collaborative spam detection and filtering. In Proceedings of Fourth IEEE International Conference on Peer-to-Peer Computing, P2P'04, pages 176–183.
7. Daniel, L. and Christopher, M. (2005): Good word attacks on statistical spam filters. In Proceedings of Second Conference on Email and Anti-Spam, CEAS'2005, 2005.
8. Delany, S., Padraig, C., Alexey, T. and Lorcan, C. (2004): A case-based technique for tracking consept drift in spam filtering. Knowledge-based systems, pages 187–195.
9. FerrisResearch (2015): The global economic impact of spam, report #409. http://www.ferris.com/get content file.php?id=364
10. Hulten, G., Penta, A., Gopalakrishnan, S. and Manav, M. (2004): Trends in spam products and methods. In Proceedings of the First Conference on Email and Anti-Spam, CEAS'2004, 2004.
11. Lai, C. and Tsai, M. (2004): An empirical performance comparison of machine learning methods for spam e-mail categorization. Hybrid Intelligent Systems, pages 44–48, 2004.
12. Lorenzo, L., Mari, M. and Poggi, A. (2005): Cafe – collaborative agents for filtering e-mails. In Proceedings of 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise, WETICE'05, pages 356–361.
13. MAAWG (2006): Messaging anti-abuse working group. Email metrics report. Third & fourth quarter 2006. Accessed on June 17, 2014 from http://www.maawg.org/about/MAAWGMetric 2006_3 4 report.pdf
14. Mikko, S. and Carl, S. (2006): Effective anti-spam strategies in companies: An international study. In Proceedings of HICSS '06,volume 6.
15. SpamDefined (2001): Spam defined. Accessed on July 1, 2014 from: http://www.monkeys.com/spamdefined
16. SPAMHAUS (2005): The definition of spam. Accessed on March 14, 2014 from: http://www.spamhaus.org/definition.html.
17. Wittel, G. and Wu, F. (2004): On attacking statistical spam filters. In Proceedings of First Conference on Email and Anti-Spam,CEAS'2004.
18. Zhou, F., Zhuang, L., Zhao, B. Huang, L., Joseph, A. and Kubiatowicz, J. (2003): Approximate object location and spam filtering on peer-to-peer systems. In Proceedings of ACM/IFIP/USENIX International Middleware Conference, Middleware.
19. Wikipedia (2015): Support Vector Machine. https://en.wikipedia.org/wiki/Support_vector_machine
20. Longe, O.B., Robert, A.B.C, Chiemeke, S.C and Ojo. F.O. (2008). Feature Outliers and Their Effects on the Efficiencies Of Text Classifiers In The Domain Of Electronic Mail. The Journal of Computer Science and Its Applications. Vol. 15, No. 2.
21. Andrew F. (2004): Investigation of Support Vector Machines for Email Classification. Dissertation submitted to the School of Computer Science and Software Engineering Monash University.