E-Mail Spam Filtering: A Review of Techniques and Trends

Alexy Bhowmick and Shyamanta M. Hazarika

Abstract We present an inclusive review of recent and successful content-based e-mail spam filtering techniques. Our focus is mainly on machine learning-based spam filters and variants inspired from them. We report on relevant ideas, techniques, taxonomy, major efforts, and the state-of-the-art in the field. The initial interpretation of the prior work examines the basics of e-mail spam filtering and feature engineering. We conclude by studying techniques, evaluation benchmarks, and explore the promising offshoots of latest developments and suggest lines of future investigations.

Keywords Spam · Spam filtering · Techniques · False positive
Machine learning

1 Introduction

E-mail or electronic-mail is a fast, effective, and inexpensive method of exchanging messages over the Internet. Whether it is a personal message from a family member, a company-wide message from the boss, researchers across continents sharing recent findings, or astronauts staying in touch with their family (via e-mail uplinks or IP phones), e-mail is a preferred means for communication. Used worldwide by 2.3 billion users, at the time of writing the article, e-mail usage is projected to increase up to 4.3 billion accounts by 2016 [1]. But the increasing dependence on e-mail has induced the emergence of many problems caused by 'illegitimate' e-mails, i.e., *spam*. According to the Text Retrieval Conference (TREC) the term

A. Bhowmick (✉)
School of Technology, Assam Don Bosco University, Guwahati 781017, Assam, India
e-mail: alexy.bhowmick@dbuniversity.ac.in

S.M. Hazarika
Deptartment of Computer Science and Engineering, Tezpur University, Tezpur 784028, Assam, India
e-mail: smh@tezu.ernet.in

A. Kalam et al. (eds.), *Advances in Electronics, Communication and Computing*, Lecture Notes in Electrical Engineering 443,
https://doi.org/10.1007/978-981-10-4765-7_61

'*spam*' is—any unsolicited e-mail that is sent indiscriminately [2]. Spam e-mails are unsolicited, un-ratified, and usually mass mailed. Spam being a carrier of malware causes the proliferation of unsolicited advertisements, fraud schemes, phishing messages, explicit content, promotions of cause, etc. On an organizational front, spam effects include: (i) annoyance to individual users, (ii) less reliable e-mails, (iii) loss of work productivity, (iv) misuse of network bandwidth, (v) wastage of file server storage space and computational power, (vi) spread of viruses, worms, and Trojan horses, and (vii) financial losses through phishing, denial of service (DoS), directory harvesting attacks, etc.

Figure 1 depicts the e-mail architecture and how e-mail works. Spam is a broad concept that is still not completely understood. In general, spam has many forms—chat rooms are subject to *chat spam*, blogs are subject to *blog spam* (splogs), search engines are often misled by *web spam* (search engine spamming or spamdexing), while social systems are plagued by *social spam*. This paper focuses on '*e-mail spam*' and its variants, and not 'spam' in general. Prior attempts to review e-mail spam filtering using machine learning have been made, the most notable ones being [2–7]; most recent empirical studies being [8–10]. We extend earlier surveys by taking an updated set of works into account. We present a content analysis of the major spam-filtering surveys over the period (2004–2015). Significant amounts of historical and recent literature, including gray literature were studied to report recent advances and findings. We believe our survey is of complementary nature and provides an inclusive review of the state-of-the-art methods in content-based e-mail spam filtering. Our work addresses the following:

- *First*, we perform an exploration of the major spam characteristics and discuss feature engineering for spam e-mails.
- *Second*, we present a qualitative summary of major surveys on spam e-mails over the period (2004–2015) and taxonomy of content-based approaches to e-mail spam filtering.

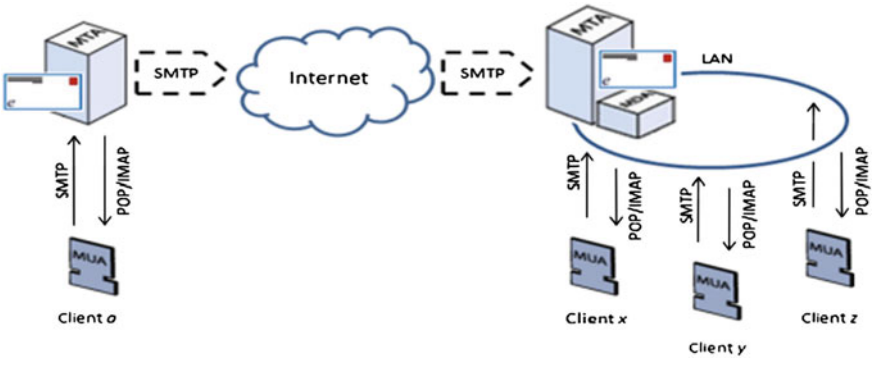

Fig. 1 The e-mail architecture

- *Third*, the article reports on evaluation measures, bench marks and new findings and suggest lines of future investigations for emerging spam types.

2 Feature Engineering

Feature selection is a key issue and has become the subject of much research. It has mainly three objectives: (i) enhancing the classifier's predictive accuracy, (ii) building effective and economical classifiers, and (iii) obtaining a better understanding of the elementary process involved in generation of data. Dimensionality reduction and feature subset selection are two preferred techniques for lowering the feature set dimension. While feature subset selection involves the extraction of a subset of the original attributes, dimensionality reduction involves linear combinations of the original feature set.

Table 1 presents a summary of feature extraction and selection in popular literature. This review article also examined a number of major earlier surveys on spam filtering over the period (2004–2015). A summary of popular machine

Table 1 A summary of feature extraction and feature selection techniques in popular literature

References	Year	Approaches
[12]	2004	Studied subject line, header, and message body. Employed information gain (IG), document frequency (DF), and chi-square test for selecting features. Found *bag of words* model quite effective on spam filtering, and header features as important as message body
[13]	2006	Extracted fixed-length character *n-grams* and *v*ariable-length character *n-grams*. Explored information gain (IG) as a feature selection technique. Character *n-grams* were noted to be richer and definitive than word-tokens
[14]	2006	Considered features of three types: *word, character, structured feature* in a feature-based *versus* feature-free comparison. Employed information gain (IG) as a feature selection technique. Noted feature-free methods to be more correct than the feature-based systems, however feature-free approaches took much longer than feature-based approach in classifying e-mails
[15]	2005	Used behavioral patterns of spammers, Meta-heuristics as features Employed term frequency, inverse document Frequency (TFIDF), SpamKANN for feature selection. Tested SVM, Decision trees, Naïve Bayes to get increased prediction accuracy than keywords
[16]	2003	Experimented on features: header (H), textual (T), handcrafted features (HH), etc. Different ways of feature selection for Decision Tree and Naive Bayes models were evaluated. The usefulness and importance of different type of features were discussed in detail in experiments
[17]	2006	Considered subject, body, header, attachment feature. Analyzed strength and weaknesses of document frequency (DF), Information Gain (IG), Chi-square test, and Mutual Information. Presented a deep analysis of feature selection methods. Found e-mail attachments to be useful when integrated with models

learning-based techniques categorizing them according to perspective (Algorithm, Architecture, Methods, and Trends) is presented in Table 2.

Articles classified under '*Algorithm*' reflect research that focused on classification algorithms and their implementations and evaluations. Articles classified under '*Architecture*' concentrated on development of spam filtering infrastructures. Articles classified under '*Methods*' refers to study of the existing filtering methods while '*Trends*' speaks of discourses concentrating on emerging methods and the adaptation of spam filtering methods over time. Limitations listed in the last

Table 2 A summary of popular machine learning-based spam filtering attempts by authors according to perspective with their strengths and limitations

References (Year)	Perspective	Strength and limitations
[18] (2004)	Naive Bayes, k-NN, ANN, SVM	Techniques benefits beginners *Does not* deal with feature selection
	Algorithms, methods	
[3] (2006)	Naive Bayes, Logitboost, SVM	Resulted in—*LingSpam* and *PU1* *Ignored* headers, HTML, attachments
	Algorithm, methods, trends	
[4] (2006)	Bayesian filtering	Broad review of implementations *Focuses primarily* on automated, filters
	Methods, architecture	
[5] (2008)	SVM, TF-IDF, boosting	Explains feature extraction methods *Does not cover* neighboring topics
	Algorithms, methods, trends	
[2] (2008)	SVM, perceptron, OSBF	Testing achieves FPR = 0.2% User feedback *difficult to simulate*
	Algorithms, methods, trends	
[6] (2009)	Regression, ensembles	Focuses on textual and image analysis. Focuses *only* on application specific aspects
	Algorithms, methods	
[2] (2010)	SVM, Naive Bayes	Proposed Matthews correlation coefficient (MCC). Need for comparison
	Algorithms, methods	
[3] (2012)	MDL principle, SVM	Uses six, well known, large public databases. Bogofilter, SpamAssassin *not considered*
	Algorithms, methods	
[15] (2012)	Signature, k-NN, ANN, SVM	Focuses on distributed computing paradigms. *Avoids* interoperability issues
	Methods, architecture	
[7] (2013)	Statistical analysis, n-grams	Investigated *topic drift* Limited datasets
	Trends	
[8] (2015)	Naïve Bayes, J48, Clustering	Comparative study of different techniques Limited datasets and tools
	Methods	

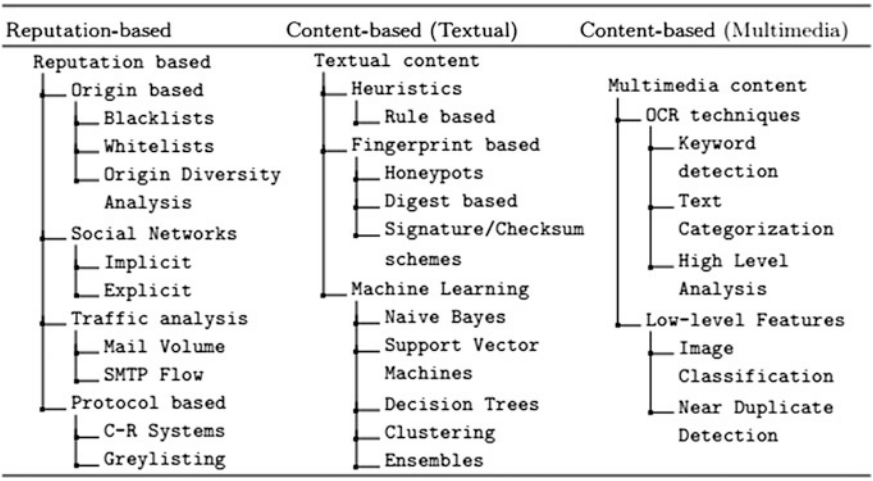

Fig. 2 A taxonomy of e-mail spam filtering techniques

column, corresponding to each article are as acknowledged by the authors themselves. Perusing the different spam techniques and the methods used by researchers to combat spam, taxonomy of spam filtering techniques is presented (Fig. 2) next.

3 Publicly Available Datasets

Most of the datasets publicly available are static datasets with very few concept drift datasets. Many authors construct their own image spam or phishing corpus. Table 3 lists public corpora with associated information used in spam filtering experiments.

4 Future Trends and Conclusion

Models built on old data become less accurate or inconsistent making the rebuilding of the model imperative (called *virtual concept drift*). Spam filtering is a dynamic problem that involves concept drift. While the understanding of an unwanted message may remain the same, the statistical properties of the spam e-mail changes over time since it is driven by spammers involved in a never-ending arms race with spam filters. Another reason for concept drift could be the different products or scams driven by spam that tends to become popular. The dynamic nature of spam is one of its most testing aspects. An effective spam filter must be able to track target concept drift, swiftly adapt to it, and have a successful mechanism to identify the drift or evolution in spam features.

Table 3 Public corpora used in e-mail spam filtering experiments

Corpus name	No. of messages (Spam|Ham)		Spam rate (%)	Year of creation	References
SpamAssassin	1897	4150	31	2002	[17]
Enron-Spam	13,496	16,545	–	2006	[19]
Ling Spam	481	2412	17	2000	[20]
PU1	481	618	44	2000	[7]
PU2	142	579	20	2003	[12]
PU3	1826	2313	44	2003	[12]
PUA	571	571	50	2003	[12]
Gen Spam	41,404		78	2005	[17]
Spambase	1813	2788	39	1999	[20]
ZH1	1205	428	74	2004	[12]
TREC 2005	52,790	39,399	–	2005	[4]
TREC 2006	24,912	12,910	–	2006	[5]
TREC 2007	50,199	25,220	–	2007	[18]
Spam archive	>2,20,000		100	1998	[3]
Biggio	8549	0	–	2005	[12]
Princeton spam benchmark	1071	0	–	–	[12]
Spam archive	>2,20,000		100	1998	[3]
Dredze dataset	3927	2006	–	2007	[3]
Phishing corpus	415	0	–	2005	[3]

Content-based spam filtering systems, though widely adopted as a successful spam defense strategy, has unfortunately substituted the spam issue with a false positive one. Such systems achieve a high accuracy but there exists some false positive tradeoff. False positives are *more severe and expensive* than spam. Reduction of false positives is another domain in email spam analysis where much work needs to be been done on leveraging existing algorithms. Future researches must address the fact that e-mail spam filtering problem is co-evolutionary, since spammers attempt to outdo the advances in predictive accuracy of the classifiers all the time.

One of the biggest spam problems today even as spam e-mail volumes associated with botnets are receding is the *snowshoe spam*. Showshoe spamming is a technique that uses multiple IP addresses, websites, and sub-networks to send spam, so as to avoid detection by spam filters. Spammers operate by distributing their spam load across a wide footprint of systems to keep from sinking, just as snowshoe wearers do. With many users today migrating to social networks as a means of communication, spammers are diversifying in order to stay in business.

E-mail prioritization is an urgent research area with not much research done. In addition to basic communication, e-mail systems are used for a wide variety of other tasks such as—business and personal communication, advertisements, reminders, management of tasks, and cloud storage, etc. There is a serious need to

address the information overload issue by developing systems that can learn personal priorities from data and identify important e-mails for each user. Prioritizing e-mail as per its importance or classifying emails into personalized folders as in [9, 11] is another desirable characteristic in a spam filter. Prioritizing e-mail or perhaps redirecting urgent messages to handheld devices could be another way of managing e-mails.

Fortunately, machine learning-based systems enable systems to learn and adapt to new threats, reacting to counteractive measures adopted by spammers. No single anti-spam solution may be the right answer. A multi-faceted approach that combines legal and technical solutions and more is likely to provide a death blow to such spam. As long as spam exists it will continue to have adverse effects on the preservation of integrity of e-mails and the user's perception on the effectiveness of spam filters. Overall remarkable advancements have been achieved and continue to be achieved, however, some outstanding problems in e-mail spam filtering as highlighted above still remain. Till more improvements in spam filtering happen, anti-spam research will remain an active research area.

References

1. Radicati: Email Statistics Report, 2012–2016 Executive Summary. Technical Report 650, Radicati (2016)
2. Cormack, G.V.: Email spam filtering: a systematic review. Found. Trends Inf. Retrieval **1**(4), 335–455 (2008)
3. Androutsopoulos, I., Paliouras, G., Michelakis, E.: Learning to filter unsolicited commercial e-mail. Technical Report in National Centre for Scientific Research Demokritos, Athens, Greece (2006)
4. Carpinter, J., Hunt, R.: Tightening the net: a review of current and next generation spam filtering tools. Comput Secur. **25**(8), 566–578 (2006)
5. Blanzieri, E., Bryl, A.: A survey of learning-based techniques of email spam filtering. J. Artif. Intell. Rev. **29**(1), 63–92 (2008)
6. Guzella, T.S., Caminhas, W.M.: A review of machine learning approaches to spam filtering. Expert Syst. Appl. **36**(7), 10206–10222 (2009)
7. Wang, D., Irani, D., Pu, C.: A study on evolution of email spam over fifteen years. In: Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Work sharing (CollaborateCom), Austin, TX, USA (2013)
8. Vyas, T., Prajapati, P., Gadhwal, S.: A survey and evaluation of supervised machine learning techniques for spam e-mail filtering. In: Proceedings of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), IEEE, 5–7 March 2015
9. Alsmadi, I., Alhami, I.: Clustering and classification of email contents. J. King Saud Univ. Comput. Inf. Sci. **27**, 46–57 (2015)
10. Li, W., Meng, W.: An empirical study on email classification using supervised machine learning in real environments. In: IEEE International Conference on Communications (ICC), IEEE, 8–12 June 2015
11. Sethi, H., Sirohi, H., Thakur, M.K.: Intelligent mail box. In: Proceedings of Third International Conference India 2016, vol. 3, pp. 441–450 (2016)

12. Zhang, L., Zhu, J., Yao, T.: An evaluation of statistical spam filtering techniques spam filtering as text categorization. ACM Trans. Asian Lang. Inf. Process. (TALIP) **3**(4), 243–269 (2004)
13. Kanaris, I., Kanaris, K., Houvardas, I., Stamatatos, E.: Words versus character n-grams for anti-spam filtering. Int. J. Artif. Intell. Tools **16**(6), 1–20 (2006)
14. Delany, S.J., Bridge, D.: Feature based and feature free textual CBR: a comparison in spam filtering. In: Proceedings of the 17th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'06), pp. 244–253 (2006)
15. Yeh, C.Y., Wu, C.H., Doong, S.H.: Effective spam classification based on meta-heuristics. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp. 3872–3877 (2005)
16. Diao, Y., Lu, H., Wu, D.: A comparative study of classification based personal e-mail filtering. In: Knowledge Discovery and Data Mining. Current Issues and New Applications, pp. 408–419 (2003)
17. M´endez, J.R., D´ıaz, F., Iglesias, E.L., Corchado, J.M.: A comparative performance study of feature selection methods for the anti-spam filtering domain. In: Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, pp. 106–120. Springer, Berlin, Heidelberg (2006)
18. Tretyakov, K.: Machine learning techniques in spam filtering. In: Data Mining Problem-Oriented Seminar, MTAT.03.177, pp. 60–79 (2004)
19. Koprinska, I., Poon, J., Clark, J., Chan, J.: Learning to classify e-mail. Inf. Sci. **177**(10), 2167–2187 (2007)
20. Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V.: Stacking classifiers for anti-spam filtering of e-mail. In: Empirical methods in Natural Language Processing, pp. 44–50 (2001)