HYBRID APPROACH FOR SPAM EMAIL DETECTION

SYED MOHD ANWAR ALHABSHI BIN SYED HAMED

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Computer Science (Information Security)

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2018

DEDICATION

This thesis dedicated to:

The sake of Allah, my Creator and Designer,
My great teacher and messenger, Mohammed (May Allah bless and grant him), who taught us the purpose of life,
My great parents: Syed Hamed, and Sharifah Hindon, who never stop supporting me in countless ways,
My dearest wife: Nurazlyna, who leads me through the valley of darkness with the light of hope and support,
My beloved brothers and sisters,
My beloved kid: Syed Darweesh, whom I cannot force myself to stop loving,
To all my family, the symbol of love and giving,
My heartfelt thanks to the respected supervisor, Dr Maheyzah for the support, guidance and for her enduring patience,
My friends who encourage and support me,
All the people in my life who touch my heart,
Thank you…!

ACKNOWLEDGEMENT

First and foremost, I feel grateful and praise to Allah for granting me knowledge and determination for me successfully achievement for this work and project.

I would like to express heartfelt gratitude to my supervisor Dr Maheyzah Binti Md Siraj for her constant support during my study at UTM. She inspired me greatly to work and guidance and for their patience.

Besides, I would like to thank the authority of Universiti Teknologi Malaysia (UTM) for providing me with a good environment and facilities such as Computer laboratory to complete this project with software who contributed to my success.

I must express my very profound gratitude and appreciation to my parents and to my wife for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

ABSTRACT

On this era, email is a convenient way to enable the user to communicate everywhere in the world which it has the internet. It is because of the economic and fast method of communication. The email message can send to the single user or distribute to the group. Majority of the users does not know the life exclusive of e-mail. For this issue, it becomes an email as the medium of communication of a malicious person. This project aimed at Spam Email. This project concentrated on a hybrid approach namely Neural Network (NN) and Particle Swarm Optimization (PSO) designed to detect the spam emails. The comparisons between the hybrid approach for NN_PSO with GA algorithm and NN classifiers to show the best performance for spam detection. The Spambase used contains 1813 as spams (39.40%) and 2788 as non-spam (60.6%) implemented on these algorithms. The comparisons performance criteria based on accuracy, false positive, false negative, precision, recall and f-measure. The feature selection used by applying GA algorithm to reducing the redundant and irrelevant features. The performance of F-Measure shows that the hybrid NN_PSO, GA_NN and NN are 94.10%, 92.60% and 91.39% respectively. The results recommended using the hybrid of NN_PSO with GA algorithm for the best performance for spam email detection.

ABSTRAK

Pada era ini, e-mel adalah cara yang mudah untuk membolehkan pengguna berkomunikasi di mana-mana di dunia yang mempunyai internet. Ia adalah kaedah komunikasi yang ekonomik dan cepat. Mesej e-mel boleh dihantar kepada pengguna tunggal atau mengedarkan kepada kumpulan. Majoriti pengguna tidak mengetahui kehidupan eksklusif e-mel. Projek ini fokus kepada untuk Spam Email. Projek ini tertumpu pada pendekatan hibrid iaitu Rangkaian Neural (NN) dan Pengoptimuman Swarm Partikel (PSO) yang direka untuk mengesan e-mel spam. Perbandingan antara pendekatan hibrid untuk NN_PSO dengan algoritma GA dan pengelas NN untuk menunjukkan prestasi terbaik untuk pengesanan spam. Spambase yang digunakan mengandungi 1813 sebagai spam (39.40%) dan 2788 sebagai bukan spam (60.6%) yang dilaksanakan pada algoritma ini. Kriteria prestasi perbandingan berdasarkan *accuracy*, *false positive*, *false negative*, *precision, recall* dan *f-measure*. Pemilihan ciri dengan menggunakan algoritma GA untuk mengurangkan ciri-ciri yang berlebihan dan tidak relevan. Prestasi f-*measure* menunjukkan bahawa hibrid NN_PSO, GA_NN dan NN masing-masing 94.10%, 92.60% dan 91.39%. Hasilnya disyorkan menggunakan hybrid NN_PSO dengan algoritma GA untuk prestasi terbaik untuk pengesanan emel spam.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	–	Accuracy
ANN	–	Artificial Neural Network
Bcc	–	Blind carbon copy
BNC	–	British National Corpus
BoW	–	Bag-of-Words
BP	–	Back-Propagation
CCERT	–	Council of Computer Education Research & Training
DK	–	Domain Keys
DKIM	–	Domain Keys Identified Mail
DMP	–	Designated Mailers Protocol
DoS	–	Denial Of Service
EMG	–	Electromyographic
EMP	–	Excessive Multi-Posting
F1	–	F-Measure
FN	–	False Negative
FP	–	False Positive
GA	–	Genetic Algorithm
GA_NN	–	Genetic Algorithm with Neural Network
HTML	–	Hypertext Markup Language
IIM	–	Identified Internet Mail
IM	–	Instant Messaging
IT	–	Information Technology
kNN	–	K-Nearest Neighbor
MLP	–	Perceptions Multilayer
MTA	–	Mail Transfer Agent

NB	–	Naive Bayes
NN	–	Neural Network
NN_PSO	–	Hybrid Neural Network with Particle Swarm Optimization
OSB	–	Orthogonal Sparse Bigrams
P	–	Precision
PID	–	Proportional-Integral-Derivative
PSO	–	Particle Swarm Optimization
R	–	Recall
RMSE	–	Root Mean Square Error
SBPH	–	Sparse Binary Polynomial Hash
SMS	–	Short Message Service
SMTP	–	Simple Mail Transfer Protocol
SPF	–	Sender Policy Framework
SVM	–	Support Vector Machines
TREC	–	Text Retrieval Conference
UBM	–	Unsolicited Bulk Mail
UCE	–	Unsolicited Commercial Email

CHAPTER 1

INTRODUCTION

1.1 Overview

In this era, email is a convenient way to enable the user to communicate globally in the world which it has the internet. It is an economical and fast method of communication. The email message can send to the single user or distribute to the group. Majority of the users does not know the life exclusive of e-mail. For this issue, it becomes an email as the medium of communication of a malicious person. The rapid growth of the internet, at the same time the spam rate is also increased. In the second week of 2014, it shows that 70% of the report statistics for traffic of emails were spam (Nizamani S. *et al.*, 2014).

The definition of the spam also referred to as unsolicited bulk mail (UBM), unsolicited commercial email (UCE), excessive multi-posting (EMP), spam mail or junk mail (Bhuleskar *et al.*, 2009). The cause of the spam can affect the legitimate email reach to the email user based on the reason it overloads to the user inbox and has the malicious code in emails (Al-Mukhtar, 2012). The researchers have made the spam assessment by conduct the survey to get the status of spam in Kingdom of Saudi Arabia (KSA). It shows that the main distribute the spam email including commercials, phishing, sexual contents, religious reasons, etc. It causes no good purposes for the bandwidth, excess use and gets resources (Abdullah Al-Kadhi and Mishaal, 2011).

Between 2005 and 2007, the worldwide cost of spam expected by Ferris Research is US$ 50 billion and US$ 100 billion respectively (Bauer *et al.*, 2008). Because of these causes to the outright violation of personal space and some requirement to prevent the spam while using the general delivery. As part of the CAN-SPAM, the U.S House of Representatives endorsed the bill on December 16, 2003, the financial punishments of $ 6 million and five-year prison to prevent the unwanted messages. (Lee, 2005a; Sivanadyan, 2003).

Recognize how spam has generated to break down the evolution of spam filters. However, that not allowed taking a gander for each type of spam filters. By performing this study, it may be possible for spam filters one-stage moving forward to find front the spammers and put the spam to ends.

1.2 Background of Problem

With the broad use of the internet and service of emails, it becomes significant in our life. At the same time, the spam also increased, cause waste time-consuming to deleting the spam. Spam also can waste the network resources, gets the virus and it not suitable to shows the under-aged recipients to inappropriate content.

For the classification email in specific algorithms, it has the different problem with wrongly describing the legitimate email as spam namely called as False Positive and wrongly classify the non-spam as legitimate email called False Negative. When users get the spam email as legitimate, the user becomes annoyed. For spam detection, the effect of the low accuracy, the false positive or the false negative is part of the issue to the datasets.

Develop the technique to categories of the spam is complex, with define the spam types and modify the classification task near impossible. The spammers also attempt to modify emails in order not to catch using the technique, adding hard to deliver accurate detection. Now, several of the effective spam filtering studies to distinguishing the spam from legitimate emails (Wei, C. P. *et al.*, 2008).

For classification analysis, several of spam filtering used email contents to classify the spam or ham, for example, Bayesian analysis (Sahami *et al.*, 1998), machine learning approaches (Guzella and Caminhas, 2009) and heuristics approaches (Cook *et al.*, 2006).

Machine learning is overall techniques been used by the researchers for detection of spam and gets successful good results. However, in the machine learning, the situation of the pre-processing should have the high scale of characteristics space on email because of it can obstruction to the classifiers. A large number of words in the message should extract because the excess of the characteristics can degrade the classification.

Many a new algorithm proposed for training feed-forward neural network classifier can get the fast convergence in the network. However, improper set a large number of hidden layers weight in the neural network can get the problem large-scale optimization. Currently, for the optimization methods can consider identifying to using the genetic algorithms.

1.3 Problem Statement

A significant issue to the neural network classifiers has the problem of error convergence. A large of the data have features contains irrelevant and redundant used in the classification of NN and hybrid of NN and PSO. The classifier parameter, initial weight, population sizes of the NN_PSO classifiers can decrease the error rate. All these influences can trap the classifier to gets the optimal solution. The use of the global optimization algorithm to gets the solution by the implement in pre-processing phase using features selection algorithm and the combination of the classification as trainer algorithm in a classifier.

1.4 Aim of the Project

This project aims to perform on the dataset for obtaining the accurate classification in line with f-measure of the balance between precision and recall and reduce the rate of the false positive, the false negative in a selected dataset for spam email detection. It will achieve by implementing the proposed hybrid of Neural Network (NN) with Particle Swarm Optimization (PSO), and use of Genetic Algorithm (GA) with the selected parameter for features selection.

1.5 Objective of the Project

This project will aim to perform on the dataset for obtaining the accurate classification in line with f-measure of the balance between precision and recall, and reduce the rate of the false positive, the false negative with using the proposed spam email detection based on the classification of the dataset. The objective of the project as the following:

i. To select significance features to represent dataset by using the Genetic Algorithm (GA).

ii. To develop a hybrid of Neural Network (NN) and Particle Swarm Optimization (PSO) for spam email classification.

iii. To evaluate benchmark the effectiveness of the proposed hybrid approach based on accuracy, the false positive, the false negative and the f-measure.

1.6 Scope of the Project

This project performs on the dataset for obtaining the accurate classification in line with f-measure of the balance between precision and recall and reduces the rate of the false positive, the false negative using proposed spam email detection based on the classification of the dataset. The scope of the project as the following:

i. The implementation of Neural Network (NN) and the hybrid approach of Neural Network (NN) and Particle Swarm Optimization (PSO) algorithm on the content of the email and use of Genetic Algorithm (GA) as a features selection.
ii. Use of Spambase datasets got from the UCI website.
iii. The performance measurement of the dataset evaluated based on classification accuracy, the false positive, the false negative and the f-measure.

1.7 Significant of the Project

The rapid growth of the internet, at the same time the spam email is also increased. It needs to prevent it using the spam email detection. The different method of the spam email detection with different impact spam email, which is to detect and remove the spam email from user inbox. This own technique can classify the email for spam or non-spam. By employing the technique in this study by using the hybrid of the NN and PSO, it can help to detect and block the spam in the user mailbox.

1.8 Organization of the Project

This chapter organised into four topics. For the first topic, contains the overview of the spam emails, problem background, and problem statement, the target of the project, the project objectives, project scope and significance of the study. The second topic describes literature review of the spam including definition, types of spam and available spam filtering technique. The third topic describes the methodology of the project used to achieve the project objectives. The fourth topic describes the feature selection of the data. The fifth chapter focuses on implementation and results from spam email detection.

REFERENCES

Abdullah Al-Kadhi, Mishaal, 2011. Assessment of the status of spam in the Kingdom of Saudi Arabia. J. King Saud Univ. Comput. Inf. Sci. 23, 45–58.

Abido, M. A. (2002). Optimal design of power-system stabilizers using particle swarm optimization. IEEE transactions on energy conversion, 17(3), 406-413.

Aghdam, M. H., & Heidari, S. (2015). Feature selection using particle swarm optimization in text categorization. Journal of Artificial Intelligence and Soft Computing Research, 5(4), 231-238.

Al-Mukhtar, M. M., & Tabra, Y. M. (2012). An effective spam filter based on a combined support vector machine approach. International Journal of Internet Technology and Secured Transactions, 4(1), 42-54.

Alba, E., Garcia-Nieto, J., Jourdan, L., & Talbi, E. G. (2007, September). Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In Evolutionary Computation, 2007. CEC 2007. IEEE Congress on (pp. 284-290). IEEE.

Alsmadi, I., & Alhami, I. (2015). Clustering and classification of email contents. Journal of King Saud University-Computer and Information Sciences, 27(1), 46-57.

Altman, N. S. (1992) “An introduction to kernel and nearest-neighbor nonparametric regression”, The American Statistician, 46 (3), 175–185.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000, July). 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 (pp. 160-167). ACM.

Aski, A. S., & Sourati, N. K. (2016). Proposed efficient algorithm to filter spam using machine learning techniques. Pacific Science Review A: Natural Science and Engineering, 18(2), 145-149.

Babatunde, O. H., Armstrong, L., Leng, J., & Diepeveen, D. (2014). Zernike moments and genetic algorithm: Tutorial and application.

Bashir, Z. A., & El-Hawary, M. E. (2009). Applying wavelets to short-term load forecasting using PSO-based neural networks. IEEE transactions on power systems, 24(1), 20-27.

Bauer, J. M., Van Eeten, M. J. G., & Chattopadhyay, T. (2008). Financial aspects of network security: Malware and spam. ITU (International Telecommunication Union). Available online at http://www. itu. int/ITU-D/cyb.

Bhuleskar, R., Sherlekar, A. and Pandit, A. (2009) 'Hybrid spam e-mail filtering', First International Conference on Computational Intelligence, Communication Systems and Networks, CICSYN '09, pp.302–307.

Bhowmick, A., & Hazarika, S. M. (2016). Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends. arXiv preprint arXiv:1606.01042.

CAPTCHA (2005) The CAPTCHA project.

http://www.captcha.net/ Accessed:31.05.06

Carpinter, J., & Hunt, R. (2006). Tightening the net: A review of current and next generation spam filtering tools. Computers & security, 25(8), 566-578.

Christina, V., Karpagavalli, S., & Suganya, G. (2010). A study on email spam filtering techniques. International Journal of Computer Applications, 12(1), 0975-8887.

Clerc M, Kennedy J (2002) The particle swarm—explosion, stability, and convergence in a multidimensional complex space. IEEE Transactions on Evolutionary Computation 6: 58–73.

Cook, D., Hartnett, J., Manderson, K., & Scanlan, J. (2006, January). Catching spam before it arrives: domain specific dynamic blacklists. In Proceedings of the 2006 Australasian workshops on Grid computing and e-research-Volume 54 (pp. 193-202). Australian Computer Society, Inc.

Csmining Group. 2010 Spam email datasets. Available at: http://csmining.org/index.php/spam-email-datasets-.html

Duan Z, Dong Y, Gopalan K (2005) Diffmail: a differentiated message delivery architecture to control spam. In: Proceedings of 11th international conference on parallel and distributed systems, ICPADS 2005. Vol 2, pp 255–259.

Foqaha, M., & Awad, M. (2017). Hybrid Approach to Optimize the Centers of Radial Basis Function Neural Network Using Particle Swarm Optimization. JCP, 12(5), 396-407.

Lai, G. H., Chen, C. M., Laih, C. S., & Chen, T. (2009). A collaborative anti-spam system. Expert Systems with Applications, 36(3), 6645-6653.

Lee, Y. (2005). The CAN-SPAM Act: a silver bullet solution?. Communications of the ACM, 48(6), 131-132.

Goodman J, Cormack GV, Heckerman D (2007) Spam and the ongoing battle for the inbox. Commun of the ACM 50(2): 25–33.

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.

Hamed, H. N. A. (2006). Particle swarm optimization for neural network learning enhancement (Doctoral dissertation, Universiti Teknologi Malaysia).

Hulten, Geoff & Penta, Anthony & Seshadrinathan, Gopalakrishnan & Mishra, Manav. (2004). Trends in Spam Products and Methods.

I. Koprinska, J. Poon, J. Clark, and J. Chan, “Learning to Classify Email,” Information Sciences Including Special Issue on Hybrid Intelligent Systems, Vol. 177(10), Elsevier, 2007, pp. 2167–2187.

Jafari-Marandi, R., & Smith, B. K. (2017). Fluid Genetic Algorithm (FGA). Journal of Computational Design and Engineering, 4(2), 158-167.

Jain, K., & Agrawal, S. (2014, September). A hybrid approach for spam filtering using local concentration based K-Means clustering. In Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference- (pp. 194-199). IEEE.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In Machine Learning Proceedings 1994 (pp. 121-129).

Khater, I. M., Al-Jarrah, O. M., & Al-Duwairi, B. Hierarchical Email Spam Filtering.

Kuipers B, Liu A, Gautam A, Gouda M (2005) Zmail: zero-sum free market control of spam. In: Proceedings of the 25th IEEE international conference on distributed computing systems workshops, ICDCS 2005. IEEE Computer Society, pp 20–26.

Lai, C. C., & Tsai, M. C. (2004, December). An empirical performance comparison of machine learning methods for spam e-mail categorization. In Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on (pp. 44-48). IEEE.

Li K, Pu C, Ahamad M (2004) Resisting spam delivery by tcp damping. In: Proceedings of the first conference on email and anti-spam, CEAS’2004.

Lichman, M. (2013). UCI Machine Learning Repository. Available At: [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Liu, Z., Lin, W., Li, N., & Lee, D. (2005, November). Detecting and filtering instant messaging spam-a global and personalized approach. In *Secure Network Protocols, 2005. (NPSec). 1st IEEE ICNP Workshop on* (pp. 19-24). IEEE.

Lugaresi N (2004) European union vs. spam: a legal response. In: Proceedings of the first conference on email and anti-spam, CEAS’2004.

Mashor, M. Y. (1999) “Some Properties of RBF Network with Applications to System Identification”, International Journal of Computer and Engineering Management, 7.

Mohamad, M., & Selamat, A. (2015, April). An evaluation on the efficiency of hybrid feature selection in spam email classification. In Computer, Communications, and Control Technology (I4CT), 2015 International Conference on (pp. 227-231). IEEE.

Mujtaba, G., Shuib, L., Raj, R. G., Majeed, N., & Al-Garadi, M. A. (2017). Email Classification Research Trends: Review and Open Issues. IEEE Access.

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In Proceedings of the 21st international conference on World Wide Web. ACM.

Nadir Omer Fadl Elssied, Otman Ibrahim, Waheeb Abu-Ulbeh, "An Improved of Spam E-mail Classification Mechanism Using K-Means Clustering", Journal of Theoretical and Applied Information Technology, Vol. 60 No.3, February 2014.

Nazirova, S. (2011). Survey on spam filtering techniques. Communications and Network, 3(03), 153.

Nizamani, S., Memon, N., Glasdam, M., & Nguyen, D. D. (2014). Detection of fraudulent emails by employing advanced feature abundance. Egyptian Informatics Journal, 15(3), 169-174.

Ou, Y. Y., Gromiha, M. M., Chen, S. A., &Suwa, M. (2008) "TMBETADISC-RBF: Discrimination ofβ-barrel membrane proteins using RBF networks and PSSM profiles", Computational Biology andChemistry, 32(3), 227–231.

P. G. Juneja, R. K. Pateriya," Survey on Email Spam Types and Spam Filtering Techniques", in International Journal of Engineering Research & Technology, March-2014, Vol. 3, Issue 3.

Radicati (2016) Email Statistics Report, 2012-2016 - Executive Summary. Tech. Rep. 650, Radicati

R. Bhuleskar, A. Sherlekar and A. Pandit, "Hybrid Spam E-Mail Filtering," 2009 First International Conference on Computational Intelligence, Communication Systems and Networks, Indore, 23-25 July 2009, pp. 302-307.

Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017, August). Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets. In IOP

Conference Series: Materials Science and Engineering (Vol. 226, No. 1, p. 012091). IOP Publishing.

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).

Saito T (2005) Anti-spam system: another way of preventing spam. In: Proceedings of the 16th international workshop on database and expert systems applications, DEXA 2005 pp 57–61.

Samha, A., Li, Y., & Zhang, J. (2014). Aspect-based opinion extraction from customer reviews. International Journal of Computer Science & Information Technology, 6 (3).

Sangwan, O. P., Bhatia, P. K., & Singh, Y. (2011) "Radial basis function neural network based approach to test oracle", ACM SIGSOFT Software Engineering Notes, 36(5), 1- 5.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.

Sharma A., Manisha, Dr.Manisha, Dr.Rekha J. (2014), A survey on spam detection techniques. International Journal of Advanced Research in Computer and Communication Engineering, 3(12).

Sivanadyan, T. (2003). Spam? not any more! detecting spam emails using neural networks. Technical report, University of Wisconsin.

Solihin, M. I., Tack, L. F., & Kean, M. L. (2011). Tuning of PID controller using particle swarm optimization (PSO). International Journal on Advanced Science, Engineering and Information Technology, 1(4), 458-461.

Sturgeon, W. (2003b) "Major victory in war on spam".

Posted Sep 09. Accessed Oct 30, 2003. <http://www.silicon.com/research/specialreports/thespamreport/0,39025001,10005930,00.htm>

Subasi, A. (2013). Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders. Computers in biology and medicine, 43(5), 576-586.

Symantec. March 2011 Intelligence Report. Available at: http://www.symantec.com/about/news/release/article.jspprid=2011032901

Thomas Claburn, Spim, Like Spam, Is On The Rise, Information Week, March 30, 2004.

Tuteja, S. K. (2016). A survey on classification algorithms for email spam filtering. International Journal of Engineering Science, 6(5), 5937-5940.

Twining RD, Williamson MM, Mowbray M, Rahmouni M (2004) Email prioritization: reducing delays on legitimate mail caused by junk mail. Technical Report HPL-2004-5R1, HP Labs

W. Sturgeon, “Mobile spam: Is the next plague upon us?”, http://www.silicon.com/research/specialreports/thespamreport/0,39025001,10004599,00.htm, 2003.

Wei, C. P., Chen, H. C., & Cheng, T. H. (2008). Effective spam filtering: A single class learning and ensemble approach. Decision Support Systems, 45(3), 491-503.

Whissell JS, Clarke CLA (2011) Clustering for Semi-Supervised Spam Filtering. In: Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS ’11), pp 125–134.

Xiang, Y., Chowdhury, M., & Ali, S. (2004, January). Filtering mobile spam by support vector machine. In CSITeA'04: Third International Conference on Computer Sciences, Software Engineering, Information Technology, E-Business and Applications (pp. 1-4). International Society for Computers and Their Applications (ISCA).

Xu, R., Xia, Y., Wong, K.-F., & Li, W. (2008). Opinion annotation in on-line Chinese product reviews. LREC.