Content-Based Concept Drift Detection for Email Spam Filtering

Morteza Zi Hayat
School of Electrical and Computer Engineering
University of Tehran,
Tehran, Iran
zihayatm@ut.ac.ir

Javad Basiri
School of Electrical and Computer Engineering
University of Tehran,
Tehran, Iran
basiri@ut.ac.ir

Leila Seyedhossein
School of Electrical and Computer Engineering
University of Tehran,
Tehran, Iran
l.seyedhossein@ut.ac.ir

Azadeh Shakery
School of Electrical and Computer Engineering
University of Tehran,
Tehran, Iran
shakery@ece.ut.ac.ir

Abstract—**The continued growth of Email usage, which is naturally followed by an increase in unsolicited emails so called spams, motivates research in spam filtering area. In the context of spam filtering systems, addressing the evolving nature of spams, which leads to obsolete the related models, has been always a challenge. In this paper an adaptive spam filtering system based on language model is proposed which can detect concept drift based on computing the deviation in email contents distribution. The proposed method can be used along with any existing classifier; particularly in this paper we use Naïve Bayes method as classifier. The proposed method has been evaluated with Enron data set. The results indicate the efficiency of the method in detecting concept drift and its superiority over Naïve Bayes classifier in terms of accuracy.**

Keywords-component; spam filtering; concept drift; KL divergence; language model.

I. INTRODUCTION

The continued growth of internet users has caused Email to be one of the most convenient and economic methods of communication. So concentrating on its challenges has been noticed in research communities. Spam Emails emails have several harmful effects such as degrading user trust to advertising emails, decreasing communication quality and also companies' creditability issues. In addition, in most cases, these emails have jobbery and fraudulent goals.

There are many proposed methods in research communities for detecting and discriminating spams from the legitimate emails. Most of them use machine learning techniques such as support vector machines, Naïve Bayes classifiers, and case base reasoning rules [1]. The evolving nature of spams and also legitimate emails leads to change in underlying distribution of emails content. These changes make the models built on old examples, inadequate for classifying new emails and the models become obsolete gradually. So detecting these changes and regular updating of the model is an important issue in the context of spam filtering. The learned model of a classifier should adapt itself to classify new emails correctly. This problem is known as concept drift. Concept drift detection methods help to find critical situations in which the model should be updated [2].

Existing methods for spam filtering have been applied successfully in offline or static environments. However, a few ones can deal with concept drift phenomena. In this paper we have proposed a new adaptive spam filtering method which concentrates on concept drift detection based on language models of emails. The main idea behind our approach is that when email contents are changed, the corresponding language model distribution is changed too. So we can detect concept drift by analyzing the language model distribution.

The rest of the paper is organized as follows: section II introduces related works in spam filtering –researches which can adapt themselves to the changes. The proposed approach is introduced in section III. Section IV presents the evaluation results. Finally the paper concludes in section V.

II. RELATED WORKS

Classification methods have been widely applied in the context of spam filtering. Also dynamic environment of this context have been considered in many researches.

For a recent survey of spam filtering, we can see the proposed method in [3]. In this research two variants of Naive Bayes classifier have been applied for spam filtering: multinomial and multi-bernoulli. Also an active learning method has been proposed to select the e-mails whose classifier's uncertainty is high and adds them to the training data. Another example is in [4] where SVM and Hamming Distance are used to classify e-mails. This research proposes a sliding window-based approach including W sections. There is one classifier in each section and they use Exceeding margin update technique to update each classifier dynamically. Each email is labeled by majority voting strategy between all W classifiers. In [5] an adaptive spam filtering method has been proposed based on information theory. For this purpose an objective function is defined to classify training and testing data. This function is aimed at minimizing the loss in mutual information between e-mail instances and words set including terms of training and testing set, before and after prediction. In [6] lazy learning approach has been applied to filter spam emails. Because of their lazy learning based method, the updating process is easy. In this work two similar strategies are considered to detect concept drift. The performance of the system is monitored by the user feedback and, if the rate of false alarms exceeds a certain threshold, the model should be

978-1-4244-8185-9/10/$26.00 ©2010 IEEE

retrained. Another example is in [7] where a new spam filtering method based on case base reasoning has been proposed which can track concept drift. An instance selection approach is used to handle concept drift which involves two stages. In the first stage, noisy cases are removed using case based editing technique. Thereafter, in the second stage, redundant cases are removed. In [8] a new approach for feature selection in spam hunting is proposed which is able to handle concept drift. It defines Achieved Information (AI) measure to compute the amount of information which can be achieved by a term. When a term in an email is not presented by any instances in the knowledge repository, it is considered as a fully predictive feature and its AI measure just depend on its probability in the email instance and its size. Therefore, it updates the current features by considering the new features in each email. In another research [9] a spam-filtering system which uses a cluster-based classification method has been proposed. This approach overcomes the problem of skewed class distribution and using an incremental learning method handles the concept drift. It groups the documents in each class into several clusters and selects features from each cluster. To handle the concept drift, when the system classifies a new email and assigns it to the nearest cluster, the coherence of the target cluster is considered. As another example for using instance-based reasoning spam filtering, we can see [10]. In this research, two approaches have been proposed to improve a previously instance-based reasoning spam filtering model to track concept drift. In the first approach, the representative terms of the information contained in each email are selected and weighted according to the window size and in the second one; those emails which are more applicable are predicted in order to propose an accurate classification.

To summarize, to the best of our knowledge in most of the recent works in this area, either the label of data or user feedback have been used to detect concept drift. In this paper, we have proposed a new approach which can detect concept drift without considering these two types of information.

III. THE PROPOSED APPROACH

Spam filtering is a binary classification method in which each email is labeled as spam or non-spam. Several machine learning algorithms have been used in spam filtering, among them Naïve Bayes is a very popular one [11]. There are several forms of Naïve Bayes [12]. The proposed spam filtering system in this paper is based on multinomial Naïve Bayes classification method which takes into account term frequencies of words in an email [13].

As mentioned before, the content of emails change over time, so the current model may become obsolete and unable to detect new spams. So the main problem for a spam filtering system is identifying a change in data distribution and then adapting the model. The main difference of the proposed approach is in the way of detecting and handling concept drift in a spam filtering system. Fig.1 shows an overview of our proposed approach. We first construct the classification model according to the training data. After each block of emails arrives, its content distribution is compared with the existing content distribution model. If two distributions are too different, concept drift has occurred. In this state, the model is updated based on combination of the existing model and the new block's model. The detailed explanation of the proposed method is as follows.

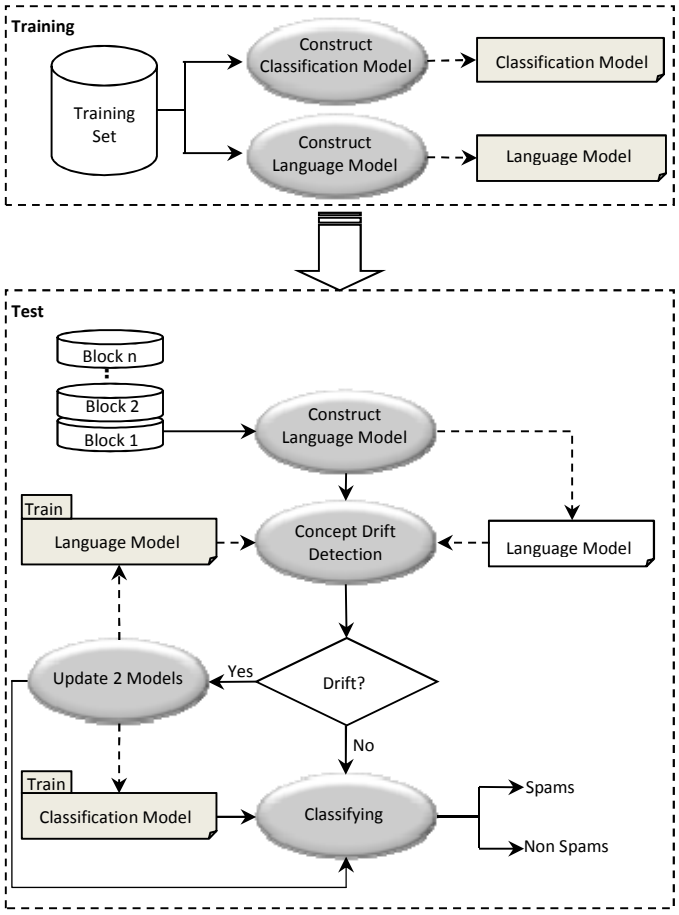

Figure 1. Overview of the proposed approach

A. Tracking the concept drift

As aforementioned above, concept drift can be tracked by detecting the deviations in the data distribution. The proposed approach uses language model technique to represent the distribution of the block content. A language model assigns a probability to each word in a document. Among different approaches for language model computation, we use the following one [13]:

$$P(W \in Block) = \frac{TF(W)}{\sum_{word \in Block} TF(word)} \quad (1)$$

Where, TF(W) is the frequency of term W in the current block.

In expression (1) each word in each block will have a weight. This weight indicates the word importance in the block. It can be seen that the probabilities are based on the block content. As a result any change in the data distribution can be regarded as change in content. According to this fact when the distribution between two consecutive blocks is very different, this change is considered as concept drift. So,

concept drift detection boils down to computing the difference between two distributions. To this end we use Kullback–Leibler divergence (KL-Divergence) measure which is a non-symmetric and none-negative measure. Let P and Q be two probability distributions; the KL divergence of Q from P is defined as follows [14]:

$$D_{\mathrm{KL}}(\mathrm{P}\,||\mathrm{Q}) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \quad (2)$$

So the formal expression of the proposed approach is as follows: When a new block of emails arrives, first its language model is computed. Then the difference between the block's language model and the existing language model is calculated using KL divergence measure. If the difference between them is more than a determined threshold, it is considered as concept drift. Therefore we can define the KL divergence as follows:

$$D_{\mathrm{KL}}(LM_{block}\,||LM_{old}) = \sum_i P(i)\log\frac{LM_{block}(i)}{LM_{old}(i)} \quad (3)$$

Where, LM_{block} and LM_{old} are the language models of the current block and the old one respectively.

When a concept drift is detected, the system should adapt itself to the changes in a way that the new emails can be classified correctly. In this situation, the document model of Naïve Bayes classifier and the language model used for concept drift detection are obsolete and should be updated. The update procedure of the proposed approach consists of three steps which are explained below:

1) Updating the term probabilities

Term probabilities of Naïve Bayes classifier should be changed according to the current block. To this end we classify M messages of the current block using the previous model. Thereafter, we compute the terms probabilities of the M messages and combine them with the previous ones to gain the new probabilities using the following formula:

$$P_{new}(W|C_i) = \alpha \times P_{block}(W|C_i) + (1-\alpha)\times P_{old}(W|C_i) \quad (4)$$

$$i \in \{Spam\,, None - Spam\}$$

Where, C_i is the class label and α is a parameter controlling the influence of each part. Higher values of α means the lead to higher influence of the current block's terms probabilities and lower influence of the previous ones on the new terms probabilities. After updating, the remaining messages in the current block will be classified using the updated model.

2) Updating the language model

The language model which is used in concept drift detection should be updated according to the new block. For this purpose we combine the old language model with the language model of the current block using the following formula:

$$P(W|LM_{new}) = \beta \times P(W|LM_{block}) + (1-\beta)\times P(W|LM_{old}) \quad (5)$$

Where, LM_{block} and LM_{old} are the language models of the current block and the old one respectively and β is a parameter controlling the influence of each language model. Higher values of β lead to higher effect of current block's language model and lower effect of the old one in the new language model.

3) Updating the Prior Probabilities of Classes

In the Naïve Bayes classifier, each document D is classified using the following formula:

$$P(C|D) = \arg max_c\, P(D|C)P(C) \quad (6)$$

Where, C is the class label and $P(\mathrm{C})$ is the prior probability of each class (spam/non-spam). As the number of messages in the legitimate and spam class for the current block may be different from the previous ones, we should update $P(\mathrm{C})$ in this equation accordingly. For this purpose we use the labels of M messages of the current block which are classified using the previous model to compute P(C) for the current block. Thereafter, we combine the previous and current $P(\mathrm{C})$ to compute the new $P(\mathrm{C})$ using the following formulas:

$$P_{block}(C_i) = \frac{count_{block}C_i)}{M} \quad (7)$$

$$P_{new}(C_i) = \frac{P_{old}(C_i)+P_{block}(C_i)}{2} \quad (8)$$

$$C_i \in \{Spam\,, legtimate\}$$

Where, $P_{block}(C)$ and $P_{old}(C)$ are the current and previous values of $P(\mathrm{C})$ in equation (6) respectively and $count_{block}(C_i)$ is the number of messages in the first M messages of the current block that belong to class Ci

IV. Experimental Results

We have evaluated our proposed method in terms of right detection of concept drift and improvement of spam detection. For this purpose the performance of our proposed method is compared with a similar model without considering concept drift. The details of dataset and the evaluation results will be explained in the following subsections.

A. Data set

For this research we need a labeled dataset which is suitable for concept drift experiments. For this purpose we use the data which is used in [12]. This data consists of 6 different datasets which is obtained from Enron corpus. Since this corpus only consists of legitimate messages, spam messages which are obtained from four different sources are imported to the data. The order and ratio of spam messages are simulated based on real scenarios. More information about this dataset can be found in [12].

In this research we use 3000 messages of this collection. Since we want to experiment concept drift, we need the content of data to change over time. So we have to simulate this scenario manually. We choose the first 1000 messages of the dataset from Enron1, the second 1000 messages from Enron2, the third 1000 messages from Enron3 to generate concept drift. This approach has been used in some other researches too as the one presented in [5]. Table I shows the details of the data set.

Combining three data sets, we have different contents in the test set. To illustrate the different content of the blocks, table

II shows the five most important words in each block based on Chi square feature selection method [13].

TABLE I. DATA SET DETAILS

Data Set Name	# Documents	#Legitimates	# Spams
Farmer	1000	715	285
Kaminski	1000	737	263
Kitchen	1000	731	269

TABLE II. BLOCKS CONTENT DETAILS BY FIVE MOST IMPORTANT WORDS

	Enron 1	Enron 2	Enron 3
Top Five Top Features for Legitimate Emails	2000	Vince	2001
	Thanks	Click	Louise
	on	Enron	Enron
	cc	2000	Original
	Enron	cc	sent
Top Five Top Features for Spam Emails	Thanks	2000	Stop
	2000	Money	2005
	http	Wish	Wish
	cc	List	Professional
	Paliourg	Investor	Email

We have considered only 500 messages as training data and the rest for testing. Because the training data have little knowledge about future emails contents, we can demonstrate the significance of updating the model. Moreover, Chi square method has been used for feature selection. 25% of the top features have been chosen in our experiment.

We assume that testing data arrive in equal sized blocks. To show the efficiency of the proposed method, we have evaluated the method using two different block sizes: 300 and 500.

B. Results

In the first step of the experiment, we show the effectiveness of KL divergence measure in concept drift detection. Figs.2 and 3 show the KL values between the blocks of the testing data and the training data without any update. These figures illustrate how much the content of each block differs from the content of the training data.

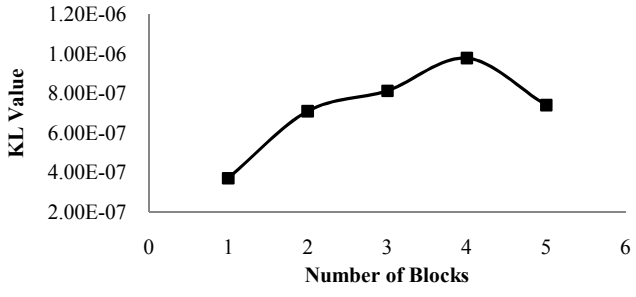

Figure 2. KL divergence values between training data and the blocks of testing data with block length 500

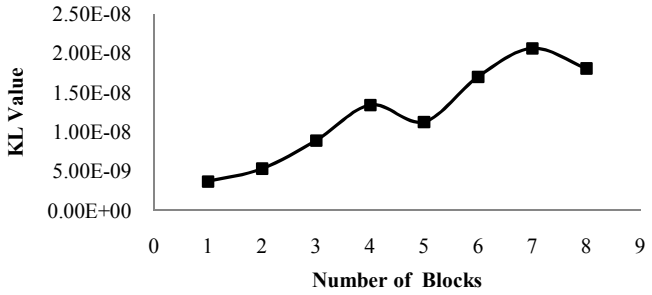

Figure 3. KL divergence values between training data and the blocks of testing data with block length 300

As we expected, KL value for the first block is less than the others; because the content of the first block is more similar to the training data. We use these values to detect concept drift. Larger values indicate more dissimilarity between the contents. For example, the content of fourth block is more different from the training data than the others. Moreover, experimental result with block length 300, demonstrate that the proposed method can work on different block sizes, because in this experiment the length of training data is 500 while the testing blocks have different block size 300.

We use a threshold value to discriminate those blocks in which concept drift occurs from the other ones. According to some experiments, we choose 6×10^{-7} for block length 500 and 5×10^{-9} for block length 300. Figs. 4 and 5 show KL divergence values for each block before and after update with specified threshold. First concept drift is detected on second block. As we can see in Fig. 4 the value of KL divergence decreases after update. In the next experiment with block length 300, we have more updates during the test as it can be seen in Fig.5.

In addition to above parameter, we have three other parameters: number of message M that should be seen by obsolete model for using in updating step, α in equation (4) and β in equation (5). Based on our experiments, we choose the value 200 and 150 for block length 500 and 300 respectively and the value 0.5 for both α and β parameters.

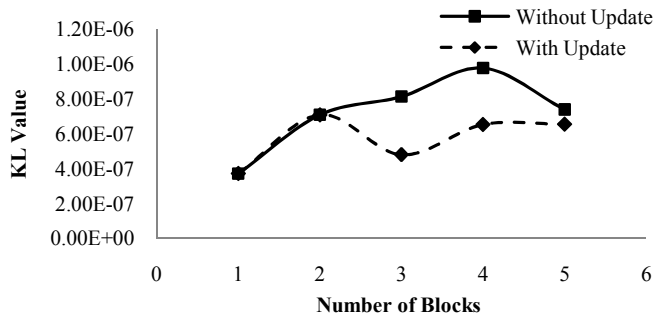

Figure 4. KL values before and after updating with block length 500

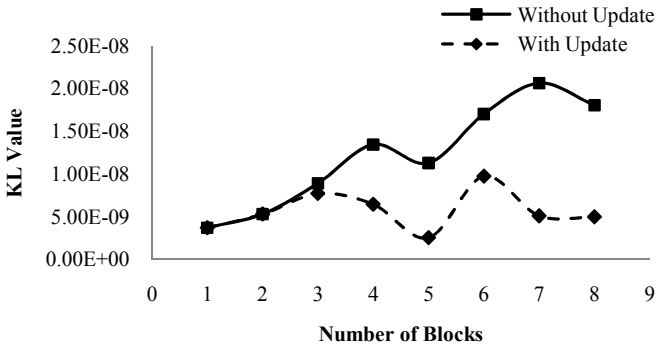

Figure 5. KL values before and after updating with block length 300

To evaluate the proposed approach, we have compared it with multinomial Naïve Bayes classifier algorithm without updating the model during the evaluation. We use a standard measure to evaluate the performance of the proposed system: accuracy, which is the percentage of test messages that are correctly classified by the model. It is calculated by the following formula [13]:

$$Accuracy = \frac{TP+TN}{P+N} \quad (9)$$

Where, TP is the number of spam messages which is labeled as spam, TN is the number of legitimate messages which are labeled as non-spam, P is the total number of spam messages and N is the total number of ham messages in test data.

Figs.6 and 7 present the performance of the proposed method against the static Naïve Bayes method in terms of accuracy on the testing set. In this figure the horizontal axis indicates the number of the current documents in the testing data and the vertical axis indicates the accuracy of each method till the current document in the test data.

As we can see in Fig. 6, in the second block of multinomial Naïve Bayes method, there is an obvious reduction over accuracy. The reason of this reduction is the difference between the content of the training data and the current block, in other words–a concept drift. The proposed approach detects this change at the beginning of the second block but based on this study we evaluate 200 first messages of this block based on obsolete model and then update our model. After updating, we can see that the accuracy of the proposed approach has been increased dramatically at the 700th documents. Although there is not a clear reduction over accuracy of the proposed approach in the fifth block but our method has detected a change and has updated the model which resulted in increasing accuracy.

Finally we can see from Figs. 6 and 7 that the proposed method could outperform the multinomial Naïve Bayes classification algorithm without update in terms of accuracy about 9% in block length 500 and 8% in block length 300. This is achieved by detecting the position of concept drift effectively and adapting the model in the correct time.

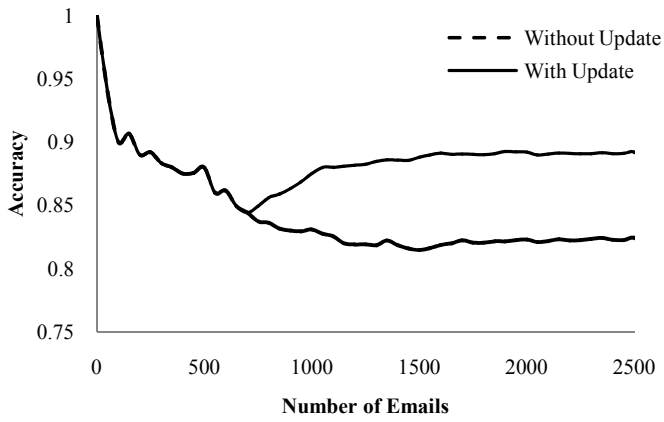

Figure 6. Effect of concept drift detection on accuracy with block length 500

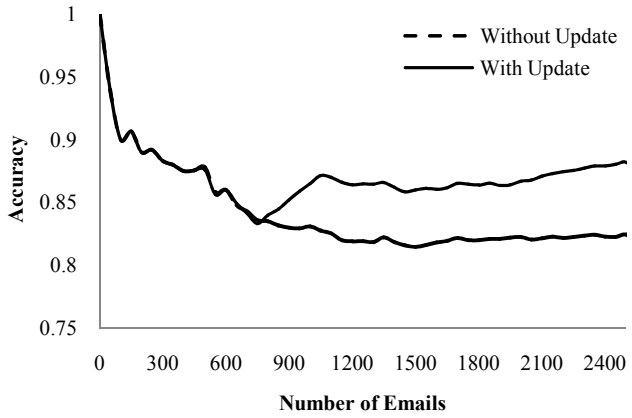

Figure 7. Effect of concept drift detection on accuracy with block length 300

I. CONCLUSION

In this paper, we have proposed a new approach to detect concept drift in a spam filtering system. The proposed method uses language model to build the content distribution model and then using KL divergence detect concept drift. After detecting concept drift it can adapt the model effectively. The proposed method has been evaluated on Enron data set and the results assert its efficiency in detecting concept drift and adapting the model.

There are several issues for future work, such as improvement of the updating strategy to adapt the model after receiving each email instead of a determined number of emails, decreasing the number of parameters which are initialized by user.

REFERENCES

[1] T.S. Guzella and W.M. Caminhas, "A review of machine learning approaches to Spam filtering," Expert Systems with Applications, vol. 36, p. 10206-10222, , 2009.

[2] J.Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," Seventh Brazilian Symp,Advances in Artificial Intelligence–SBIA 2004, Springer, p. 66–112, 2004.

[3] A. Kosmopoulos, G. Paliouras, and I. Androutsopoulos, "Adaptive spam filtering using only naïve bayes text classifiers," In Proceedings of the Conference on Email and Anti-Spam, 2008.

[4] G. Ruan and Y. Tan, "Intelligent detection approaches for spam," In Proc Int Conf on Nat Comput, 2007.

[5] X. Zhang, W. Dai, G. Xue, and Y. Yu, "Adaptive Email Spam Filtering Based on Information Theory," Web Information Systems Engineering–WISE, . 159-170, 2007.

[6] C. P., N. N., D. S.J., and H. M., "A case-based approach to spam filtering that can track concept drift," Proc. ICCBR-2003 Workshop on Long-Lived CBR Systems, 2003.

[7] S.J. Delany, P. Cunningham, A. Tsymbal, and L. And Coyle, "A case-based technique for tracking concept drift in spam filtering," Knowledge-Based Systems, p. 187-195, 2005.

[8] J. Méndez, F. Fdez-Riverola, F. Díaz, E. Iglesias, and J. Corchado, "Tracking Concept Drift at Feature Selection Stage in SpamHunting: an Anti-Spam Instance-Based Reasoning System," Proc. of the 8th European Conference on Case-Based Reasoning, 2006.

[9] W.F. Hsiao and T.M. Chang, "An incremental cluster-based approach to spam filtering," Expert Systems with Applications, vol. 34, p. 1599-1608, 2008.

[10] F. Fdez-Riverola, E. Iglesias, F. Diaz, J. Mendez, and J. Corchado, "Applying lazy learning algorithms to tackle concept drift in spam filtering," Expert Systems with Applications, vol. 33, p. 36-48, 2007.

[11] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," AAAI 98 Workshop on Text Categorization, 1998.

[12] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes," Third Conference on Email and Anti-Spam (CEAS), Citeseer, p. 125–134, 2006.

[13] C.D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[14] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," Conference on Information and Knowledge Management, p. 403, 2001.