The Impact of Feature Extraction and Selection on SMS Spam Filtering

A. K. Uysal[1], S. Gunal[1], S. Ergin[2], E. Sora Gunal[2]
[1]*Department of Computer Engineering, Anadolu University, Eskisehir, Turkiye*
[2]*Department of Electrical and Electronics Engineering, Eskisehir Osmangazi University, Eskisehir, Turkiye*
akuysal@anadolu.edu.tr

***Abstract*—This paper investigates the impact of several feature extraction and feature selection approaches on filtering of short message service (SMS) spam messages in two different languages, namely Turkish and English. The entire feature set of filtering framework consists of the features originated from the bag-of-words (BoW) model along with the ensemble of structural features (SF) specific to spam problem. The distinctive BoW features are identified using information theoretic feature selection methods. Various combinations of the BoW and SF are then fed into widely used pattern classification algorithms to classify SMS messages. The filtering framework is evaluated on both Turkish and English SMS message datasets. For this purpose, as part of the study, the first publicly available Turkish SMS message collection is constituted as well. Comprehensive experimental analysis on the respective datasets revealed that the combinations of BoW and SFs, rather than BoW features alone, provide better classification performance on both datasets. Effectiveness of the utilized feature selection methods however slightly differs in each language.**

***Index Terms*—Feature extraction, feature selection, SMS, spam filter.**

I. INTRODUCTION

In recent years, Short Message Service (SMS) has become one of the most common communication methods due to rapid increase in the number of mobile phone users worldwide. This increase has unavoidably attracted spammers and caused SMS spam (unsolicited) message problem just as in the case of spam e-mails. Today, majority of SMS messages received by mobile phones are unfortunately disturbing spam messages such as credit opportunities of banks, promotion and discount announcements of stores, and new tariffs of communications service providers.

Simple techniques including white and black list methods fail to categorize SMS messages without user intervention. Even worse, a phone number inserted into the black list may send legitimate messages beside spam, e.g., a bank may send a spam message including new credit opportunities and a legitimate message containing online banking password as well. In this case, smarter methods such as content based classification are needed.

Though the problem of SMS spam is not as old as of email spam [1], [2], there have been several efforts in the literature to detect SMS spam messages. Some examples to those efforts are as follows. Bayesian filtering techniques were employed in [3]. Feature-based and compression-model-based filters were evaluated in [4]. Another filter system using support vector machine and a thesaurus was proposed in [5]. A framework utilizing the content based filtering and challenge-response was introduced in [6]. Another SMS anti-spam system combining behavior-based social network and temporal analysis was presented in [7]. Performances of a number of classifiers in SMS spam filtering were compared in [8]. Bayesian learning and support vector machine classification were used in [9]. Local-concentration-based [10] and stylistically motivated features [11] were employed for the filtering process. Bayesian based classifiers were utilized together with the distinctive features determined by information theoretic feature selection methods in [12]. Finally, a number of recent studies on SMS spam filtering are reviewed in [13].

In regard to the abovementioned studies, this paper extensively analyses the effects of several feature extraction and feature selection methods together on filtering SMS spam messages in two different languages, namely Turkish and English. The entire feature set of the filtering scheme is composed of the features originated from the bag-of-words (BoW) model [14], and also an ensemble of structural features (SF) adopted for the spam problem. The distinctive features based on the bag-of-words model are determined using chi-square and Gini index based feature selection methods. The selected features are then combined with the structural features, and fed into two distinct pattern classification algorithms, namely k-nearest neighbor and support vector machine, to classify SMS messages as either spam or legitimate. The filtering framework is evaluated on two separate SMS message datasets consisting of Turkish and English messages, respectively. For this purpose, as part of the study, the first publicly available Turkish SMS message collection is constituted whereas an existing dataset in English is employed as well. Extensive experimental analysis on both datasets revealed that the combinations of

Manuscript received June 07, 2012; accepted November 06, 2012.
This work was supported by Anadolu University, Fund of Scientific Research Projects under grant number 1103F054.

BoW and SFs, rather than BoW features alone, provide better classification performance. Nevertheless, efficacy of the feature selection methods slightly differs in each language.

The remainder of the paper is organized as follows: First of all, the SMS message datasets used in the study are explained. Next, the feature extraction approaches applied on SMS messages are presented. Then, mathematical backgrounds of the feature selection methods are described. Afterwards, the pattern classification algorithms are discussed. Subsequently, the experimental analysis and related results are provided. Finally, some concluding remarks are given.

II. DATASETS

Although numerous e-mail datasets, or collections, [15]–[17] have been offered for the use of researchers, there are just a limited number of publicly available SMS message collections in the literature. Therefore, as part of this study, a brand new SMS message collection is constituted in Turkish, which is one of the widely used agglutinative languages worldwide. This is the first Turkish SMS message collection within the academic literature. The collection consists of 420 spam and 430 legitimate messages that are collected from volunteers. The collection, namely TurkishSMS, is publicly available at http://ceng.anadolu.edu.tr/par/ so that researchers may use it freely for academic purposes. Additionally, another SMS message collection in English [18], which is a good example of non-agglutinative languages, is utilized in the experimental study as well. This collection contains 425 spam and 450 legitimate messages.

Since both Turkish and English datasets are balanced and their sizes are almost equal, the experimental results on these datasets can be fairly compared to each other.

III. FEATURE EXTRACTION

Detection of SMS spam messages is actually a subset of spam e-mail detection problem. While an e-mail may contain text, graphics, hyperlinks, and even attached files [19], an SMS message contains only text limited with only 160 characters [20]. Consequently, detection of spam messages corresponds to a 2-class text classification problem where the classes are defined as “spam” and “legitimate”.

Vast amount of text classification studies make use of the bag-of-words model [21] to represent text documents where the exact ordering of words, or terms, in the documents is ignored but the number of term occurrences is considered. Each distinct term in a document collection consequently constitutes an individual feature. Terms are assigned particular weights representing their importance in a given document [22]. The most common weighting scheme is Term Frequency - Inverse Document Frequency (TF-IDF) that scales down the number of occurrences of a term in a document by considering the number of documents in the collection containing that term [23]. Thus, a document is represented by multi-dimensional feature vector where each dimension of the vector corresponds to the weighted value for a distinct word within the document collection, which is also known as the vector space model [24].

Even if SMS spam filtering can be treated as conventional text classification task, the structure of spam messages can be significantly different than that of formal texts. Since the size of an SMS message is limited with just 160 characters, both the message length and number of terms have of great importance. Also, the usage of upper or lower case characters can be indicator of spam. Similarly, some non-alphanumeric characters (e.g., “!”, “$”) and numeric characters (e.g., phone numbers) are commonly encountered in spam messages. Finally, URL links are usually observed in SMS spam as well. Considering all those characteristics, in this paper, an ensemble of structural features is adopted along with the features originated from the bag-of-words model. The structural features (SF) extracted from a given SMS message are summarized in Table I.

TABLE I. LIST OF STRUCTURAL FEATURES.

No	Name	Description
SF1	Message length	Number of all characters
SF2	Number of terms	Number of terms obtained using alphanumeric tokenization
SF3	Uppercase character ratio	Number of uppercase characters normalized by the message length
SF4	Non-alphanumeric character ratio	Number of non-alphanumeric characters normalized by the message length
SF5	Numeric character ratio	Number of numeric characters normalized by the message length
SF6	Presence of URL	Presence of “http” and/or “www” terms

It should be also noted that only stemming and lower case conversion are carried out as the preprocessing steps during the feature extraction. Since two different languages, namely Turkish and English, are in consideration within the scope of this work, the stemming stage is specific to the language. In case of Turkish messages, fixed-prefix stemming algorithm [25] is employed, whereas well-known Porter stemming algorithm [26] is utilized for the messages in English. Stopword removal is not applied due to relatively short length of the messages.

IV. FEATURE SELECTION

Though there exist filters, wrappers, and embedded feature selection methodologies, researchers prefer the filters to select distinctive features particularly in text categorization problems due to classifier independency and relatively low computation time of the filters [27]. The filter methods utilized within this work are based on chi-square (CHI2) and Gini index (GI) metrics. Both methods were proven to be quite successful in previous text categorization studies [21], [28], [29].

In statistics, CHI2 test is applied to examine independence of two events. The events, A and B, are assumed to be independent if

$$P(AB) = P(A)P(B), \tag{1}$$

where $P(AB)$ is the joint probability of A and B, while $P(A)$ and $P(B)$ are the probabilities of these two events, respectively. For selection of text features, these two events correspond to the occurrence of particular term and class, respectively. CHI2 information can be computed using

$$CHI2(t,C) = \sum_{t\in\{0,1\}} \sum_{C\in\{0,1\}} \frac{(N_{t,C} - E_{t,C})^2}{E_{t,C}}, \quad (2)$$

where N is the observed frequency and E is the expected frequency for each state of term t and class C [23]. CHI2 is a measure of how many expected counts and observed counts deviate from each other. A high value of CHI2 indicates that the hypothesis of independence is not correct. If the two events are dependent, then the occurrence of the term makes the occurrence of the class more likely. Consequently, the regarding term is relevant as a feature. CHI2 score of a term is calculated for individual classes. This score can be globalized over all classes in two ways. The first way is to compute the weighted average score for all classes while the second one is to choose the maximum score among all classes. In this paper, the former approach was preferred.

GI is another feature selection method which is an improved version of the method originally used to find out the best split of attributes in decision trees [30]. It has relatively simpler computation [31] as given below

$$GI(t) = \sum_{i=1}^{M} P(t \mid C_i)^2 P(C_i \mid t)^2 . \quad (3)$$

In this formulation, $P(t \mid C_i)$ is the probability of term t given presence of class C_i, and $P(C_i \mid t)$ is the probability of class C_i given presence of term t.

Once the importance scores of all terms within a text collection are obtained by either of the abovementioned methods, top-T terms with the highest scores are selected.

V. CLASSIFICATION

Two distinct pattern classification algorithms, namely k-nearest neighbour (kNN) and support vector machine (SVM), are employed in this work.

kNN algorithm classifies feature vectors based on the closest training examples in the feature space [32]. More specifically, an unknown feature vector is assigned to the class that is the most common amongst its k nearest neighbours where k is a positive integer. The value of k is determined empirically, e.g., it may be optimized with respect to the classification error on training dataset. In the special case that k = 1, the feature vector is simply assigned to the class of its nearest neighbour.

On the other hand, SVM, which is one of the state-of-the-art pattern classification algorithms, aims to find out maximum-margin hyperplane in a transformed feature space using the kernel trick [32]. Though there are several kernel types, linear kernel was preferred in this study due to its proven performance in text classification research before [33].

VI. EXPERIMENTAL WORK

The impacts of various feature extraction, feature selection, and pattern classification methods on filtering SMS spam messages in Turkish and English were analyzed in the experimental work. For this purpose, eight different feature sets were considered. Those sets are listed in Table II. The first feature set contains only BoW features. The sets between two and seven contain BoW features and a single structural feature. The last feature set is composed of BoW features and all six structural features together. From now on, the last feature set (BoW + SF1 + SF2 + SF3 + SF4 + SF5+ SF6) will be represented by (BoW + SF1:SF6) for convenience.

TABLE II. LIST OF FEATURE SETS.

No	Feature Set
1	BoW
2	BoW + SF1
3	BoW + SF2
4	BoW + SF3
5	BoW + SF4
6	BoW + SF5
7	BoW + SF6
8	BoW + SF1 + SF2 + SF3 + SF4 + SF5+ SF6

During the experiments, selection of BoW features were carried out using CHI2 and GI methods, where the number of selected features ranged from 1% to 100% of the entire BoW features. As an example, Top-10 terms determined by CHI2 and GI methods are listed in Table III for each dataset. It should be noted that several stopwords specific to Turkish (e.g., “ve”, “ile”, “icin”) and English languages (e.g., “i”, “to”, “that”, “your”) are surprisingly present in these lists. Total numbers of pre-processed distinct terms in Turkish and English datasets are 2.690 and 3.179, respectively.

TABLE III. TOP-10 DISCRIMINATIVE TERMS IN (A) TURKISH (B) ENGLISH DATASET.

Selection	Terms
Part A	
CHI2	com, ve, gonde, icin, tl, tr, sadec, hemen, kazan, ile
GI	com, ve, icin, indir, tl, firsa, gonde, tr, ozel, sadec
Part B	
CHI2	call, your, i, txt, stop, free, 1, to, that, now
GI	to, call, i, your, now, you, a, txt, stop, for

The feature sets were then fed into kNN and SVM classifiers. Since both datasets are balanced (i.e., the number of SMS messages in legitimate and spam classes are almost equal), well known Micro-F1 score [23] was employed to assess the classification performance. The classification results are presented in Fig. 1 and Fig. 2 for Turkish dataset and in Fig. 3, Fig. 4 for English dataset, respectively. The results were obtained using 3-fold cross validation to evaluate the datasets objectively.

In general, rather than BoW features alone, combinations of BoW (regardless of the utilized feature selection method) and SFs provided higher scores in most cases. Particularly, the contributions of SF1, SF2, and SF1:SF6 to classification performance were more obvious than that of the other SFs.

In case of Turkish messages, the highest Micro-F1 score was approximately 0.98. This score was obtained using SF2 (or, SF1:SF6), and 50% of BoW features selected by CHI2, which were together applied on SVM classifier. On the other hand, the maximum score achieved by kNN classifier was around 0.95 with the combination of SF2 and 50% of BoW features selected by GI.

In case of English messages, the highest Micro-F1 score was around 0.96. This value was achieved using SF1:SF6, and 100% of BoW features, which were together applied on

SVM classifier. Since all BoW features were employed to attain the highest score, no particular feature selection method was superior to another. In contrast, the maximum score achieved by kNN classifier was around 0.90 with the combination of SF3 and just 1% of BoW features selected by CHI2.

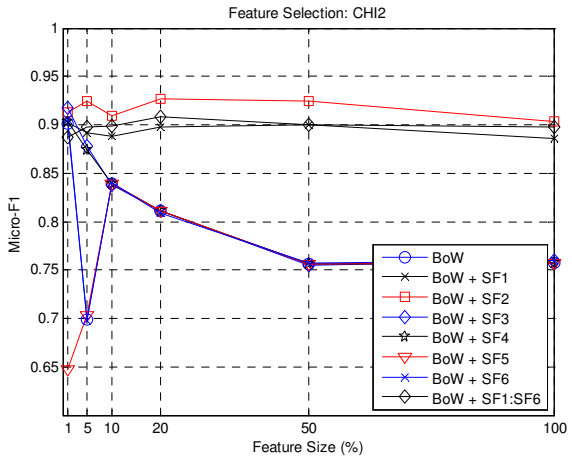
In addition to the classification performance, dimension reduction rate is another important aspect of recognition process. Consequently, an analysis for dimension reduction was also carried. In order to compare efficacy of the feature combinations in terms of dimension reduction rate and Micro-F1 values, a dimension reduction (DR) scoring scheme [34] was adopted for this work. This scheme favors higher Micro-F1 scores at lower feature dimensions as formulated in

$$\text{DR Score} = \frac{1}{d}\sum_{i=1}^{d}\frac{\dim_D}{\dim_i}\times S_i\,, \tag{4}$$

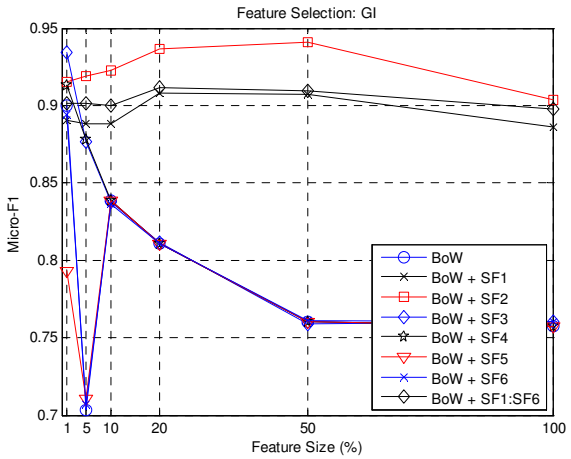
where $\dim_D$ is the maximum feature size utilized, d is the number of trials, $\dim_i$ is the feature size at the ith trial, and S_i is the Micro-F1 score of the ith trial. Since the classification results of SVM classifier were better than that of kNN in all cases as illustrated by Fig. 1–Fig. 4, the scores attained only by SVM were considered during this analysis. Top-5 DR scores for both datasets were computed and listed in Table IV. One can easily note from this table that the feature set (BoW + SF1:SF6) and GI based selection method surpass the other combinations for Turkish messages. In case of English messages, on the other hand, CHI2 based selection method replaces GI whereas the feature set remains the same. Another interesting finding from these results was that while the better feature selection method enrolled in obtaining the highest Micro-F1 score was CHI2 in Turkish messages, GI took the first place in terms of DR performance.

TABLE IV. TOP-5 RESULTS OF DIMENSION REDUCTION ANALYSIS FOR (A) TURKISH (B) ENGLISH DATASET.

No	DR Score	Feature Set	Feature Selection
Part A			
1	21.866	BoW + SF1:SF6	GI
2	21.692	BoW + SF1:SF6	CHI2
3	21.678	BoW + SF3	CHI2
4	21.661	BoW + SF2	GI
5	21.648	BoW + SF1	GI
Part B			
1	21.871	BoW + SF1:SF6	CHI2
2	21.838	BoW + SF1:SF6	GI
3	21.629	BoW + SF1	CHI2
4	21.451	BoW + SF1	GI
5	21.441	BoW + SF2	CHI2

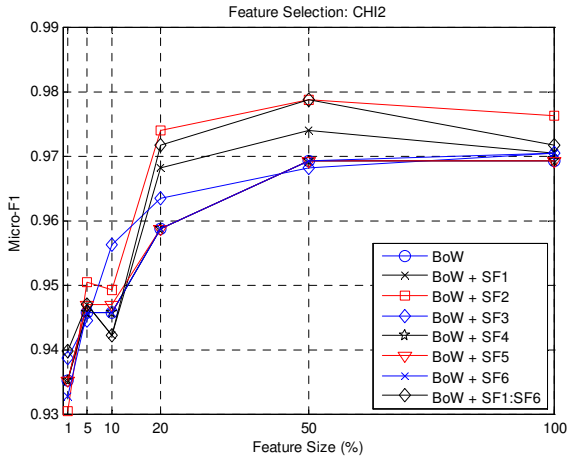


a)

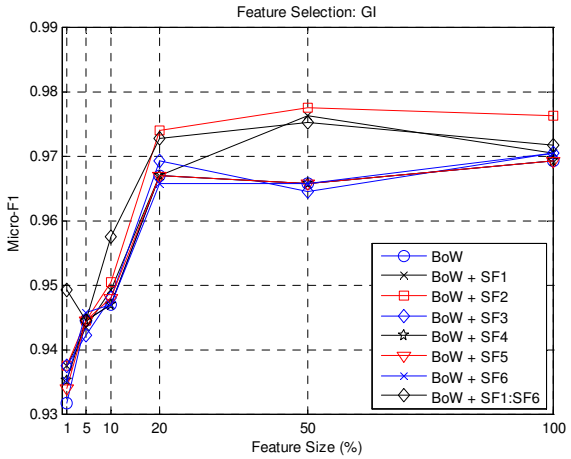


b)

Fig. 1. kNN classification results for Turkish dataset.

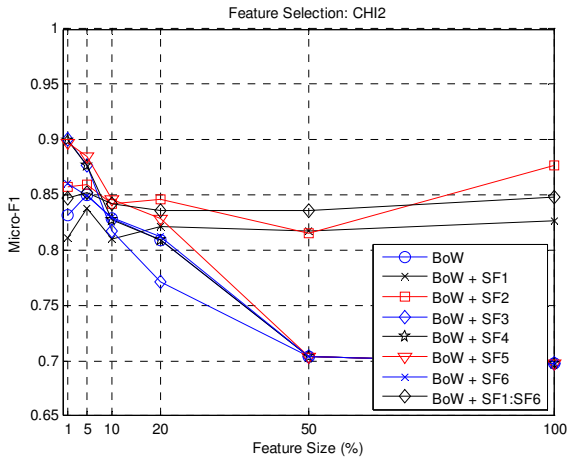


a)

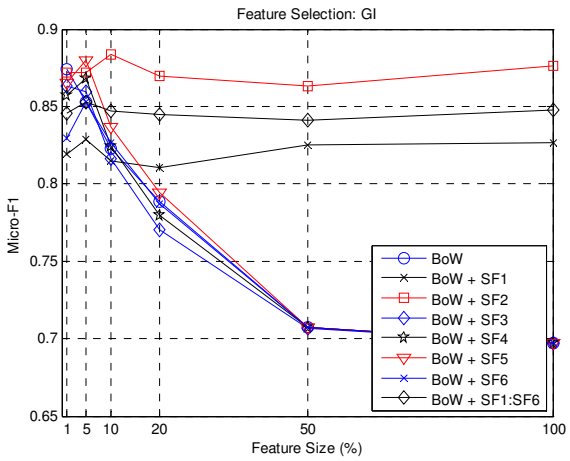


b)

Fig. 2. SVM classification results for Turkish dataset.

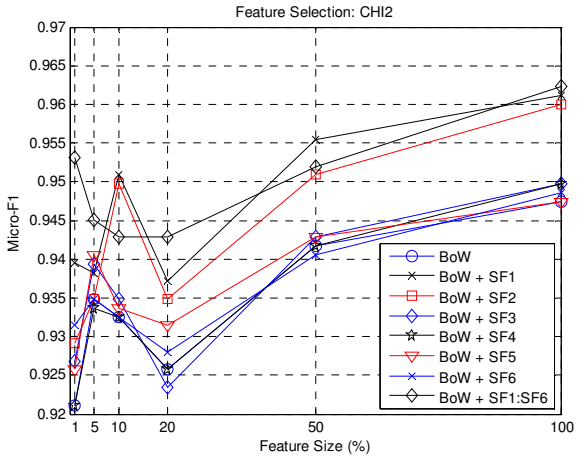


a)

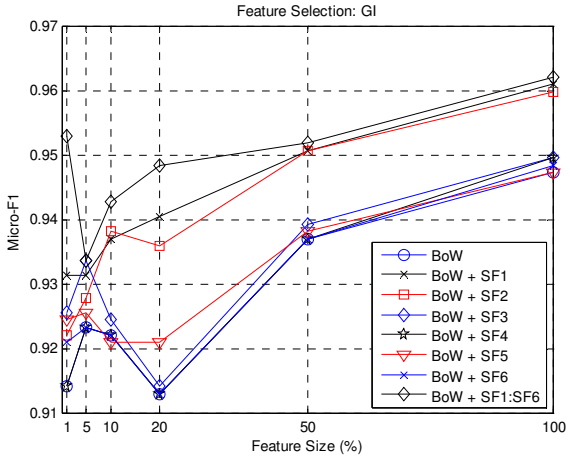


b)

Fig. 3. kNN classification results for English dataset.



a)



b)

Fig. 4. SVM classification results for English dataset.

VII. CONCLUSIONS

In this paper, the impact of various feature extraction and selection methodologies on SMS spam filtering, particularly for Turkish and English languages, was thoroughly examined in terms of classification accuracy and dimension reduction rate. Outcome of an in-depth experimental work indicated that the combinations of BoW and structural features, rather than BoW features alone, offer better classification performance most of the time. On the other hand, efficacy of the utilized feature selection strategies was not significantly superior to each other for both languages. Since Turkish and English are the leading examples of agglutinative and non-agglutinative languages respectively, the outcome of this study can also be an indicator for the other languages with similar characteristics as well.

Inspection of new structural features, assessment of other feature selection and classification methods on SMS spam filtering problem remain as interesting future works. Furthermore, this work may be extended to MMS (Multimedia Messaging Service) spam filtering task by incorporating appropriate signal processing techniques as well.

REFERENCES

[1] D. Puniškis, R. Laurutis, and R. Dirmeikis, "An artificial neural nets for spam e–mail recognition", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 5, pp. 73–76, 2006.

[2] D. Puniškis, R. Laurutis, "Behavior statistic based neural net anti–spam filters", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 6, pp. 35–38, 2007.

[3] J. M. G. Hidalgo, G. C. Bringas, E. P. Sanz, F. C. Garcia, "Content based SMS spam filtering", in *ACM Symposium on Document Engineering*, Amsterdam, The Netherlands, 2006, pp. 107–114.

[4] G. Cormack, J. M. G. Hidalgo, E. P. Sanz, "Spam filtering for short messages", in *Proc. of the 16th ACM Conference on Information and Knowledge Management*, Lisbon, Portugal, 2007, pp. 313–320. [Online]. Available: http://dx.doi.org/10.1145/1321440.1321486

[5] I. Joe, H. Shim, "An SMS spam filtering system using support vector machine", *Lecture Notes in Computer Science,* vol. 6485, pp. 577–584, 2010. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-17569-5_56

[6] J. W. Yoon, H. Kim, J. H. Huh, "Hybrid spam filtering for mobile communication", *Computers & Security,* vol. 29, no. 4, pp. 446–459, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.cose.2009.11.003

[7] C. Wang, Y. Zhang, X. Chen, Z. Liu, L. Shi, G. Chen, F. Qiu, C. Ying, W. Lu, "A behavior–based SMS antispam system", *IBM Journal of Research and Development,* vol. 54, no. 6, pp. 651–666, 2010. [Online]. Available: http://dx.doi.org/10.1147/JRD.2010.2066050

[8] T. A. Almeida, J. M. G. Hidalgo, A. Yamakami, "Contributions to the

study of SMS spam filtering: new collection and results", in *Proc. of the 11th ACM Symposium on Document Engineering*, Mountain View, California, USA, 2011, pp. 259–262. [Online]. Available: http://dx.doi.org/10.1145/2034691.2034742

[9] K. Yadav, P. Kumaraguru, A. Goyal, A. Gupta, V. Naik, "SMSAssassin: crowdsourcing driven mobile–based system for SMS spam filtering", *the 12th Workshop on Mobile Computing Systems and Applications*, Phoenix, Arizona, 2011.

[10] Z. Yuanchun, T. Ying, "A local–concentration–based feature extraction approach for spam filtering", *IEEE Transactions on Information Forensics and Security,* vol. 6, no. 2, pp. 486–497, 2011. [Online]. Available: http://dx.doi.org/10.1109/TIFS.2010.2103060

[11] D. N. Sohn, J. T. Lee, K. S. Han, H. C. Rim, "Content–based mobile spam classification using stylistically motivated features", *Pattern Recognition Letters,* vol. 33, no. 3, pp. 364–369, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2011.10.017

[12] A. K. Uysal, S. Gunal, S. Ergin, E. Sora Gunal, "A novel framework for SMS spam filtering", in *Proc. of* the *IEEE International Symposium on Innovations in Intelligent Systems and Applications*, Trabzon, Turkiye, 2012.

[13] S. J. Delany, M. Buckley, D. Greene, "SMS spam filtering: methods and data", *Expert Systems with Applications,* vol. 39, no. 10, pp. 9899–9908, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2012.02.053

[14] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with tfidf for text categorization", in *Proc. of the 14th International Conference on Machine Learning*, 1997, pp. 143–151.

[15] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, C. D. Spyropoulos, "An evaluation of naive Bayesian anti–spam filtering", in *Proc. of the Workshop on Machine Learning in the New Information Age*, 2000, pp. 9–17.

[16] V. Metsis, I. Androutsopoulos, G. Paliouras, "Spam filtering with naive Bayes – which naive Bayes?", in *Proc. of the 3rd Conference on Email and Anti–Spam*, 2006, pp. 28–69.

[17] The Spamassassin public mail corpus. [Online]. Available: http://spamassassin.apache.org/publiccorpus/

[18] M. T. Nuruzzaman, C. Lee, D. Choi, "Independent and personal SMS spam filtering", in *Proc. of the IEEE 11th International Conference on Computer and Information Technology*, 2011.

[19] S. Gunal, S. Ergin, M. B. Gulmezoglu, and O. N. Gerek, "On feature extraction for spam e–mail detection", *Lecture Notes in Computer Science,* vol. 4105, pp. 635–642, 2006. [Online]. Available: http://dx.doi.org/10.1007/11848035_84

[20] *Technical realization of the Short Message Service – Point to Point*, ETSI, GSM 03.40, 1992.

[21] S. Gunal, "Hybrid feature selection for text classification", *Turkish Journal of Electrical Engineering & Computer Sciences,* vol. 20, no. 2, pp. 1296–1311, 2012.

[22] G. Salton, C. Buckley, "Term–weighting approaches in automatic text retrieval", *Information Processing and Management,* vol. 24, no. 5, pp. 513–523, 1988. [Online]. Available: http://dx.doi.org/10.1016/0306-4573(88)90021-0

[23] C. D. Manning, P. Raghavan, H. Schutze, *Introduction to Information Retrieval*. New York, USA: Cambridge University Press, 2008.

[24] G. Salton, A. Wong, C. S. Yang, "A vector space model for automatic indexing", *Communications of the ACM,* vol. 18, no. 11, pp. 613–620, 1975. [Online]. Available: http://dx.doi.org/10.1145/361219.361220

[25] F. Can, S. Kocberber, E. Balcik, C. Kaynak, H. C. Ocalan, O. M. Vursavas, "Information retrieval on Turkish texts", *Journal of the American Society for Information Science and Technology,* vol. 59, no. 3, pp. 407–421, 2008. [Online]. Available: http://dx.doi.org/10.1002/asi.20750

[26] M. F. Porter, "An algorithm for suffix stripping", *Program,* vol. 14, no. 3, pp. 130–137, 1980. [Online]. Available: http://dx.doi.org/10.1108/eb046814

[27] G. Forman, "An extensive empirical study of feature selection metrics for text classification", *Journal of Machine Learning Research,* vol. 3, pp. 1289–1305, 2003.

[28] Y.–T. Chen and M. C. Chen, "Using chi–square statistics to measure similarities for text categorization", *Expert Systems with Applications,* vol. 38, no. 4, pp. 3085–3090, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.08.100

[29] H. Ogura, H. Amano, and M. Kondo, "Comparison of metrics for feature selection in imbalanced text classification", *Expert Systems with Applications,* vol. 38, no. 5, pp. 4978–4989, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.09.153

[30] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, Z. Wang, "A novel feature selection algorithm for text categorization", *Expert Systems with Applications,* vol. 33, no. 1, pp. 1–5, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2006.04.001

[31] H. Ogura, H. Amano, M. Kondo, "Feature selection with a measure of deviations from Poisson in text categorization", *Decision Support Systems,* vol. 36, no. 3, pp. 6826–6832, 2009.

[32] S. Theodoridis, K. Koutroumbas, *Pattern Recognition 4th ed*., Academic Press, 2008.

[33] W. Zhang, T. Yoshida, X. J. Tang, "Text classification based on multi–word with support vector machine", *Knowledge–Based Systems,* vol. 21, no. 8, pp. 879–886, Dec. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2008.03.044

[34] S. Gunal, R. Edizkan, "Subspace based feature selection for pattern recognition", *Information Sciences,* vol. 178, no. 19, pp. 3716–3726, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2008.06.001