Effect of Untrained Strange Words in Mail Filtering

Seiya　TEMMA*　　Hiroshi　MATSUNO**

ABSTRACT

Numerous email filtering approaches based on machine learning have been proposed, enabling classifications approaching complete filtering. These methods categorize newly received emails based on features extracted from past received emails; thus their classification performance decreases for emails with altered features. To accurately classify such emails, it is imperative to extract features from newly emerged words untrained by machine learning. Therefore, upon examining the occurrence tendencies of these words, we discovered a trend that over time, spam emails contain more out-of-dictionary words. To utilize this trait for classification, we proposed a method to facilitate categorizing emails with many out-of-dictionary untrained words as spam. Consequently, applying this approach reveals that by assigning a relatively high spam probability around 0.7 to untrained words untrained, the filtering performance improves.

Keywords: Email filtering, Strange words, untrained words, Changes in occurrence tendencies.

1. Introduction

With the growth of the Internet, various spam filtering methods have been employed to classify the increasing amount of spam emails.

These approaches performance is the more words appearing in a similar tendency to the mails used in machine learning a test mail contains, the higher its classification performance, and conversely, the more words exhibiting different tendencies, the lower its performance. Among these words, those appearing for the first time in a test mail, namely untrained words that have not been machine learned, are particularly difficult to use for classification, deteriorating performance.

To intentionally decrease this performance, spam senders use symbols, spaces, and HTML tags in words as shown below [1].

- price$ for be$t drug$!
- Sym8oL
- priceC I A L I S
- <font>se</font>xu<font>al</font>

*福山大学工学部情報工学科　　**山口大学大学院創成科学研究科

Such words are not listed in dictionaries (hereafter called strange words) and can be easily altered by changing said combinations, creating new ones daily, many becoming untrained words. This is one reason why spam features temporally change easily.

We have previously confirmed that the classification accuracy by trained strange words is higher than other words like nouns, verbs, and adjectives [2], indicating this new viewpoint of utilizing trained strange words can be a technique to bring us closer to perfect filtering, but we have not dealt with untrained strange words.

In this paper, we newly propose a technique for applying untrained strange words to email filtering. Specifically, to further verify the utility of strange words, we analyze the characteristics of untrained strange words appearing in email subjects and bodies, present a model for utilizing the extracted features in classification, and by applying this to bsfilter [3], experimentally confirm improved classification performance.

2. Experimental Setup

2.1 Email dataset

To enable reproduction and comparison by third parties, publicly available datasets were used in this study, comprised of relatively old emails but still widely utilized in recent studies, containing the strange words with integrated symbols and HTML tags.

2.1.1 SpamAssassin public corpus [4]

Used extensively in many prior email filtering studies (hereafter, SpamAssassin). Comprised of emails received over approximately 2 years from January 2002 to December 2003 (4,150 ham emails, 1,987 spam emails, total 6,173 emails), not including emails received by spam traps.

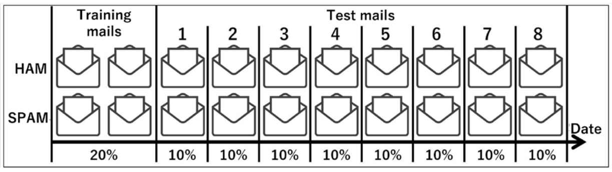


Figure 1: The categorization of training emails and 8 pattern test emails.

2. 1. 2 TREC 2007 spam corpus [5], [6]

Employed in the Text REtrieval Conference and numerous other email filtering studies (hereafter, TREC). Comprised of emails received by a specific server over about 3 months from April 8 to July 6, 2007 (25,220 ham emails, 50,199 spam emails, total 75,419 emails), including emails obtained by spam traps.

2.2 Data Set Partitioning and Usage

To compare experimental results using two data sets with different numbers of emails and reception periods, as shown in Figure 1, the oldest 20% of emails were assigned as the learning set and the remainder as the test mail set for both ham and spam emails. These were ordered chronologically and divided evenly into 8 sections, numbered 1 through 8.

In actual filter operation, newly received emails would normally be added to update machine learning. However, to focus on feature changes over time, fixed training emails were analyzed.

To limit information to that actually needed by recipients, only subjects and bodies were used (headers and attachments were removed and emails without bodies were excluded).

Words split by NLTK's tokenizer were used as words, with those not registered in WordNet [7] defined as strange words.

3. Analysis of Untrained Word Characteristics

3.1 Temporal Changes in Ham and Spam Email Classification Performance

Ham email features also change due to shifts in content related to activities, topics, and personal relationships. In addition to trend changes, spam email features alter via the aforementioned creation and replacement of strange words.

As these feature changes accumulate over time, classification performance can be expected to decrease for both ham and spam emails.

To closely examine this, classification experiments were conducted using SpamAssassin and TREC described in Section 2.1 with bsfilter. Results are shown in Figure 2.

The horizontal axis denotes the test email set number, with larger numbers indicating longer elapsed times from the training email set. The vertical axis shows the mean spam probability value calculated for each test email set, presented separately for ham and spam emails.

It can be seen that SpamAssassin and TREC demonstrate the same tendency, with ham emails maintaining a low spam probability around 0.0, indicating time-independent and highly accurate classification.

In contrast, even for test email set 1 with the shortest elapsed time, spam email spam probability is low at around 0.85, decreasing to around 0.6 thereafter, suggesting deteriorating classification accuracy. This may be attributed to the previously mentioned trend changes and creation/replacement of strange words increasing untrained words.

3.2 Potential Applicability of Untrained Words to Classification

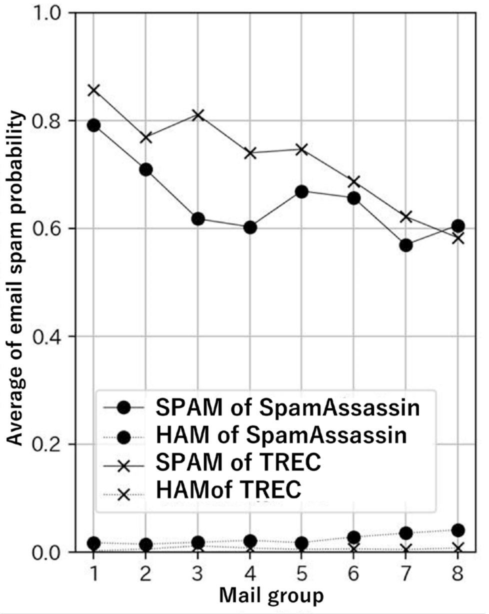


Figure 2: The classification performance of bsfilter.

To investigate the potential for utilizing untrained words, temporal changes in the proportion of untrained word types were examined. Results are shown in Figure 3.

The horizontal axis denotes the test email set number, with larger numbers indicating longer elapsed times from the training email set. The vertical axis shows the proportion of untrained word types out of all word types appearing in an email, averaged over each test email set. Results are separated into strange and in the dictionary words, for ham and spam emails respectively.

From this figure, it can be seen that for both SpamAssassin and TREC, the proportion of untrained strange words is higher and exhibits an increasing trend over time for spam emails compared to ham emails.

In other words, when the proportion of untrained strange words in an email is high, that email tends to be spam, with this tendency strengthening over elapsed days.

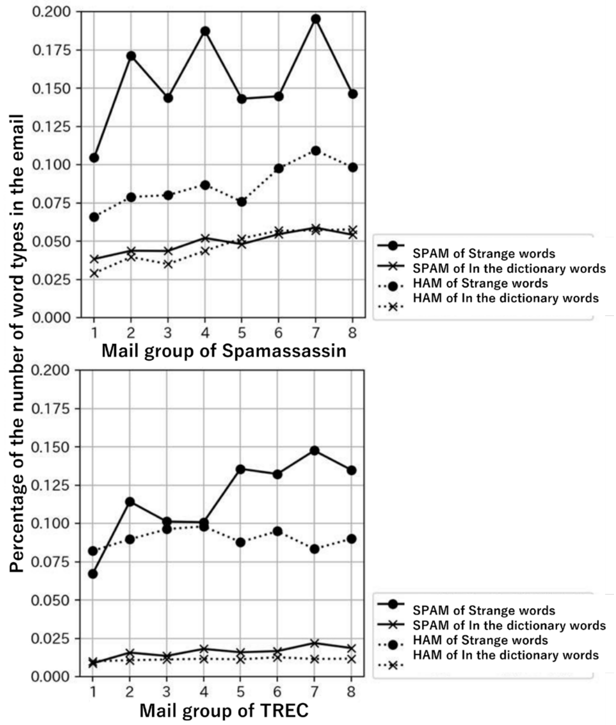


Figure 3: The percentage of untrained words in the email body.

4. Applying Untrained Strange Words to Classification

4.1 Incorporating Untrained Strange Words into Existing Filters

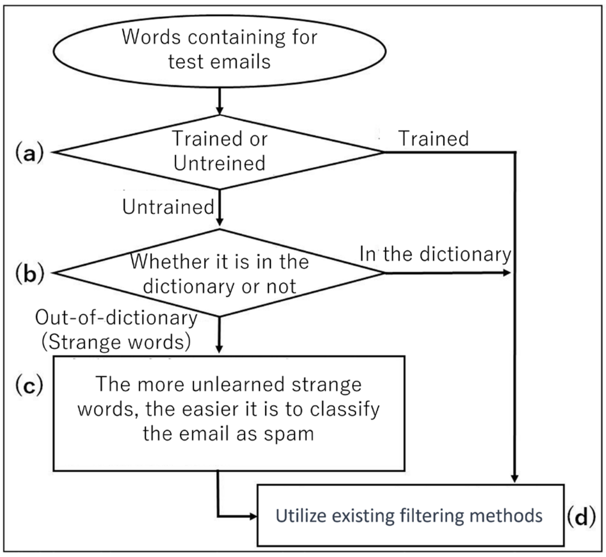
To apply untrained strange words to existing methods, classification can be biased to more easily categorize emails containing greater numbers of them as spam. Specifically, the following processing flow shown in Figure 4 is used:

(a) For each word in a test email, divide into untrained words (not contained in the training email set) and trained words (contained in the training email set).

(b) Divide untrained words into strange and in the dictionary words.

(c) Count untrained strange words, and incorporate processing into the existing filtering method to bias classification towards spam for larger numbers.

(d) Classify the email using trained words with the existing filtering method.



Words containing for test emails
(a)
Trained or Untreined
Trained
Untrained
(b)
Whether it is in the dictionary or not
In the dictionary
Out-of-dictionary (Strange words)
(c)
The more unlearned strange words, the easier it is to classify the email as spam
Utilize existing filtering methods
(d)

Figure 4: The process of handling words in the proposed methodology.

In this study, the utility of applying untrained strange words was experimentally investigated by incorporating them into the currently widely used bsfilter. Specifically, Figure 4 (c) was implemented by uniformly setting a spam probability for untrained strange words, and classification results were compared with the original bsfilter.

Experiments similar to above were conducted across spam probabilities spanning 0.0 to 1.0 to identify values improving classification accuracy. Results are presented in Figure 5.

The horizontal axis shows the set spam probability. The vertical axis gives values averaged over test email sets 1 through 8 for:

(Revised spam probability) - (Original spam probability)

Positive values indicate increased and negative values indicate decreased email spam probabilities.

From the figure, setting untrained strange word spam probability around 0.7 yielded improved classification for both SpamAssassin and TREC data sets.

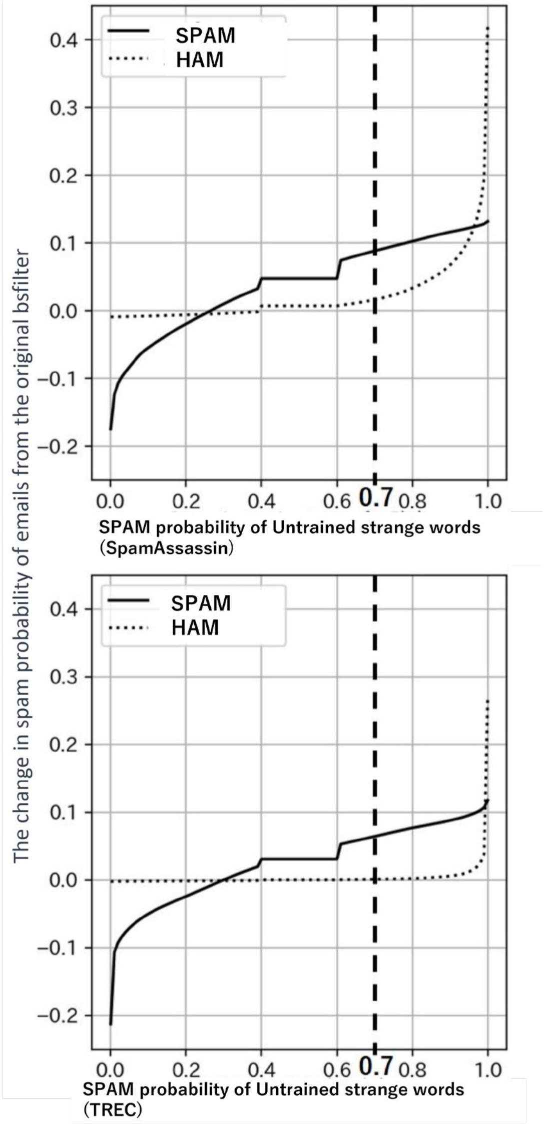


Figure 5: The relationship between the spam probability assigned to untrained words and classification performance.

4.2 Classification Performance Changes from Utilizing Untrained Strange Words

To alter the training email set, in addition to Pattern 1 used thus far in analysis, Pattern 2 employing email sets 1 and 2, Pattern 3 with email sets 3 and 4, and Pattern 4 using email sets 5 and 6 for learning were prepared as shown in Figure 6. Test emails were newer than learning set emails.

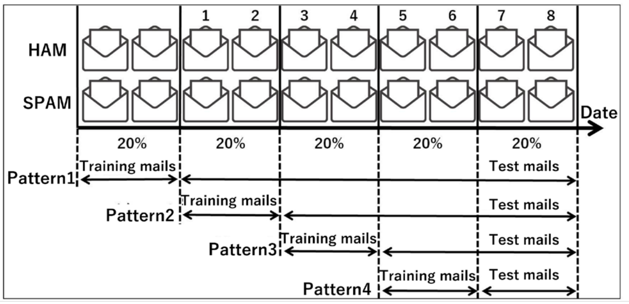


Figure 6: 4 patterns of email dataset.

Classification accuracy was determined using TREC and SpamAssassin for Patterns 1 through 4. Results are presented in Figure 7 with pattern numbers on the horizontal axis and AUC values on the vertical axis; higher pattern numbers indicate more recent training email reception dates.

An additional comparative test with bsfilter fixed to not use untrained strange words (unused) to examine utility of the 0.7 spam probability revised bsfilter and original bsfilter in leveraging untrained strange words.

It can be seen in the figure that with a 0.7 spam probability setting, classification accuracy remained high for SpamAssassin at 0.98 or above and TREC at 0.99 or above regardless of training email reception date.

For TREC results, classification accuracy converged across all methods when using newer training emails. This likely owes to the short approximately 3 month TREC reception span, resulting in minor emergence of new features and decreases in untrained words due to proximity of learning and test email reception dates.

5. Conclusion

This paper focused on spam senders' creation of untrained strange words and examined potential applicability to email classification.

First, classification experiments employing the original bsfilter confirmed decreasing spam email classification performance over time.

Investigating this cause revealed greater quantities of

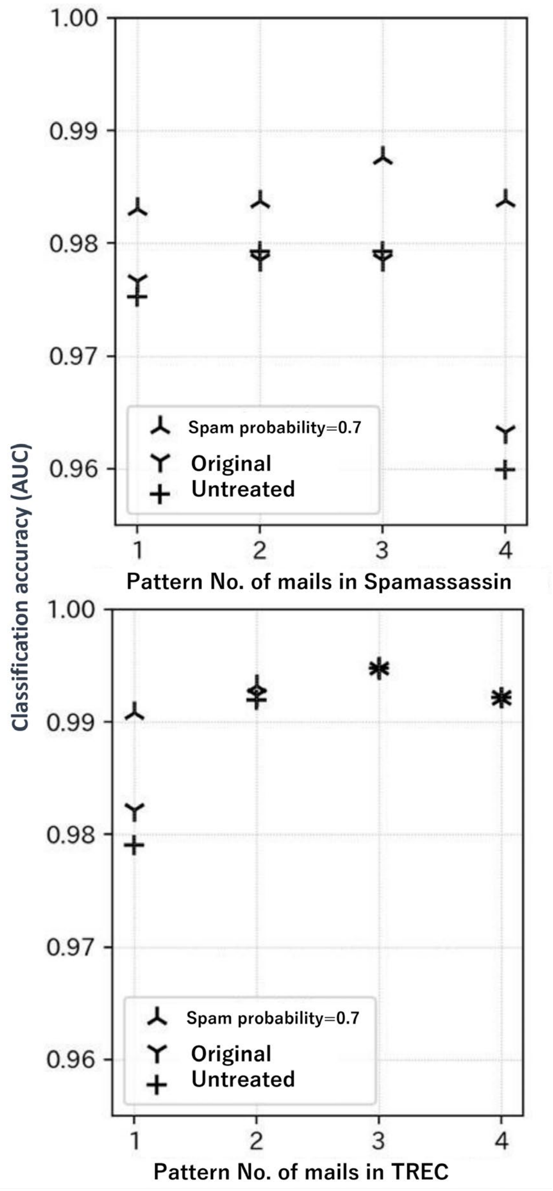


Figure 7: The improvement in classification accuracy by setting a spam probability of 0.7 for untrained words.

untrained strange words in spam versus ham emails, with the number increasing over time.

To incorporate this tendency into existing techniques, a model was proposed to facilitate categorizing emails with more untrained strange words as spam (Figure 4).

To experimentally validate efficacy of the proposed model, uniform spam probability assignment to untrained strange words was tested in bsfilter, confirming optimal classification improvement with a value of 0.7.

Email filtering has seen ongoing enhancements, reaching performance limits. Further accuracy gains approaching perfect filtering necessitate new perspectives, with this paper introducing the previously unexploited viewpoint of leveraging untrained strange words.

Acknowledgments

This research was supported by the JST Exploratory Research for Next Generation Researchers Program, under the grant number JPMJSP2111.

References

[1] Fawcett, T. "In vivo" spam filtering: a challenge problem for KDD. ACM SIGKDD Explorations Newsletter, Vol. 5, No. 2, pp. 140–148, 2003.

[2] S.Tenma, and H.Matsuno, "Feature Analysis of Strange Words for More Effective Mail Filtering" The Institute of Electronics, Information and Communication Engineers, Vol. J105-D, No. 11, 2022. In Japanese.

[3] bsfilter, available from ⟨https://ja.osdn.net/projects/bsfilter/⟩ (accessed 2021-10-01).

[4] SpamAssassin public corpus, available from ⟨https://spamassassin.apache.org/old/publiccorpus/⟩ (accessed 2022-05-23).

[5] Spam Track, available from ⟨https://trec.nist.gov/data/spam.html⟩ (accessed 2021-10-01).

[6] Cormack, G.V. "TREC 2007 spam track overview," Proc. 16th Text REtrieval Conference, TREC 2007.

[7] Miller, G.A. "WordNet: A lexical database for English," Comm. ACM, Vol.38, No.11, pp.39–41, 1995.