The Contribution of Stylistic Information to Content-based Mobile Spam Filtering

Dae-Neung Sohn and **Jung-Tae Lee** and **Hae-Chang Rim**
Department of Computer and Radio Communications Engineering
Korea University
Seoul, 136-713, South Korea
{danny,jtlee,rim}@nlp.korea.ac.kr

Abstract

Content-based approaches to detecting mobile spam to date have focused mainly on analyzing the *topical* aspect of a SMS message (*what* it is about) but not on the *stylistic* aspect (*how* it is written). In this paper, as a preliminary step, we investigate the utility of commonly used stylistic features based on shallow linguistic analysis for learning mobile spam filters. Experimental results show that the use of stylistic information is potentially effective for enhancing the performance of the mobile spam filters.

1 Introduction

Mobile spam, also known as SMS spam, is a subset of spam that involves unsolicited advertising text messages sent to mobile phones through the Short Message Service (SMS) and has increasingly become a major issue from the early 2000s with the popularity of mobile phones. Governments and many service providers have taken various countermeasures in order to reduce the number of mobile spam (*e.g.* by imposing substantial fines on spammers, blocking specific phone numbers, creating an alias address, *etc.*). Nevertheless, the rate of mobile spam continues to rise.

Recently, a more technical approach to mobile spam filtering based on the content of a SMS message has started gaining attention in the spam research community. Gómez Hidalgo et al. (2006) previously explored the use of statistical learning-based classifiers trained with lexical features, such as character and word n-grams, for mobile spam filtering. However, content-based spam filtering directed at SMS messages are very challenging, due to the fact that such messages consist of only a few words. More recent studies focused on expanding the feature set for learning-based mobile spam classifiers with additional features, such as orthogonal sparse word bi-grams (Cormack et al., 2007a; Cormack et al., 2007b).

Collectively, the features exploited in earlier content-based approach to mobile spam filtering are topical terms or phrases that statistically indicate the spamness of a SMS message, such as *"loan"* or *"70% off sale"*. However, there is no guarantee that legitimate (non-spam) messages would not contain such expressions. Any of us may send a SMS message such as *"need ur advise on private **loans**, plz call me"* or *"mary, abc.com is having **70% off sale** today"*. For current content-based mobile spam filters, there is a chance that they would classify such legitimate messages as spam. This motivated us to not only rely on the message content itself but incorporate new features that reflect its "style," the manner in which the content is expressed, in mobile spam filtering.

The main goal of this paper is to investigate the potential of stylistic features in improving the performance of learning-based mobile spam filters. In particular, we adopt stylistic features previously suggested in authorship attribution studies based on stylometry, the statistical analysis of linguistic style.[1] Our assumption behind adopting the features from authorship attribution are as follows:

- There are two types of SMS message senders, namely *spammers* and *non-spammers*.
- Spammers have distinctive linguistic styles and writing behaviors (as opposed to non-spammers) and use them consistently.
- The SMS message as an end product carries the author's "fingerprints".

[1] Authorship attribution involves identifying the author of a text given some stylistic characteristics of authors' writing. See Holmes (1998) for overview.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 321–324,
Suntec, Singapore, 4 August 2009.

Although there are many types of stylistic features suggested in the literature, we make use of the ones that are readily computable and countable from SMS message texts without any complex linguistic analysis as a preliminary step, including word and sentence lengths (Mendenhall, 1887), frequencies of function words (Mosteller and Wallace, 1964), and part-of-speech tags and tag n-grams (Argamon-Engelson et al., 1998; Koppel et al., 2003; Santini, 2004).

Our experimental result on a large-scale, real world SMS dataset demonstrates that the newly added stylistic features effectively contributes to statistically significant improvement on the performance of learning-based mobile spam filters.

2 Stylistic Feature Set

All stylistic features listed below have been automatically extracted using shallow linguistic analysis. Note that most of them have been motivated from previous stylometry studies.

2.1 Length features: *LEN*

Mendenhall (1887) first created the idea of counting word lengths to judge the authorship of texts, followed by Yule (1939) and Morton (1965) with the use of sentence lengths. In this paper, we measure the overall byte length of SMS messages and the average byte length of words in the message as features.

2.2 Function word frequencies: *FW*

Motivated from a number of stylometry studies based on function words including Mosteller and Wallace (1964), Tweedie et al. (1996) and Argamon and Levitan (2005), we measure the frequencies of function words in SMS messages as features. The intuition behind function words is that due to their high frequency in languages and highly grammaticalized roles, such words are unlikely to be subject to conscious control by the author and that the frequencies of different function words would vary greatly across different authors (Argamon and Levitan, 2005).

2.3 Part-of-speech n-grams: *POS*

Following the work of Argamon-Engelson et al. (1998), Koppel et al. (2003), Santini (2004) and Gamon (2004), we extract part-of-speech n-grams (up to trigrams) from the SMS messages and use their frequencies as features. The idea behind their utility is that spammers would favor certain syntactic constructions in their messages.

2.4 Special characters: *SC*

We have observed that many SMS messages contain special characters and that their usage varies between spam and non-spam messages. For instance, non-spammers often use special characters to create emoticons to express their mood, such as ":-)" (smiling) or "T_T" (crying), whereas spammers tend to use special character or patterns related to monetary matters, such as "$$$" or "%". Therefore, we also measured the ratio of special characters, the number of emoticons, and the number of special character patterns in SMS messages as features.[2]

3 Learning a Mobile Spam Filter

In this paper, we use maximum entropy model, which have shown robust performance in various text classification tasks in the literature, for learning the mobile spam filter. Simply put, given a number of training samples (in our case, SMS messages), each with a label Y (where $Y = 1$ if spam and 0 otherwise) and a feature vector $\overline{x}$, the filter learns a vector of feature weight parameters $\overline{w}$. Given a test sample X with its feature vector $\overline{x}$, the filer outputs the conditional probability of predicting the data as spam, $P(Y = 1|X = \overline{x})$. We use the L-BFGS algorithm (Malouf, 2002) and the Information Gain (IG) measure for parameter estimation and feature selection, respectively.

4 Experiments

4.1 SMS test collections

We use a collection of mobile SMS messages in Korean, with 18,000 (60%) legitimate messages and 12,000 (40%) spam messages. This collection is based on one used in our previous work (Sohn et al., 2008) augmented with 10,000 new messages. Note that the size is approximately 30 times larger than the most previous work by Cormack et al. (2007a) on mobile spam filtering.

4.2 Feature setting

We compare three types of feature sets, as follows:

[2]For emoticon and special pattern counts, we used manually constructed lexicons consisting of 439 emoticons and 229 special patterns.

- *Baseline*: This set consists of lexical features in SMS messages, including words, character n-grams, and orthogonal sparse word bigrams (OSB)[3]. This feature set represents the content-based approaches previously proposed by Gómez Hidalgo et al. (2006), Cormack et al. (2007a) and Cormack et al. (2007b).

- *Proposed*: This feature set consists of all the stylistic features mentioned in Section 2.

- *Combined*: This set is a combination of both the baseline and proposed feature sets.

For all three sets, we make use of 100 features with the highest IG values.

4.3 Evaluation measures

Since spam filtering task is very sensitive to false-positives (*i.e.* legitimate classified as spam) and false-negatives (*i.e.* spam classified as legitimate), special care must be taken when choosing an appropriate evaluation criterion.

Following the TREC Spam Track, we evaluate the filters using ROC curves that plot false-positive rate against false-negative rate. As a summary measure, we report one minus area under the ROC curve (1−AUC) as a percentage with confidence intervals, which is the TREC's official evaluation measure.[4] Note that *lower* 1−AUC(%) value means *better* performance. We used the TREC Spam Filter Evaluation Toolkit[5] in order to perform the ROC analysis.

4.4 Results

All experiments were performed using 10-fold cross validation. Statistical significance of differences between results were computed with a two-tailed paired t-test. The symbol † indicates statistical significance over an appropriate baseline at $p < 0.01$ level.

Table 1 reports the 1−AUC(%) summary for each feature settings listed in Section 4.2. Notice that *Proposed* achieves significantly better performance than *Baseline*. (Recall that the smaller, the better.) An even greater performance gain is obtained by combining both *Proposed* and *Baseline*. This clearly indicates that stylistic aspects of SMS messages are potentially effective for mobile spam filtering.

Feature set	1−AUC (%)	
Baseline	10.7227	[9.4476 - 12.1176]
Proposed	**4.8644**†	[4.2726 - 5.5886]
Combined	**3.7538**†	[3.1186 - 4.4802]

Table 1: Performance of different feature settings.

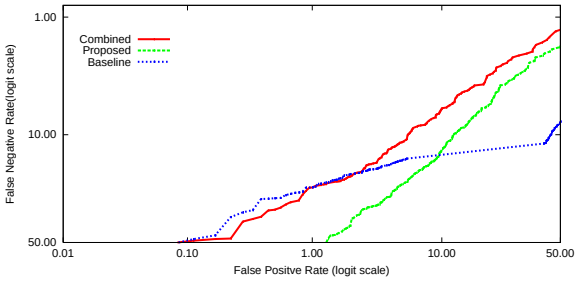

Figure 1: ROC curves of different feature settings.

Figure 1 shows the ROC curves of each feature settings. Notice the tradeoff when *Proposed* is used solely with comparison to *Baseline*; false-positive rate is worsened in return for gaining better false-negative rate. Fortunately, when both feature sets are combined, false-positive rate is remained unchanged while the lowest false-negative rate is achieved. This suggests that the addition of stylistic features contributes to the enhancement of false-negative rate while not hurting false-positive rate (*i.e.* the cases where spam is classified as legitimate are significantly lessened).

In order to evaluate the contribution of different types of stylistic features, we conducted a series of experiments by removing features of a specific type at a time from *Combined*. Table 2 shows the detailed result. Notice that *LEN* and *SC* features are the most helpful, since the performance drops significantly after removing either of them. Interestingly, *FW* and *POS* features show similar contributions; we suggest that these two feature types have similar effects in this filtering task.

We also conducted another series of experiments, by adding one feature type at a time to *Baseline*. Table 3 reports the results. Notice that *LEN* features are consistently the most helpful. The most interesting result is that *POS* features continuously contributes the least. We carefully

[3]OSB refers to words separated by 3 or fewer words, along with an indicator of the difference in word positions; for example, the expression *"the quick brown fox"* would induce following OSB features: *"the* (0) *quick"*, *"the* (1) *brown"*, *"the* (2) *fox"*, *"quick* (0) *brown"*, *"quick* (1) *fox"*, and *"brown* (0) *fox"* (Cormack et al., 2007a).

[4]For detail on ROC analysis, see Cormack et al. (2007a).

[5]Available at http://plg.uwaterloo.ca/.trlynam/spamjig/

Feature set	1−AUC (%)	
Combined	3.7538	[3.1186 - 4.4802]
− LEN	**4.7351**†	[4.0457 - 5.6405]
− FW	3.9823†	[3.3048 - 4.5930]
− POS	4.0712†	[3.4057 - 4.8630]
− SC	**4.7644**†	[4.1012 - 5.4350]

Table 2: Performance by removing one stylistic feature set from the *Combined* set.

Feature set	1−AUC (%)	
Baseline	10.7227	[9.4476 - 12.1176]
+ LEN	**5.5275**†	[4.0457 - 6.6281]
+ FW	6.0828†	[5.1783 - 6.9249]
+ POS	9.6103†	[8.7190 - 11.0579]
+ SC	7.5288†	[6.6049 - 8.4466]

Table 3: Performance by adding one stylistic feature set to the *Baseline* set.

hypothesize that the result is due to high dependencies between *POS* and lexical features.

5 Discussion

In this paper, we have introduced new features that indicate the written style of texts for content-based mobile spam filtering. We have also shown that the stylistic features are potentially useful in improving the performance of mobile spam filters.

This is definitely a work in progress, and much more experimentation is required. Deep linguistic analysis-based stylistic features, such as context free grammar production frequencies (Gamon, 2004) and syntactic rewrite rules in an automatic parse (Baayen et al., 1996), that have already been successfully used in the stylometry literature may be considered. Perhaps most importantly, the method must be tested on various mobile spam data sets written in languages other than Korean. These would be our future work.

References

Shlomo Argamon and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of ACH/ALLC '05*.

Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-based text categorization: What newspaper am i reading? In *Proceedings of AAAI '98 Workshop on Text Categorization*, pages 1–4.

H. Baayen, H. van Halteren, and F. Tweedie. 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Gordon V. Cormack, José María Gómez Hidalgo, and Enrique Puertas Sánz. 2007a. Spam filtering for short messages. In *Proceedings of CIKM '07*, pages 313–320.

Gordon V. Cormack, José María Gómez Hidalgo, and Enrique Puertas Sánz. 2007b. Feature engineering for mobile (sms) spam filtering. In *Proceedings of SIGIR '07*, pages 871–872.

Michael Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of COLING '04*, page 611.

José María Gómez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sánz, and Francisco Carrero García. 2006. Content based sms spam filtering. In *Proceedings of DocEng '06*, pages 107–114.

David I. Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

Moshe Koppel, Shlomo Argamon, and Anat R. Shimoni. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of COLING '02*, pages 1–7.

T. C. Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214):237–246.

A. Q. Morton. 1965. The authorship of greek prose. *Journal of the Royal Statistical Society Series A (General)*, 128(2):169–233.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Marina Santini. 2004. A shallow approach to syntactic feature extraction for genre classification. In *Proceedings of CLUK Colloquium '04*.

Dae-Neung Sohn, Joong-Hwi Shin, Jung-Tae Lee, Seung-Wook Lee, and Hae-Chang Rim. 2008. Contents-based korean sms spam filtering using morpheme unit features. In *Proceedings of HCLT '08*, pages 194–199.

E. J. Tweedie, S. Singh, and D. I. Holmes. 1996. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30:1–10.

G. Udny Yule. 1939. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. *Biometrika*, 30(3-4):363–390.