A Corpus-based Study on Differences in Lexical Styles of Genuine and Spam SMS Messages

Vedant Ghavate

Abstract

This paper presents a lexical study on SMS messages being categorized as spam and genuine and attempts to differentiate between the two through lexical metrics. In the growing world of mobile phones, SMSs have reduced use but important functionality like one-time passwords still mark their significance. However, a major pain point in SMS is the human-like spam SMS messages that accumulate in the inbox. Through this study, we seek to discover the semantics of a spam SMS. The data is assembled through three different sources from the UK to University to Singapore. A corpus is created from this data and further analyzed using different metrics like TTR and ARI. The results of this study can help us establish the lexical quality of a spam SMS message. The comparison presented in this study can help us differentiate between spam and genuine SMS message.

1. Introduction

The text communication service component of phone, online, and mobile communication systems, known as Short Message Service (SMS), uses standardized communication protocols to enable the transmission of brief text messages between fixed lines and mobile phone devices. The International Telecommunication Union (ITU) estimates that as of 2006, the global market value of SMS was over 81 billion dollars.

The drawback is that an increasing number of marketers are employing text messages to target customers, making mobile phones the next target of electronic junk mail. Any unwanted text message sent to a mobile phone is referred to as spam SMS spam. Despite being uncommon in North America, this practice has been widespread in several regions of Asia.

The fundamental issue with SMS spam is that, in addition to being annoying, it may also be costly because some individuals pay for text message delivery. Additionally, software for spam filtering on mobile phones is not widely available. Another issue is the possibility of

critical, genuine messages being censored because they are deemed urgent. However, many carriers give their customers ways to reduce unwanted SMS texts.

Academic researchers working in this area are having issues, much like carriers that are coping with a large amount of SMS spam. For instance, a lack of legitimate, accessible datasets might make it difficult to compare various strategies.

Unlike email spam, which has a wide range of datasets at its disposal, mobile spam filtering currently still has a small number of corpora. Another issue is that because normal SMS transmission is restricted to 140 bytes, or 160 characters of the English alphabet, established email spam filters may perform significantly worse when used to deal with mobile spam.

Therefore, a separate analysis is required to deal with SMS Spam. Through this study, I have attempted to evaluate Spam SMS on various metrics against genuine SMS to understand if Spam SMS has a distinct lexical identity.

2. Methodologies

First, the dataset obtained was in a CSV format, where the first column described the categories ham or spam, and the second column contained the actual SMS text. To make use of the nltk.corpus.PlaintextCorpusReader() function, we need to create and store every row in a separate Txt file. The nltk.corpus.PlaintextCorpusReader() can only accept text as the parameter. So using CSV processing tools every SMS in the second file is stored in a separate txt file in a folder of the category to which it belongs (ham or spam). (In the dataset and throughout this paper, genuine texts are referred to as ‘Ham’).

Now, these text file folder is ready to be fed to the nltk.corpus.PlaintextCorpusReader() to create a corpus. After creating a corpus, we evaluate the two corpora on the same metrics and compare their results.

Lemmatizing and stemming have not been performed for this corpus as analysis ahead relieved that lemmatizing and stemming wipe out the key differences between the SPam and genuine Datasets.

3. Findings and Discussion

Count of Words and Sentences in SMS Corpus

	Ham	Spam
Number of Words	54056	146677
Number of Sentences	4825	747

The dataset is highly imbalanced as we have a limited number of Spam SMSs as compared to genuine (Ham) SMS. However, further analysis is scaled accordingly to balance the less volume of spam SMS.

MSTTR Comparision for Ham and Spam SMS

The lexical complexity of the corpus is evaluated by the mean-segmental token-type ratio (MSTTR). Claucitng MSTTR involves two steps, First Divide the corpus into non-overlapping standard-sized chunks. Finally, calculate TTR for each chunk and average the TTR for each chuck. Here, we have chosen 2000 as the chunk size.

	Ham	Spam
MSTTR	0.4645581395348838	0.3959

Here, we see a clear difference in the MSTTR values as genuine SMS seem to have a much higher MSTTR value. This is explained by the personalized variety of different individuals sending these texts. In the case of Spam SMS, the leXciality complexity observed is low as the messages are intended to be easy to read and most often written by bots and algorithms incapable of constructing lexically complex language.

Comparing Lexical Richness through the FK Readability Test

The Flesch Kincaid Grade Level is a widely used readability formula that assesses the approximate reading grade level of a text.

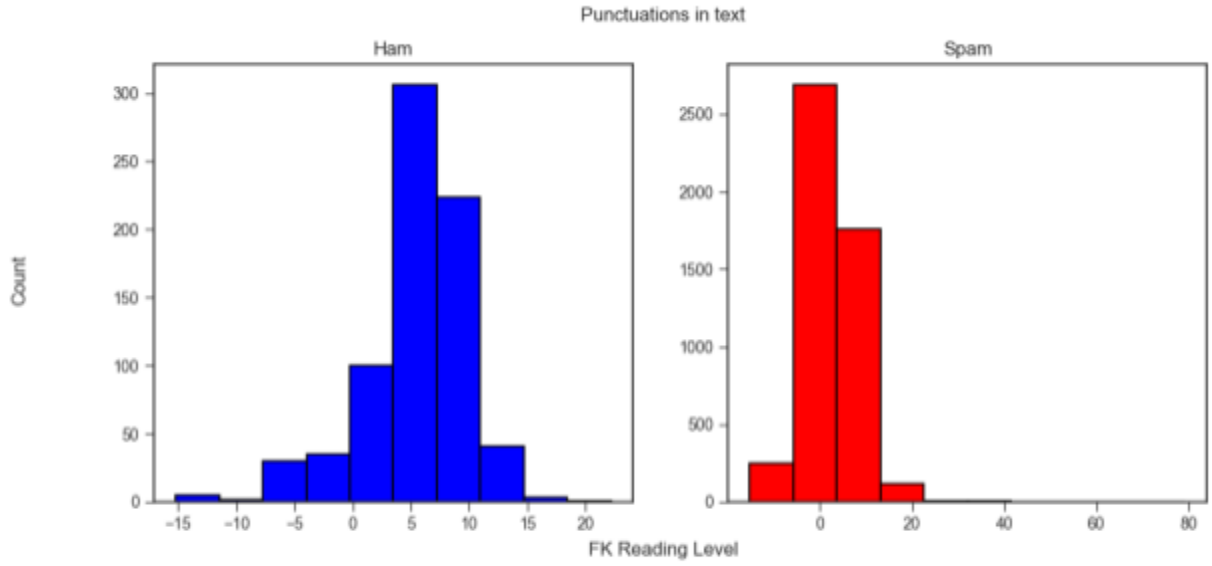


Figure : Histogram of FK Readability Level

The histogram above highlights the poor readability of genuine texts and the high readability of Spam SMS. This can be understood as spam SMS are carefully and intentionally made easy to read and easy to understand. While genuine SMS will involve a high level of personalized and customized (eg. two friends will talk to each other with their own personalized short form). Thus, the Spam SMS outperforms the Genuine SMS on the FK readability test.

Frequency Distribution

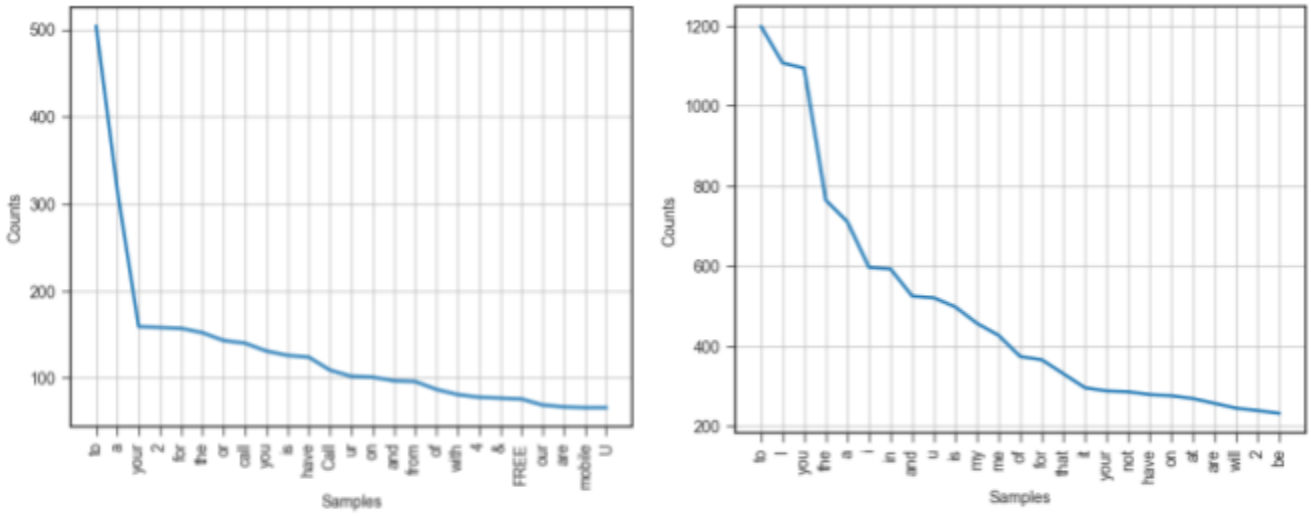


Figure : Sample vs Frequency Graph (Spam - *right*, Ham - *Left*)

The genuine texts have pronouns and basic prepositions are the most frequently occurring words while the spam texts show a high frequency of words like 'free' and 'call'. The intention of spam messages is clearly explained in this graph while the genuine texts show a generalized language.

Top 5 Ham Dataset		Top 5 Spam dataset	
i	559.96	call	411.73
my	173.61	free	336.49
me	119.73	claim	291.74
but	101.29	txt	271.09

Concordance

To find the difference in the use of words and to highlight the intent of the overall text, we check the use of the 'free' in both the corpus. This reveals that free is used to convey offers and deals in the case of Spam SMS while its use is very regular and mainly indicated the availability of an individual.

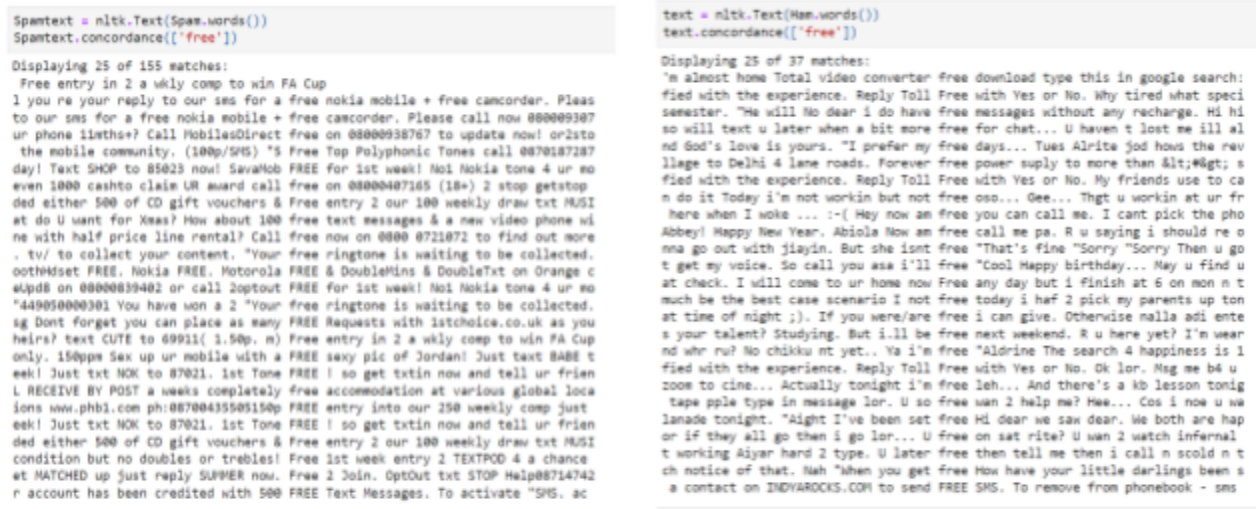
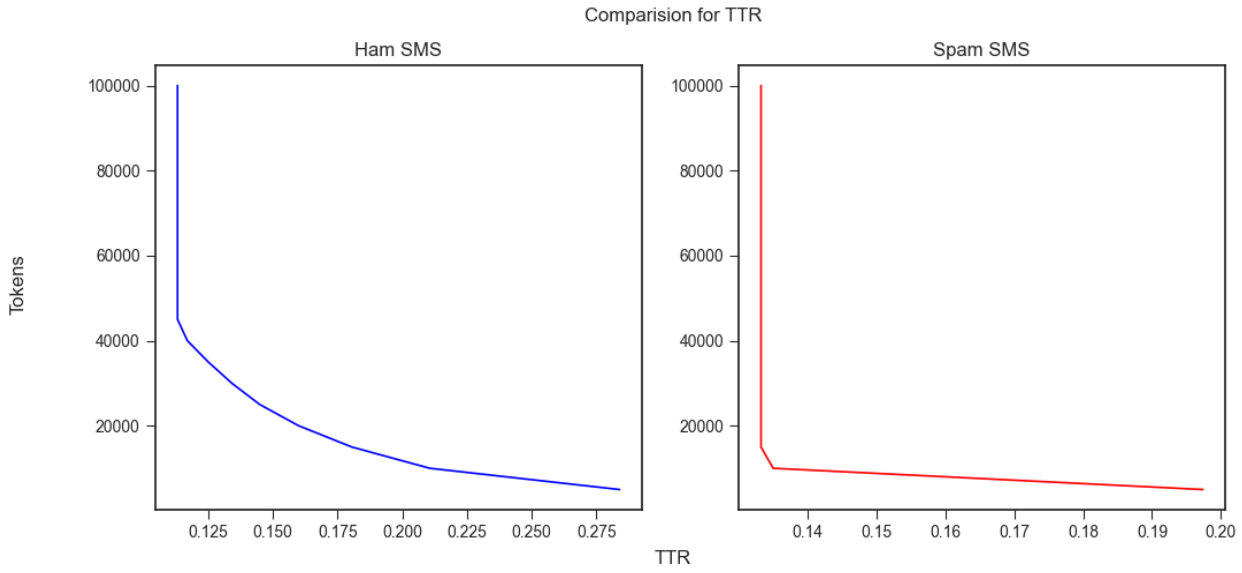


Figure : Concordance in KYIC view for the word 'free'

Comparing TTR of Spam and Ham SMS



s the English language w

Automated Readability Index

The automated readability index (ARI) is a readability test for English texts, designed to gauge the understandability of a text.

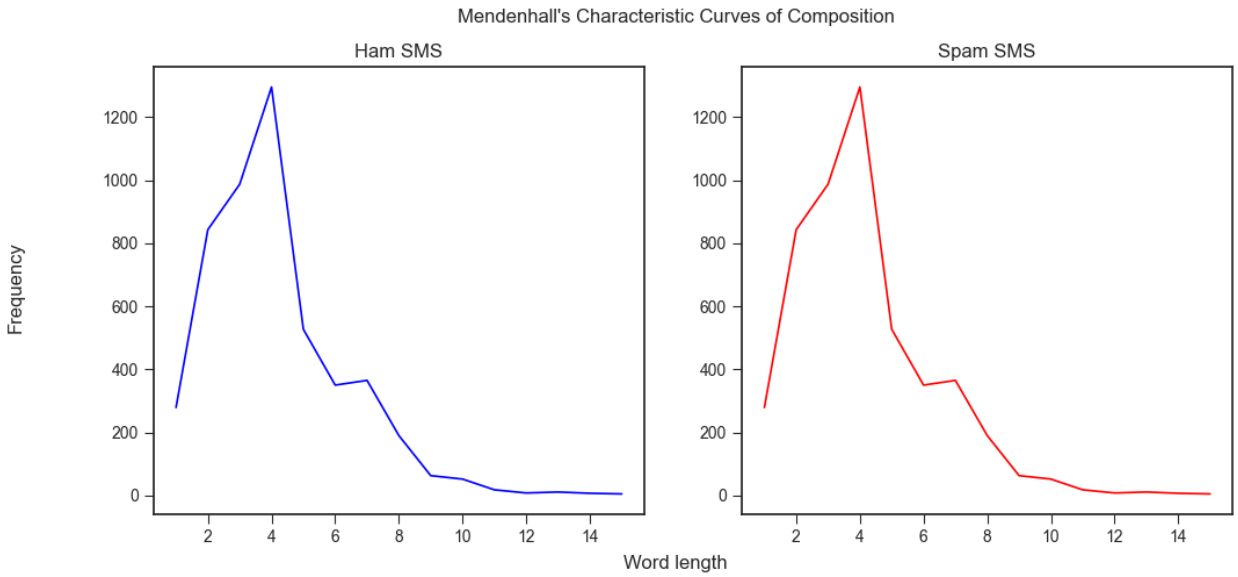
	Ham	Spam
ARI	3.2858620047450273	10.813959606065907

Table : ARI value for Ham and Spam SMS

From our analysis, the ARI value for Ham and Spam shows a stark difference and indicates that Spam is more readable than genuine text. This may hold true as spam texts are designed by companies to be better understood by a large audience and are meant to convey the message clearly. Also, genuine texts will lack the readability level as typing patterns are evolving and do not retain the same level of English language as before, along with this genuine texts will be personalized and will have readability that is meant to be understood by only the recipient.

Mendenhall's Characteristic Curves of Composition

Mendenhall's Characteristic Curves of Composition (1887) plots the overall distribution of word lengths as a signature of an author's style.



Plotting the mehedale's curve for both ham and spam, we can observe that the plot line are almost identical indicating that the linguistic style of ham and spam remains the same. Hence, Mehedale's curve cannot be considered a metric for differentiation between genuine and spam SMS messages.

Collocations

A Collocation is a series of words that occur more than the chance frequency of their occurrence.

```
[('have', 'won'),
 ('you', 'have'),
 ('please', 'call'),
 ('won', 'a'),
 ('call', 'from'),
 ('your', 'mobile'),
 ('this', 'is'),
 ('u', 'have'),
 ('to', 'contact'),
 ('call', 'now')]
```

```
[('are', 'you'),
 ('i', 'am'),
 ('going', 'to'),
 ('have', 'a'),
 ('want', 'to'),
 ('will', 'be'),
 ('i', 'will'),
 ('do', 'you'),
 ('a', 'great'),
 ('how', 'much')]
```

Figure 4: Spam (on the right) Ham (on the left)

From the collocations obtained for both Ham and Spam datasets, we can conclude that Spam SMS will have more collocations appearing like 'call', 'contact', 'won', and 'mobile' which indicated the SMS is regarding a lottery or is the asking to call back. In the case of genuine SMS, we observe, the top collocations have pronouns, this is due to the high level of personalization that genuine texts tend to have.

4. Conclusion

In conclusion, after applying multiple metrics of linguistic analysis, we observe a series of metrics showing a significant difference between genuine and spam text.

This establishes our initial thought that spam SMS messages do have a distinct lexical identity. The Spam SMS consists of specific collocations and keywords, thus having a particular form and lexical style. The overall intent of the Spam SMS is very obvious and is explained very clearly through the metrics above.

Further analysis will reveal more insight into the corpus and models can be trained accordingly using this analysis to build a SPam SMS classifier. The relevant metric discovered through this study can be used as a data point to train the model. However, a larger dataset would present a much more comprehensive analysis.

References

Almeida, T.A., Hidalgo, J.M., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: new collection and results. ACM Symposium on Document Engineering.

G. V. Cormack, J. M. Gómez Hidalgo, and E. Puertas Sanz. Spam Filtering for Short Messages. In Proc. of the 16th ACM CIKM, pages 313--320, Lisbon, Portugal, 2007.

M. Gómez Hidalgo. Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization. In Proc. of the 17th ACM SAC, pages 615--620, Madrid, Spain, 2002. Google ScholarDigital Library

L. Zhang, J. Zhu, and T. Yao. An Evaluation of Statistical Spam Filtering Techniques. ACM TALIP, 3(4):243--269, 2004. Google ScholarDigital Library

http://www.grumbletext.co.uk/

http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/SMSCorpus/

http://www.esp.uem.es/jmgomez/SMSspamcorpus/

http://www.esp.uem.es/jmgomez/SMSspamcorpus/