Intelligent Spam Filtration Based on SMS Linguistics

Richard Mfitumukiza
Introduction to Machine Learning
Vector Institute
mfitumukiza@protonmail.com

Abstract

The amount of Short Message Service (SMS) spam is increasing. According to (Joe, I., & Shim, H. 2010) people classify SMS spam as violating personal privacy (21.3%), a wast of time (24.8%) and as annoying (32.3%). Current solutions like email spam filtration offer unsatisfactory performance (Almeida, Hidalgo, J. M., & Yamakami, A.; 2011). Email spam filters fails because they are context-based filtration applied on SMS messages that offer very little context. Thus I propose a new approach; SMS spam filtration that is based on unique linguistic characteristic of SMS messages. I parsed UCI SMS spam collection to extract features such as punctuation, usage of symbols and unique emotional expressions. Results shows a high accuracy and precision of spam filtration based on SMS linguistic features; consequently the usual Natural Language Processing (NLP) data cleansing is detrimental when applied to SMS messaging.

Introduction

For the last 4 decades of mobile phone development, we have witnessed one of the greatest technology penetration; it is estimated to be about 91% adopted globally (Deloitte, 2017). This adoption gave rise to SMS, which is evolving into Instant Messaging (IM).

SMS spams have been on the rise as well, generally what qualifies as SMS spam are unsolicited SMS, often of commercial nature and sent in bulk indiscriminately to multiple recipients. These SMS spams proved to be an increasing threat, that needs to be addressed. This problem is closely related to Cyber security, even thought out of scope for this paper, but proposed approach would be beneficial, based on obtained results.

In general SMS spam problem is insufficiently handled, due to multiple factors. The main factors are legal , academic and technical factors (Almeida, Hidalgo, J. M., & Yamakami, A.; 2011). In this paper I will focus on technical factors, specifically SMS spam filtering software.

One of the reason it is not sufficiently resolved, is because it is not perfectly compatible with existing solutions. The main solution available, that is closely related to SMS spams is email spam filtration. The main reasons that email spam filtration doesn’t work well as SMS spam filtration are on multiple levels.

On linguistic level, SMS messages are characteristically different from email messages. The main differences are observed on message length, frequency, punctuation and so on.

On academic level, for very long time, the progress was stalled by lack of data. There were no significant public data, ready and available for researchers to work on and address the problem. This had domino effect, where lack of data led to lack of research which led to lack of extensive development of SMS spam filtration systems. This changed in 2011, when a collection of over 5500 SMS messages was published for research purposes (Almeida, Hidalgo, J. M., & Yamakami, A.; 2011).

This SMS collection presented opportunity for multiple research. In this paper I intend to use the collection for the purpose of building an SMS spam filtration system. In particular I want to explore the usage of machine learning models, to address unique linguistic characteristics of SMS messages.

Problem definition

SMS pricing is getting cheaper and cheaper, in most cases, it already got to zero charges. This encourages a large volume of SMS spams, which overwhelms the traditional capacity of telecommunication service providers to handle SMS spams. Telecommunication service providers mostly offer to block all unsolicited SMS messages without checking if the message was actually a spam or not. This is obviously problematic.

Email spam filters are not effective on SMS messages and generally demonstrate a significant drop in performance when applied to SMS messages. This is mainly due to differences between email and SMS messages. Emails are relatively longer, less frequent and more formal than SMS messages. Consequently context-based email spam filters don't perform well when applied to SMS messages; which are relatively shorter, more frequent and less formal than emails.

Since the release of the UCI SMS spam collection dataset in 2011, It became easier to do research for SMS spam filters. UCI SMS spam collection is a dataset of about 5500 SMS messages, It was collected across multiple sources and across different geographical locations. It was collected with consent and with knowledge of research usage (Almeida, Hidalgo, J. M., & Yamakami, A.; 2011).

In this research I would like to explore the effect of a machine learning based SMS spam filter that takes into consideration unique linguistic characteristics of SMS. My hypothesis is that SMS spam filtration will be greatly improved in terms of accuracy, precision and recall metrics.

My approach will consist of parsing SMS messages to extract features that are potentially unique to SMS messages, like quantity of punctuation, quantity of capitalization, length of message, usage of symbols and so forth.

A secondary impact I hope to get from this paper is to demonstrate the need for a new text corpus that is specific to SMS and IM messaging. This will need cooperation from across academia, mainly from linguistics.

In addition to aforementioned feature extraction, I will employ an array of different tools that have demonstrated great performance in Natural Language Processing (NLP) like text vectorization and ML algorithms.

Among popular text vectorization; I would explore the effectiveness of Term Frequency Inverse Document Frequency (TF-IDF) because it captures both the frequency of a word in one document and the frequency of the same word across multiple documents, and it condense it down to a ratio, one number. TF-IDF vectorization will be contrasted with other feature extracted from the messages.

After extracting features and vectorizing the messages, The resulting data will be used to create a classification model. This model will be based on contrasting different ML algorithms like support vector machine, random forest and gradient boosting.

Data exploration and description

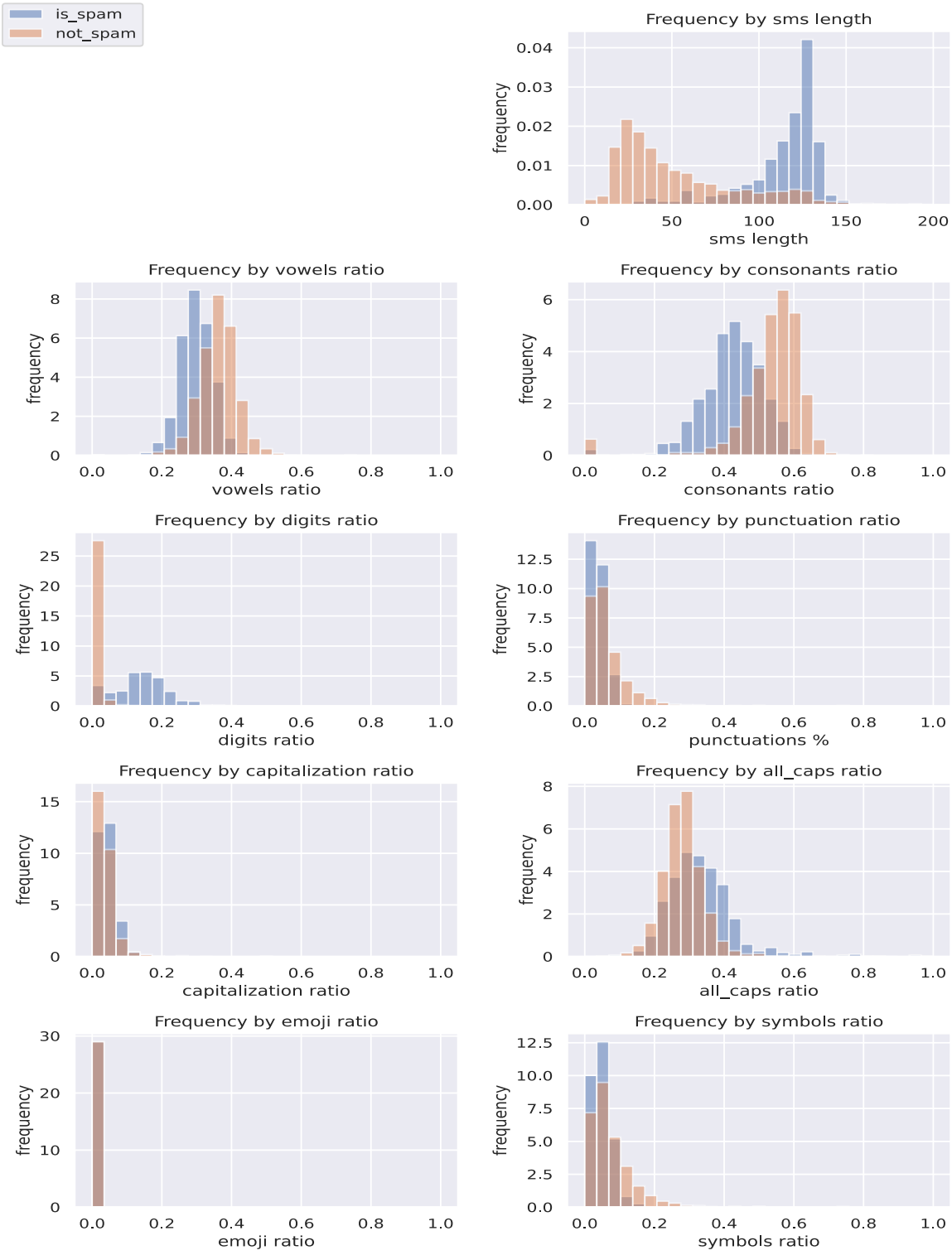
As aforementioned this collection consist of 5576 rows of individual messages and 2 columns one for messages and the second for type of message, which can be spam or not.

After initial clean up and exploration the following features were extracted from the text:

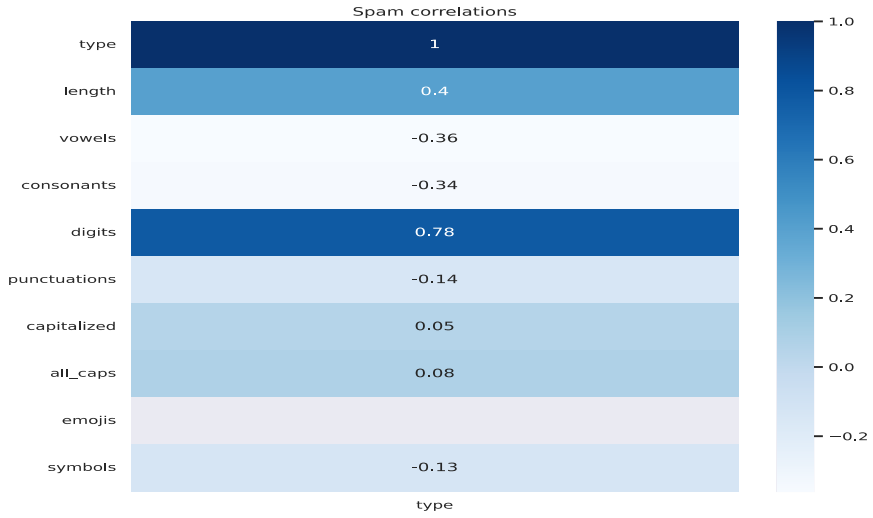
- length of the message (char)
- ratio of vowels to length of the message
- ratio of consonants to length of the message
- ratio of digits to length of the message
- ratio of punctuation to length of the message
- ratio of capitalized word to length of the message
- ratio of all-capital words to length of the message
- ratio of symbols to length of the message

These features are based on unique characteristics of SMS messages; I will refer to them as sms-linguistics features or simply linguistic features.

The following figure shows distribution of spam and non-spam messages, depending on each of feature among the linguistic features. Some features show a clear distinction for spam messages.


is_spam
not_spam
Frequency by sms length
frequency
sms length
Frequency by vowels ratio
vowels ratio
Frequency by consonants ratio
consonants ratio
Frequency by digits ratio
digits ratio
Frequency by punctuation ratio
punctuations %
Frequency by capitalization ratio
capitalization ratio
Frequency by all_caps ratio
all_caps ratio
Frequency by emoji ratio
emoji ratio
Frequency by symbols ratio
symbols ratio

The following figure shows correlation of spam messages with the linguistic features. Top features to show a strong correlation is the number of digits (relative to the length of the message). Followed by actual length of the messages. Ratio of vowels and consonants also plays a discernible role is distinguishing spam messages.



SMS messages were vectorized. The vectorization method used is Term Frequency – Inverse Document Frequency. This was preferred over vectorization such as count vector or N-gram because they only capture the importance of a word as it relates to one message (Giannakopoulos, G., & Karkaletsis, V. 2009). TF-IDF was preferred over vectorization such as Word2vec and Doc2vec because they heavily rely on context in the documents, which SMS messages lack (Wang, Q., Han, X., & Wang, X. 2009).

Stemming and lemmas were not used before vectorization, because abbreviated and SMS messages would loose a big part of their components. This was found to be the case according to Cornack in his review (G. Cormack, 2008).

Both the linguistic features and vectorized messages were combined into one big matrix. The combined matrix extended to a total of 7009 features.

Finally, the total collection of SMS were divided into training 60%, validation 20% and testing 20% subsets. Which was used on each model.

Model description

A total of 3 models were used: support vector classifier (SVM), random forest and gradient boosting. Support vector machine was chosen because it is established as a benchmark for this dataset, it was found to be the best model in the initial research as per Almeida (Almeida, T.A., et al, 2011). Random forest was chosen to compare the effectiveness of bootstrapping and aggregating on this dataset. Gradient boosting is used to contrast boosting with aforementioned algorithms.

A grid search was used to optimize and tune hyper-parameters for each algorithm. SVM was optimized on: kernel and c value. Random forest was optimized on: number of estimators and maximum depth. Gradient boosting was optimized on: number of estimators and learning rate. Each grid search was cross validated on a five folds.

In total I evaluated 3 types of data extraction: linguistic features, TF-IDF and both combined; across 3 models : SVM, Random forest and gradient boosting. That makes a total of 9 evaluations. On each evaluation 3 metrics were taken to compare their performances. The 3 metrics used are : accuracy, precision and recall scores.

Accuracy informs us of a ratio that was predicted correctly; precision informs us of a ratio that was predicted to be spam messages which were truly spam messages and recall informs us of the ratio that was predicted to be spam messages out of all actual spam messages.

The main focus is on a high accuracy to make sure that the model is accurate as often as possible. Precision will be prioritized over recall; because we want our model to label a message as spam only if it is truly a spam message.

Results and findings

This table shows all results from the 9 evaluations.

Extraction type	Algorithm	Accuracy	Precision	Recall
SMS linguistics	Support Vector Machine	**0.98**	0.99	0.82
	Random Forest	0.97	0.95	0.88
	Gradient Boosting	0.97	0.9	0.89
TF-IDF	Support Vector Machine	0.89	**1**	0.28
	Random Forest	0.87	**1**	0.13
	Gradient Boosting	0.96	0.99	0.84
Combined	Support Vector Machine	0.93	0.98	0.52
	Random Forest	0.85	0	0
	Gradient Boosting	0.96	0.89	**0.9**

Considering all 3 extraction methods: linguistic features, TF-IDF and their combination. Linguistic features gave the best accuracy at 98%, TF-IDF gave the best precision at 100% and combined gave the best recall score at 90%.

Considering all 3 models: SVM, RandomForest and GradientBoosting. With a trade-off that favours accuracy and precision; SVM gave the best results when applied to linguistic features. Random forest gave the worst results especially when applied to combined data. Gradient boosting gave the best recall score but at the expense of a loss in either accuracy or precision.

Conclusion

It became very clear that extracting SMS unique characteristics is as important and effective as other well known NLP methods. It also demonstrated that the usual NLP cleaning such as dropping punctuation and transforming all letters to lowercase is a loss of important features. I therefor propose that a new corpus is needed; a corpus that takes into consideration the unique linguistic features of SMS and IM messages.

References

Deloitte (2017) Global mobile consumer trends: Second edition

Almeida, Hidalgo, J. M., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: new collection and results. Proceedings of the 11th ACM Symposium on Document Engineering, 259–262. https://doi.org/10.1145/2034691.2034742

Taufiq Nuruzzaman, Lee, C., Abdullah, M. F. A. bin, & Choi, D. (2012). Simple SMS spam filtering on independent mobile phone. Security and Communication Networks, 5(10), 1209–1220. https://doi.org/10.1002/sec.577

G. Cormack. Email Spam Filtering: A Systematic Review. Foundations and Trends in Information
Retrieval, 1(4):335–455, 2008.

Wang, Q., Han, X., & Wang, X. (2009, September). Studying of classifying junk messages based on the data mining. In *2009 International Conference on Management and Service Science* (pp. 1-4). IEEE.

Orkphol, K., & Yang, W. (2019). Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet*, *11*(5), 114.

Joe, I., & Shim, H. (2010, December). An SMS spam filtering system using support vector machine. In *International Conference on Future Generation Information Technology* (pp. 577-584). Springer, Berlin, Heidelberg.