# Detecting SMS Spam in the Age of Legitimate Bulk Messaging

Bradley Reaves, Logan Blue, Dave Tian, Patrick Traynor, Kevin R. B. Butler
{reaves, bluel, daveti}@ufl.edu {traynor, butler}@cise.ufl.edu

Florida Institute for Cybersecurity Research
University of Florida
Gainesville, Florida

## ABSTRACT

Text messaging is used by more people around the world than any other communications technology. As such, it presents a desirable medium for spammers. While this problem has been studied by many researchers over the years, the recent increase in legitimate bulk traffic (e.g., account verification, 2FA, etc.) has dramatically changed the mix of traffic seen in this space, reducing the effectiveness of previous spam classification efforts. This paper demonstrates the performance degradation of those detectors when used on a large-scale corpus of text messages containing both bulk and spam messages. Against our labeled dataset of text messages collected over 14 months, the precision and recall of past classifiers fall to 23.8% and 61.3% respectively. However, using our classification techniques and labeled clusters, precision and recall rise to 100% and 96.8%. We not only show that our collected dataset helps to correct many of the overtraining errors seen in previous studies, but also present insights into a number of current SMS spam campaigns.

## 1. INTRODUCTION

Text messaging has been one of the greatest drivers of subscriptions for mobile phones. From the simplest clamshells to modern smart phones, virtually every cellular-capable device supports SMS. Unsurprisingly, these systems have been targeted extensively by spammers. The research community has, in turn, responded with a range of filtering mechanisms. However, this ecosystem and the messages it carries have changed dramatically in the past few years.

The most significant change in this ecosystem is the widespread interconnection with non-cellular services. Specifically, a wide range of web applications now use text messaging to interact with their customers. From second factor authentication (2FA) to account activation, the volume of legitimate messages with very little variation in their content is on the rise [2]. While a critical part of overall security for users, this shift in the makeup of traffic is having a major impact on the efficacy of SMS spam filtering. Because legitimate bulk messages have characteristics similar to spam, including the ubiquity of a number (like a short code or one-time password) or a URL, as well as a call to action ("click here"), we hypothesize that SMS spam filters will need to change to account for a new messaging paradigm.

In this paper, we leverage a dataset of nearly 400,000 messages collected over the course of 14 months. We obtain such data by crawling public SMS gateways. Users rely on these public gateways to receive legitimate SMS verification messages as well as to avoid having their actual phone numbers exposed to lists that receive spam. We rely on this data to make the following contributions:

- **Release Largest Public Dataset:** We release a labeled dataset of bulk messaging and SMS spam, which is larger than any previously published spam dataset by nearly an order of magnitude.
- **Weaknesses in Previous Datasets:** We show that existing SMS spam/ham corpora do not sufficiently reflect the prevalence of bulk messages in modern SMS communications, preventing effective SMS spam detection. Specifically, we demonstrate that previously proposed mechanisms trained on such datasets exhibit extremely poor results (e.g., 23% recall) in the presence of such messages.
- **Characterization of SMS Spam Campaign:** We provide deeper insight into ongoing SMS spam campaigns, including both topic and network analysis. We find that the number of messages sent in a campaign is best explained by the volume of sending numbers available to the campaign.

## 2. RELATED WORK

Text messaging has become the subject of a wide range of security research. For instance, many services now rely on SMS for the delivery of authentication tokens for use in 2FA systems [1,5,9,23]. Recent work has demonstrated that many such systems are vulnerable to attack for a range of reasons including poor entropy [12,25] or susceptibility to interception [16]. Text messaging has also been analyzed as the cause of significant denial of service attacks [17,28–30] and a medium for emergency alerts [27].

SMS spam has received significant attention from the community. Researchers have developed a range of techniques for detecting such spam, with significant focus on message content [4,6,8,10,13,21,22,26,31,33]. This class of mitigation has by far been the most popular in the research commu-

*WiSec'16, July 18–20, 2016, Darmstadt, Germany.*

ISBN 978-1-4503-4270-4/16/07...$15.00

DOI: http://dx.doi.org/10.1145/2939918.2939937

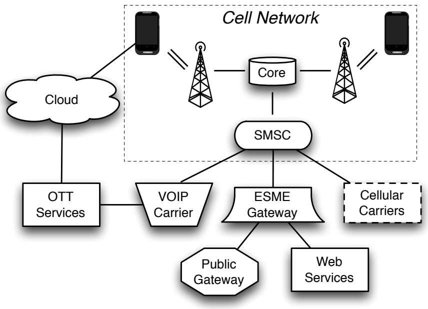


Figure 1: A high-level overview of the SMS ecosystem.

nity as collecting SMS spam can be done without special access to carrier-level data. The research community has relied almost exclusively on publicly available datasets, like those made available by Chen and Kan [7] or Almeida et al. [4]. Unfortunately, these datasets are quite limited, with only a few hundred actual spam messages. Other efforts have instead focused on network behaviors, such as volumes, sources and destinations [11,14,15,18–20,32]. Unfortunately, this latter class of analysis is generally limited to network providers, making independent validation difficult.

## 3. BACKGROUND

Text messaging within the traditional closed telephony ecosystem works as follows: a user generates a message on their phone and transmits it to their local base station, which delivers the SMS to the Short Messaging Service Center (SMSC). With the aid of other nodes in the network, the SMSC forwards the SMS to its destination for delivery.

Modern telephony networks accept text messages from a far larger set of sources. In addition to the SMSC receiving text messages from users served by other cellular providers, many VoIP providers (e.g., Vonage, Google Voice) also allow their users to send text messages. Messaging apps transported by Over the Top (OTT) connections now deliver messages via the public Internet. Lastly, a wider range of External Short Messaging Entities (ESMEs) such as web services used for two-factor authentication (e.g., Google Authenticator, Duo Security). Within this class also lies entities known as Public Gateways. These public websites allow anyone to *receive* a text message online by publishing telephone numbers that can receive text messages, and posting such messages to the web when they are received. These services are completely open — they require no registration or login, and it is clear to all users that any message sent to the gateway is publicly available.

It is through these Public Gateways that we are able to conduct our measurement study. Because these interfaces publish text messages for destinations that span a range of providers and continents, our work provides the first global picture into SMS spam (especially that which bypasses the spam filters of providers).

## 4. DATA CHARACTERIZATION

This paper makes use of several previously compiled datasets. First, we use two existing SMS spam and ham corpora. We use a spam corpus compiled by Almeida and Hidalgo [4] that contains 747 messages. For legitimate messages, we use a corpus of 55,835 messages collected by Chen and Kan [7] from submissions of personal text messages from volunteers. We refer to these two corpora as the "public corpus." To the best of our knowledge, these messages are the largest publicly available collection of SMS ham and spam.

Many of the insights of this paper are made possible by a collection of SMS from another source: public SMS gateways. Public SMS gateways are websites that purchase a public phone number and post all text messages received by that number to a public website visible to anyone. These websites claim to exist for various reasons, including to avoid SMS spam by not revealing a user's true phone number, but the majority of messages (over 67.6%) received by these gateways consist of account verification requests or one-time passwords (i.e., legitimate bulk SMS). This means that the message type distribution of our data may not be representative of messages seen by a traditional mobile carrier. Even though this data may have fewer personal messages than typical, it is still a valuable data source for understanding the effects of bulk messaging on SMS spam classification. These gateways provide complete message content, sender and receiver numbers, and the time of message. The message data that we use was collected by scraping these websites, resulting in a dataset of 386,327 messages sent to over 400 numbers in 28 countries over a period of 14 months. Many of these messages are duplicates, or are syntactically or semantically identical (e.g., "Hello Alice" and "Hello Bob").

In a prior study [25], this data was grouped by ordering messages lexically and identifying boundaries where Levenshtein distance fell below 90%. The largest of these groups were manually labeled to identify message intent, including indicating if a message appeared to be unsolicited bulk advertising (i.e., spam). Only 1.0% of this labeled data consisted of spam messages. Note that messages sent by individuals are systematically excluded from analysis because they are not self-similar and do not form large groups.

For our experiments, we carved the gateway data into two distinct datasets. The first was one message from every labeled group (called "labeled gateway data"). This dataset is intended to train a machine learning classifier, and accordingly overwhelmingly similar messages are removed to avoid overfitting the classifier. This dataset consists of 754 messages, including 31 (4.1%) spam messages. The second dataset was all messages that were previously unlabeled, called the "unlabeled gateway data". This dataset consists of 99,363 messages of an unknown mixture of personal messages, legitimate bulk messages, and spam.

We have released both the labeled gateway training data and confirmed spam discovered in the unlabeled gateway dataset (details provided in subsequent sections). This dataset contains 1316 unique bulk messaging ham messages and 5673 spam messages. It is available at http://www.sms-analysis.org.

**Ethical Considerations** We note that there are ethical questions that must be considered in collecting this data. First, the data is publicly available, and therefore under United States regulations an institutional review board does not need to oversee experiments that collect or use this data. Furthermore, we note that users who expect to receive messages at these messages are aware that they will be publicly available, and accordingly must reasonably have low privacy expectations. However, senders of messages may not be aware that these messages will be public. Because of this, we seek to focus our use of this data on bulk messaging, where message content is unlikely to be confidential to either

Table 1: Classifier Performance

| Training | P | P | P + LGW | P + LGW |
|---|---|---|---|---|
| Testing | P | LGW | LGW | UGW |
| Precision | 94.1% | 23.8% | 100% | 84.6% |
| Recall | 88.8% | 61.3% | 96.8% | — |
| FP | 0.1% | 8.1% | 0.0% | 1.3% |
| FN | 0.1% | 1.6% | 0.1% | — |

Key: P — Public Corpora, LGW — Labeled Gateway Data, UGW — Unlabeled Gateway Data

the sender or recipient. Our methods are designed so that we systematically exclude messages between individuals, and in the event that any personally identifiable information (PII) is disclosed, we do not further analyze, extract, or make use of that information in any way. We note that any PII in this data was *already publicly leaked* before we collect and analyze it, so our use of this data does not further damage any individual's privacy. Finally, our corpora have been scrubbed of personally identifiable information by replacing sensitive information with fixed constants. We replaced every instance of names, physical addresses, email addresses, phone numbers, dates/times, usernames, passwords, and URLs that contain potentially unique paths or parameters. Every released message was examined by two researchers.

# 5. EVALUATING SMS SPAM CLASSIFIERS

As discussed in earlier sections, prior SMS spam corpora were collected by researchers who solicit volunteers to provide examples of SMS spam or legitimate messages. We believe that these corpora, under which the bulk of SMS spam research has been conducted, are fundamentally limited. For example, SMS has increasingly become a means of contact for many online services to provide information to users and to provide security related services like two-factor authentication. However, the existing corpora for SMS spam research do not account for such messages. Accordingly, we hypothesize that existing SMS spam detection research based on the corpora available will fail to accurately classify legitimate messages as benign.

We designed several experiments to test this hypothesis. The following subsections detail these experiments and their findings. Existing literature on machine learning for content-based SMS spam classification has exhaustively examined choices of machine learning algorithm [8] and feature selection [26], finding that while there is an optimal-accuracy design, other choices lead to only minor degradations in performance. We then implement and evaluate this classifier against gateway data to evaluate the effect of the spam corpus on the detection of SMS spam in the face of legitimate bulk messaging. *Our aim in doing so is demonstrate the impact on spam classification of changes in legitimate SMS messaging, not to establish an empirically optimal classifier.*

We conclude by retraining and applying this classifier to identify SMS spam in unlabeled gateway data.

## 5.1 Classifier Selection and Implementation

To evaluate the question of how bulk SMS would be classified, we needed to implement an SMS spam classifier. After reviewing the literature, we found that the best performing classifiers (taking into account accuracy, precision, and recall on cross-validated evaluation) use a support vector machine (SVM) [8]. SVM classifiers permit the use of kernels that allow an expansion of input data into a higher-dimensional space to improve classification performance. The kernels used in prior work were unspecified, so we use a linear kernel as it is the simplest possible kernel. We confirmed this provided the best performance compared to other kernels, but omit a full analysis for space reasons. Regarding features, prior work has investigated a naive binary bag-of-words model, using only counts of keywords common in spam, n-grams, and more complicated feature sets. Prior work found that a simple binary vector indicating the presence of a word in the message performed best [26], so we also use this approach. Like Tan et al. [26], we preprocess the data to remove features that could induce classification on non-semantically meaningful features, including making all words lower case and replacing all URLs, email addresses, stand-alone numbers, and English days of the week with a fixed string. As in prior work, we do not remove stop words[1] from the feature vector. We use the scikit-learn Python library [24] for feature analysis and classifier implementation. Several other classifiers were evaluated using a variety of feature selection techniques. We found that results were consistent with those found in prior work, and omit further discussion for space.

With this classifier implemented, we train the classifier and evaluate its performance on the existing public corpora, then train and test the classifier using 5-fold cross validation to ensure consistency with previous work. The vocabulary in this dataset results in a feature vector with 39,558 words. After training, we see an overall accuracy of 99.8%. Precision (a measure of how many messages identified as spam are actually spam) was 94.1%, while recall (a measure of how much spam was correctly identified) was 88.8%. These results are consistent with the findings of Tan et al. [26], who found an F1 score of 93.6%, comparable to our classifier's F1 score of 91.4%. In summary, the classifier performance seems quite good.

## 5.2 Evaluating Classifier with Training Data

Having trained and validated a classifier, we can test our hypothesis that the classifier will fail to properly categorize legitimate bulk SMS messages, instead labeling it as spam. After classifying the data, we find that the classifier's performance significantly declines, confirming our hypothesis. Precision falls from 94.1% to 23.8%. Recall also declined from 88.8% to 61.3%. The practical impact of this classifier's poor performance on the user is best reflected by the overall false positive rate. In total, 8.1% of legitimate bulk messages would be miscategorized by the classifier, providing a frustrating user experience. In particular, dropping account verification messages will make new services inaccessible, and dropping SMS authentication messages would make services effectively unavailable for users.

To understand these results, we investigated the feature weights learned by our classifier. Feature weights indicate the relative importance of a particular feature in determining if a message is spam; positive weights indicate that a feature is indicative of spam, while weights close to 0 do not strongly indicate spam or ham. For example, the feature indicating the presence of a number has a weight of 0.637, while the word "rain" has a weight of -0.628. This indicates that the presence of a number (like a phone number or a price) is a strong indicator of "spamminess."

[1] Stop words are extremely common words, like "the", "and", etc., often removed during natural language data analysis.

To better understand our false positives, we examined the weights of the 20 most frequent words in our false positives. We find that words that are prevalent in legitimate bulk SMS like “code” or “verify” have weights with low absolute value (0.046 and 0.000 for these words). The words that are frequently used in these messages have weights that contribute almost nothing to the decision of the classifier.

As a result, the following message from the GW dataset is mislabeled as spam due to the effect of large positive weights provided by the features “has number” and “has URL.”

```
WhatsApp code 351-852.  You can also tap on
this link to verify your phone:
v.whatsapp.com/351852
```

## 5.3 Evaluating Classifier on Labeled Data

Machine learning classifier performance is governed by many factors regarding model selection; however, experience shows that small datasets are often a bottleneck for classifier performance [3]. We hypothesized that better data, not a better model, was required to rectify the performance issues we found. To test this hypothesis, we retrain the classifier mentioned above to include the labeled gateway messages.

After running a cross validation analysis, we find that classifier performance increases to numbers comparable or better than those in the first experiment. We see an overall accuracy of 99.9%, with precision and recall of 100% and 96.8%. It is thus possible to distinguish legitimate and unsolicited bulk messages, at least in a cross-validation setting.

We again examined the feature weights of our messages, and we found that the features like “code” and “verify” have acquired strong weights: -0.402 and -0.706 respectively. This shows that the public corpus fails to provide enough data samples to fully cover the domain of legitimate messages, but this can be rectified using gateway data.

## 5.4 Evaluating classifier on unlabeled data

While cross validation is a standard technique for evaluating a classifier given a finite data set, it loses predictive value compared to using a true testing data set. To further evaluate our new retrained-classifier, we apply it to 99,363 unlabeled gateway messages. Because our gateway labeling data focused on messages that were highly similar or repeated to a high degree, we felt confident that there was spam in the unlabeled data as well.

To evaluate the new retrained classifier, we classified these messages, finding 8179 messages of unlabeled gateway data (8.2%) labeled as spam by the classifier. However, this does not tell us how many messages are legitimate bulk messages (i.e. false positives) and many are actually unsolicited. To answer this question, we manually label the messages marked as spam by our classifier.

Fortunately, many of these messages are similar in content, so they can be grouped together to label them. To facilitate clustering, we describe each message using a common technique in text data known as latent semantic analysis (LSA). LSA describes high dimensional text data as a low-dimensional feature vector that groups semantically similar messages together. LSA computes a term frequency – inverse document frequency matrix of the corpus, then applies a singular value decomposition to select the most important singular vectors, reducing the document space. We then cluster documents using the DBSCAN clustering algorithm. DBSCAN identifies clusters by specifying a minimum cluster density and finding elements that form regions with density greater than the threshold. Unlike k-means, it does not make assumptions about cluster shape, or the number of clusters. After clustering, we identified 475 clusters of spam in the gateway data. We evaluate the effectiveness of our clustering algorithm by computing the average silhouette score of each message. Briefly, this score indicates the similarity of objects within each cluster (as opposed to a neighboring cluster), and our score of 0.644 indicates a good clustering structure. We characterize these clusters in more detail in the following section.

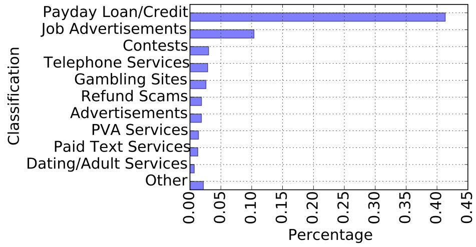


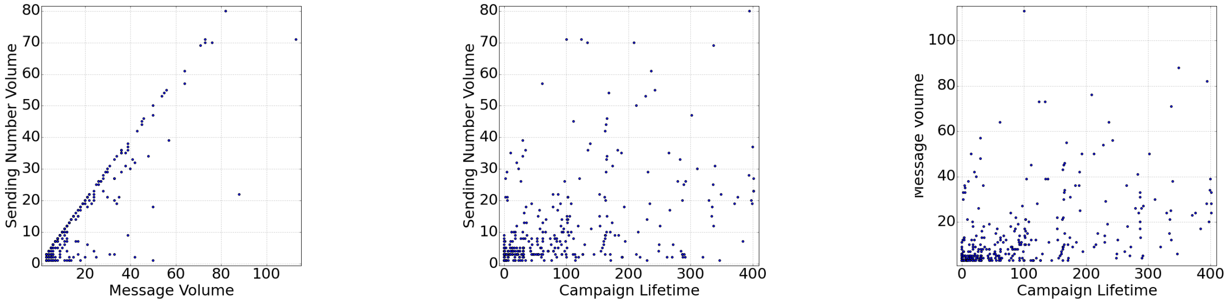
Figure 2: The top spam categories in gateway data

We then manually labeled these clusters for topic (e.g., pharma, payday loans, etc.) and whether the messages were actually spam (e.g., false positives). Unfortunately, determining if a message is solicited is not a perfect science, and there are some limitations to this approach. First and foremost, a message sent to some users may be solicited while the same message sent to others could be unwanted by others. Furthermore, we were not the intended recipients of these messages, and in some messages context is not always available to us when labeling. In situations where doubt was warranted, we erred on the side of assuming a message was indeed solicited (i.e. not spam). For example, we labeled any message as “not spam” if it seemed to be the response to a user inquiry or if it seemed to be part of an exchange in which a user could have prompted the message. Therefore, we believe that our reported results are conservative. Second, we ignore messages that were not clustered, so ground truth is unavailable for 13.1% of messages labeled as spam. Additionally, we did not have the resources to examine messages that were not classified as spam. Therefore, we cannot definitively measure recall or false negatives.

With labeled classification results, we can evaluate the performance of a SMS spam classifier trained with awareness of legitimate bulk messages. We found in total that 1261 messages appeared to be messages that could have been legitimate bulk messages. This corresponds to a corresponding precision of 84.6% – a substantial increase over the expected 23.8% that would be seen without training for legitimate bulk messages. This classifier also drastically reduces the false positive rate. We see a false positive rate of only 1.3% as opposed to the earlier 8.1%.

# 6. CLUSTERED SPAM DATA

The previous section described how it is necessary to include legitimate bulk messages in order to effectively classify messages from a modern SMS corpus. Those experiments produced a labeled dataset of over 8179 labeled messages grouped into 475 clusters, and this set provides a great example of the utility of using public data to develop content-based SMS spam classifiers. In this particular case, this data set is unique because it spans many countries, carriers, and



(a) `Sender Vol. vs Message Vol.`

(b) `Sender Vol. vs Lifetime`

(c) `Message Vol. vs Lifetime`

Figure 3: Campaign message volume is strongly correlated with sending message volume, while campaign lifetime is less related to the amount of messages sent or numbers used by a campaign.

months of time, unlike prior works that have studied only victim-submitted messages or spam in a single network.

## 6.1 Content Analysis

The gateway data included source and destination phone numbers. We used the Twilio phone number lookup service to provide information on the destination phone numbers (i.e. numbers controlled by the gateways), including the destination country and carriers. The United Kingdom received an overwhelming majority of the spam messages — 72.1%. This is even a disproportionate share considering that the UK received only 11.4% of the total messages in the gateway dataset. Australia, China, and Belgium also had disproportionally high spam message volumes as well.

These clusters were categorized into 18 distinct categories, and the top 10 categories are also shown in Figure 2. Messages offering payday loans or other forms of credit comprised 41.3% of all labeled spam in this message — dwarfing all other categories. Following loan spam was job advertising messages. 97.5% of these messages — 827 — were sent from a single number in a 7 hour period. Each message was personalized with a unique name and address; we believe that these messages were sent to a gateway as a test run for a bulk messenger service before sending the messages to their intended recipients. Because gateways collect a number of account verification requests, it was unsurprising to find advertising for telephony services ("obtain a phone number") or phone verification services. We also found the standard contests, online gambling opportunities, and a small number (57) of adult-oriented services common in spam data. However, we did find some more interesting schemes. One example was messages claiming to offer refunds or payouts for reasons as varied as unclaimed tax refunds, unclaimed injury settlements, or unfairly levied bank fees.

## 6.2 Network Analysis

By combining content analysis with network features like sending numbers, we can gain additional insights into SMS spam activity not available to earlier studies. In particular, we can study the activity of a given spam *campaign* — messages that may come from many different phone numbers but delivering a similar message to many users. For our analysis, we treat each spam cluster as a campaign. These campaigns are extensive in scope. They can have lifetimes of over a year (402 days) with a median lifetime of 53 days, transmit messages to up to 12 countries, and send from up to 80 numbers with a median of 5.

We hypothesized that if networks take any sort of proactive measure to prevent nuisance bulk messaging, that spam campaigns with high message volumes and long lifetimes would need to use many sending numbers to deliver high message volumes over time. We also hypothesized that long-lived campaigns would have have high message volumes. Figure 3c visualizes the relationship between these variables, with each data point representing a single spam campaign. We also compute the Spearman correlation coefficients[2] between these variables. As expected, we found that the message volume and the number of sending phone numbers was strongly correlated ($\rho = 0.761$), as shown in Figure 3a. Surprisingly, we found a lower correlation ($\rho = 0.530$) between message volume and campaign lifetime; as shown in Figure 3b, low-volume campaigns are present across the lifetime range. Finally, we see that while many short-lived campaigns have low numbers of sending messages, many long-lived campaigns are successful using a small number of messages. These variables also share a weak correlation ($\rho = 0.473$). Overall, this data implies that spammers who want to send at high volumes must use many numbers to do so, but apart from many campaigns that send only a few messages over a short time scale, campaign lifetime seems unrelated to either the sending number volume nor the messaging volume.

# 7. CONCLUSION

As text messaging has evolved from a closed system where every message was generated within the cellular network to one where a wide variety of non-cellular services can send these messages, the nature of SMS data has substantially changed. The rise of legitimate bulk messages, which may syntactically resemble spam but provide valuable services such as two-factor authentication to users, means that traditional approaches to characterizing SMS spam are no longer adequate for classification. We address these problems in this paper by releasing the largest corpus of publicly available labeled bulk messages and SMS spam. Based on our classification techniques, we demonstrate that compared to

[2] Spearman correlations, represented as $\rho$, measure with a value from $-1$ to 1 whether a monotonic function (not a strictly linear function, as in the case of a Pearson correlation) relates two variables.

previous work, we raise precision across the public corpus from 23.8% to 100%, and raise recall from 61.3% to 96.8%. Even in the absence of manual labeling, we raise precision to 84.6% with a 1.3% false positive rate compared to 8.1% using previous techniques. We also find substantial amounts of SMS spam are related to finance, and certain countries are disproportionately targeted by spam. Our results demonstrate that new approaches to spam classification, and adequately sized SMS corpora, are essential to ensure the accurate classification of text messages as their form and function evolve and diversify.

## Acknowledgments

The authors thank our shepherd, Emiliano De Cristofaro, and our anonymous reviewers for their helpful guidance. This work was supported in part by the National Science Foundation under grant numbers CNS-1526718, CNS-1464087, CNS-1540217, CNS-1542018, and CNS-1464088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## 8. REFERENCES

[1] Mobile Authentication. https://www.duosecurity.com/product/methods/duo-mobile.

[2] Massive growth in A2P SMS expected. http://www.telecompaper.com/news/massive-growth-in-a2p-sms-expected-dimoco--1129833, 2016.

[3] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning From Data*. AMLBook, United States, Mar. 2012.

[4] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami. Contributions to the Study of SMS Spam Filtering: New Collection and Results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, DocEng '11, pages 259–262, New York, NY, USA, 2011. ACM.

[5] F. Aloul, S. Zahidi, and W. El-Hajj. Two factor authentication using mobile phones. In *IEEE/ACS International Conference on Computer Systems and Applications, 2009. AICCSA 2009*, pages 641–644, May 2009.

[6] L. Aouad, A. Mosquera, S. Grzonkowski, and D. Morss. SMS Spam — A Holistic View. In *Proceedings of the 11th International Conference on Security and Cryptography*, 2014.

[7] T. Chen and M.-Y. Kan. Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation*, Aug. 2012.

[8] G. V. Cormack, J. M. Gomez Hidalgo, and E. P. Sanz. Spam filtering for short messages. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge management*. ACM, 2007.

[9] D. DeFigueiredo. The Case for Mobile Two-Factor Authentication. *IEEE Security & Privacy*, Sept. 2011.

[10] S. J. Delany, M. Buckley, and D. Greene. SMS Spam Filtering. *Expert Syst. Appl.*, Aug. 2012.

[11] S. Dixit, S. Gupta, and C. V. Ravishankar. Lohit: An Online Detection & Control System for Cellular SMS Spam. *IASTED Communication, Network, and Information Security*, 2005.

[12] A. Dmitrienko, C. Liebchen, C. Rossow, and A.-R. Sadeghi. On the (In)Security of Mobile Two-Factor Authentication. In *Financial Cryptography and Data Security*. Springer, Mar. 2014.

[13] J. M. Gomez Hidalgo, G. C. Bringas, E. P. Sanz, and F. C. Garcia. Content Based SMS Spam Filtering. In *Proceedings of the 2006 ACM Symposium on Document Engineering*, New York, NY, USA, 2006. ACM.

[14] N. Jiang, Y. Jin, A. Skudlark, and Z.-L. Zhang. Understanding SMS Spam in a Large Cellular Network. In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '13, New York, NY, USA, 2013. ACM.

[15] A. Mosquera, L. Aouad, S. Grzonkowski, and D. Morss. On Detecting Messaging Abuse in Short Text Messages using Linguistic and Behavioral patterns. *arXiv preprint arXiv:1408.3934*, 2014.

[16] C. Mulliner, R. Borgaonkar, P. Stewin, and J.-P. Seifert. SMS-based One-Time Passwords: Attacks and Defense. In *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2013.

[17] C. Mulliner, N. Golde, and J.-P. Seifert. SMS of Death: From Analyzing to Attacking Mobile Phones on a Large Scale. In *Proceedings of the USENIX Security Symposium (SECURITY)*, 2011.

[18] I. Murynets and R. P. Jover. Analysis of SMS Spam in Mobility Networks. *International Journal of Advanced Computer Science*, May 2013.

[19] I. Murynets and R. Piqueras Jover. Crime Scene Investigation: SMS Spam Data Analysis. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, New York, NY, USA, 2012. ACM.

[20] Nan Jiang, Yu Jin, A. Skudlark, and Zhi-Li Zhang. Greystar: Fast and Accurate Detection of SMS Spam Numbers in Large Cellular Networks using Grey Phone Space. In *Proceedings of the 22nd USENIX Security Symposium.*, Washington DC, USA, 2013. USENIX Association.

[21] A. Narayan and P. Saxena. The Curse of 140 Characters: Evaluating the Efficacy of SMS Spam Detection on Android. In *Proceedings of the Third ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, SPSM '13, New York, NY, USA, 2013. ACM.

[22] M. T. Nuruzzaman, C. Lee, and D. Choi. Independent and Personal SMS Spam Filtering. In *2011 IEEE 11th International Conference on Computer and Information Technology (CIT)*, Aug. 2011.

[23] F. S. Park, C. Gangakhedkar, and P. Traynor. Leveraging Cellular Infrastructure to Improve Fraud Prevention. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, 2009.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011.

[25] B. Reaves, N. Scaife, D. Tian, L. Blue, P. Traynor, and K. Butler. Sending out an SMS: Characterizing the Security of the SMS Ecosystem with Public Gateways. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2016.

[26] H. Tan, N. Goharian, and M. Sherr. $100,000 Prize Jackpot. Call Now!: Identifying the Pertinent Features of SMS Spam. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, New York, NY, USA, 2012. ACM.

[27] P. Traynor. Characterizing the Security Implications of Third-Party EAS Over Cellular Text Messaging Services. *IEEE Transactions on Mobile Computing (TMC)*, 11(6):983–994, 2012.

[28] P. Traynor, W. Enck, P. McDaniel, and T. La Porta. Exploiting Open Functionality in SMS-Capable Cellular Networks. *Journal of Computer Security (JCS)*, 16(6):713–742, 2008.

[29] P. Traynor, W. Enck, P. McDaniel, and T. La Porta. Mitigating Attacks On Open Functionality in SMS-Capable Cellular Networks. *IEEE/ACM Transactions on Networking (TON)*, 17(1), 2009.

[30] P. Traynor, P. McDaniel, and T. La Porta. On Attack Causality in Internet-Connected Cellular Networks. In *Proceedings of the USENIX Security Symposium (SECURITY)*, 2007.

[31] A. K. Uysal, S. Gunal, S. Ergin, and E. S. Gunal. A Novel Framework for SMS Spam Filtering. In *2012 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, July 2012.

[32] Q. Xu, E. W. Xiang, Q. Yang, J. Du, and J. Zhong. SMS Spam Detection Using Noncontent Features. *IEEE Intelligent Systems*, 27(6):44–51, 2012.

[33] K. Yadav, P. Kumaraguru, A. Goyal, A. Gupta, and V. Naik. SMSAssassin: Crowdsourcing Driven Mobile-based System for SMS Spam Filtering. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, HotMobile '11, pages 1–6, New York, NY, USA, 2011. ACM.