“Sorry, I was hacked”
A Classification of Compromised Twitter Accounts

Eva Zangerle
Databases and Information Systems
Institute of Computer Science
University of Innsbruck, Austria
eva.zangerle@uibk.ac.at

Günther Specht
Databases and Information Systems
Institute of Computer Science
University of Innsbruck, Austria
guenther.specht@uibk.ac.at

ABSTRACT

Online social networks like Facebook or Twitter have become powerful information diffusion platforms as they have attracted hundreds of millions of users. The possibility of reaching millions of users within these networks not only attracted standard users, but also cyber-criminals who abuse the networks by spreading spam. This is accomplished by either creating fake accounts, bots, cyborgs or by hacking and compromising accounts. Compromised accounts are subsequently used to spread spam in the name of their legitimate owner. This work sets out to investigate how Twitter users react to having their account hacked and how they deal with compromised accounts.
We crawled a data set of tweets in which users state that their account was hacked and subsequently performed a supervised classification of these tweets based on the reaction and behavior of the respective user. We find that 27.30% of the analyzed Twitter users change to a new account once their account was hacked. 50.91% of all users either state that they were hacked or apologize for any unsolicited tweets or direct messages.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues—*Abuse and Crime Involving Computers*; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Security, Measurement, Human Factors, Experimentation

Keywords

Microblogging, Twitter, Social Media, Account Compromising, Abuse, Spam, Machine Learning

SAC 2014 March 24–28 2014, Gyeongju, Republic of Korea.

1. INTRODUCTION

Online social networks have become important means of communication within the last decade, enabling users to reach out to other users to communicate and spread information. The microblogging platform Twitter is among the most popular platforms, serving approximately 200 million users [20] and issuing a total of 400 million tweets per day [30]. The vast amount of reachable users and the amount of interchanged messages on the Twitter platform also attracts cyber-criminals, whose objective is to spread spam messages containing URLs of affiliate websites. This results in the fact that 8% of all URLs posted on Twitter lead to scam, malware or phishing websites [12]. Furthermore, the fact that the click-through rate (number of URLs clicked by users) of spam on Twitter is two orders of a magnitude higher than for email spam makes Twitter even more attractive to cyber-criminals. An analysis of 200 million tweets revealed that 50% of spam tweets were used to promote free music, games, books, jewelry, electronics and vehicles. Also, gambling and financial products, such as loans, are promoted via spam on Twitter [12].
Within the last years, four main approaches for spreading spam on social networks have been observed [9, 10]: (i) setting up a fake account which is solely used for spreading spam messages, (ii) setting up a bot (a program automatically performing a certain task, i.e., sending tweets), (iii) setting up a cyborg (either a bot-assisted human or a human-assisted bot [9]) or (iv) compromising accounts of human users. Currently, the primary strategy for spam attacks is the compromising of accounts [12]. We define a Twitter account as *compromised* if the according account was hacked by a third party and subsequently used for spreading tweets, direct messages or following and unfollowing users without the knowledge of the original account owner. Chronologically, the compromising of an account can be depicted as follows: over a certain period of time, the user's tweets exhibit a characteristic behavior when tweeting (i.e., language used, URLs and hashtags added to the tweet). At a certain point of time, the user's account gets hacked and cyber-criminals hijack the account in order to spread spam and wrong information to the user's followers pretending that these tweets originate from the legitimate account owner. Hence, compromised accounts exploit the trust relationship between the original owner of the compromised account and the user's followers and followees. After a certain amount of time, the user detects that the account was hacked, requests a password reset from the Twitter platform and re-

captures the account. Often, the user subsequently sends out a tweet to all followers stating e.g. “If I sent you spams via DM, I’m really sorry - my account got hacked.” in order to repudiate from possible malicious tweets or direct messages (DM) sent from the account. According to Twitter’s help center, the compromising of accounts can be lead back to various reasons: “Accounts may become compromised if you’ve entrusted your username and password to a malicious third-party application or website, if your Twitter account is vulnerable due to a weak password, if viruses or malware on your computer are collecting passwords, or if you’re on a compromised network.” [28].

This work sets out to analyze the behavior of Twitter users in the case of having their account hacked and compromised. To get a better understanding of how Twitter users deal with getting their accounts hacked, we analyze tweets of users who state that their account was hacked and classify these tweets in order to infer behavioral information. Given this overall goal, we address the following research questions in this paper:

- How do Twitter users react after having recognized that their account was compromised?
- Which actions do these users take after having been hacked?

The remainder of this paper is structured as follows. Section 2 describes the microblogging platform Twitter, the data set underlying our analysis and the according crawling procedure. Section 3 covers the empirical evaluation including general data set statistics and Section 4 describes the classification approach we facilitated and the metrics we made use of for the evaluation. Section 5 contains the results of the behavioral classification of compromised Twitter accounts and we discuss the results in Section 6. Section 7 features research related to our work and we conclude our paper in Section 8.

2. BACKGROUND AND DATA

In the following section, we briefly sketch the Twitter platform and its characteristics. Subsequently, we describe the crawling procedure facilitated for the data set underlying our analysis.

2.1 Twitter

Twitter is a microblogging platform which allows its users to post *tweets* which are at most 140 characters long. User A may *follow* user B which ensures that user A receives all of user B’s tweets. Such relationships are directed and user A is the *follower* of user B, whereas user B is the *followee.* Users can also directly address other users by making use of *mentions.* By stating a username within a tweet, the tweet is directly delivered to the specific user. Another characteristic feature of the Twitter platform are *retweets*, where users can propagate tweets originally posted by other users by redistributing these tweets to their own followers. Retweeted messages can be identified by “RT @username” followed by the original tweet text, where @username is the name of the user who originally posted this tweet. Furthermore, *hashtags* can be used for a manual categorization of tweets by stating the topic preceded by a hash-sign within the tweet (e.g., #android or #syria).

Twitter handles abuse based on a set of strict rules [26] which form the basis of an automated detection algorithm which deactivates accounts exhibiting behavior which might indicate spam [24]. Twitter’s detection algorithm is based on various behavioral indicators as e.g., number of unsolicited mentions, issuing duplicated messages or gaining a large number of followers within a short period of time. Thomas et al. found that 77% of all spamming accounts are suspended within one day and 92% are suspended within three days [24].

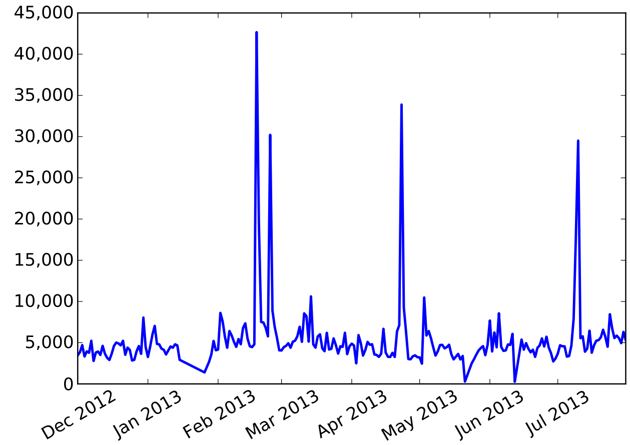

Figure 1: Distribution of Crawled Tweets Per Day

2.2 Twitter Data Collection

In order to gather a representative data set underlying our analysis, we facilitate the following data collection methodology. We crawled Twitter via the Twitter Filter API [25] which allows for filtering the public Twitter stream for given keywords. To particularly filter for tweets related to hacked accounts, we searched the Twitter stream for tweets containing the strings “hacked” or “compromised” and the string “account”. In total, we were able to gather 1,231,468 tweets between December 1^{st} 2012 and July 30^{th} 2013. The Filter API delivers all tweets matching the given query up to a rate limiting equal to the rate limiting of the public streaming API (approximately 1% of all tweets). As the number of tweets matching our query constantly was below this limit (maximum number of tweets crawled per day: 42,670), we were able to crawl all tweets matching the given filter keywords during the given time period.

3. DATA SET ANALYSIS

In the following section, we present a first analysis of the crawled data set. The data set comprises a total of 1,231,468 tweets published by 839,013 distinct users. Figure 1 features a plot of the number of tweets gathered per day for the given crawling period. As for the three peaks pictured, the peak around 2013/02/18 and 2013/02/19 is related to the hack of Burger King’s Twitter account. The peak on 2013/04/23 is concerned with the hack of the Twitter account of Associated Press. The peak around 2013/07/08 and 2013/07/09 is related to Niall Horan’s hacked account where the message of this hack was widely spread by his fans.

Attribute	Value	Attribute	Value
Tweets	1,231,468	Original tweets	384,466
Distinct users	839,013	Avg. tweets/d	5,331
Retweets	339,824	Min. tweets/d	262
Hashtag occur.	179,994	Max. tweets/d	42,670
URLs	125,603	Mentions	1,165,524

Table 1: Data Set Characteristics

Hashtag	Occurr.	Occurr. %
#hacked	17,123	9.51%
#twitter	8,506	4.73%
#facebook	3,788	2.10%
#instagram	3,328	1.84%
#lulz	2,838	1.58%

Table 2: Hashtag Occurrences

Table 1 features the most important facts about the data set. The data set features a total of 179,994 hashtag usages with an average of 0.15 hashtags per tweet (111,964 hashtags when not considering retweets). The most popular hashtags can be found in Table 2 where #hacked is featured in 1.39% of all tweets in the data set and responsible for 9.51% of all hashtag occurrences. In total, 30,806 distinct hashtags have been facilitated. Hashtags are often used by users to refer to the social network in which the user's account was hacked (i.e., #twitter, #facebook or #instagram). The hashtag #lulz can be defined as "...the plural of lol (i.e., lols) that is either misspelled or intentionally changed to not resemble another internet geek slang word" according to the TagDef website [1]. The data set furthermore features 27.59% retweets which we extracted based on the retweet syntax identified by Boyd et al. [5]. For our analysis, we additionally adapt the notion of original tweets which neither feature a retweet nor a mention or reply to another user [6]. Our data set features 384,466 original tweets which amounts to 31.22% of all tweets. Moreover, the data set contains a total of 1,165,524 mentions where 846,673 tweets feature at least one mention. Interestingly, 1,105 tweets within our data set are directed to Twitter's support center's account (@support) asking for help due to a compromised or hacked Twitter account.

4. CLASSIFICATION

In this work, we propose to view the analysis of a user's behavior in case of a compromised Twitter account as a classification problem. Thus, we performed a classification task based on the data set previously described in Section 3. We applied a supervised learning algorithm aiming at finding classes of users reacting similarly to having their Twitter account compromised. We relied on Support Vector Machines (SVM) for the classification of tweets. SVMs are suited for the classification of texts and basically represent texts as n-dimensional feature vectors based on a bag-of-words representation [15]. Based on this vector-representation in an n-dimensional space, a SVM aims at finding virtual hyperplanes which divide the classes best based on a previous training phase. In our case, we made use of a linear SVM kernel. We performed experiments with a set of other classification methods (including non-linear SVM, Naive Bayes' classification and kNN classification). In particular, we made use of the scikit-learn Python toolkit (which internally relies on libSVM [8] for SVM classification), which provides implementations for each of these classification methods [22]. In order to tune the parameters of the individual classification methods, we made use of a grid search which spans a grid of parameters for the given classification methods and subsequently cross-validates all possible combinations of parameters aiming at finding optimal parameters [14]. The results of the grid search process over all proposed classification methods showed that SVMs performed best for our prerequisites.

As for the input data for the training-, test- and the actual classification phase, we relied on the text of the tweets and did not incorporate any further meta data as i.e., the time the tweet was sent, user details or the user's tweeting history. Hence, we aim at classifying single, isolated tweets. In a first step, we performed the following preliminary tasks on the given data. These tasks included removing all non-English tweets and all retweets. Furthermore, we treated direct messages and replies (featuring a mention of one or more users) equal to original messages as also direct messages and replies are used to state that a particular account was hacked. Syntactically, these preliminary tasks also included lower-casing all tweets and removing punctuation. Additionally, we performed tests to assess whether further preprocessing steps (stemming, removal of stopwords, URLs, hashtags or user mentions from the input tweets) would influence the classification quality. Hence, we also included these preprocessing steps into the grid search process in order to find optimal SVM classification parameters. This enabled us to evaluate which of the aforementioned preprocessing steps have a positive affect on the classification process and outcome (see Section 5). As for the stemming process, we made use of the Python implementation of the Porter stemming algorithm for the English language [23] provided by the Natural Language Toolkit (NLTK) [4]. The result of these evaluations showed that the removal of URLs from tweets as a preprocessing step improves the quality of classifications. However, removing mentions of users, hashtags or stopwords or performing stemming on our data did not increase the performance of the SVM classification process and hence, we did not include these steps in the preprocessing phase.

4.1 Metrics

For the evaluation of the performed classification tasks and also for the comparison of various classification approaches, we relied on the traditional IR-metrics recall, precision, F_1-measure and accuracy [32]. We present the results of the classification as confusion matrices, an example of a confusion matrix can be seen in Table 3. In this matrix, the rows marked as True denote the actual behavioral classes (in this example termed as class1 and class2), whereas the columns marked as Predicted denote the classes computed by the SVM classifier. The entries a and d represent the number of correctly classified items of class 1 resp. class 2, b and c represent the number of falsely classified items for class 1 resp. class 2.

The recall for a given class can be defined as the ratio of the number of tweets which were correctly classified to the

		Predicted	
		Class1	Class2
True	Class1	a	b
	Class2	c	d

Table 3: Confusion Matrix

number of actual tweets in the class. In order to illustrate this metric, we can also define recall based on a confusion matrix as shown in Table 3 as $r = a/(a + b)$. In contrast, precision can be defined as the ratio of the number of tweets which were correctly classified to the total number of tweets predicted in this class. In terms of the confusion matrix, precision can be computed as $p = a/(a+c)$. The F_1-measure combines both precision and recall into one measure and is the harmonic mean of recall and precision ($F_1 = (r * p * 2)/(r + p)$). Accuracy is defined as the total number of correctly classified items over all classes, i.e., $acc = (a + d)/(a + b + c + d)$.

The classification experiments were performed using a 5-fold cross-validation in order to provide reliable results on the quality of the classification process. For k-fold cross validation, we randomly split the training set into k complementary folds. $(k - 1)$ folds are used as a training set for the classifier whereas one fold is retained and serves as the test set for the classifier evaluation. This procedure is repeated k times with each fold used once as the test sample. The metrics resulting from the evaluation runs of the k folds are then averaged.

4.2 Input Data

In a first step—prior to the actual classification of how users react to a hacked account—we needed to filter out those tweets which actually state that the according Twitter account has been hacked. This is due to the fact that crawling for the keywords "hacked" and "compromised" also returns tweets of users stating that e.g., a friend's account was hacked (e.g. "@JuliaMWB I think your account was hacked, friend!") or that the user's email or other social network account was hacked (e.g., "If you get an e-mail from me asking how are you, don't click the link. My gmail account was hacked and they blasted an email with the link"), which can also be inferred from the hashtags used (cf. Table 2). However, our goal is to analyze tweets where the users state that their own Twitter account has been hacked, as e.g. in "Apparently my account was hacked. I haven't sent anyone a direct message. #Stupidhackers". Hence, we performed a binary classification task for filtering these relevant messages. We performed this classification by applying a linear SVM classifier using the same procedure as for the actual final classification task (as described in Section 4). This classification resulted in a set of 358,639 tweets relevant for our further analysis (out of 859,214 input tweets which resulted from stripping the original data set from retweets and duplicates).

4.3 Training and Test Data

For the training and test phases of the SVM classifier, we randomly selected 2,500 tweets from within the data set and manually classified these tweets for each of the two classification processes. These classified sets of tweets serve as input for the k-fold cross-validation.

5. CLASSIFICATION RESULTS

In the following we present the results of the classification task as previously described. In a first step, we performed the preliminary classification as depicted in Section 4.2. The confusion matrix for the preliminary classification of tweets can be seen in Table 4. We achieved an overall accuracy of 82.51% and an overall F_1-score of 82.33%. In terms of classification performance, these results are comparable with other binary classification tasks for Twitter data, e.g. when classifying spam or credible messages on Twitter [2, 7]. As we are subsequently performing the behavioral classification process on all true positives (entry a in Table 3), a high value for the class MyAccount (86.21%) is crucial whereas accuracy for the other class is negligible.

		Predicted	
		MyAccount	Other
True	MyAccount	**86.21%**	13.79%
	Other	21.29%	**78.71%**

Table 4: Confusion Matrix for Preliminary Classification

After a manual exploration of the tweets, we decided to employ the following classes for the behavioral classification task:

1. Users who state that their account was hacked (e.g., "ooh looks like I've been hacked! That explains the inability to get into my account! Will be putting that right").

2. Users stating that their account has been hacked who immediately apologize for any unsolicited tweets (e.g., "My Account was hacked pls ignore all the tweets I sent today. I apologize for the inconvenience").

3. Users stating that their account has been hacked who immediately apologize for any unsolicited direct messages (e.g., "If I sent you spams via DM, I'm really sorry - my account got hacked.").

4. Users who state that they were hacked and moved to a new account (e.g., "Hey guys, go follow my new account because this one is hacked and is sending out spam").

5. Users who have been hacked and state that they now changed their password (e.g., "Very sorry everyone. my account was hacked. password changed, hopefully that does the trick.").

6. Users who state that they were "hacked" by a friend or relative, where hacked refers to e.g., leaving a device unattended (e.g., "my brother hacked my account sorry").

7. Other tweets, not belonging to any of the above described classes (mostly related to wrongly classified tweets in the preliminary classification step).

Given these classes, we performed the classification using the procedure described in Section 4. In the following, we present and discuss the results of the classification which can be seen in the pie chart pictured in Figure 2. The most

		Predicted						
		Class1	Class2	Class3	Class4	Class5	Class6	Class7
True	Class1	**80.89%**	0.89%	4.44%	9.33%	0.00%	4.00%	0.44%
	Class2	38.10%	**42.86%**	14.29%	0.00%	0.00%	4.76%	0.00%
	Class3	25.58%	6.98%	**65.12%**	0.00%	0.00%	0.00%	2.33%
	Class4	0.00%	0.00%	0.00%	**98.68%**	0.00%	1.32%	0.00%
	Class5	36.36%	0.00%	18.18%	0.00%	**36.36%**	9.09%	0.00%
	Class6	25.00%	0.00%	8.33%	0.00%	0.00%	**66.67%**	0.00%
	Class7	8.33%	0.00%	0.00%	33.33%	0.00%	0.00%	**58.33%**

Table 5: Confusion Matrix for Behavioral Classification

prominent category are users stating that their account has been hacked and that they moved to a new account, this class amounts to 27.30%. Within these tweets, the users state that they created a new account and ask their followers to follow their new account in order to receive their updates. 23.36% of all tweets in the analyzed data set simply state that the according account has been hacked with no further information given (except for occasional cursing). 13.87% of the users within our data set apologize for unsolicited tweets sent from their account and 13.68% apologize for unsolicited direct messages sent to their followers during the compromising of their account. Furthermore, 3.87% state that their account was hacked and that they now changed their Twitter password which conforms to Twitter's advice for compromised accounts [28].

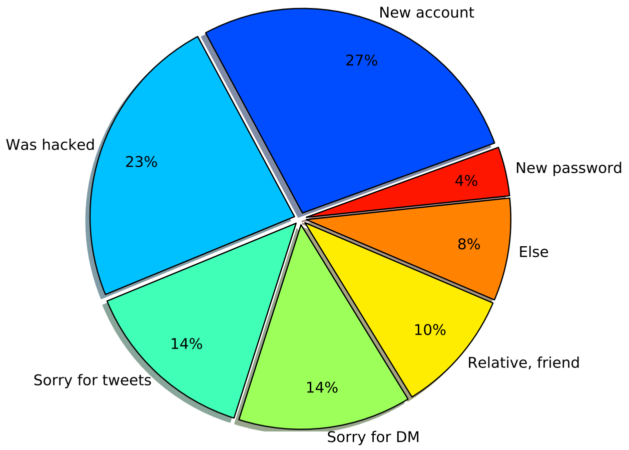

Figure 2: Classification Results

The confusion matrix for the classification can be seen in Table 5. The overall accuracy for all classes is 78.25%, the F_1 score achieves 77.96%. To investigate where the classifier did not perform well, we examined the according cases manually. In particular, we observed that the classification for classes 2 and 5 lead to unsatisfactory results. The performance of both classes can be lead back to only marginal textual differences between tweets of these classes (e.g., the usage of "tweet" instead of "direct message" within tweets contained in classes 2 and 3).

To confirm that the overlap between the classes we facilitated was kept at a minimum, we further examined the classified tweets. Therefore, we extracted the most discriminant features for each class from the trained classifier (as described in Section 4). This results in a list of terms most characteristic for each of the seven classes. I.e., the most discriminant features for class 3 (sorry for direct messages) are e.g., 'dm', 'direct', 'message' and 'sorry'. Based on these lists of features, we searched for tweets within our data set which contain features from multiple classes in order to check whether tweets could be assigned to multiple classes. The (semantically) most similar classes are classes 2 (sorry for tweets) and 3 (sorry for direct messages), therefore we show the overlap for these two classes in the following. Our analysis showed that a total of 737 tweets can be assigned to either class 2 or 3. When further examining these tweets manually, we observe that this overlap is related to users apologizing for having sent unsolicited tweets and direct messages within one single tweet. Naturally, this tweet can be assigned to both of these classes and hence, leads to an overlap within these two classes which can hardly be avoided. Still we believe that the distinction between class 2 (sorry for tweets) and 3 (sorry for direct messages) is important as it also reveals information about how cyber-criminals spread information to the followers of the compromised user account.

6. DISCUSSION

In the following, we aim at getting a closer look at the results and discuss the implications of our findings. Based on the results of the classification, we performed a manual exploration of the classes to find further evidence of how Twitter users react to having their account hacked. As for class 1, users who state that they got hacked within a tweet, 20,975 tweets are posted to ask a particular other user to follow back as this particular following-relationship has been lost during the compromising of the account. This amounts to 25.04% of all tweets within this class. The fact that 27.30% of all tweets state that the respective user created a new account is remarkable as the Twitter Support Page advises users to (i) change the password, (ii) revoke connection to third-party applications and (iii) update the passwords in the trusted third-party applications [28]. The deactivation and subsequent deletion of an account, resp. the creation of a new account is not mentioned on these support pages. This user behavior leaves the according account to cybercriminals who continue to tweet via this account. However, 77% of spam accounts are detected within one day. Therefore, a second strategy applied by cyber-criminals is to not make use of the account for a certain amount of time before starting to spread spam (also referred to as the dormancy period) [24]. Within our data set, a total of 3,236 users who tweet that they moved to a new account also state that they deleted (or will delete) their old account (3.36% of all users within this class). Apparently, users rarely seek for help by directly contacting Twitter's @support account, as our data set shows only 1,105 tweets directed at this account. Knowing that abandoned accounts can still be used by cyber-criminals and that Twitter users invest considerable effort for redirecting followers to their new account, it appears likely that Twitter users hardly know how to appropriately handle a compromised account. Furthermore, Twitter also provides guidelines for safe tweeting [27]. These guidelines advise users to (i) use a strong password, (ii) use login verification, (iii) be careful about (untrusted) third-party apps and suspicious links (phishing) and (iv) be cautious about spyware and viruses on their computer. However, still a large number of accounts get compromised and users still seem miss- and underinformed. Therefore, it seems advisable to further promote help and support mechanisms and provide more information about how to deal with compromised accounts, i.e., that a compromised account can be restored.

7. RELATED WORK

Work related to our research can be grouped into two streams of research: (i) spam and its detection within social networks, in particular on Twitter and (ii) the analysis of behavior and relationships of cyber-criminals within social networks.

The abuse of social websites aiming at spreading spam to users has been widely investigated within the last years. Heymann et al. [13] distinguish three different categories of countermeasures to cope with spam on social networks: (i) detection, (ii) demotion and (iii) prevention. The detection of spam can be accomplished automatically by pattern-based classification or users who actively report spam. The strategy of demotion is related to ranking spam contributions lower in e.g., search results whereas prevention is concerned with limiting automated interaction with social platforms (e.g., by using captchas).

Grier et al. find that spammers primarily make use of compromised accounts in order to spread spam [12]. The authors analyzed URLs within tweets and found that 8% of all URLs posted are related to phishing or spam attacks. The automatic detection of compromised accounts has recently been studied by Egele et al. [10]. The authors aim at creating a behavioral profile for each user for which they focus on predefined features and detect abnormal behavior by comparing the behavioral profile to the features of a new tweet t. Subsequently, a tweet database is searched for messages similar to t where similarity is defined as the textual similarity between tweets and the similarity of URLs mentioned in these tweets. If the system detects more than ten highly similar messages, the account is flagged as compromised. The authors argue that one single abnormal tweet does not necessarily imply a compromised account, as the user e.g. might have changed the Twitter software or app she uses for tweeting.

Thomas et al. analyzed the lifespan of Twitter spam accounts and found that Twitter is able to detect 77% of all spam accounts within the first day of creation and 92% of all spam accounts are detected within three days [24]. Generally, the detection of spam has been studied widely. Lee et al. create social honeypots on Twitter to gather information about social spam behavior and subsequently propose methods for identifying spam [17, 18]. Lee and Kim analyze redirect chains of URLs used in spam tweets in order to detect spam [19]. McChord and Chuah propose to facilitate traditional classifiers employing content-based features to classify spam [21], whereas Benevenuto et al. make use of content- and behavioral features to identify spammers [2]. The use of topics of collective attention on Twitter (e.g., viral videos, memes or breaking news) for disseminating spam has been studied in [16]. Bilge et al. observed that spam attacks get more popular with a rising popularity of online social networks [3]. In this particular work, the authors focus on identity theft attacks where profiles of users in online social networks are copied and then used to send out contact or friendship requests to users who already connected to the original user.

The social relationships of spam accounts has been studied in [11, 31]. The authors find that cyber-criminals on Twitter form a small-world network and also strive for non-criminal followers in order to reach a wider audience. Chu et al. introduce the distinction between human, bot and cyborg users of Twitter [9]. The authors present a classification method for these three categories of users which also incorporates a spam detection mechanism based on Bayesian classification. Also, Wagner et al. present an approach aiming at identifying bots based on a feature-based classification method [29].

8. CONCLUSION AND FUTURE WORK

In this work we ask how Twitter users deal with having their account hacked and compromised. We analyzed a data set of 1,231,468 tweets and presented a classification of such tweets aiming at finding classes of users reacting similarly to a compromised account. In particular, 23.36% of all users simply state that their account was hacked, whereas a total of 27.55% apologize for unsolicited tweets or direct messages which have been sent via their compromised account. Furthermore, we find that 27.30% of all analyzed tweets state that the user changed to a new account after having had her Twitter account compromised supposedly due to a lack of information in regards to how to recapture a compromised account.

As for future work, we intend to conduct a survey of Twitter users who had their accounts hacked aiming at identifying how account compromising influences user behavior on the Twitter platform. Additionally, we aim at getting a better understanding of which actions users take after having their account hacked, e.g. how users manage hacked accounts to prevent additional damage. Furthermore, we plan to investigate how user trust in the Twitter platform is influenced by account compromising and which implications a possibly changed trust level has on the user's behavior.

9. REFERENCES

[1] #TagDef hashtag definition directory. `http://tagdef.com`.

[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. In *Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS)*, volume 6, 2010.

[3] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your Contacts are Belong to us: Automated Identity Theft Attacks on Social Networks. In *Proceedings of the 18th International Conference on WWW*, pages 551–560, 2009.

[4] S. Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[5] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *43rd Hawaii International Conference on System Sciences (HICSS)*, pages 1–10, 2010.

[6] A. Bruns and S. Stieglitz. Quantitative Approaches to Comparing Communication Patterns on Twitter. *Journal of Technology in Human Services*, 30(3-4):160–185, December 2012.

[7] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *Proceedings of the 20th Intl. Conference on WWW*, pages 675–684, 2011.

[8] C.-C. Chang and C.-J. Lin. LIBSVM: A Library for Support Vector Mchines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, May 2011.

[9] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is Tweeting on Twitter: Human, Bot, or Cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 21–30, 2010.

[10] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. COMPA: Detecting Compromised Accounts on Social Networks. In *ISOC Network and Distributed System Security Symposium (NDSS)*, 2013.

[11] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and P. K. Gummadi. Understanding and Combating Link Farming in the Twitter Social Network. In *Proceedings of the 21st International Conference on WWW*, pages 61–70, 2012.

[12] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @ spam: the Underground on 140 Characters or Less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, pages 27–37, 2010.

[13] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. *Internet Computing, IEEE*, 11(6):36–45, 2007.

[14] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[15] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Machine Learning: ECML-98*, pages 137–142, 1998.

[16] K. Lee, J. Caverlee, K. Y. Kamath, and Z. Cheng. Detecting Collective Attention Spam. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 48–55, 2012.

[17] K. Lee, J. Caverlee, and S. Webb. Uncovering Social Spammers: Social Honeypots + Machine Learning. In *Proceedings of the 33rd InternationalACM SIGIR Conference*, pages 435–442, 2010.

[18] K. Lee, B. D. Eoff, and J. Caverlee. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *Intl. AAAI Conference on Weblogs and Social Media*. AAAI Press, 2011.

[19] S. Lee and J. Kim. WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream. *IEEE Transactions on Dependable and Secure Computing*, 10(3):183–195, 2013.

[20] Mashable. Twitter now has more than 200 million monthly active users `http://mashable.com/2012/12/18/twitter-200-million-active-users/`.

[21] M. McCord and M. Chuah. Spam Detection on Twitter Using Traditional Classifiers. In *Proceedings of the 8th International Conference on Autonomic and Trusted Computing*, volume 6906 of *Lecture Notes in Computer Science*, pages 175–186. Springer, 2011.

[22] 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*, 12:2825–2830, 2011.

[23] M. F. Porter. An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, 14(3):130–137, 1980.

[24] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended Accounts in Retrospect: an Analysis of Twitter Spam. In *Proc. of the 2011 ACM SIGCOMM Conference on Internet Measurement*, pages 243–258, 2011.

[25] Twitter. Filter API Documentation. `https://dev.twitter.com/docs/api/1.1/post/statuses/filter`.

[26] Twitter Rules. `https://support.twitter.com/articles/18311-the-twitter-rules#`.

[27] Twitter Support. Keeping your account secure. `https://support.twitter.com/articles/76036-safety-keeping-your-account-secure#`.

[28] Twitter Support. My account has been compromised. `https://support.twitter.com/articles/31796-my-account-has-been-compromised#`.

[29] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier. When Social Bots Attack: Modeling Susceptibility of Users in Online Social Networks. In *2nd workshop on Making Sense of Microposts at WWW2012*, 2012.

[30] Washington Post. Twitter turns 7: Users send over 400 million tweets per day. `http://articles.washingtonpost.com/2013-03-21/business/37889387_1_tweets-jack-dorsey-twitter`.

[31] C. Yang, R. C. Harkreader, J. Zhang, S. Shin, and G. Gu. Analyzing Spammers' Social Networks for Fun and Profit: a Case Study of Cyber Criminal Ecosystem on Twitter. In *Proceedings of the 21st International Conference on WWW*, pages 71–80, 2012.

[32] Y. Yang. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1(1-2):69–90, May 1999.