A Hybrid Approach for Spam Detection for Twitter

Malik Mateen
NU FAST, Islamabad, Pakistan.
mateen2454@gmail.com

Muhammad Aleem
Capital University of Science and Technology,
Islamabad, Pakistan.
aleem@cust.edu.pk

Muhammad Azhar Iqbal
Capital University of Science and Technology,
Islamabad, Pakistan.
azhar@cust.edu.pk

Muhammad Arshad Islam
Capital University of Science and Technology,
Islamabad, Pakistan.
arshad.islam@cust.edu.pk

***Abstract*—Online social networks (OSNs) are becoming extremely popular among Internet users as they spend significant amount of time on popular social networking sites like Facebook, Twitter and Google+. These sites are turning out to be fundamentally pervasive and are developing a communication channel for billions of users. Online community use them to find new friends, update their existing friends list with their latest thoughts and activities. Huge information available on these sites attracts the interest of cyber criminals who misuse these sites to exploit vulnerabilities for their illicit benefits such as advertising some product or to attract victims to click on malicious links or infecting users system just for the purpose of making money. Spam detection is one of the major problems these days in social networking sites such as twitter. Most previous techniques use different set of features to classify spam and non-spam users. In this paper, we proposed a hybrid technique which uses content-based as well as graph-based features for identification of spammers on twitter platform. We have analysed the proposed technique on real Twitter dataset with 11k uses and more than 400k tweets approximately. Our results show that the detection rate of our proposed technique is much higher than any of the existing techniques.**

I. Introduction

It has becomes quite simple to access information from all around the world with the help of Internet. Increase in popularity of social networking sites allows us to gather enormous amount of data and information about users, their relationships, friends and family. Large amount of data present on these sites attracts also malicious users. Such users use autonomous programs that act like human to steal the user's personal information, spreading misinformation and propaganda. These special programs are called social bots. These social bots are capable of sending friend requests, posting messages and can also influence the opinion of user by posting deceptive information on different Websites[1]. Preliminary social bots were designed to have positive impact on online social communities. They are used for awareness and collaboration of human on different social issues but due to expansion of social structure they are misused. These social bots maintain human like profiles with randomly selected pictures and name. These bots may send friend requests or follow randomly to the user that are selected from the list thus gaining the trust of user that accepts request of these social bots.

Spam detection is critical task for security of social media. It is very important to identify spam in online social network in order to protect users from various kinds of attacks. In social networking sites, information is shared on trust relationships. Usually it is shared among personal friends and it might be public as well, i.e., it can be accessed by everyone. Some users have tendency to accept unknown friend request to become popular at the cost of his privacy. Network trust is very important as far as security perspectives are concerned. Different campaigns are run by social networking sites for the awareness of user from various kinds of threats. Unfortunately there is no mechanism in online social networking sites for providing secure verification of users.

A. Online Social Network Vulnerabilities

Large number of users and huge amount of information being shared increases security and privacy issues in online social networking sites (OSNs). According to statistics released by Facebook 655 million user log on to this site and share 4.75 billion pieces of information with each other [3]. Large amount of data present on these sites attract the malicious groups. These groups use autonomous programs that act like human to seal the user's personal information, spread misinformation and propaganda. These spacial programs are called social bots. These social bots are capable of sending friend requests, posting messages and can also influence the opinion of user by posting deceptive information on different Websites [4]. User can face various type of attacks while using OSNs that include viruses which can be send by spammers to harm user's system, personal information can be stolen by trust worthy third party to use for their own interests, Fake accounts can also harm user by gaining trust to effect user's reputation. These bots send friend request randomly to the user selected from the list. If victim accepts their request then they start communicating with these victim friends which if accept request increase bots acceptance rate. For example, someone using social engineering to hack computer network might try to gain the confidence of an official user and get them to disclose information that compromises the network's

978-1-4673-9073-6/17/$31.00 ©2017 IEEE

security. Social engineers often rely on the natural helpfulness of people as well as on their weaknesses. They may call the authorised employee with some kind of urgent problem that requires immediate network access.The Social Engineering Attack Cycle is shown in Fig. 1. Clone attacks are also bigger threats in which attacker creates clone profiles of user's friends, if person accept the request then all the information will be accessible to the attacker.

Twitter has two million users in one month which share 8.3 million tweets per hour [2]. Twitter limit user to post their messages (tweets) up to 140 characters. A tweet may contain some textual highlights for better user connection experience, which are likewise to be misused by spammers. A hashtag is a specific word or an expression prefixed with the #symbol. it is utilised for gathering tweets about a specifics topic. Spammer use the hashtag to spread their tweets using a spamming trap called hashtags hijacking. By Using with username can transfer tweet direct to the user which encourages spammer to specifically send spam. Spammer incorporate spam messages and making spam web content accessible to maximum lot. In previous couple of years social networking sites have gained popularity due to which these sites also attracted spammer to exploit social trust of users, and they have achieved a much higher success rate than traditional spam sending methods. Spammers aim to incorporate spam content in tweet by implanting an outside URL in a tweet to appeal a user to the spam sites. For protection of users, Twitter provides rules against spammer[1]. One can identify spammers bots by the less number of followers. However they try to follow large number of users in order to spread misinformation. Followers to Following ratio is generally greater than one in case of legitimate users and less than one for Spammers. There are certain regularities in spam posting behaviour as they post messages at regular interval unlike legitimate users. Spammer bots also use different techniques to evade detection by paying people to follow them, by mixing spam tweets with normal tweets this motivates researcher to develop a new technique for spammers detection.

In this paper we have evaluated content based and graph based features to detect twitter spam. We have used a comprehensive dataset that contains more than 11 million tweets. Our results show that the hybrid approach significantly improves the precision as well as recall for the data set used. In next section we have discussed state of the art existing approaches for the detection of spam in online social networks. In section III we have presented our technique for merging the two kinds of features. Section V describes the details of our experiments have been presented and conclusion and future work is discussed in section VI.

II. Literature Review

Online social networking sites are built on principles of trust and has attracted many researchers due to its popularity. One experiments performed using Facebook dataset showed

[1]http://www.twitter.com/rules

Fig. 1. The Social Engineering Attack Cycle

the 41% users accepted unknown friend request[5]. Research showed that users are likely to click on the links posted by unknown persons[6]. Yang et. al[7] gives complete details analysis on evasion tactics that were used by spammers. They also purposed different features in tweet that can be used for spam detection. This paper analyses different techniques that are used by spammer to avoid detection also proposed different set of features to classify spam users. After extracting features different machine learning classifier are used random forest outperform all other classier with F measure 0.9

Lee et. al. [8] created honeypots in MySpace and Twitter for the detection of spam profile. They extracted the features like number of followers, behaviour of posting features. They used machine learning classifier LIBSVM for detection of spammers. This technique has been trained on two data sets but validation is done on one data set. Gee et. al [9] have analysed, to which degree spam has entered social networking sites. Author created honey-profiles on different social networking sites. The basic purpose is to determine behaviour and interaction pattern of spammer and then purposing different techniques to detect and avoid spam behaviour.

Egele et al. [10] proposed a technique for detection of compromised account. They tested this technique on twitter and Facebook data set. This technique identify compromised accounts by determining sudden irregularities in users account. M. McCord et. al [11] also used tweet features for spam detection. They collected tweets and extracted features like timing of tweets posted, retweets length, mentions and keywords. Different classification algorithms are used for spam detection. Random forest outperforms all other classification algorithms in this work. In [12] authors presented another technique for spam detection in twitter. Authors collected approximately twenty six thousands of users and their tweets for spam detection and various features are extracted from the tweets these features includes URL rate and interaction ratio. They used J48 classifier for spam classification.

In [13] authors proposed a technique that collectively examines spam messages and group them in to a single campaign.

The collective behaviour of spammers is analysed and categorised them as spam campaign. Twitter data set is collected using API method provided by twitter. After collection of data labelling is done manually by using human expertise and by following twitter policy to detect spam from data set. Different machine learning classifier were used and their accuracy was measured. Random Forest outperform all these classifier with low false positive rate in above mention circumstances.

In [14], authors presented a new application for Facebook users known as "MyPageKeeper" that detects spams. It protects users from spam attacks. MyPageKeeper uses content of profile instead of using other user information. SVM is used for detection of spammer from non spammers post using various features. MyPageKeeper considers post to be spam if some specific device is used for posting messages, post contain different false promises, post is created by using specific person without knowledge of that person, by clicking URLs in the post ask for some survey etc.

Graph based methods have been reported in literature to be used for social network analysis[15], topic classification[16], and sentiment analysis[17]. H.Gao et al [18] used graph features like Sender Social Degree, Interaction History, Cluster Size, Average URL Number per Message, Unique URL Number, they apply these techniques to two data set Facebook and twitter for spam detection. SVM and Decision trees are used for classification.

III. Proposed Technique

Most of previous research has been done in detecting spam and identifying profile of spammer. Previous papers use different techniques that use various methods for spam profile detection and spam detection individually. Each paper uses its own dataset and its features for classification. As discussed earlier, various kinds of features are used for spam detection such as *user-based*, *content-based* and *graph-based* with each of them having its own pros and cons. Inspired form these techniques we proposed a technique that uses combination of all of the three above mentioned techniques. We use these features to create model that distinguish well between human's and spam profile. We evaluate our techniques using twitter dataset, which is one of the most popular social networking sites. Such hybrid features for spam profile detection are effective against evasion tactics. We aim to achieve higher accuracy by combing all these features. After classification we will find correlation between features and will eliminate correlated features.

A. User Based Features

User-based features are based on a users relationships and properties of user account. In online social networks, users can build their own social networks by following uses and allowing others to follow them. As any spammer has to reach large number of profiles to spread misinformation so they try to follow large number of users. According to twitter policy if number of people following is more as compare to number of people following him then account might be consider as spam. It is very significant to add user related features in the model for spam identification. As User features are related to users account so we extract all the attributes that are related to uses account.

a) Number of followers: Follower define the popularity of someone profile. Spammers bots have generally less followers as they don't exits physically.

b) Number of following: On the Twitter following someone means you will see their tweets (Twitter updates) in your personal time-line. Twitter lets you see who you follow and also who is following you. Followers are people who receive other people's Twitter updates.

c) Age of account: Age of account is the date when account has been created according to twitter policy (mentioned in section I) spammer are newly created accounts.

d) FF ratio: It is the ratio of follower to following of any user account. Spammer bots don't received friends request as no one knows him personally in real life thus there is huge difference between number of friend request send to number of friend request received. FF ratio is very less for real user and in the case of bots it will be very high. In Twitter following and followers are public by default. Mathematically it can be represented as

$$FF = \frac{No.of Following}{No.of Followers}$$

e) Reputation: It is ratio between number of followers to sum of following and followers.

$$Reputation = \frac{Followers}{Followers + Following}$$

B. Content-based Features

Content-based features are related to tweets posted by user as spam bots post lot of duplicate content as compare to normal user that don't post duplicate tweets. Content-based features are based on content or messages that users write. These are most important features as spam users post most of tweets to spread misinformation and these tweet contains lot of malicious URLs to advertise their product. We use content-based features and compute their mean, median, minimum and maximum values of these features. Content base features are as follows

f) Total number of tweets : This feature include total number of tweets that user posted in his/her lifespan. Spammers have low life span as they use busty property to post tweets. So it is very vital to use this kind of feature to detect spam accounts.

g) Hashtag ratio: It is the ratio between the tweets containing hashtags to total tweets posted and tweets contains unique hashtag.

$$Total\#tags = \frac{Duplicatehashtags}{UniqueHashtags \times tweetcount}$$

h) URL's ratio: It is the ratio between duplicate URLs to number of distinct URLs in tweets and sum of Tweets.

$$TotalURLs = \frac{\#DuplicateURLs}{\#UniqueURLs \times tweetcount}$$

i) Mentions ratio: Users are identified by uniques username *@username* format user can also reply with @username and send messages. User can replay to other users whether or not user is in his friend list or not. *@username* can be written everywhere in the tweet. Spammers also misuse this feature to send spams to other users so according to twitter policy. If users message contains large number of mention and reply tags than user is consider to be spam user. It is ratio between tweets contain @ to sum of tweets.

$$@tweets = \frac{Tweetscontaining@}{TotalNumberOftweets}$$

j) Tweet frequency: As spammers post tweets robotically by using twitter API or by using web interface at regular intervals as research shows that spam users are active on a specific time day. Moreover tweet frequency is greater than a genuine twitter user. The basic idea behind including this feature is to detect automatic behaviour of spam users while normal user shows have random behaviour.

k) Spam words: We use 100 most popular spam words and count the number of occurrence of these spam words in tweets of users. As spammer uses these popular spam words to spread misinformation and to advertise their products, this feature can be vital to identify spam tweets.

C. Graph-based Features

We are using Graph-based features to overcome the evasion tactics performed by spammers. Spammer use different dodging techniques to avoid being detected. They can purchase followers form third party website for Internet black market also they can exchange their followers with other users to look like legitimate users. Different websites are used to purchase followers. Most common websites include *BuytwitterFriends.com, twittersource .com, Usocial.net* and *Tweetcha.com*. We included graph-based features to identify the spammer behaviour that uses these strategies to avoid being detected. Even twitter can change their behaviour of sending messages but it is very difficult for them to alter the position in social graph.

l) in/out degree: Suppose if two account follow each other than there will be bi directional edge between them. As spammer follow large number of unknown users but they don't force them to follow back. Also spammers have no physical exisitance in the real world so out-degree for spammer would be high as compared to in-degree.

m) Betweenness: Betweenness centrality is defined as number that measures the influence of a node based on its presence on the number of shortest path in a social network [19] . Betweenness centrality of the selected node is a fraction of the number of shortest paths from all nodes to all others that pass through the selected node.

Mathematically, the betweenness of a network node v (represented as $BC(v)$) is the sum over all pairs of nodes, of the fraction of the shortest path between u and w that pass through v.

$$BC(v) = \sum_{u \neq w \neq v} \frac{\sigma_{uw}(v)}{\sigma_{uw}} \quad (1)$$

IV. Experimental Setup

In order to evaluate our approach, we need a labeled collection of user dataset. To best our knowledge there is no such data publicly available due to twitter policy so we had to build one or use the dataset that are used in previous techniques. This section describes the challenges for data collection and labelings data for spam bots detection.

A. Challenges in Collecting Dataset

There is no dataset available with tweets publicly due to twitter privacy policy. Most of researchers in past have used their own data set by extracting it using API . However, constructing a ground truth is very complicated task. Most of the researchers have done it manually by evaluating each and every feature of account to mark it as spam and non spam users. We use data set that is provided by Guofei Gu.[2] Data set provided is in raw form so we have to preprocess in order to get data in filteered form. Provided dataset is in csv from but tweets and other information also contain commas in their text which distort csv files. The dataset consists of $10,256$ users and 467480 tweets.

V. Results Discussion

We have evaluated the results suing three most common classier, i.e., J48, Decorate and Naive-Bayes. Fig. 2, 3, 4 show that J48 outperformed all other classifiers in all three sub classes of features. Decorate has same true positive and precision as of J48 however, false positive rate of J48 is one percent more than Decorate. Naive-Bayes has performed poorly as dataset among all in all three sub classes. We also compare the result of classifier by using user based and Graph based features. Results show that content based and Graph based features classification results outperforms the user based classification results. This suggests that user based feature do not play a significant role in the identification of spam users. User-based and graph-based techniques show 90% of correct classification results while content-based and graph-based features shows 92 % accuracy. it is noteworthy that the features used in this paper have very low false negative as compared to existing techniques.

We have analysed our results in the perspective of the earlier works (A[20] that use the content-based features shown in Fig. 2, user-based features as shown in Fig. 3, graph-based features as shown in Fig. 4. We have tested the existing techniques on the dataset we extracted as discussed above in this paper. We observe a significant improvement in the results from the features used in [20] and [7]. However, in the case of [12] the results are competitive. We acknowledge that the number of features used in this paper is large and we intend to reduce this number by finding correlation among the feature so that we can obtain the minimum possible feature list without compromising on the results. Fig. 3 shows that J48 outperforms all other classifier in term of accuracy but have high false positive rate J48 has precision of 94%.

[2]http://faculty.cs.tamu.edu/guofei/.

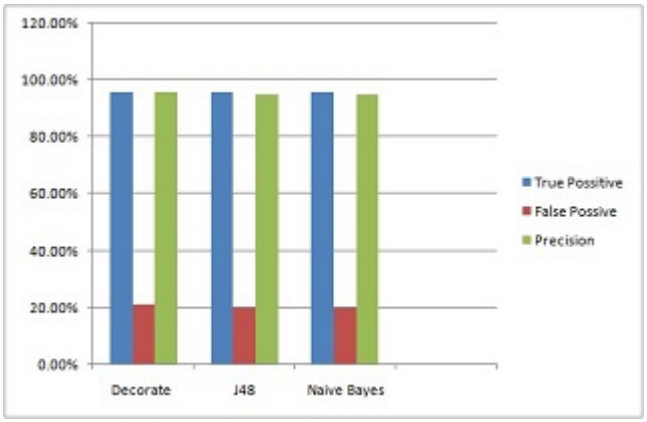

Fig. 2. Classification results using User-based features

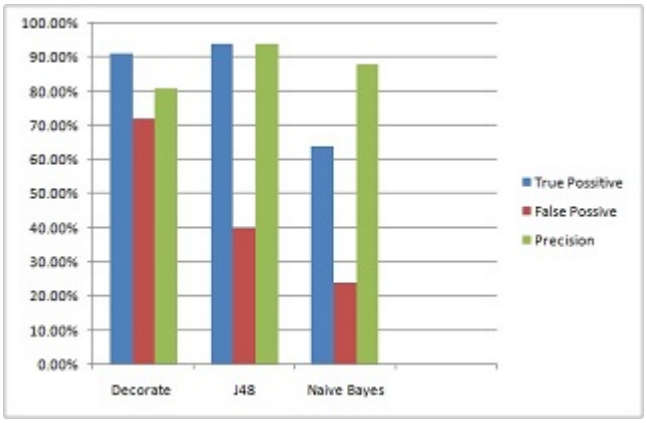

Fig. 3. Classification results using Contect-based features

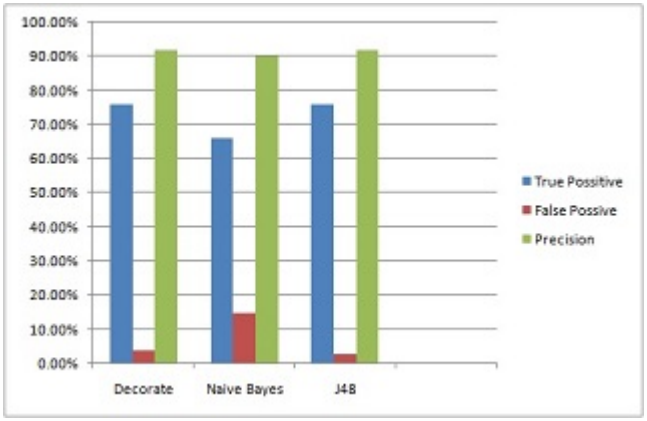

Fig. 4. Classification results using Graph-based features

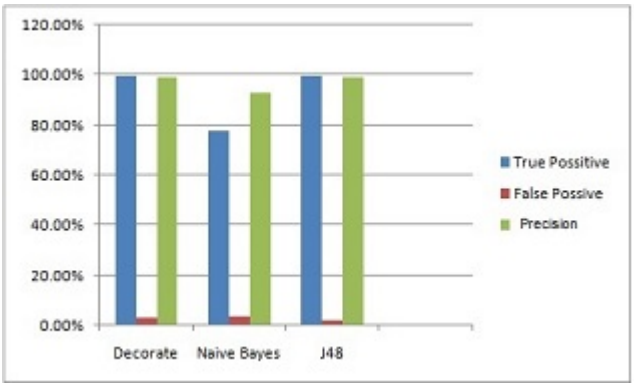

Fig. 5. Classification Results using Features proposed in this paper

Results of Fig. 2 show that decorate outperform other two classifier. Results show that other two classifiers are unable to detect spammers that are included in this dataset with high precision. It might be due to quality of features that are used in the paper. Fig. 4 shows that Decorate perform better as compared to other classifier. Moreover, the number of false positive is also less form previously described techniques. Fig. V shows the results of classification using the features in the hybrid technique, i.e., combining *user-based, content-based* and *graph-based* features. We can observe a significant improvement in precision for decorate and j48, i.e., up to 97.6%.

VI. Conclusion

In this paper we presented the a hybrid set of features for detecting spammers on social networking site. We used twitter dataset of almost 11K users. We proposed a technique that used *user-based, content-based* and *graph-based* features for spam profile detection. Our experiments show high classification accuracy with low false positive. In future we will extend the evaluation of the proposed hybrid features by testing on other social networking sites like Facebook, MySpace also work should be done in developing a new techniques based on tweet level spam detection which can be done by categorising text in the tweets.

References

[1] T. C. Marshall, "Facebook surveillance of former romantic partners: associations with postbreakup recovery and personal growth," *Cyberpsychology, Behavior, and Social Networking*, vol. 15, no. 10, pp. 521–526, 2012.

[2] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," 2010.

[3] V. Namen and K. Kinnison, "Facebook facts and twitter tips-prosecutors and social media: An analysis of the implications associated with the use of social media in the prosecution function."

[4] D. J. Watts and P. S. Dodds, "Influentials, networks, and public opinion formation," *Journal of consumer research*, vol. 34, no. 4, pp. 441–458, 2007.

[5] F. Nagle and L. Singh, "Can friends be trusted? exploring privacy in online social networks," in *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*. IEEE, 2009, pp. 312–315.

[6] J. C. McElwee, J. M. Stevens, and A. Thiruppathi, "Method for mitigating web-based one-click attacks," 2009.

[7] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 8, pp. 1280–1293, 2013.
[8] S. Lee and J. Kim, "Warningbird: Detecting suspicious urls in twitter stream," in *Symposium on Network and Distributed System Security (NDSS)*, 2012.
[9] M. Verma and S. Sofat, "Techniques to detect spammers in twitter-a survey," *International Journal of Computer Applications*, vol. 85, no. 10, 2014.
[10] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao, "Follow the green: growth and dynamics in twitter follower markets," in *Proceedings of the 2013 conference on Internet measurement conference*. ACM, 2013, pp. 163–176.
[11] M. McCord and M. Chuah, "Spam detection on twitter using traditional classifiers," in *Autonomic and Trusted Computing*. Springer, 2011, pp. 175–186.
[12] P.-C. Lin and P.-M. Huang, "A study of effective features for detecting long-surviving twitter spam accounts," in *Advanced Communication Technology (ICACT), 2013 15th International Conference on*. IEEE, 2013, pp. 841–846.
[13] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *Dependable and Secure Computing, IEEE Transactions on*, vol. 9, no. 6, pp. 811–824, 2012.
[14] G. Zhou, Y. Wu, T. Yan, T. He, C. Huang, J. A. Stankovic, and T. F. Abdelzaher, "A multifrequency mac specially designed for wireless sensor network applications," *ACM Trans. Embed. Comput. Syst.*, vol. 9, no. 4.
[15] R. Asif and M. A. Islam, "Finding most collaborating mathematicians a co-author network analysis of mathematics domain," in *2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, April 2016, pp. 289–293.
[16] J. Hussain and M. A. Islam, "Evaluation of graph centrality measures for tweet classification," in *2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, April 2016, pp. 126–131.
[17] N. Nodarakis, S. Sioutas, A. Tsakalidis, and G. Tzimas, "Large scale sentiment analysis on twitter with spark."
[18] W. Guan, H. Gao, M. Yang, Y. Li, H. Ma, W. Qian, Z. Cao, and X. Yang, "Hot social events on sinaweibo," *arXiv preprint arXiv:1304.3898*, 2013.
[19] L. Freeman, "A set of measures of centrality based upon betweenness," *Sociometry*, vol. 4, no. 12, pp. 35–41, 1977.
[20] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, 2010, p. 12.