Connections between Twitter Spammer Categories

Gordon Edwards
School of Informatics
University of Edinburgh
Edinburgh, Scotland, UK
g.n.edwards@sms.ed.ac.uk

Amy Guy
School of Informatics
University of Edinburgh
Edinburgh, Scotland, UK
Amy.Guy@ed.ac.uk

ABSTRACT

Twitter has become a viable platform for spammers, who often form networks to further their reach. Troublesomely, targeted users become increasingly frustrated, or worse, view content resulting in computer virus infection. We build on previous work around detecting spam on Twitter, proposing that subcategorising spammers can increase our understanding of their connections in spammer networks and aid detection. After defining five subcategories of spammers and classifying users accordingly, correlations between the categories of spammers and the categories of their followers and followees are explored. We also find that all spam subcategories follow a higher share of non-spam accounts than any individual spam subcategories, and, unexpectedly, that every spammer subcategory is followed by non-spammers more than by individual counterparts.

Keywords

Twitter, spammer categories, spam, social media, microposts, machine learning

1. INTRODUCTION

Twitter's popularity attracts spammers, providing them with a very publicly-accessible user base. It reported that less than 5% of its users are spammers, but that figure is likely to be higher in reality [2], especially with the more wide-ranging criteria for spam adopted in this paper. Spam can pose a security threat to users, or just cause annoyance — either way leaving them disillusioned with Twitter.

Users are not compelled to follow accounts they deem to be spam. However, the ability to quickly determine if a new follower is a spammer is useful in deciding whether to follow back. Automatic detection could save users from wasting time checking each new follower, and spare them from potentially dangerous spam. Spammers can also reach users via a mention or a direct message; in this case investigating the tweet author safeguards against spam.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

It is suggested in [7] that spammers collude within Twitter networks — that if each account is a node in a graph, then from each node a spam account can be reached by traversing five edges with probability $p = 0.63$. Working together in networks helps spammers proliferate, as it is unlikely a whole network will successfully be taken down. Adding new accounts to their network as others are removed, each can rely on follows from accounts within the network. A desirable but false impression of popularity is thus given. Detecting and classifying whole spammer networks at once could enable more efficient elimination of spam, compared to assessing on a continual basis all individual accounts on the site.

Previous work considers various machine learning techniques for detecting spam, such as Random Forest and Naïve Bayes, either from live feeds or from research corpora [1, 4]. Broadly, it refers to two sets of features upon which users can be classified: content-based, such as mean number of hashtags per tweet, and user-based, such as number of followers of the authoring user [4].

The preceding literature frames spam classification as a binary process (not spam/spam). However, further investigation reveals recurring subtypes of spam—for example users advertising products, or users disseminating pornography—providing a novel approach to classification. Aside from academic interest, classifying into subtypes means users could engage in more refined decisions about blocking of content or users than Twitter's spam filtering currently allows. It also facilitates pinpointing of the most harmful spam, such as tweets concealing viruses and phishing attacks.

Emergent trends, which we will examine, in the distribution of an account's followers and those they follow between the categories may increase confidence that it belongs to a particular category. Finding that one spammer is commonly connected to a particular type yields a fast way to discover accounts of that type, potentially to block or suspend. Connections between different spammer categories are not very dangerous in themselves—though could lure a user to viewing further spam accounts—but they form a potential means of detecting spammer networks.

This paper, part of an ongoing research project, lays the groundwork for investigating the extent to which different categories of spammers are connected to others, and to genuine users. It establishes that these connections result from

spammers' collusion within networks. We build on the work of [7], but contrastingly not confining ourselves to just one trending topic. In Section 2 we describe our defined subcategories of spam, training set, features, and classifier. We then summarise our findings in Section 3 and their limitations in Section 4.

2. CLASSIFICATION

2.1 Spam Subcategories

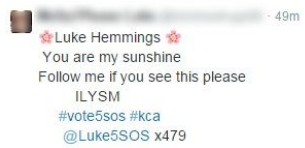
The Twitter API [6] offers the means to collect a sample of 1,420 users to form a training set, to subsequently hand-label as *spam* and *not spam*. During this annotation process spam subcategories become apparent. Whilst not necessarily definitive, they are reasonably defensible. Though applicable to users and tweets, we only use the categories in relation to users. They are defined below with example tweets typical from the type of spammer. Their distribution is displayed in Figure 1.

- *advertising*: users who tweet extremely frequently, mostly, if not always, advertising products, or tweets advertising a product authored by such a user. Normally the tweets contain links, often shortened using a URL shortener.

- *explicit*: users who post exclusively, or almost so, photos, videos, and links, perhaps shortened with a URL shortener, to websites of a pornographic or adult nature, or tweets that that contain this kind of content.

- *follower gain*: users claiming the ability to boost other users' follower bases, frequently, in most of their tweets, asking users for retweets and to follow certain accounts. A tweet in this category claims that retweeting or following a mentioned (via *@username*) account will result in the receipt of followers.

- *celebrity*: users who tweet plead relentlessly for the follow back of a public figure in their tweets. Ascertaining whether an individual tweet falls into this category is generally harder. Examining the authoring user should be indicative — ascertaining whether a suspect tweet is a unique occurrence for that user and therefore not representative.



- *bot*: accounts whose tweets are generated by a bot that auto-posts content from some source, or details usage of an online app. Tweets that fall into this category often contain a URL, but, again, to be certain in classification the authoring account may need to be examined.

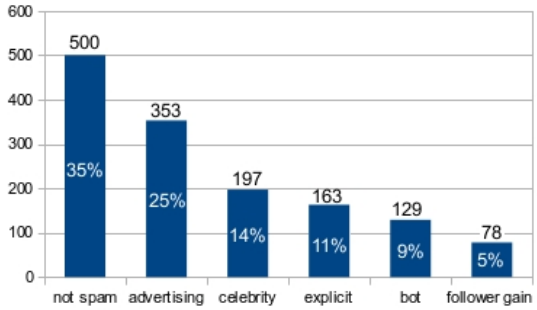


Figure 1: Class distribution of the dataset

2.2 Features

Feature representations of Twitter users can be formed, as per previous work, using content-based and user-based features [4]. Fifty features, 15 user-based and 35 content-based, sufficiently represent users. The content features require the tweet history of the user: their latest 200 tweets, or fewer if they do not have that many. Some features unique to this paper are:

User-based Features
Screen name and description Levenshtein similarity[1]
Percentage of non-alphanumeric characters in description

Content-based Features
Mean number of new lines in the user's tweets
Relative standard deviation of the number of new lines in the user's tweets

2.3 Classifier

The `Random Forest` classifier implementation in the `Weka` Java library [5] provides the basis for implementing a classifier tailored to the spam subcategory classification task. Maximising the spam recall desirably increases the probability of classifying a spammer's spam followers and followees[2] into the subcategories correctly. Thus, the classifier first binarily classifies users as *not spam* and *spam*, using the `Random Forest` classifier — considering all instances labelled as one of the spam subcategories as labelled *spam*. Then, if the outputted classification is *not spam* and the associated confidence is not less than a set threshold[3], *not spam* is returned. Otherwise, the instance is reclassified, again with

[1]Description of Levenshtein similarity: `www.cs.tufts.edu/comp/150GEN/classpages/Levenshtein.html`

[2]For the purposes of this paper "followees" refer to the accounts which a user is following.

[3]Given threshold α, instances initially classified with the binary classifier *not spam*, with confidence c, $c \leq \alpha$, are

the `Random Forest` classifier, applied to dataset with the *not spam* instances filtered out, so one of the spam subcategories is necessarily returned. Conveniently, using Weka's `AdaBoostM1` implementation furthers reduces misclassification due to class imbalance.

Ten-fold cross-validation, provided through `Weka`, allows the classifier to be evaluated, with the collected sample of 1,420 users forming the validation set:

	Recall	Precision	F-Measure
not spam	0.74	0.80	0.77
explicit	0.77	0.83	0.80
advertising	0.84	0.64	0.72
follower gain	0.56	0.90	0.69
bot	0.36	0.56	0.44
celebrity	0.78	0.74	0.76

The classifier performs poorly on the class *bot*, most often misclassifying as *advertising*, so there can be no confidence in conclusions made regarding that class. The misclassification is probably due to the inherent similarity between the behaviours of spammers in each category.

2.4 Results Reporting

For each class, given a sample of 70 contained users the tailored classifier can be used to attain the mean class percentages of followers and followees — 500 (or as many as there are) are sampled for each. Given more time and computational resources, a larger dataset could be formed. All the percentages are rounded to the nearest integer.

Contingency tables are also constructed given the counts of *(category, follower category)* pairs and *(category, followee category)* pairs. These help reveal the extent to which spammers are connected to their followers and to their followees.

3. DISCUSSION OF RESULTS

Possible inaccuracies in classifications detailed in Section 4 mean care should be taken in drawing conclusions, and it is unlikely all of them will be infallible. The results report that genuine users have 73% *not spam* followers on average, 20% higher than the *not spam* followers share of *advertising* and *bot* accounts. Tallying with our intuition, the fair conclusion to draw here given the classifier performance on these follower classes for *not spam* is that genuine users will have a noticeably higher share of *not spam* followers than spammers, a trait that can increase the confidence that a user classified as *not spam* is indeed so. With a fair degree of confidence the results show that genuine users are likely to follow back around half of their genuine followers. The reported number of followers and followees for accounts that spammers follow back is usually higher than for accounts they do not, implying that spammers target their connections to popular accounts.

The average share of *not spam* accounts followed across the *advertising*, *bot*, *celebrity*, and *follower gain* categories, 60%, is notably higher than that of any of the spam subcategories, showing their persistent efforts to gain genuine users' attention. However, perhaps surprisingly, on average 50% of

assumed to be *spam*, to further increase the *spam* recall.

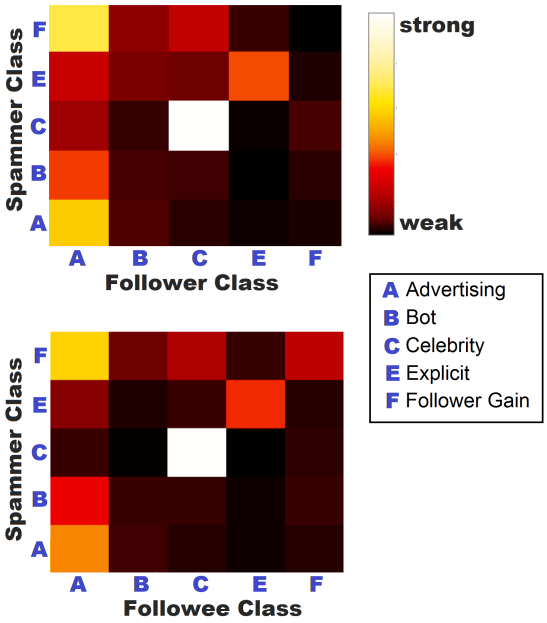


Figure 2: Heat maps showing respectively the strength of connection between spammer subcategories and their follower subcategories, and between spammer subcategories and their followee subcategories.

a spammer's followers are genuine users for each subcategory. Users are either consciously following spammers—perhaps *advertising* accounts hoping to find good deals or *celebrity* accounts because they are interested in the associated celebrity—or through ignorance, lacking a tool to warn them. No one spam category is a landslide winner in attaining genuine followers though.

On average, about 30% of *advertising* followers belong to the same category — a share much higher than any other spam subcategory. Also around 23% of the accounts followed are *advertising*—again, a share much higher than the other spam subcategories—suggesting a significant degree of connection between *advertising* accounts, confirmed later.

Other subcategories appearing to have a high degree of intra-connection are *explicit* and *celebrity*. Accounts in the former have a higher share of *explicit* followers than any other follower subcategory, averaging at 20%, and also follow more accounts of the same subcategory than the others, with a share averaging around 42%. Accounts in the latter have a higher share of *celebrity* followers than the other follower subcategories, averaging at 33%. Such accounts also follow more accounts of the same subcategory than the others, with a share averaging around 42%.

However, accounts in the *bot* category have a higher share, averaging at 26%, of *advertising* followers than *bot* followers (averaging at only 6%) or any other subcategory of follower. Likewise the followees share is higher for *advertising*, averaging at 18%, than *bot* (averaging at only 4%) and the other subcategories. This discrepancy could be due to the categories' inherent similarity; arguably both have the same

Class	Follower	Recall	Precision	F1
advertising	*advertising*	0.51	0.60	0.59
	bot	0.43	0.74	0.55
	not spam	0.57	0.43	0.49
bot	*advertising*	0.56	0.63	0.59
	bot	0.44	0.55	0.49
	not spam	0.48	0.58	0.52
celebrity	*celebrity*	0.63	0.50	0.56
	not spam	0.37	0.93	0.53
explicit	*explicit*	0.43	1.0	0.6
follower gain	*not spam*	0.39	0.83	0.53
not spam	*follower gain*	1.0	0.50	0.67
	not spam	0.44	0.98	0.61

Table 1: For each subcategory of spammer the performance when the classifying each subcategory of follower.

aim— to direct users to content—so there is incentive for them to connect with each other. As previously warned, given the categories are not definitive, *advertising* and *bot* could reasonably be merged into one category, probably reducing the classification error.

We confirm the hypothesised relationships in the connections between spammers of the same subcategory using Cramér's V correlation ϕ_c [3]. Measuring the correlation between two categorical random variables given a constructed contingency table, it ranges from 0, where the two random variables are independent, to 1, where they are equal. Letting `X = Subcategory of spammer` and Y = `Subcategory of follower`, $\phi_c = 0.39$, showing that there is some association between a spammer subcategory and their follower subcategory. Similarly, if `Subcategory of spammer` and Y = `Subcategory of followee`, then $\phi_c = 0.47$, showing there is an analogous correlation between a spammer subcategory and their followee subcategory.

The fairly strong positive correlations and attained percentage shares aforementioned evidence the degree of collusion between spammers, and that those in the same subcategories are deliberately connecting to form networks — notable relationships are present. Predicated on these correlations, the heat maps in Figure 2 show the strength of spammer connections. Because it is a hallmark of spam, establishing the presence of such connections aids spammer network detection and individual account classification.

4. LIMITATIONS

When the classifier is further tested by classifying a sample of followers of users from each of the categories, the performance reported in Table 4 is worse than the cross-validation in Section 2.3, likely due to large variations in the distribution as the sample is more deterministic than the validation set. Thus in Section 3 only sound conclusions respecting these figures were drawn, but improvements made in future work could allow further conclusions regarding the connections between some of the combinations of categories not considered. A larger test sample, perhaps yielding different figures, would clearly be preferable but was not practicable given the time constraints.

5. SUMMARY AND FUTURE WORK

This paper presents the findings of new research. By forming a training set of users and implementing a classifier tailored to the task, underpinned by Random Forest, users can be classified into the defined classes. Analysing the distribution of these classes in users' followers and followees allows inferences to be made about the relationships between users, crucially between spammers. We observe that many genuine users are falling into the trap of connecting with a range of types of spammer.

We reveal that spammers mainly have their largest share of connections devoted to non-spammers and their second largest to spammers of the same subcategory. However there are exceptions, with some subcategories connecting with a proportionally very much smaller number of spammers from the same category. Correlations are found between spammer subcategories and their follower and followee subcategories, showing that spammers are colluding with each other in networks, with a significant degree of connection between spammers of the same category.

Establishing connections between subcategories in a large contiguous network, starting from one account and branching outwards, recursively analysing the followers and followees, could be a future extension. Visualising this network would be interesting, allowing clusters of spammers of different subcategories to be determined. Also the subcategories could usefully be refined, and perhaps more introduced.

6. ACKNOWLEDGMENTS

We thank Krzysztof Jerzy Geras, School of Informatics, University of Edinburgh, for explaining to us how to find correlations, which we subsequently found and included in this paper.

7. REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[2] J. Brustein. Twitter's bot census didn't actually happen. `http://www.businessweek.com/articles/2014-08-12/twitters-bot-population%-remains-a-mystery-and-a-problem`. [Online; accessed 14/11/2014].

[3] P. Dattalo. Nominal association: Phi and cramer's v. `http://www.people.vcu.edu/~pdattalo/702SuppRead/MeasAssoc/NominalAssoc.%html`, 2002. [Online; accessed 10/03/2015].

[4] M. McCord and M. Chuah. Spam detection on twitter using traditional classifiers. In *Proceedings of the 8th International Conference on Autonomic and Trusted Computing*, ATC'11, pages 175–186, Berlin, Heidelberg, 2011. Springer-Verlag.

[5] N. Z. The University of Waikato. Weka. `http://www.cs.waikato.ac.nz/ml/weka/`.

[6] Twitter4j. Twitter4j. `http://twitter4j.org/`.

[7] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a twitter network. In *Volume 15, Number 1 - 4 January 2010, First Monday peer-reviewed journal*. First Monday, 2010.