Badly Evolved? Exploring Long-Surviving Suspicious Users on Twitter

Majid Alfifi(✉) and James Caverlee

Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA
{alfifi,caverlee}@tamu.edu

Abstract. We study the behavior of long-lived eventually suspended accounts in social media through a comprehensive investigation of Arabic Twitter. With a threefold study of (i) the content these accounts post; (ii) the evolution of their linguistic patterns; and (iii) their activity evolution, we compare long-lived users versus short-lived, legitimate, and pro-ISIS users. We find that these long-lived accounts – though trying to appear normal – do exhibit significantly different behaviors from both normal and other suspended users. We additionally identify temporal changes and assess their value in supporting discovery of these accounts and find out that most accounts have actually being "hiding in plain sight" and are detectable early in their lifetime. Finally, we successfully apply our findings to address a series of classification tasks, most notably to determine whether a given account is a long-surviving account.

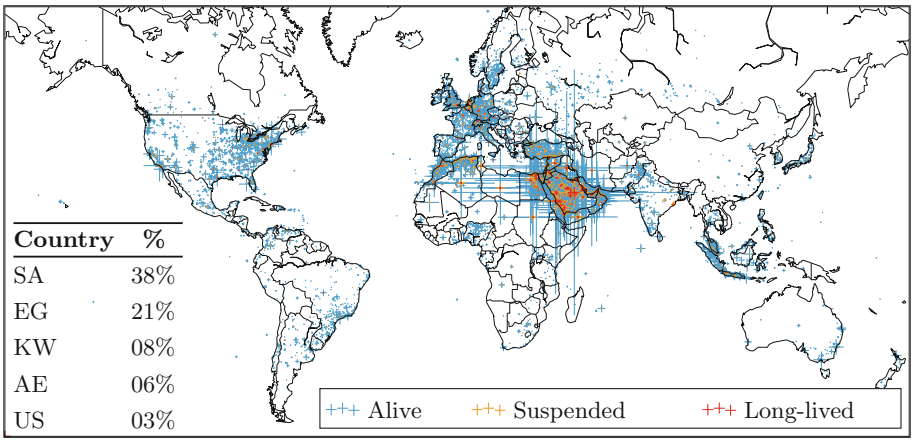
1 Introduction

Social media enables an unprecedented shift in how we communicate to others and how we consume content. And yet, this change has also enabled innovative new methods to manipulate at scale: examples include spreading propaganda and misinformation [18,19], polluting information streams with spam [1,10], and strategically distracting populations by fabricating social media posts [13,23]. For example, [6] found that politically motivated individuals provoked interaction by injecting partisan content into information streams whose primary audience consists of ideologically-opposed users.

While these suspicious efforts can affect social media users worldwide, of special importance is their impact in Arabic social media. The past few years have seen social media as an effective tool for facilitating uprisings and enticing dissent in the Middle East [2,16,21]. The embrace of social media in the region has made it a battle ground for ISIS and similar groups and existing regimes, all spreading propaganda, recruiting sympathizers, and even undermining rivals [2]. Of these different movements, what do they post? How do they evolve? Do they engage in particular strategies to evade detection?

Towards tackling these questions, we focus in this paper on an initial investigation into the behaviors of a special type of users in Arabic social media – long-surviving content polluters who engage in pro-terrorism, spam, and other negative behaviors. Specifically, we analyze all of Arabic Twitter from 2015 (Fig. 1)

G.L. Ciampaglia et al. (Eds.): SocInfo 2017, Part I, LNCS 10539, pp. 218–233, 2017.
DOI: 10.1007/978-3-319-67217-5_14



Dataset	Size
Tweets	9,285,246,636
Accounts	26,711,275
Tweets from Suspended Accounts	1,960,160,536
Suspended Accounts	6,175,113

Fig. 1. Arabic tweets in 2015. 21% of the tweets were generated by eventually suspended accounts which represent 23% of all active accounts in 2015.

to identify the complete set of active[1] long-surviving content polluters, totaling 17,909 accounts and 42,630,795 tweets. A previous work [15] manually identified 816 spam accounts on Twitter, monitored them until later suspended, and then explored effective features for detecting such long-surviving accounts. Our work takes a different approach as we already have at our disposal all Arabic tweets generated in 2015, including those of suspended accounts, and therefore can look back at long-surviving accounts in retrospect, enabling us to study long-surviving accounts at a much larger scale taking into considerations different types of accounts.

In contrast to many efforts that have focused on short-lived accounts engaging in "extreme" (and easily detectable behavior) [10,11,20,24], our interest is on the activities and behaviors of long-lived accounts. Do these accounts always engage in bad behaviors over time? Or do they begin as somewhat legitimate accounts before evolving into bad ones? Is the rate of change gradual? Or do we observe abrupt changes?

Hence, we conduct a threefold study of these long-lived accounts to complement existing studies of traditional anti-spam efforts [14,24,26]: (i) first, we study the content of what these accounts post, finding multiple classes of account types including traditional spammers, pro-ISIS groups, and other politically-motivated

[1] We consider an account to be active and long-surviving if it had tweeted at least once on at least six different months in 2015.

groups; (ii) second, we analyze the evolution of linguistic patterns of these accounts – including an examination of their self-similarity and cross-entropy versus the rest of Arabic Twitter; (iii) third, we examine the activity evolution of these users, to explore their differences with legitimate and short-lived accounts. Finally, we apply our findings to address several classification tasks about the different groups most notably determining if an account is a long-surviving account.

2 Data and Preliminary Analysis

We obtain a large dataset of 9.3 billion tweets (Fig. 1) representing all tweets generated in the Arabic language in 2015 through a private full access to the Twitter Firehose. Of these 9.3 billion tweets, about 2 billion were from suspended accounts. Using all the geotagged tweets in our dataset as a proxy to estimate the level of contribution from different regions in the world, we see in Fig. 1 that most tweets were generated from the Middle East and Africa, but with a global footprint. We find that only 50% of the accounts are suspended after their first tweet (Fig. 2a), with many accounts posting 100s and even 1000s of tweets. This stands in contrast to previous studies [24], reported in 2011, finding 77% of Twitter spam accounts being suspended within a day of their first tweet. Spammers might have developed more sophisticated methods over time or maybe Twitter spam control isn't as efficient on Arabic Twitter although we notice a dramatic improvement over time in fighting spam (Fig. 3).

Moreover, we find that while 90% of spam accounts are suspended within 40 days of their creation (Fig. 2b), a large group of accounts live much longer, in some cases for years.

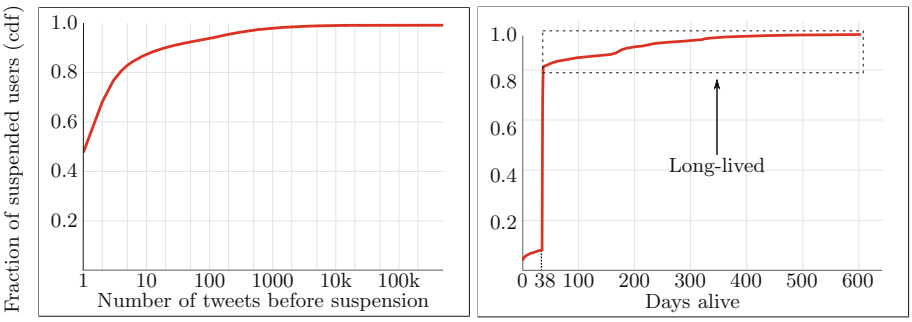


Fig. 2. Lifespan of suspended accounts. (a) 50% of accounts are suspended after their first tweet, but many accounts post 100s and even 1000s of tweets. While most accounts get suspended within 38 days of creation (b), 10% of accounts live for more than 40 days with some for years.

To support our study of the characteristics of such long-lived (but ultimately suspended) accounts, we focus on four types of users extracted from this large dataset:

Suspended Long-Lived Accounts (Long-Lived): We first identify our base set of accounts who lived for at least six months, were active, and were ultimately suspended. We focus on accounts that were created in 2015 and tweeted in at least six different months. In total, we identify 17,909 long-lived accounts who created 42,630,795 tweets.

Suspended Short-Lived Accounts (Short-Lived): To complement this collection of long-lived accounts, we identify an equal-size randomly sampled set of 17,909 short-lived accounts. We consider a short-lived account to be one that lived for 30 days or less in 2015 before being suspended. In total, these randomly sampled accounts generated 14,129,870 tweets.

Legitimate Accounts (Legit): In contrast to these two collections of suspended accounts, we identify a set of legitimate accounts as a point of comparison. We randomly sample 17,909 accounts from all legitimate accounts that were created in 2015 and were still active in November 2016. We further require these accounts to have stopped tweeting the last two months in our dataset as a potential end-of-life signal. This latter restriction is useful when we later compare the evolution of different groups from birth to death. These legitimate accounts posted 9,772,176 tweets.

ISIS-Related Accounts (ISIS): Finally, we identify a collection of ISIS accounts, based on the Anonymous group initiative called LuckyTroll: this effort originally identified more than 25,000 ISIS sympathizers and supporters through crowdsourced reporting[2]. We only use accounts that have actually been suspended by Twitter, indicating multiple users have reported those accounts. Similar to other groups, we consider ISIS accounts that were born in 2015 and were suspended before 2016. This leaves us with 17,518 ISIS accounts responsible for 11,849,065 tweets[3]. We find only 7 common accounts between this ISIS dataset and long-surviving accounts mainly because we use all ISIS accounts without further filtering by level of activity and length of life. A previous work [9] used this same ISIS users dataset but, due to Twitter limitations, they were only able to recover 10% of their tweets through the Truthy project at Indiana University [8]. By contrast, in this work we were able to recover 100% of the content they generated in 2015 with our private full access to the Twitter Firehose.

Having considered all active long-lived accounts we decided to sample same-size sets of the much larger sets of legit and short-lived accounts for manageability and to avoid class imbalance when we later explore automatic detection of long-lived accounts.

[2] The website hosting this dataset has been taken offline but we were able to recover accounts from http://archive.is/A6f3L.

[3] Contact the first author for access to the ISIS dataset.

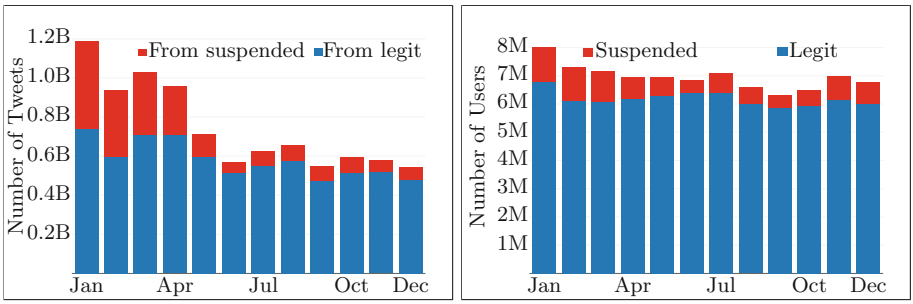


Fig. 3. Twitter cracks down on spam in 2015 resulting in 50% reduction in content (a) while roughly maintaining the same number of monthly active users (b).

3 Investigation

Given these four distinct user groups, we focus here on the behaviors of long-lived suspended accounts.

3.1 What Do Long-Lived Accounts Post?

We begin by examining the types of long-lived accounts through an application of LDA [3] over all the tweets posted by these accounts. With the help of a native Arabic speaker on the team, we identify six popular account types among these long-lived accounts, as illustrated in Table 1. We additionally highlight the potential violation of Twitter rules after inspecting accounts in each group[4]. We see that the most popular group is dominated by ostensibly legitimate topics like sports and news, while also inserting low-quality spam URLs. Similarly, the fourth group (e.g., Muhammed, Prophet) posts innocuous religious posts while also inserting spam URLs. The second group of accounts focuses on account manipulation, by posting offers for buying retweets and followers. In an interesting direction, we do find multiple groups of accounts engaged in potentially politically motivated posts: the third group (e.g., ISIS, Syria, Aleppo) is pro-ISIS; from inspection these accounts appear to have been suspended for promoting terrorism. The sixth group (e.g., Saudi, support, Yemen) posts about the Saudi-Yemen war from a pro-Saudi perspective. Finally, we find the fifth group (e.g., retweet, clips, bypass) posts mainly adult content. These findings indicate that long-lived accounts are diverse, with many different goals.

We further apply LDA to each month's tweets posted by these accounts. We find consistent recurring topics over the months giving a first hint that long-lived accounts exhibit the same behavior throughout their lifetime but have managed to evade detection. This finding suggests that these accounts have not engaged in legitimate behavior for most of their lives before engaging in some "extreme" behavior (and being suspended).

[4] https://support.twitter.com/articles/18311-the-twitter-rules.

Table 1. LDA topics for long-lived suspended accounts.

#	Keywords	English Translation	Potential Violation	%
1	شاهد، الخبر، مدريد، الهلال مباراة، ريال، الاتحاد، برشلونة	watch, News, Real Madrid, Al-Hilal, match, Ittihad, Barcelona	Spam	37%
2	للبيع، الف، تبادل، للطلب رتويت، واتس، اليوم، الرياض	For sale, thousand, exchange, orders, retweet, WhatsApp, today, Riyadh	Selling or purchasing account interactions	19%
3	الدولة، جبهة، الشيخ، النصرة سوريا، حلب، داعش، الشام	State, Front, Sheikh, Nusrah Syria, Aleppo, Daesh(ISIS), Levant	Promoting terrorism	12%
4	وسلم، صلى، محمد، النبي قروب، البخاري، مسلم، رسول	Peace, pray, Muhammed, Prophet, group, Bukhari, Muslim, Messenger	Spam	11%
5	محارم، افلام، الحجب، للمشاهده خاص، تتناك، بنت، عرض	Incest, clips, blocked, watch private, fucked, girl, show	Graphic content	11%
6	السعودية، دعم، اليمن، ايران الداخلية، الرياض، الحوثيين، الجنوب	Saudi, support, Yemen, Iran, Interior, Riyadh, Houthis, south	Spam	10%

3.2 How Do Long-Lived Accounts Evolve?

In this section, we investigate the evolution of each of our four user groups from three perspectives: self-similarity, linguistic evolution, and behavioral evolution.

User Lifecycles. To compare the lifecycle of different accounts and since they all may land on different set of months, we split the lifespan of each user into 10 stages and distribute their activities over these stages [7]. A life-stage of 1% corresponds to birth and a life-stage of 100% corresponds to death (either getting suspended for the case of suspended users, or never tweeting for more than 2 months for legit users). We then can compare how different users evolved by comparing their corresponding life-stages regardless of the actual months they were active in and regardless of the real length of their lifetime.

Self-similarity. We first measure how much accounts mimic their own previous tweets. We expect legitimate normal accounts to be more innovative in their use of language while accounts of spammy nature may repeat themselves more often since they usually have a specific agenda or target to achieve. For this purpose we measure the lexical overlap between the set of words used in each tweet and the previous 10 tweets produced by the same account using Jaccard similarity. For example, given a tweet (t) after the tenth tweet and the previous 10 tweets t_j ($1 \leq j \leq 10$), the self-similarity SS at t is calculated as follows:

$$SS = \frac{1}{10} \sum_{j=1}^{10} \frac{|t \cap t_j|}{|t \cup t_j|}$$

Then, we average the calculated self-similarities at each stage in a user lifespan.

Figure 4a shows that short-lived accounts repeat themselves the most; these users clearly engage in the most copy-paste "extreme" behavior, indicating why

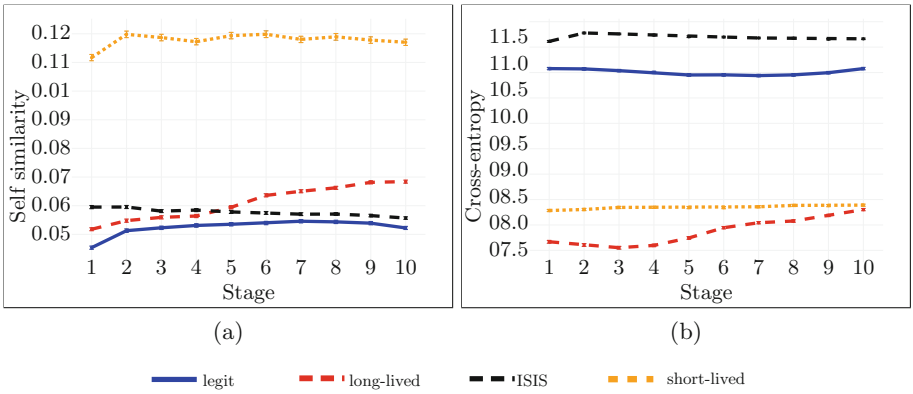


Fig. 4. Self-similarity and cross-entropy: long-lived accounts tend to repeat themselves more at their last stages (a) and also increasingly deviate more from the community (b), hinting that they may have been used as a "sleeper-cell" accounts. The usually extreme content of ISIS pushes them the furthest on the linguistic cross-entropy scale (b). Throughout this paper, error bars indicate the standard error of the mean.

they may have been detected so quickly. In contrast, the legitimate, long-lived, and ISIS accounts engage in much less self-similarity. The ISIS accounts behave most similarly to the legitimate accounts, suggesting that these pro-ISIS posters are mostly real users and not bots. And notice the uptick in self-similarity for the long-lived accounts; this finding offers evidence that these accounts engage in more repeated posting later in the lifespans. Perhaps they are "sleeper cells", who have behaved in a legitimate-like way for much of their lifespan before becoming "activated" and behaving in a more extreme (hence detectable) fashion.

Linguistic Evolution. The examination of self-similarity gives some hints that long-lived accounts are fundamentally different than our other user groups. Here, we extend this investigation to study how these four groups differ from the overall Twitter community over their lifetime. Do these accounts reflect the overall evolution of the Twitter community? Or is there a sign of "sleeper cell" behavior where accounts keep generating some common text and at some point start to post on a different topic? For this purpose we build a series of bigram language models [7] with Katz back-off smoothing [12] one for each month in 2015 that represents the overall background language used by all accounts in that month. Then we quantify how the language of an individual tweet (t) differs from the background language model (BLM) of the month (m) it was produced in by calculating the cross entropy between t and BLM_m:

$$H_t(t, \text{BLM}_m) = -\frac{1}{N} \sum_i \log P_{\text{BLM}_m}(b_i)$$

where b_i are the bigrams of t and $P_{\text{BLM}_m}(b_i)$ is the probability of b_i based on the corresponding month's language model. Cross entropy captures how surprising a

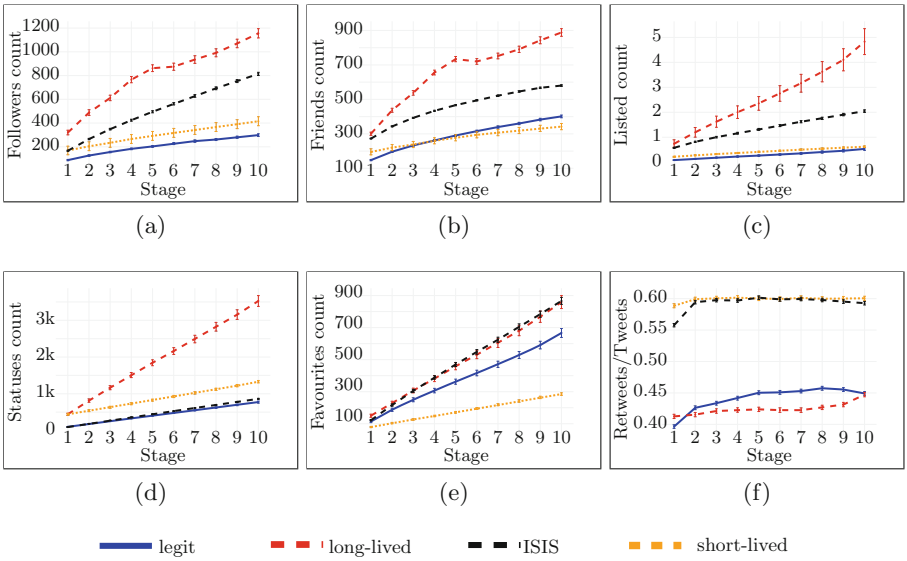


Fig. 5. Evolution of user features. Overall, long-lived accounts tend to be more active. On average, they have more followers (a), more friends (b), manage to get listed more (c), generate more tweets (d), favorite more tweets (along with ISIS accounts) (e), but they might have reduced their retweeting ratio to evade detection (f).

tweet t is with respect to the language used by the rest of the Twitter community: higher values indicate that a tweet differs more. We then calculate the average tweet cross entropy of a user (u) at each stage (i):

$$H_u(u, S_i) = \frac{1}{|S_i|} \sum_{t \in S_i} H_t(t, \text{BLM}_m)$$

where S_i is the set of all tweets generated by a user u at a stage i. Note that a stage may span multiple months but the cross-entropy for each tweet is calculated based on the month the tweet was posted in irrespective of the latent stage it ends up on.

Figure 4b shows that different groups start off at a certain distance from the overall community that basically defines their social character for the rest of their lifetime. We observe that ISIS accounts and legitimate accounts are linguistically the most innovative (high values of cross-entropy compared to the background language model). Reinforcing our finding that ISIS accounts engage in little self-similarity, these two findings suggest that ISIS accounts are managed by a sophisticated human-in-the-loop command-and-control with fundamentally different posting tactics than traditional spam accounts. We may also attribute the high cross-entropy for ISIS accounts to their unique extreme messaging which may be rejected by the majority of other users (the background language model) making them appear more "surprising" and therefore furthest

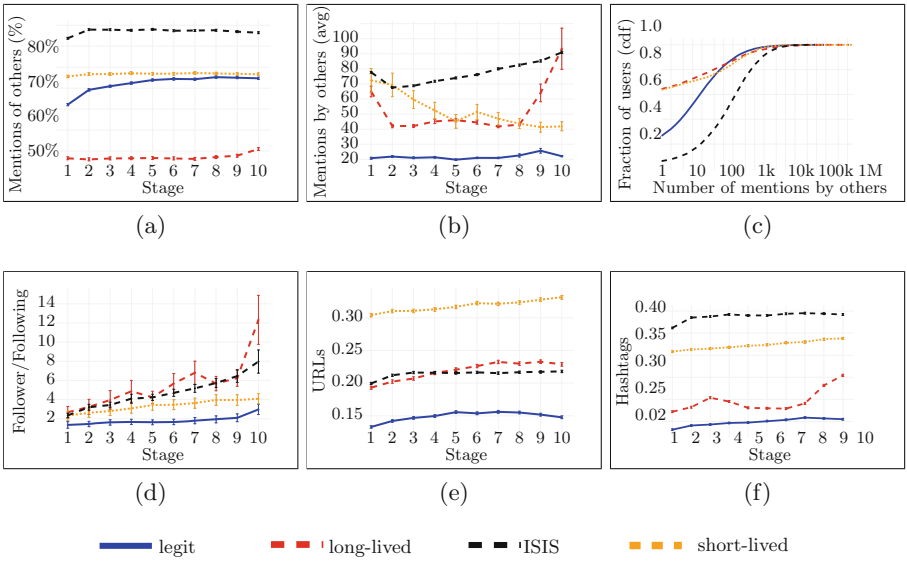


Fig. 6. Evolution of user features cont'd. (a) ISIS accounts are mainly used for interacting with others with 85% of their tweets having at least one mention; long-lived are less interactive. (b) Long-surviving accounts get activated and exposed to the Twitter community in the last stages of their life resulting in their suspension. (c) ISIS accounts do a good job getting high fraction of their accounts mentioned by the community; about 60% of long-lived and short-lived accounts are never mentioned at all. (d) Over time, ISIS accounts improve their follow-back ratio as opposed to long-lived accounts who greatly fluctuates over time - a sign of manipulation. Short-lived accounts spread more URLs (e) and hashtags (f) than other groups. Long-lived accounts engage in increased hashtag sharing (f) in the last stages of their lives.

from the Twitter community. Both long-lived and short-lived accounts are linguistically the least innovative, giving some counter-evidence to our suspicion that long-lived accounts are engaging in a "sleeper cell" behavior. Indeed, the long-lived accounts are the least innovative throughout their lifetimes. This indicates that these accounts, though surviving for a long time, may reveal clear signals that could be used for early detection.

Activity Evolution. Finally, we study the activity signals, other than text, over the lifetime of these four different groups. How sharply does a group gain (or lose) followers? does the community lose (or gain) interest in a group more than others over time? Overall, as shown in Fig. 5, we find that long-lived accounts tend to be more active compared to the other groups in our study. For example, they have more followers, more friends, get listed more often, and generate more tweets. However, long-surviving accounts avoid excessive retweeting possibly a detection evading strategy. In addition, Fig. 6d shows that the followers/following ratio of long-lived and ISIS accounts tend to increase over time indicating that they succeed in getting more users to follow them back. However, the

followers/following ratio of long-surviving accounts fluctuates hinting at a potential use of fake accounts pool constantly suspended by Twitter in bulk and hence the sudden decrease. ISIS accounts show a much stable curve another sign that those accounts are managed differently. We also notice a sudden interest by the Twitter community in the long-lived accounts towards the end of their life (Fig. 6b) indicating that those accounts were activated and also exposed to detection and hence got suspended. Figure 6c shows that the ratio of ISIS accounts that get mentioned by other users is higher than any other group indicating a better outreach compared to long-lived accounts where 60% of their accounts never get mentioned.

Turning to the activity patterns of the tweets themselves, we see that on average short-lived accounts use more URLs than any other group, indicating a common theme of using the platform to spread external content (Fig. 6e). In contrast, ISIS accounts use more hashtags (Fig. 6f) than any other group and make more use of mentions (Fig. 6a) showing a preference of interacting directly with others rather than communicating one-way, potentially an effort to spread agenda or seek support and sympathy from the community.

4 Automatic Long-Survivors Detection

This initial investigation has highlighted several dimensions in which long-lived content polluters differ from traditional (short-lived) spammers, legitimate users, and even highly-focused pro-ISIS users. We are now in a position to apply these findings by building machine-learned classifiers to automate some important decisions revolving around these suspicious accounts. We consider the following questions about long-surviving and pro-terrorism accounts:

1. Is an account a long-surviving suspicious account?
2. Is an account pro-terrorism?
3. Is a long-surviving account also pro-terrorism?
4. Are long-surviving and pro-terrorism accounts sleeper cells?

Features. Previous studies have evaluated different features and their effectiveness in detecting overall spam accounts on Twitter [1,5,10,14,22,25,27]. We consider the following three types of features (See Table 2 for more details):

- **Language features:** Our results have shown that cross-entropy of the language of an account against the overall Twitter community and self-similarity of an account are two powerful differentiators of the different types making them first candidate to use for our classification tasks.
- **Behavior:** We have also seen that mentions by others is a good discriminator so we use it in addition to the follow-back ratio.
- **Content:** We noticed how long-surviving accounts use hashtags less and more URLs so we add these two features for comparison purposes.

Table 2. Classification features

Feature	Group	Source	Description
Cross-entropy	Language	Ours	How surprising is the language of an account?
Self-similarity	Language	Ours	How much does an account repeat itself?
Interaction (self)	Behavior	[25]	Tweets with mentions/all tweets
Interaction (community)	Behavior	ours	Mentions by user/mentions of user by others
Follow-back ratio	Behavior	[14]	Followers/following
Hashtags ratio	Content	[25]	Tweets with hashtags/all tweets
URLs ratio	Content	[14]	Tweets with URLs/all tweets

Classification Algorithm. We experimented with a variety of classification algorithms available in the Apache Spark machine learning package [17,28] - logistic regression, decision trees, naive Bayes, and random forests - and found the latter to work best. Hence all results reported here were obtained using random forests [4] implementation available in Apache Spark.

We use balanced training and test sets containing equal numbers of positive and negative examples, so random guessing results in an accuracy and area under the receiver operating characteristic (ROC) curve (AUC) of 50%. We performed all classification experiments using 10-fold cross validation.

4.1 Classification Tasks

We answer the above questions with the following specific three tasks along with a fourth orthogonal task that evaluates tasks performances as accounts evolved throughout their lifetime.

Task 1: Is an Account a Long-Surviving Account? Here the objective is to detect if a given Twitter account is currently running under the radar and should be suspended. Is it possible to distinguish a long-surviving account from other suspicious groups and also from normal users?

Here the set of positive examples consists of all 17,909 long-surviving accounts and the negatives examples are the 17,909 legit active accounts randomly sampled from the much larger set of legitimate accounts.

Task 2: Is an Account Pro-extremism? This is similar to Task 1 but here we would like to find out if accounts such as ISIS supporters are different from normal accounts.

Here the positive examples are the 17,518 community reported ISIS accounts and the negative examples are 17,518 randomly sampled legitimate accounts.

Task 3: Is a Long-Surviving Account also Pro-terrorism? This should give some insight on how ISIS operates its social networking campaigns. Do they hire the same teams that might be operating other long-surviving accounts? Or do they have their own methods?

Similarly to Task 2, we use our set of 17,518 ISIS accounts as the positive examples but we now sample 17,518 long-surviving accounts from the 17,909 long-surviving accounts.

How Early Can We Classify Accounts? Finally, we add an orthogonal task to the above three tasks that considers the evolution of accounts over their lifetime. Does detection accuracy improve as accounts progress in their lifespan? i.e. has an account always be suspicious or was there a moment when it turned bad and hence got suspended? Did an ISIS account exist for the purpose of supporting terrorism or was it recruited after being a normal account for some time.

Here we use the same sets of positive and negative examples set up for each of the first three tasks but we now train a series of classifiers, one for each stage of the accounts' lifespan. Does classifier performance improve as we learn more about the accounts?

4.2 Results

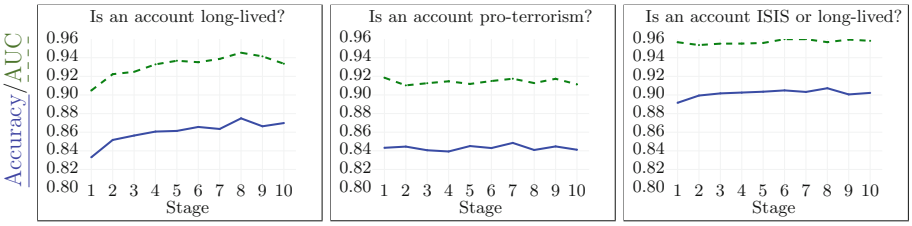
Table 3 reports the performance on our first three tasks when using all the features. Surprisingly distinguishing ISIS accounts from other long-surviving accounts (Task 3) is the easiest task, with an accuracy (AUC) of 88% (94%), easier than distinguishing them from legitimate users (Task 2), yet another indicator that ISIS supporters are being operated at a different level and are slightly more successful at appearing normal. Distinguishing long-surviving accounts (Task 1) and ISIS supporters (Task 2) from legitimate accounts is equally challenging to our classifier with accuracy of 85% (93%) and 86% (92%) respectively.

Table 3. Classification results for all tasks.

Task	Accuracy	AUC
Is a Twitter account a long-surviving account?	85%	93%
Is a Twitter account pro-extremism?	86%	92%
Are ISIS accounts different from other long-surviving accounts?	88%	94%

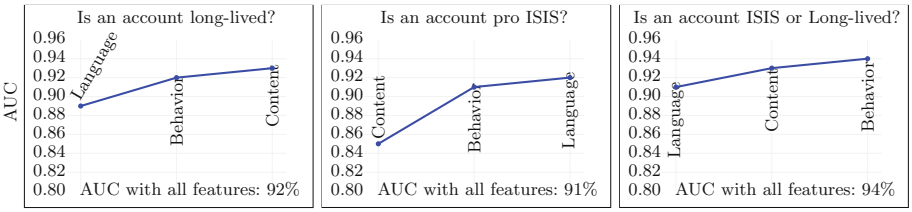
Figure 7 shows performance results for all tasks by considering the whole lifetime of accounts. We notice that all tasks can be answered with an accuracy of 82% or more by only knowing the first few weeks (10% of lifespan) of an account. Interestingly we also see that there are no signs of recruitment of ISIS supporters (i.e. most ISIS accounts existed for that purpose from the beginning). This is evident in the classifier's ability to detect ISIS supporters early in the accounts lifetime. It's also interesting to see that except for Task 1 the classifier didn't improve by knowing more historical information about long-surviving accounts strongly suggesting that most of those accounts have been hiding in plain sight!

Feature Importance. In order to understand which features are important for which task, we evaluate smaller models that consist of only one of the three feature groups (Table 2). We used feature forward selection strategy to order



(a) Task 1 (b) Task 2 (c) Task 3

Fig. 7. Tasks performance over accounts lifetime. All classification tasks achieved 82% or better from even at the first few weeks of accounts creation strongly suggesting such accounts have always been suspicious. Distinguishing long-lived accounts from legit accounts slightly improves over accounts lifetime (a) while other tasks remain almost the same. (c) Better classification performance hints that long-surviving ISIS accounts are very different from other long-surviving accounts.



(a) Task 1 (b) Task 2 (c) Task 3

Fig. 8. Results of forward feature selection for first three tasks. Different feature groups contribute their share with language group being the most discriminative most of the time.

feature groups by importance. Figure 8 shows the results for the three tasks. The conclusion is that all feature groups contribute their share, but with diminishing returns and that the discriminating power of different feature groups changes based on the task with language being the most discriminative most of the time.

5 Conclusion

We have explored a unique subset of content polluters on social networks, namely long-surviving suspicious accounts on Twitter. We utilized large-scale private access to all Arabic tweets generated in 2015 to study this group, comparing it to normal Twitter users, pro-terrorism users, and the prevalent short-lived spam. We found that the majority of these long-lived accounts have been successfully evading detection for long time mainly by avoiding behaviors that lead to detection such as mainly posting URLs or participating in many trending hashtags. We uncovered characteristic differences in terms of linguistic character, self-similarity, and other behavioral signals.

We have also exploited a labeled dataset of more than 25k ISIS accounts to further investigate the role of those long-surviving accounts in terrorism. We found that ISIS users are quite different from other long-surviving content polluters and that they are hungry for interaction with the Twitter community rather than a one-way communication style adopted by other spam accounts, giving insight into ISIS's sophisticated use of social media. However, we similarly find that most of these accounts have been ISIS supporters for the majority of their life and are hence detectable early in their lifetime.

We relied on our findings to build an automatic classification system to determine whether an account is a long-surviving or a pro-terrorism. By combining features from the language these accounts use, the content they post, and their behavioral signals we were able to classify users with an accuracy of 84% or more and AUC/ROC of 91% or more. Furthermore, knowing only the first 10% of an account lifetime, we were able to classify an account with 82% accuracy and 90% AUC/ROC strongly suggesting many of these accounts could have been detected earlier.

We are eager to extend this work to explore the community structure underlying these long-lived users – do we find tight clusters of inter-connected users with similar patterns of linguistic evolution? Although we have found that the majority of accounts are created for a purpose that doesn't change, we are still exploring methods to identify "change points" in the lifespan of a small fraction of users that have shown signs of bad evolution to identify users whose behaviors are shifting (e.g., to find ISIS sympathizing accounts).

In this work, we only focused on Arabic Twitter and we only considered the Twitter platform utilizing our access to all Arabic tweets in 2015. However, we are yet to find if our results will transfer to different platforms and different demographics.

Acknowledgments. This work was supported in part by AFOSR grant FA9550-15-1-0149. Majid Alfifi is partially funded by a scholarship from King Fahd University of Petroleum and Minerals. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors. We'd like to also thank the anonymous reviewers for their helpful feedback.

References

1. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on Twitter. In: Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS), vol. 6, p. 12 (2010)
2. Berger, J.M., Morgan, J.: The ISIS Twitter census: defining and describing the population of ISIS supporters on twitter. In: The Brookings Project on US Relations with the Islamic World, vol. 3, no. 20 (2015)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

5. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684. ACM (2011)
6. Conover, M., Ratkiewicz, J., Francisco, M.R., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on Twitter. ICWSM **133**, 89–96 (2011)
7. Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., Potts, C.: No country for old members: user lifecycle and linguistic change in online communities. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 307–318. International World Wide Web Conferences Steering Committee (2013)
8. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botornot: a system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273–274. International World Wide Web Conferences Steering Committee (2016)
9. Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A., Galstyan, A.: Predicting online extremism, content adopters, and interaction reciprocity. In: Spiro, E., Ahn, Y.-Y. (eds.) SocInfo 2016. LNCS, vol. 10047, pp. 22–39. Springer, Cham (2016). doi:10.1007/978-3-319-47874-6_3
10. Grier, C., Thomas, K., Paxson, V., Zhang, M.: @ spam: the underground on 140 characters or less. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, pp. 27–37. ACM (2010)
11. Hu, X., Tang, J., Zhang, Y., Liu, H.: Social spammer detection in microblogging. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)
12. Katz, S.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. Acoust. Speech Sig. Process. **35**(3), 400–401 (1987)
13. King, G., Pan, J., Roberts, M.E.: How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. Harvard University (2016)
14. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 435–442. ACM (2010)
15. Lin, P.C., Huang, P.M.: A study of effective features for detecting long-surviving twitter spam accounts. In: 2013 15th International Conference on Advanced Communication Technology (ICACT), pp. 841–846. IEEE (2013)
16. Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., et al.: The Arab spring—the revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int. J. Commun. **5**, 31 (2011)
17. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al.: Mllib: machine learning in apache spark. J. Mach. Learn. Res. **17**(34), 1–7 (2016)
18. Mustafaraj, E., Metaxas, P.T.: From obscurity to prominence in minutes: political speech and real-time search. In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line (2010)
19. Ratkiewicz, J., Conover, M., Meiss, M.R., Gonçalves, B., Flammini, A., Menczer, F.: Detecting and tracking political abuse in social media. ICWSM **11**, 297–304 (2011)
20. Song, J., Lee, S., Kim, J.: Spam filtering in Twitter using sender-receiver relationship. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) RAID 2011. LNCS, vol. 6961, pp. 301–317. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23644-0_16

21. Starbird, K., Palen, L.: (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 7–16. ACM (2012)
22. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1–9. ACM (2010)
23. Thomas, K., Grier, C., Paxson, V.: Adapting social spam infrastructure for political censorship. In: LEET (2012)
24. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of Twitter spam. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 243–258. ACM (2011)
25. Wang, A.H.: Don't follow me: spam detection in Twitter. In: Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT), pp. 1–10. IEEE (2010)
26. Wei, W., Joseph, K., Liu, H., Carley, K.M.: The fragility of Twitter social networks against suspended users. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 9–16. ACM (2015)
27. Yang, C., Harkreader, R.C., Gu, G.: Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) RAID 2011. LNCS, vol. 6961, pp. 318–337. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23644-0_17
28. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, p. 2. USENIX Association (2012)