Revealing the Blackmarket Retweet Game: A Hybrid Approach

Shreyash Arya(✉) and Hridoy Sankar Dutta

Indraprastha Institute of Information Technology, Delhi, India
{shreyash15097,hridoyd}@iiitd.ac.in

Abstract. The advent of online social networks has led to a significant spread of important news and opinions. In the case of Twitter, the popularity of a tweet is measured by the number of retweets it gains. A significant number of retweets help to broadcast a tweet well and makes the topic of the tweet popular. Individuals and organizations involved in product launches, promotional events, etc. look for a broader reach in their audience and approach blackmarket services. These services artificially provide a gain in retweets of a tweet as the retweets' natural increase is difficult and time-consuming. We refer to such tweets as collusive tweets. Users who submit their tweets to the blackmarket services gain artificial boosting to their social growth and appear credible to the end-users, leading to false promotions and campaigns. Existing methods are mostly centered around the problem of detection of fake, fraudulent, and spam activities. Thus, detecting collusive tweets is an important yet challenging problem that is not yet well understood.

In this paper, we propose a model that takes into account the textual, retweeters-centric, and source-user-centric characteristics of a tweet for an accurate and automatic prediction of tweets submitted to blackmarket services. By conducting extensive experiments on collusive tweets' real-world data, we show how our model detects tweets submitted to blackmarket services for collusive retweet appraisals. Moreover, we extract a meaningful latent representation of collusive tweets and their corresponding users (source users and retweeters), leading to some exciting discoveries in practice. In addition to identifying collusive tweets, we also analyze different types of collusive tweets to evaluate the impact of various factors that lead to a tweet getting submitted to blackmarket services.

Keywords: Collusion detection · Classification · Twitter

1 Introduction

Online media leads the current age of information (specifically the online social networks), being a significant source of daily content dispersion and consumption. It has been perceived as having both positive and negative impacts in various domains such as politics, organizations, governments, content creation,

T. Chakraborty et al. (Eds.): CONSTRAINT 2021, CCIS 1402, pp. 30–41, 2021.
https://doi.org/10.1007/978-3-030-73696-5_4

source of information news, business, and health care [1]. It has been driving the contemporary society where people are open to publicly (privately as well) share their opinions and become influential and popular in terms of social media currency such as likes, comments, subscribers, shares, and views on these platforms. Having reach to a wider audience leads to monetary benefits, better listing in recommendations, and even influencing and polarizing the significant issues such as political outcomes. To gain popularity, individuals and organizations have been using blackmarket services which helps boost the reach of the content artificially (in terms of social currency). This inorganic behavior affects social media's organic behavior, driving people's attention to artificial boosting of social reputation which is known as *collusion*.

All online platforms such as social networks, rating/review platforms, video streaming platforms, recruitment platforms, discussion platforms, music sharing platforms and development platforms are susceptible to blackmarket/collusive activities and being collusively affected by boosting the appraisals present in the platforms artificially. Entities present in blackmarket services shows both organic and inorganic behaviors. These are humans only employed by these services and hence challenging to track down by the already present literature on social bots detection, fake/spam detection, anomaly detection etc. but still being closely related [2–10]. There are two types of blackmarket services: *Premium* and *Freemium*. *Premium services* are the paid services with customers and suppliers, whereas *freemium services* are barter-based services where customers are also suppliers for other customers [11].

There have been attempts to detect these collusive identities on various social platforms such as Twitter and YouTube by employing majorly feature-based, graph-based, and deep learning-based approaches [12–18]. However, collusive entity detection is still in its infancy due to the unusual behavior exhibited by them. Collusive users perform collusive activities in an asynchronous manner.

This paper devises a hybrid feature-based model that uses user features, tweet features, user-user interaction features and user-tweet interaction features for collusive tweet detection. We further analyze and detect a potential core group of collusive users. Section 2 discusses the dataset; Sect. 3 describes the modeled framework. Finally, Sect. 4 contains all the experiments conducted. We conclude the paper in Sect. 5.

2 Dataset

The data is the main success of this task as the datasets from the blackmarket services are neither publicly available nor have official APIs to fetch the data. In the case of Twitter, API is publicly available to fetch the data, but it has several rate limits. We collected the data using the official Twitter API and a customized web scraping tool to collect data from the blackmarket services.

The tweet and user ids were gathered from blackmarket services, which denotes the collusive sets. Using these ids, metadata and timeline information were extracted from Twitter. Specifically, we extracted the text present inside the tweet, tweet metadata such as retweets count, retweeter ids, retweeters timeline data, tweeters timeline data, and temporal data of tweets and retweets. For the genuine users set, the data was collected from the verified accounts (following [23]) on Twitter. Note that only English tweets were extracted using the `‘lang’` parameter in the API and later manually verified[1].

For optimizing the data collection process, we used the Parallel version of the well-known Tweepy framework (a framework to collect data from Twitter)[2]. The final dataset contains 1539 collusive and 1500 genuine tweets. For the user-user and user-tweet interactions, 13,000 collusive and 13,000 genuine users are considered. The whole user-user interaction matrix is used as an input adjacency matrix for the graph-based analysis, and a subset of 3,000 (randomly selected) equal collusive and genuine users is used in the classification task.

3 A Hybrid Detection Framework

In a blackmarket service, the tweets are submitted to gain popularity by increasing their retweets, which helps the tweet to broadcast well. The blackmarket (freemium) works based on a barter system where a user earns credit by retweeting other users' tweets who have submitted their tweets on the blackmarket website and use the earned credits to buy blackmarket services for themselves. Hence, due to earning credits' greed, the users show erratic behavior (collusive behavior) that is not demonstrated by a genuine user (who has not submitted the tweet to the blackmarket service to gain credits). So, we aim to predict whether a tweet is collusive or not using tweet and source (users and retweeters) indicators.

3.1 Indicators

Here, we present the indicators for our classification model which is composed of the following two parts: (i) tweet-level indicators, and (ii) source (users and retweeters) indicators[3].

Tweet Indicators. These indicators capture the implicit features of tweets submitted to blackmarket services:

Retweet Count: This indicator captures the most fundamental aspect of tweets submitted to blackmarket services. The change in the retweet count is observed as the tweets are forwarded to any blackmarket service. If the retweet count of a tweet increases by more than 99 retweets on the same day, the indicator is marked as one (else zero).

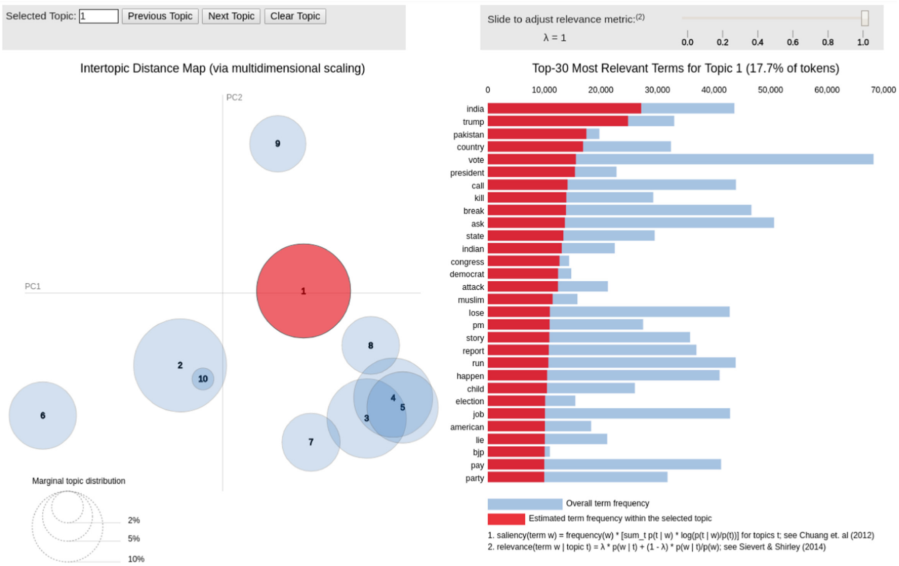
[1] The data is manually verified and validated by three experts in the domain.

[2] https://github.com/shrebox/Parallel-Tweepy.

[3] Indicators are parameterized at best values found after experimentation.

Tweet2vec: This indicator is generated using the publicly available Tweet2vec encoder [19], which encodes a tweet's character level embedding into vector-based representation. This feature helps capture tweets' linguistic intricacies, which contain out-of-vocabulary words and unusual character sequences.

LDA Similarity: This indicator captures the random retweeting behavior of blackmarket users who retweet to earn blackmarket credits. The retweeters of a collusive tweet are random users, i.e., are not in the source tweeter's follower-followee network. These collusive retweeters also have different tweet content (diverse topical interest) from the source tweeter. The timeline of all the retweeters is compared with the source tweeter's timeline using the similarity between the topics discussed. Latent Dirichlet Allocation (LDA) extracts the distribution of topics among the timeline tweets and represent them in term of vectors. Finally, the cosine similarity scores between LDA generated vectors is used as a threshold (kept as a parameter at 0.25). If the content matches above this threshold, the tweet is marked non-collusive (0); else, it is marked collusive (1).



Selected Topic: 1

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 1. Visualization of LDA modeled topics of a tweeter's timeline content: the circle represents the topics, the area of the circle defines the importance of the topic in the entire corpus, and the distance between the centers of circles represents the similarity between topics. Image on the right side shows the top-30 relevant terms for the selected topic.

Source Indicators. These indicators capture the user's retweeting behavior and their interactions:

Retweeter Aggression: This indicator is used to capture the users' greedy aspect where a user in blackmarket service tries to increase his/her credit by retweeting other users' tweets in that service. Tweets are extracted from the retweeter's timeline and are marked (0 or 1) as collusive if the 50% retweeters of tweet retweet more than 50 times in the time frame (t − 2) days to the current day.

Top Collusive Tweets: This indicator aims to capture the credit-based barter system of blackmarket services. The blackmarket user will try to retweet as many tweets as possible quickly to gain the credits used in blackmarket services. Hence, the top tweets on a collusive user's timeline should be populated mostly by the tweets that belong to the blackmarket service. The tweet id of the top tweets in a user's timeline is checked and marked as collusive (0 or 1) if 80% top are present in the blackmarket service's database.

User-User Interaction Matrix: The user-user interaction matrix is a 2D matrix with users on both rows and columns. Each cell consists of the frequency of retweets that a user has done to another user's tweet. This matrix captures the retweet interaction behavior between the users.

4 Experiments

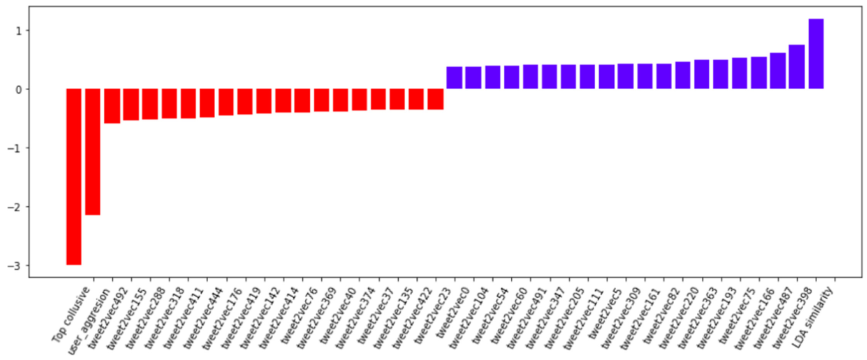


Fig. 2. Top 20 most contributing features for both ends of classification with feature weight on the Y-axis (.coef_ parameter) and feature name on the X-axis: Linear SVM

The experiments below are divided into three sections. The first section, 4.1, presents the details and results for the classification model designed using the previously mentioned custom features. In Sect. 4.2, a quantitative analysis of the collusive tweets and users is performed to give a high-level overview of our assumptions and results. The last section, 4.3, takes a graph-based approach to detect the blackmarket service's core-members of the blackmarket service contributing to such collusive networks' effective working.

4.1 Classification

The indicators mentioned in previous sections are used as feature input vectors to the supervised classifiers to detect collusive tweets. For classification and evaluation: Linear SVM, Thresholded[4] R^2 scores, Logistic Regression, Gradient-boosted Decision Trees (XGBoost), and Multi-layer Perceptron (MLP) from the scikit-learn package are used considering individual features and combinations of all the features. Default parameters for all the classifiers are used except MLP with three hidden layers (150, 100, and 50) and max iterations of 300. All the features are concatenated together (6,039 rows and 503 columns) and trained on a 70-30 train-test split. Also, the user-user interaction matrix is reduced using TruncatedSVD as done in [20].

Table 1. Classification test accuracy scores

Features	Linear SVM	Thresholded R^2	Logistic Reg.	XGBoost	MLP
Tweet2vec	0.784	0.791	0.805	0.805	0.800
Retweet aggression	0.772	0.811	0.811	0.777	0.783
Top collusive tweet	0.762	0.777	0.765	0.752	0.783
LDA similarity	0.745	0.743	0.761	0.745	0.772
User-user interaction	0.964	0.965	0.959	0.967	0.975
Combined (expect user)	0.920	0.891	0.936	0.945	0.923
Combined (total)	0.961	0.963	0.961	0.962	0.974

Table 2. Classification metrics (Precision, Recall, F1-score; Macro)

Classifiers	Except interaction matrix	Interaction matrix	Combined
Linear SVM	0.92, 0.92, 0.92	0.97, 0.96, 0.96	0.96, 0.96, 0.96
Thresholded R^2	0.89, 0.89, 0.89	0.97, 0.96, 0.97	0.96, 0.96, 0.96
Logistic regression	0.94, 0.93, 0.94	0.96, 0.96, 0.96	0.97, 0.96, 0.96
XGBoost	0.95, 0.94, 0.94	0.97, 0.97, 0.97	0.97, 0.96, 0.96
MLP	0.93, 0.92, 0.92	0.98, 0.96, 0.97	0.97, 0.97, 0.97

Results. Classification accuracy on the test set are shown in Table 1. Table 2 contains the values for the classification metrics - Precision, Recall, and F1-score (macro scores are reported). Also, Fig. 2 shows the feature importance as predicted by the SVM classifier[5]. As compared to the binary classification

[4] A decision threshold of 0.5 on the regressed R^2 score from linear regression is used for predicting the labels (0 or 1).

[5] SVM is shown due to comparable accuracies with other classifiers; MLP performs the best, but due to underlying neural network-based architecture, it does not have intrinsic feature importances rather complex network weights.

metrics (macro) given in Table IV of [12], our combined feature set are able to correctly classify the two classes. It shows how selecting a hand-picked feature set can help capture the inherent collusive signals. Also, the MLP classifier with an underlying three-layered neural network works better than other supervised classifiers in most cases. It shows how the network captures inherent feature structures and with better fine-tuning, it can help achieve better classification accuracy. Although, more data will be required to work with neural network-based architectures.

4.2 Quantitative Analysis

Retweet Count Change Pattern. The increasing saw-tooth behavior is captured (Fig. 3a) when we analyze the changing pattern in the retweet count of tweets submitted to a blackmarket service. It also shows the users' aggressive behavior of blackmarket service users to retweet the other user's tweets to gain credit in the network. In Fig. 3a, the x-axis denotes the timeline (8 h/unit), and the y-axis represents the retweet count change. It is extracted as a feature and used in classification (Check Sect. 3.1 - Tweet indicators).

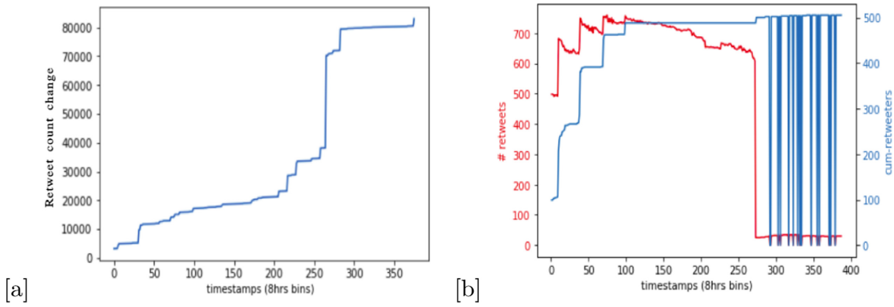
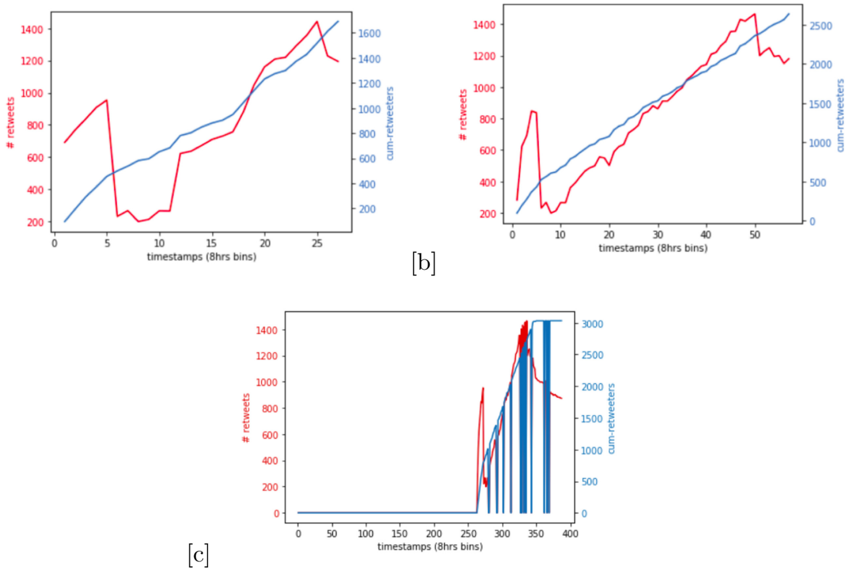


Fig. 3. (a) Retweet count change over time when tweet is submitted to a blackmarket service and (b) Number of retweets and cumulative retweeters over different timestamps.

Change in Retweeters in Two Blackmarket Services. Two blackmarket services, Like4Like (L4L) and YouLikeHits (YLH), are considered for this analysis. Retweeters with their tweets appearing in both the networks are considered. The subset of the dataset considered for this analysis contains 22,612 L4L users, 42,203 YLH users, and 10,326 intersecting users from both the networks.

Figure 3b shows that, in general, the retweet count decreases over the period, and the cumulative retweeters increase. The decrease indicates the case of deletion of the retweets after getting the credit from blackmarket services.

A similar trend can be noticed in all cases of Fig. 4, which shows the same tweet submitted to both the blackmarket services (YLH and L4L). Also, it can be seen that for the period when the cumulative retweeters remain constant or increase slowly, the number of retweets decreases. When there is a steep increase in retweeters, retweets increases accordingly.



[a] [b]

[c]

Fig. 4. (a) YouLikeHits (YLH), (b) Like4Like (L4L) and (c) YLH + L4L combined: Number of retweets and cumulative retweeters over different timestamps for the same tweet.

Discontinuity Score. Discontinuity is defined as to retweet tweets in discrete-continuous time frames by a blackmarket user, not to be captured or flagged as bots by the Twitter system. Users with a gap in retweeting days with a retweeting threshold above 50 retweets per day are considered in the analysis.

In Fig. 5, x-axis denotes the number of days after which the collusive user retweeted the blackmarket tweets, and the y-axis shows the fraction of such retweeters. The maximum collusive users retweeted after a week with a 0.35 fraction of such retweeters. It shows how the users try to evade the generic retweeting pattern and remain unfiltered from automated bot detection systems.

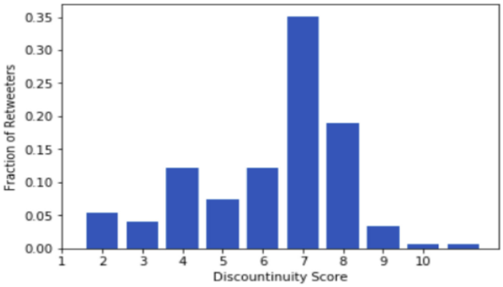


Fig. 5. Fraction of retweeters vs. discontinuity score.

4.3 Graph-Based Analysis

The user-user retweeting interaction adjacency matrix is used to generate the analysis graph. The graph's nodes correspond to the collusive users, and the edge is formed if one user retweets another user's tweet. The weight on edge is the number of retweets shared between users.

Core Component Analysis. In a blackmarket network, the core users are the fuel on which the network runs. These users contribute towards the major collusive retweeting behavior and are more prone to give away erroneous signals such as bot behavior. Bots often imitate or replace a human user's behavior. Typically they do repetitive tasks, and they can do them much faster than human users could. Hence, we did a focussed analysis of core component detection using k-core algorithm [21] and bot analysis using Botometer [22][6].

A k-core is a maximal subgraph that contains nodes of degree at least k. It's a recursive algorithm that removes all the nodes with a degree less than k until no vertices are left. K-core decomposition identifies the core user groups in the input network. The NetworkX package[7] is used to find the central core, which is the largest node degree subgraph possible with k values: Like4Like - 2635 and YouLikeHits - 2352. These central cores were extracted using the collusive user-user interaction (retweet) network as input and further analyzed for bot behavior.

Figure 6 shows that most users from the core-component lie in the bin with a 75–100% bot behavior bin range. It validates our claim that core-collusive users tend to show-bot behavior. Figure 7 shows an example of a suspected user account analyzed using Botometer. The different features such as temporal, network, and language-independent scores high on the bot score with an overall 4.3 out of 5, indicating the bot behavior.

[6] https://botometer.osome.iu.edu/.

[7] https://networkx.github.io/documentation/stable/reference/algorithms/core.html.

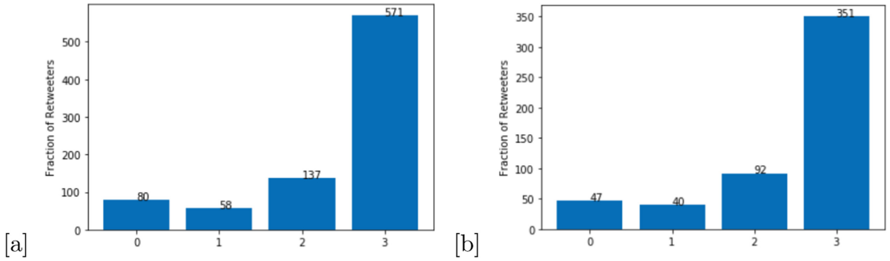
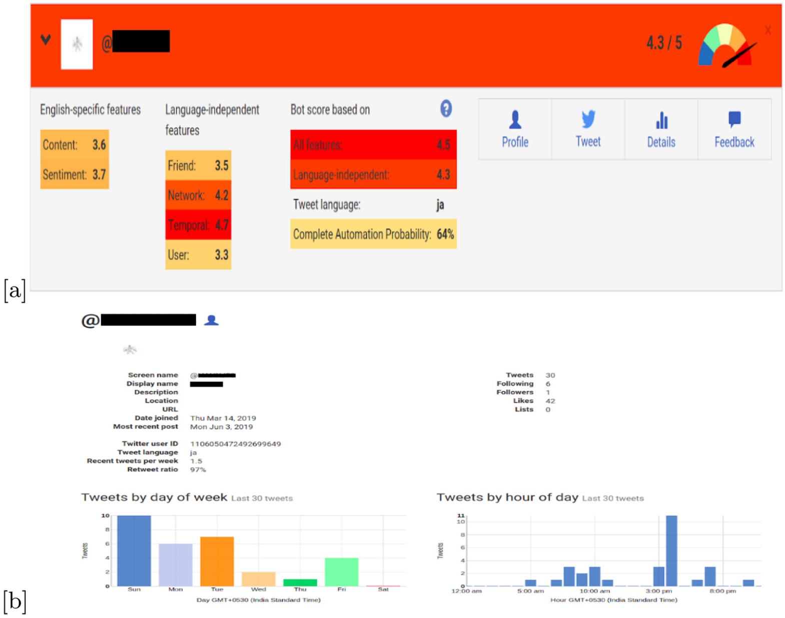


Fig. 6. (a) YouLikeHits (YLH), (b) Like4Like (L4L): Fraction of retweeters from k-core maximal subgraph vs. bins indicating the Botometer score ranges. 0: 0–25%, 1: 25–50%, 2: 50–75%, 3: 75–100%.



@[redacted] 4.3 / 5

English-specific features

Content: 3.6
Sentiment: 3.7

Language-independent features

Friend: 3.5
Network: 4.2
Temporal: 4.7
User: 3.3

Bot score based on

All features: 4.5
Language-independent: 4.3
Tweet language: ja
Complete Automation Probability: 64%

@[redacted]

Screen name @[redacted]
Display name [redacted]
Description
Location
URL
Date joined Thu Mar 14, 2019
Most recent post Mon Jun 3, 2019

Twitter user ID 1106050472492699649
Tweet language ja
Recent tweets per week 1.5
Retweet ratio 97%

Tweets 30
Following 6
Followers 1
Likes 42
Lists 0

Tweets by day of week Last 30 tweets

Tweets by hour of day Last 30 tweets

Fig. 7. Botometer analysis of a suspicious account (name of the user redacted).

5 Conclusion and Future Work

Online media platforms have become the primary source of information and hence susceptible to fall prey to malicious activities. In the race of becoming more popular and influential on these platforms, the individuals and organizations have started artificially gaining an unfair advantage for their social growth in terms of likes, comments, shares, and subscribers, using blackmarket services. This act is known as collusion, and activities are known as collusive activities as mentioned by [12]. This paper aims to discuss a hybrid approach to detect such collusive retweeting behavior on Twitter and further check its impacts on social networks' organic working. For the detection, features engineered using the tweets, users, user-user interactions, and user-tweet interactions are fed as input to supervised classifiers. Very high accuracy of around 97% and F1-score of 0.9 on the test set for binary detection is achieved by combining the intricate features. These results may contain bias, and hence further quantitative and graph-based analyses are performed, which proves our detection claims. Also, a novel dataset has been curated using a custom optimized data extraction pipeline for the task. Future directions can increase the dataset size and use deep-learning-based classification mechanisms to eliminate any present bias. The core user components detected by the k-core decomposition can be further analyzed and used to detect the core users in the collusive network, which drives the blackmarket services.

References

1. Arya, S.: The influence of social networks on human society (2020). https://doi.org/10.13140/RG.2.2.18060.54408/1
2. Ross, B., et al.: Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. Eur. J. Inf. Syst. **28**, 394–412 (2019)
3. Stieglitz, S., Brachten, F., Ross, B., Jung, A.K.: Do social bots dream of electric sheep? A categorisation of social media bot accounts (2017)
4. Ross, B., et al.: Social bots in a commercial context – a case study on SoundCloud. In: Proceedings of the 26th European Conference on Information Systems (ECIS2018) (2018)
5. Bruns, A., et al.: Detecting Twitter bots that share SoundCloud tracks. In: Proceedings of the 9th International Conference on Social Media and Society (SMSociety 2018), pp. 251–255. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3217804.3217923
6. Sharma, A., Arya, S., Kumari, S., Chatterjee, A.: Effect of lockdown interventions to control the COVID-19 epidemic in India. arXiv:2009.03168 (2020)
7. Guille, A., Favre, C.: Mention-anomaly-based event detection and tracking in Twitter. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, pp. 375–382 (2014). https://doi.org/10.1109/ASONAM.2014.6921613
8. Liu, Z., Huang, Y., Trampier, J.R.: Spatiotemporal topic association detection on tweets. In: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPACIAL 2016), pp. 1–10. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2996913.2996933. Article 28
9. Fani, H., Zarrinkalam, F., Bagheri, E., Du, W.: Time-sensitive topic-based communities on Twitter. In: Khoury, R., Drummond, C. (eds.) AI 2016. LNCS (LNAI), vol. 9673, pp. 192–204. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-34111-8_25
10. Mukherjee, A., Liu, B., Wang, J., Glance, N., Jindal, N.: Detecting group review spam. In: Proceedings of the 20th International Conference Companion on World Wide Web (WWW 2011), pp. 93–94. Association for Computing Machinery, New York (2011). https://doi.org/10.1145/1963192.1963240

11. Dutta, H.S., Chakraborty, T.: Blackmarket-driven collusion on online media: a survey (2020)
12. Dutta, H.S., Chetan, A., Joshi, B., Chakraborty, T.: Retweet us. Spotting collusive retweeters involved in blackmarket services, we will retweet you (2018)
13. Chetan, A., Joshi, B., Dutta, H.S., Chakraborty, T.: CoReRank: ranking to detect users involved in blackmarket-based collusive retweeting activities. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 330–338 (2019)
14. Dutta, H.S., Chakraborty, T.: Blackmarket-driven collusion among retweeters-analysis, detection and characterization. IEEE Trans. Inf. Forensics Secur. **15**, 1935–1944 (2019)
15. Dutta, H.S., Chetan, A., Joshi, B., Chakraborty, T.: Retweet us, we will retweet you: spotting collusive retweeters involved in blackmarket services. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 242–249 (2018)
16. Dutta, H.S., Dutta, V.R., Adhikary, A., Chakraborty, T.: HawkesEye: detecting fake retweeters using Hawkes process and topic modeling. IEEE Trans. Inf. Forensics Secur. **15**, 2667–2678 (2020)
17. Dutta, H.S., Jobanputra, M., Negi, H., Chakraborty, T.: Detecting and analyzing collusive entities on YouTube. arXiv preprint arXiv:2005.06243 (2020)
18. Arora, U., Dutta, H.S., Joshi, B., Chetan, A., Chakraborty, T.: Analyzing and detecting collusive users involved in blackmarket retweeting activities. ACM Trans. Intell. Syst. Technol. **11**(3), 1–24 (2020). Article 35
19. Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., Cohen, W.W.: Tweet2Vec: character-based distributed representations for social media. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (2016)
20. Ruchansky, N., Seo, S. Liu, Y.: CSI: a hybrid deep model for fake news detection, pp. 797–806. (2017)https://doi.org/10.1145/3132847.3132877
21. Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux, G., Vaught, T., Millman, J. (eds.) Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA, USA, pp. 11–15 (2008)
22. Yang, K.-C., Varol, O., Davis, C., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with artificial intelligence to counter social bots. Hum. Behav. Emerg. Technol. **1**, 48–61 (2019). https://doi.org/10.1002/hbe2.115
23. Shah, N., Lamba, H., Beutel, A., Faloutsos, C.: The many faces of link fraud. In: 2017 IEEE International Conference on Data Mining (ICDM), pp. 1069–1074 (2017)