The Journal of Open Source Software

BotSlayer: real-time detection of bot amplification on Twitter

Pik-Mai Hui[1], **Kai-Cheng Yang**[1], **Christopher Torres-Lugo**[1], **Zachary Monroe**[1], **Marc McCarty**[2], **Benjamin D. Serrette**[2], **Valentin Pentchev**[2], **and Filippo Menczer**[1, 2]

1 Center for Complex Networks & Systems Research, Indiana University **2** Indiana University Network Science Institute

DOI: 10.21105/joss.01706

Software

- Review
- Repository
- Archive

Submitted: 20 August 2019
Published: 18 October 2019

Summary

As social media became major platforms for political campaigns and discussions of other important issues, concerns have been growing about manipulation of the information ecosystem by bad actors. Typical techniques used by the bad actors vary from astroturf (Ratkiewicz, Conover, Meiss, Gonçalves, et al., 2011; Ratkiewicz, Conover, Meiss, Gonçalves, et al., 2011) and amplification of misinformation (Shao, Ciampaglia, et al., 2018; Shao, Hui, et al., 2018) to trolling (Zannettou et al., 2018). Attempts to manipulate discussions may and often does involve real humans; examples include trolls from Russia (Badawy, Addawood, Lerman, & Ferrara, 2019; Kim, Graham, Wan, & Rizoiu, 2019; Zannettou et al., 2018) and Iran (Zannettou et al., 2019). Recent reports show that malicious social bots — inauthentic accounts controlled in part by software — have been active during the U.S. presidential election in 2016 (Bessi & Ferrara, 2016), the 2017 Catalan referendum in Spain (Stella, Ferrara, & De Domenico, 2018), the French Presidential Election of 2017 (Ferrara, 2017), and the 2018 U.S. midterm election (Deb, Luceri, Badaway, & Ferrara, 2019).

Detecting such manipulation presents serious research challenges. Firstly, one needs to collect and analyze data, which requires significant storage and computing resources (Davis, Ciampaglia, et al., 2016). Secondly, finding patterns and signals of suspicious behaviors from huge amounts of data requires advanced computational skills (Ferrara, Varol, Menczer, & Flammini, 2016; Varol et al., 2017b). In fact, most studies on this phenomenon are disseminated months or even years after the events. Detecting potential manipulations from real-time social media data streams remains an open challenge.

To address this challenge, we developed a tool to detect and track potential amplification of information by likely coordinated bots on Twitter in real time. The tool is called `BotSlayer`. Here we introduce `BotSlayer-CE`, the open-source Community Edition of the tool. There is also a free but proprietary version that includes more sophisticated bot detection algorithms.

BotSlayer-CE is easy to install and can be customized to any topics of interest. Its embedded algorithms and user-friendly interface make it possible for experts as well as journalists and citizens to study online manipulation.

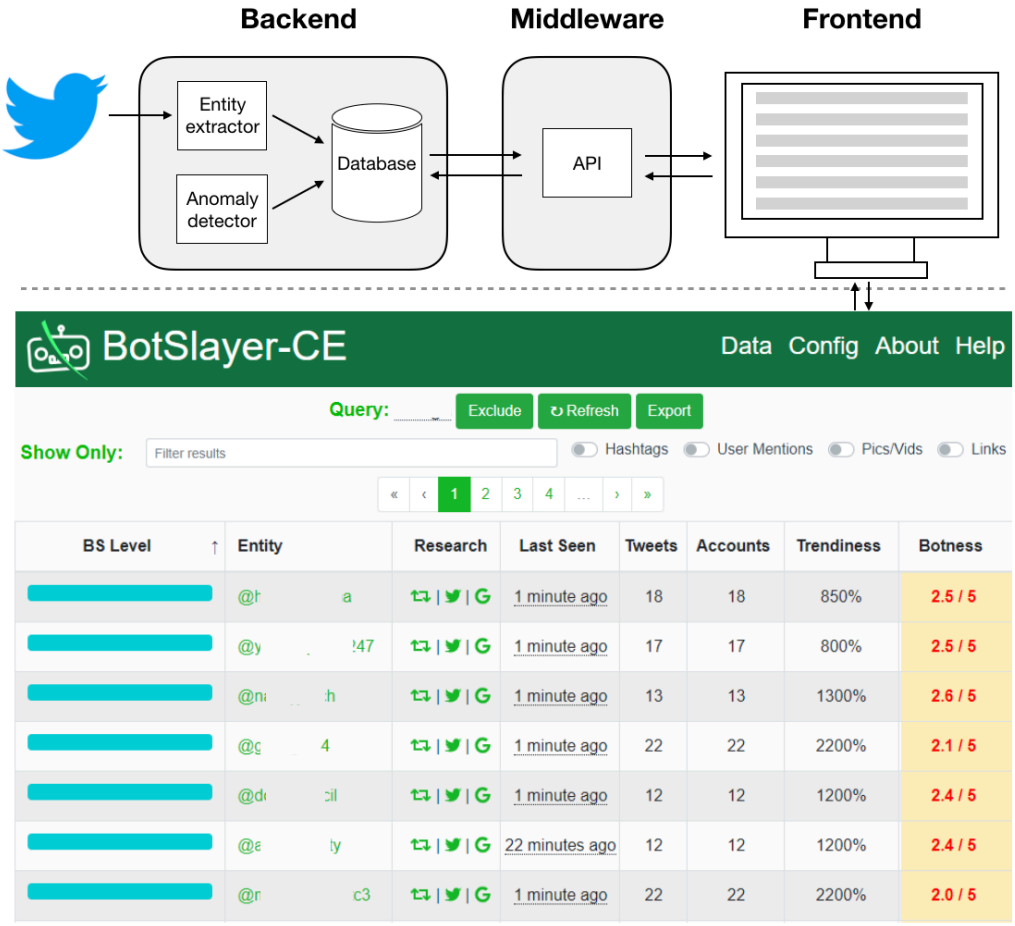

Figure 1: System architecture of BotSlayer-CE.

Figure 1 shows an overview of the BotSlayer-CE system architecture with its backend, middleware, and frontend. The backend collects and analyzes tweets, while the frontend renders a dashboard that reports suspicious content to users. The backend and frontend communicate with each other through the middleware APIs.

Data collection is query-driven and requires a Twitter app key. The user-defined query is a set of keywords of interest, see Twitter's documentation for details. These keywords are fed to Twitter's filtering API to fetch a stream of related tweets. The software extracts entities (hashtags, user handles, links, and media) for further analysis.

Entities are stored in a PostgreSQL database, interfaced with the tweet collector and the middleware using `asyncpg` and `asyncio` in Python3. All statistical and machine learning calculations are implemented in SQL query to leverage database concurrency on the server machine. The whole backend is wrapped inside a Docker container to allow flexible and portable deployment.

BotSlayer-CE provides users with a dashboard that is accessible through any web browser. The frontend allows users to set up the app key and change query of interest through a configuration page. The main page displays statistics of entities related to the query, ordered from the most suspicious to the least by a metric called `BS Level`. Users can also re-order the entities by different metrics like botness and trendiness, or filter them by keywords or types to explore potentially suspicious behaviors. For each entity, the dashboard provides links for users to go back to Twitter to check the original discussion or search on Google. Users

can also visualize the discourse around each entity on Hoaxy (Shao, Ciampaglia, Flammini, & Menczer, 2016). Finally, users can export aggregated statistics as spreadsheets.

To calculate the BS level of an entity, we extract four features: volume, trendiness, diversity, and botness in 4-hour sliding windows and apply logistic regression based on a manually annotated training set. For the volume, we count the number of tweets containing each entity during the focal time window. Trendiness is calculated as the ratio between the entity volume in two consecutive time windows. The diversity is the ratio between the number of unique users and the number of tweets they post. Finally, botness measures the level of bot-like activity. The intuition for the BS level is that entities with intermediate diversity and high volume, trendiness, and botness tend to be more suspicious.

To measure the botness, BotSlayer-CE is equipped with a simple rule-based bot scoring function. The bot scoring function uses simple heuristics based on high friend growth rate, high friend/follower ratio, high tweeting frequency, and default profile image to calculate bot scores. These heuristics yield about 0.70 AUC (Area Under the receiver operating characteristic Curve) when tested on annotated accounts (Yang et al., 2019). They may be appropriate to detect some bots and not others. Depending on the research domain, different bot detection algorithms may be advisable. One can plug their favorite bot detection system into the `BotRuler` class. One could implement simpler heuristics based on high tweet rate (Howard & Kollanyi, 2016) or default profile image (Forelle, Howard, Monroy-Hernández, & Savage, 2015), use state-of-the-art machine learning bot detection tools (Davis et al., 2016; Varol et al., 2017a), or train their own classifier. For example, the "Pro" version of BotSlayer uses a proprietary bot detection software. Accounts that display the suspicious behaviors mentioned above will have scores close to 1.

References

Badawy, A., Addawood, A., Lerman, K., & Ferrara, E. (2019). Characterizing the 2016 russian ira influence campaign. *Social Network Analysis and Mining*, *9*(1), 31. doi:10.1007/s13278-019-0578-6

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, *21*(11).

Davis, C. A., Ciampaglia, G. L., Aiello, L. M., Chung, K., Conover, M. D., Ferrara, E., Flammini, A., et al. (2016). OSoMe: The iuni observatory on social media. *PeerJ Computer Science*, *2*, e87. doi:10.7717/peerj-cs.87

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to evaluate social bots. In *In proc. 25th intl. Conf. Companion on world wide web* (pp. 273–274).

Deb, A., Luceri, L., Badaway, A., & Ferrara, E. (2019). Perils and challenges of social media and election manipulation analysis: The 2018 us midterms. In *Companion proceedings of the 2019 world wide web conference*, WWW '19 (pp. 237–247). San Francisco, USA: ACM. doi:10.1145/3308560.3316486

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, *22*(8). doi:10.5210/fm.v22i8.8005

Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2016). Detection of promoted social media campaigns. In *Proc. Tenth international aaai conference on web and social media (icswm)*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13034

Forelle, M., Howard, P., Monroy-Hernández, A., & Savage, S. (2015). Political bots and the manipulation of public opinion in venezuela. *SSRN*. doi:10.2139/ssrn.2635800

Howard, P. N., & Kollanyi, B. (2016). Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum. *Available at SSRN 2798311*.

Kim, D., Graham, T., Wan, Z., & Rizoiu, M.-A. (2019). Tracking the digital traces of russian trolls: Distinguishing the roles and strategy of trolls on twitter. *arXiv preprint arXiv:1901.05228*.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Detecting and tracking political abuse in social media. In *Proc. 5th international aaai conference on weblogs and social media (icwsm)*.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on world wide web* (pp. 249–252). ACM.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web* (pp. 745–750). International World Wide Web Conferences Steering Committee.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*, 4787. doi:10.1038/s41467-018-06930-7

Shao, C., Hui, P.-M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLoS ONE*, *13*(4), e0196087.

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 201803470. doi:10.1073/pnas.1803470115

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017a). Online human-bot interactions: Detection, estimation, and characterization. In *Proc. Intl. AAAI conf. On web and social media (icwsm)*.

Varol, O., Ferrara, E., Menczer, F., & Flammini, A. (2017b). Early detection of promoted campaigns on social media. *EPJ Data Science*, *6*(13). doi:10.1140/epjds/s13688-017-0111-y

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. doi:10.1002/hbe2.115

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. *arXiv preprint arXiv:1801.09288*. doi:10.1145/3308560.3316495

Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Who let the trolls out?: Towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science* (pp. 353–362). ACM. doi:10.1145/3292522.3326016