Names:	Richard Kavanagh
Student Numbers:	17147930
Email:	x17147930@student.ncirl.ie
Course:	(PgD) Postgraduate Diploma in Science in Data Analytics
Module:	(H9ADM) Advanced Data Mining
Project:	CA1 - Semester 2
Date:	07/08/2018

A Voight-Kampff Test for Social Media
Classification & Detection of Bots on Twitter

Motivation:

Automated agents, computer programs that perform functions intended for regular users – commonly known as bots – have been active since the very early days of the internet. However with the rise of Web 2.0 platforms, especially the likes of social media sites like Twitter and Facebook the scope and potential impact of such bots activity has increased dramatically. While some bots have good intentions and perform useful functions like provide news feeds, customer support help, and aggregate content, they can, however, still spread false and misleading information unknowingly. A much more concerning type of bot is those specifically designed with malicious intent in mind. These bots purposely mislead and manipulate online discourse with spam, abuse, fake information or personal accusations and insults. The fallout from these bots activities can range from individual harassment to widespread misinformation campaigns attempting to influence political elections and referendums. These issues reached the mainstream over the alleged influence of Russian sponsored bots in the 2016 US Presidential Election and the 2016 Brexit referendum but have been noted previously as attempting to influence the public discourse on matters such as military conflicts, terrorist attacks and the immediate aftermath of natural disasters.

As more of the population of the world becomes connected to the internet and we become more reliant on services like Twitter for our information and news the activities and targets of these bots is only going to increase. According to official figures[2] Twitters active user count stood at 336 million at the end of the first quarter of 2018. Four years earlier Twitter reported that 13.5 million or 5% of its user base was a bot or spam account[3]. This shows the pressing need for more research and attention in this area.

Research Question:

Is it possible to accurately predict whether a given Twitter account, by examining its meta-data and posting activity is being run by a bot or human?

Related Work:

As a result of the increased awareness of the presence of bots on the web there has been a corresponding surge in research on their activities, detection and the effects they have on social media and Twitter. [7] attempted to quantify the impact of bots on Twitter and the web and found that in some specific sub-domains bots play a significant role in a piece of contents popularity and reach. It can also be seen in [11] that the susceptibility of users to be influenced by bots does not diminish with increased activity or experience on Twitter but actually increases. This goes against the intuitive idea that users would develop an ability to spot bots naturally. We also note that [13] studied the behaviour and impact of bots on Twitter and found that current counter measures fails to detect the majority of bots, over a one month period (only 38/120 bots were successfully identified and removed by Twitter). These last two papers both reinforce the need for improved automated bot detection.

[6] notes that the anecdotal claims of the rising count and influence of Twitter bots is backed up by current analysis and that these bots are becoming increasingly complex and nuanced in their design and application. They conclude that the already difficult problem of bot detection is only going to get worse and that while current academic work in similar fields can be applied here, more domain specific research is needed. In line with this many papers recommend that any classifier dynamic enough to keep up with this increasing complexity will have to develop several sub-models to describe the popular subcategories bots tend to fall into.

[8] trained several variations of its classifier by restricting the feature set it has access to, this was to imitate future bot designs where features that are known to reveal a bots presence have been removed or obfuscated. They also simulated scenarios where key features are no longer available and how a high level of accuracy could still be achieved. This paper was very useful for identifying possible features we would use as well as novel ways of combining or aggregating them. While [9] divides its training set of Twitter profiles into four popularity bands, with the justification that the behaviour of bots changes a significant amount depending on the popularity context it operates in. It then finds which features sets are most effective at each popularity level through ablation tests as well finding which of these features perform well across all levels. Although we can see the efficacy and results of this combined granular/abstract approach we have decided that it lies outside the scope of this project and will treat all Twitter profiles the same.

In [12] we see many of the trends and assumptions in feature selection present in other papers such as [7] [14] and [15] confirmed and as a result have selected a similar feature set to begin with. We also see that the most successful approaches are those models that utilize these feature sets on a sufficiently large dataset. A large variety of models and supervised learning techniques such as linear regression, neural networks, naive-bayes, decision trees etc can all achieve a high level of accuracy. As a result we will be exploring multiple types of thses classifiers in our analysis.

As discussed above the majority of papers follow supervised learning techniques. [10] however explores an entirely unsupervised clustering model. This novel approach uses activity correlation as the only indicator of a bots presence. These activities can include posting, liking or sharing content. This approach works on the understanding that multiple accounts that have highly correlated activities over time are very likely to be bots, as bots tend to be operated in large groups under a single master controller. While this paper has encouraged us to examine the potential of unsupervised techniques it omits the technical details so we will not attempt to implement correlation classification ourselves.

Data Sources:

To train our classification model we plan to make use of the publicly available 'caverlee-2011' dataset available from the bot-repository website[1]. This dataset was collected from December 30, 2009 to August 2, 2010. It contains 22,223 confirmed bot profiles, their user-follow history, 2353,473 of their tweets, and 19,276 valid users, their user-follow history, and 3,259,693 of their tweets.

The data is split between six tab-separated text files

- 'content_polluters.tx'
- 'content_polluters_followings.txt'
- 'content_polluters_tweets.txt'
- 'legitamate_users.txt'
- 'legitimate_users_followings.txt'
- 'legitimate_users_tweets.txt'

and contains the following values

Twitter profiles

UserID, CreatedAt, CollectedAt, NumerOfFollowings, NumberOfFollowers, NumberOfTweets, LengthOfScreenName, LengthOfDescriptionInUserProfile

Tweets

UserID, TweetID, Tweet, CreatedAt

Following

UserID, SeriesOfNumberOfFollowings

Proposed Methodology:

Given the above values available in our dataset we have identified the following core features we can either extract directly or calculate from combining one or more values. They can be divided into two groups, PF (profile features) and CF (content features).

PF can include

- ProfileCreatedAt
- LengthOfScreenName
- LengthOfUserDescription
- NumberOfTweets
- FollowerCount
- FollowingCount
- Following/FollowerRatio

CF can include

- LinksPerTweet
- UserMentionsPerTweet
- SimalarityOfTweets

We may add additional features as we continue to explore our dataset. From our research into related work we have identified three types of models that have been shown to achieve high levels of accuracy for similar tasks

- Naive Bayes
- Decision Trees / Random Forest
- Linear Regression

The three models above are supervised learning techniques, we will also investigate the unsupervised clustering technique KNN(K Nearest Neighbours). These four models will be compared and contrasted with each other against a number of various standard metrics such as accuracy, precision, recall and F β score. Our model will ideally minimize the false-positive rate. It is preferable to miss actual bot profiles than it is to incorrectly label a valid user as a bot. This suggests metrics such as precision will be an important evaluation metric. We will then also need to compare our models accuracy to the accuracy achieved from human classification.

Potential Beneficiaries:

It is in the best interest of all parties – Twitter itself as a platform, Twitter users and the general public – that the effects of bots be identified and curtailed as much as possible. Twitter itself can only achieve long-term success through it advertising revenue, an income source that cannot work with large numbers of bots on the platform. If bots ever represented a non-negligible portion of Twitters interactions then advertisers would be much less likely to believe that their advertisements would be worth the investment, a serious concern for a company that has consistently struggled to generate profits. Twitter users, like all social media users, will only visit a platform if it has a sufficiently large human audience to engage and interact with. Bots can seriously harm the experience of users and discourage them from further use of the platform. Twitter has already experienced significant reputation damage[4] from the activities of bots on its platform.

As mentioned previously the general public is now aware of just how influential Twitter can be in establishing narratives, trends, and other social movements. Twitter is under increasing pressure from politicians and interest groups to fix the issue. It has also been noted that common methods for detecting Twitter bots can be applied to traditional spam bots and vice versa[5] so that traditional spam research stands to benefit.

Bibliography:

[0] Voight-Kampff Test, [Online]. Available: https://www.youtube.com/watch?v=Umc9ezAyJv0
[Accessed October 07, 2018].

[1] bot-repository, [Online]. Available: https://botometer.iuni.iu.edu/bot-repository/index.html.
[Accessed October 05, 2018].

[2] Washington Post [Online]. Available:
https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/?noredirect=on&utm_term=.bd46556dbc41.
[Accessed October 05, 2018].

[3] Twitters 2014 Q2 SEC filing, [Online]. Available: http://bit.ly/1kBx4M8.
[Accessed October 05, 2018]

[4] Bloomberg, [Online]. Available:
https://www.bloomberg.com/news/articles/2016-10-17/disney-said-to-have-dropped-twitter-pursuit-partly-over-image. [Accessed October 05, 208]

[5] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi.
"*Fame for sale: efficient detection of fake twitter followers*,"
CoRR, vol. abs/1509.04098, 2015. [Online]. Available: http://arxiv.org/abs/1509.04098

[6] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi.
"The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race."
CoRR, abs/1701.03017, 2017.

[7] Z. Gilani, R. Farahbakhsh, and J. Crowcroft. “*Do bots impact twitter activity?*”
In Proceedings of the 26th International Conference on World Wide Web Companion,
WWW ’17 Companion, pages 781–782, Republic and Canton of Geneva, Switzerland, 2017.
International World Wide Web Conferences Steering Committee.

[8] K. Lee, B. D. Eoff, and J. Caverlee. “*Seven months with the devils*: *A long-term study of content polluters on twitter*”. In ICWSM, 2011.

[9] Gilani, Z., Kochmar, E., Crowcroft, J.: “*Classification of twitter accounts into automated agents and human users*”. In: ASONAM (2017)

[10] Chavoshi, N.; Hamooni, H.; and Mueen, A. 2016. “*Identifying correlated bots in twitter*”.
In Social Informatics: 8th Intl. Conf., 14–21

[11] C. Wagner, S. Mitter, C. K¨orner, and M. Strohmaier. “*When social bots attack: Modeling susceptibility of users in online social networks*”. Making Sense of Microposts (# MSM2012), 2, 2012.

[12] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. 2017. “*Of Bots and Humans (on Twitter)*”. In Proceedings of the 9th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'17).

[13] C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso. “*Reverse engineering socialbot infiltration strategies in twitter*”. In Proceedings of the IEEE/ACM ASONAM, pages 25–32, 2015.

[14] Kudugunta Sneha, Emilio Ferrara, “*Deep Neural Networks for Bot Detection*”, 2018 [Online] Available http://arxiv.org/abs/1802.04289

[15] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, *Detecting automation of twitter accounts: Are you a human, bot, or cyborg?* , Dependable and Secure Computing, IEEE Transactions on v9, n6 pp 811-824 (2012).