IADIS International Journal on WWW/Internet
Vol. 16, No. 2, pp. 50-65
ISSN: 1645-7641

THE EFFECT OF ENGAGEMENT INTENSITY AND LEXICAL RICHNESS IN IDENTIFYING *BOT* ACCOUNTS ON TWITTER

Isa Inuwa-Dutse[1], Bello Shehu Bello[2], Ioannis Korkontzelos[1] and Reiko Heckel[2]
[1]*Department of Computer Science , Edge Hill University, UK*
[2]*Department of Informatics, University of Leicester, UK*

ABSTRACT

The rise in the number of automated or bot accounts on Twitter engaging in manipulative behaviour is of great concern to studies using social media as a primary data source. Many strategies have been proposed and implemented, however, the sophistication and rate of deployment of bot accounts is increasing rapidly. This impedes and limits the capabilities of detecting bot strategies. Various features broadly related to account profiles, tweet content, network and temporal patterns have been utilised in detection systems. Tweet content has been proven instrumental in this process, but limited to the terms and entities occurring. Given a set of tweets with no obvious pattern, can we distinguish contents produced by social bots from those of humans? What constitutes engagement on Twitter and how can we measure the intensity of engagement among Twitter users? Can we distinguish between *bot* and *human* accounts based on *engagement intensity?* These are important questions whose answer will improve how detection systems operate to combat malicious activities by effectively distinguishing between human and social bot accounts on Twitter. This study attempts to answer these questions by analysing the engagement intensity and lexical richness of tweets produced by human and social bot accounts using large, diverse datasets. Our results show a clear margin between the two classes in terms of engagement intensity and lexical richness. We found that it is extremely rare for a social bot to engage meaningfully with other users and that lexical features significantly improve the performance of classifying both account types. These are important dimensions to explore toward improving the effectiveness of detection systems in combating the menace of social bot accounts on Twitter.

KEYWORDS

Twitter bots, Automated Accounts, Engagement, Lexical Richness, Bot Detection, Social Network Analysis

1. INTRODUCTION

As a social being, a human's behaviour is largely influenced by close associates. Today social networking sites are important avenues for socialisation, providing instant information sharing and interaction on a large scale (Gilani et al., 2017; Wilson et al., 2012). Modern social media platforms, such as Twitter and Facebook, enable various forms of interaction among diverse users. This capability results in a huge amount of data waiting to be utilised by researchers. However, the credibility of such content is being questioned by the growing activities of automated accounts otherwise known as bots. In particular, *social bots* are automated accounts designed to interact with users either to amplify or generate contents in social media platforms. A more elaborate definition, specifically of Twitter bots, is given by Akomoto (2011):

> "*Twitter bots are small software programs that are designed to mimic human tweets. Anyone can create bots, though it usually requires programming knowledge. Some bots reply to other users when they detect specific keywords. Others may randomly tweet pre-set phrases such as proverbs. Or if the bot is designed to emulate a popular person (celebrity, historic icon, anime character etc.) their popular phrases will be tweeted. Not all bots are fully machine generated, however, and interestingly the term "bot" has also come to refer to Twitter accounts that are simply "fake" accounts*"

Some social bot accounts, such as *@congressedits*[1], are legitimate, with clear distinguishing features whereas the majority are created to mislead in various ways, such as by creating superficial popularity (Varol et al., 2017) or influencing public opinion (Howard and Kollanyi, 2016). Some bots are obvious to identify, because they use the word "*bot*" explicitly in their Twitter handle (Dutse et al., 2018). Autonomous accounts contribute a sizeable part of social media content. It was estimated that 9% - 15% of active Twitter accounts are social bots accounts (Varol et al., 2017) and require effective methods to be detected.

How can we detect if a given tweet was posted by a social bot or a human user? We investigate this question based on a collection of tweets sampled from Twitter. In particular, we investigate how effective lexical features are for the detection of social bot accounts. In contrast to previous work (Benevenuto et al., 2010; Cai t al., 2017; Dockerson et al., 2014; Lee and Kim, 2012; Thomas et al., 2011; Varol et al., 2017; Wang, 2010), our study focuses on comprehensive linguistic analysis to define lexical features effective enough for accurate detection of social bot accounts on Twitter. As shown in Table 1, the study explores a large number of tweets from social bots and humans to understand the difference between the two in terms of lexical richness and distribution of emoticons, further discussed in section 3. Our analysis highlights the distinguishing characteristics of automated accounts and how lexical features can improve detection of bots. Furthermore, we analyse the level of engagement exhibited by social bot and human accounts and its role in building detection system. Our analysis reveals the rarity of meaningful engagement with bot accounts. In particular, human users show higher numbers of replies while social bots show higher number of retweets. Further, we show how lexical features play a role in the detection and how engagement and lexical features combine to develop a powerful detection system for Twitter bots. Our dataset consists of both English and non-English bots for an evaluation of our approach. Using the lexical features only we achieve an accuracy of 86% and AUC score of 87%. We achieve AUC score of 95% by combining our

[1] A bot that tweets anonymous Wikipedia edits made from IP addresses in the US Congress.

proposed features and features utilised in a baseline method which is a significant improvement over the 71% achieved by the baseline features utilised in (Gilani et al., 2017).

Contribution: This is the first study to analyse the role of engagement intensity and lexical features for the detection of automated accounts. The study contributes a powerful set of features useful in distinguishing between humans and social bots on Twitter. We provide the first comprehensive analysis of both engagement intensity and lexical richness of tweets computed on various accounts and investigate how their incorporation improves the performance of the detection system. Specifically, engagement intensity and features based on *lexical diversity, type-token ratio, usage of contractions and emoticons* are powerful lexical signals in distinguishing between humans and social bots. The study provides a new set of distinctive features and a dataset to support the research community in identifying bot accounts.

The remaining of the paper is structured as follows. Section 2 and Section 3 present a review of related works and propose engagement & lexical features, respectively. Section 4 presents our experiments and Section 5 presents the results and a detailed discussion. Finally, Section 6 concludes the study and proposes some future work.

2. RELATED WORK

Social bot accounts are instrumental in spreading fake and malicious news on Twitter, employed to skew analysis results and opinion of users. The demand for effective detection systems has prompted a surge of various research approaches. The early work of Wang (2010) focuses on a machine learning for the detection of spam bot accounts on Twitter using features that range from graph-based to content-based. Since this publication, *bots* gained sophistication to evade detection. Early social bot accounts have been reported to lack basic account information such as meaningful screen names or profile pictures (Varol et al., 2017; Lee et al., 2011). This is no longer the case, as social bots grow in sophistication, making it difficult to identify distinguishing features from human accounts. Some approaches involve the analysis of accounts as far as their position in a *network*, their *temporal* metadata and the *content of their tweets*, to define a new set of features. The following review focuses on related studies that utilised aspects of these features to detect bot accounts.

Network and Temporal Features: Motivated by the rise in automated accounts on Twitter, which often confuse users regarding their legitimacy, Chu et al. (2012) presents a method to detect fully automated accounts and partially automated accounts orchestrated by a human user (known as *Cyborgs).* The study relies on the account activities in terms of *posting behaviour, tweet content* and *account properties* to extract useful features for classification. *Tweet content features* should not be confused with *lexical features*, which have been shown to significantly improve the performance of detection systems (Dutse et al, 2018). This study improves on the effectiveness of *lexical features* by focusing on *engagement features.* The work of Davis et al. (2016) and (Ferrara et al. (2016) developed a detection system that leverages features related to both network structure and tweeting behaviour exhibited by accounts. Subrahmanian et al. (2016) reports how researchers utilised many features related to network, temporal behaviour, tweet syntax, tweet semantics and user profile for the detection of *influence bots.* Influence bots are a category of social bot accounts that aim at influencing the opinion of other users. However, despite the wide spectrum of features considered in the study, analysis of the lexical richness was not covered. The study of Chavoshi et al. (2017) analyses the behavioural patterns of

accounts by focusing on features related to the network structure, such as local motifs, i.e. repeating behaviour, and discords, i.e. anomalous behaviour. The study also shows how temporal behaviour is useful as a means to distinguish bot from human user accounts. Haustein et al. (2017) analyses the impact of a category of automated accounts on Twitter that focuses on distributing links to scientific articles to promote research impact leading to misuse due to superficial generation of huge contents. Such accounts were found to generate large number of tweets that could skew analysis results.

Tweet Content: Contents generated by users have been analyse to extract meaningful features such as *entropy* for detection. *Entropy* measures the complexity of a system in terms of randomness and *social bots* have been shown to exhibit less randomness (Sutton et al. 2008). However, malicious bots are increasingly becoming more sophisticated and disruptive. A typical bot account has been shown to generate many benign tweets pertaining to news and updates on feeds or employing a cunning way to balance *follower-followee* ratio (Chu et al., 2012). Many detection systems have been developed by leveraging the content of tweets posted by account holders on Twitter. This approach was adopted in Dockerson et al. (2014) to detect social bot accounts on Twitter based on sentiment features, such as topic sentiment and the transition frequency of tweet's sentiment, to train a machine learning classifier. Similarly, to the *tweet content* approach that relies on linguistic analysis, Inuwa-Dutse et al. (2018) utilised features based on lexical richness to detect spam accounts on Twitter. The study of (Cai t al., 2017) proposed a deep learning approach that incorporates content and behavioural information to detect social bots.

The large body of *bot detection* work exploring various detection strategies attests to the significance of the problem. Related literature pays little attention to the analysis of lexical richness of various users' tweets and how they can inform the detection of social bots on Twitter. In contrast to previous work, this study is based on in-depth analysis of engagement intensity and lexical richness as basis for building a detection system. This study contributes effective features and strategy to help in combating the menace of malicious automated accounts on Twitter.

3. FEATURES: LEXICAL AND ENGAGEMENT

This section describes the lexical and engagement features utilised in the study, able to improve detection systems.

3.1 Lexical Features

Lexical richness is a broad concept, expressed in various forms and metrics to assess the quality of text. Metrics such as lexical diversity, lexical sophistication and lexical density are commonly used in linguistic analysis (Šišková, 2012; Templin, 1957). Our approach is based on lexical richness of tweets from various Twitter accounts to detect social bots.

3.1.1 Type-Token Ratio

Type-Token Ratio (TTR) is a simple, yet powerful metric used commonly in quantitative linguistics to measure the richness of vocabulary in a given context (Tweedie and Baayen, 1998). TTR can be expressed as $V(N)/N$, i.e. the size of vocabulary in N divided by the total size

of N. In the context of this study, TTR is the ratio of unique tokens in a tweet to the total number of tokens in the tweet. It can be argued that computing TTR over all tweets of an account may lead to a better result. However, the small size of individual tweets may skew the result due to the sparsity of unique tokens relative to the total number tokens.

3.1.2 Lexical Diversity

Lexical Diversity is an important metric in the analysis of lexical richness. It is useful in assessing the distribution of different content words[2] utilised in a textual corpus or in speech (Tweedie and Baayen, 1998). Lexical sophistication is similar to lexical diversity and focuses on understanding the distribution of advanced words. This study focuses on lexical diversity. The rationale behind using it as a feature is to assess its levels in bot and human accounts and its predictive power in detection of social bot accounts. While the contents produced by social bots were shown to be different across various social bots (Morstatter et al., 2016), they tend to exhibit similarities in terms of widespread use of URLs. In view of this, we computed lexical diversity as the total number of tokens in a tweet without URLs, user mentions and stopwords divided by the total number of tokens in the tweet.

3.1.3 Usage of Contractions

Text or speech in English can be shortened by ignoring some letters or phonetics. These kinds of contracted words are a form of lexical sophistication, useful to measure the fluency of users. Various forms of contracted words are widespread on Twitter. However, we focus on a predefined list of contractions[3] for our analysis. The expectation is for human users to use diverse contractions, while it would be difficult for a bot to use contraction unless generating its tweet from pre-existing sources, e.g. a book or a structured document.

3.1.4 Emoticons

Emoticons[4] are collections of pictorial representations of facial expressions or *emojis* in form of various characters (letters, punctuation, and numbers) to convey emotional mood. Emoticons are popular on social media, especially on Twitter, where tweets are of limited length, are useful indicators or users' sentiment. Common examples of *emoticons* are the smiley, *:-)*, and the sad face, *:-(*. Sentiment-related features have been shown to contribute in detection systems (Dockerson et al., 2014). We leverage this to understand the role of *emoticons* in detection of bot accounts using a comprehensive list of emoticons[5]. The goal is to understand how human and social bot users apply emoticons in tweets and utilise the insight to build classification models. We hypothesise human users will use emoticons in a larger proportion in comparison to social bot accounts.

3.2 Engagement Intensity

Many forms of interactions happen on Twitter at various level of granularity. Interaction with other users comes in the form of simple retweet, likes, reply and direct messages. We defined

[2] Words with meaning in a text; not in stopwords.
[3] Wikipedia list of contractions: en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions
[4] A portmanteau of *emotion* and *icon*.
[5] Available from: en.wikipedia.org/wiki/Emoticon

the engagement of users on Twitter into four different levels and analyses the levels exhibit by both humans and automated accounts.

The following features (also detailed in Table 2) are used to capture the engagement levels: *retweet (RT) count, tweet favourite count, reply* and *user mention.* The *engagement levels* are defined as follows:

- *unengaged*: none of the engagement feature
- *low engagement*: only one engagement feature. For instance, *mentioning* many users (which may or may not imply engagement)
- *moderate engagement*: any two engagement features
- *high engagement*: all engagement features i.e. *retweet (RT) count, tweet favourite count, reply* and *user mention.*

4. EXPERIMENT

This section describes the datasets utilised in this study, including the collection procedure and pre-processing. This is followed by feature selection and building the classification framework.

4.1 Dataset

We utilized three different datasets. The first two are publicly available, *Dataset1* is obtained from (Morstatter et al., 2016) and *Dataset2* from (Gilani et al., 2017. The last dataset is collected for the purpose of this study. Table 1 summarises the datasets. *Dataset2* is classified under five different groups based on popularity and volume of contents generated by the accounts. We maintain the groupings in the study and analyse the lexical richness in each group to understand how the lexical richness will vary across the different groups. The two datasets contain some non-English accounts, which were removed to facilitate proper lexical analysis of tweets. *Dataset1* provides only account *ids* and *Dataset2* provides screen-names and account features. We used the Twitter API to crawl tweets from each account. Our analysis experience with *Dataset1* and *Dataset2* reveals that most of the bot accounts are suspended and many accounts are not in English. In order to ensure genuine representation both from humans and bots accounts, we collected an additional dataset as follows.

Human accounts: We collected human user accounts who directly engage with the Twitter handles of organisations such as university and have correspondences in terms of replies to the users' queries. This is a useful technique to discount for bot accounts since bots may find it difficult to engage in meaningful conversations. Noting the dominance of influential users on Twitter (Dutse, 2018), the data collection strategy employed in this study ensures a balance proportion of users are collected. To the best of our knowledge, this is the first study to employ this approach to ensure the genuineness of human users.

Social bots accounts: Here, we collected 500 bot accounts using a publicly available bot detection system known as *Botometer*[6] (Davis et al., 2016). The *Botometer* returns a probable bot account which may result in many false positives. To mitigate that, we manually annotated the results of *Botometer*. The annotators scanned through the accounts and labelled account as bot based on the following criteria: (1) the account should be active, not suspended or deleted

[6] botometer.iuni.iu.edu/#!/api

and posting tweets in English only (2) if the account's screen name appears to be auto-generated e.g. *2jo120, 2jo24 and 37Hkyjdtytyhjgh* (3) if the profile picture of the account shows no obvious relationship with the account's posts e.g. account tweeting on Brexit but with *storm-trooper* profile picture (4) number of URLs or hashtags: if the tweets mostly consist of URLs or hashtags exceeding 70% of the content (5) activity interval: posting at least 15 tweets per minute.

The annotation process is quite laborious which explain the small size in *Dataset3.* We are not particularly interested in collecting a very high number of accounts but a high number of tweets from real human and bot accounts. We used the Twitter API to collect a maximum of 1000 tweets from each account.

Table 1. Summary of datasets utilised in the study. Dataset2 is categorised in groups based on the number of the followers in each group (*k* and *m* denote thousand and million respectively)

Category	***Bot Accounts***	***Human Accounts***	***Bot Tweets***	***Human Tweets***
Dataset1	685	641	27,766	5,341
Dataset2 1k	75	76	22,432	116,576
Dataset2100k	266	343	112,387	98,271
Dataset2 1m	137	184	45,605	53,700
Dateset2 10m	25	25	9,062	11,869
Dataset3	100	100	83,976	74,483
Dataset4	1763	1832	693655	1337394

Table 2 describes the various features utilised in the study. The proposed engagement intensity features are valid across all datasets. Table A1 in appendix shows examples of common languages in the non-English dataset (*Dataset4).*

4.2 Use of Engagement and Lexical Features for Classification

For learning purposes, machine learning models have been applied to distinguish between *bot* and human based on the set of handcrafted features (Table 2). Deep learning approach have been applied in Cai et al. (2017) for *bot* detection and the method automatically extracts the features to use. Noting the growing sophistication level of *bot*, we view this approach as limiting the features space to explore. By carefully analysing contents from both humans and bot accounts, a more effective detection mechanism can be achieved using handcrafted features. This is evidenced in Section 5. Various classification models have been trained to measure the extent at which these features aid in identifying bot accounts. Classifiers utilised for training include K-nearest neighbour (KNN), Naive Bayes, Support Vector machine for Classification (SVC) and Random Forest on the three datasets. The classifiers were built and trained using *scikit-learn*[7] (Pedregosa et al., 2011), a machine learning toolkit supported by Google and *INRIA*[8]. Stratified 10-fold cross-validation was used to measure the overall *accuracy, precision, recall and roc-auc score* of each classifier. The Random Forest classifier performs best as shown

[7] scikit-learn.org/stable
[8] inria.fr/en

in Table 4. As shown in Table 3, three experiments were conducted for *Dataset2*: (1) using our lexical features, denoted as *L*, (2) using the features in Table 2 from (Gilani et al., 2017), denoted as *F, and* (3) a combination of (1) and (2), denoted as *FL*. The proposed lexical features (L) in this study are *TTR, lexical diversity, average number contractions and emoticons.*

Table 2. Description of features used in the study

Features	Description
Age of account, Favourites-to-tweets ratio, Lists per user, Followers-to-friends ratio, User favourites, Likes/favourites per tweet, Retweets per tweet, User replies, User tweets, User retweets, Tweet frequency, URLs count, Activity source type[S1= browser, S2= mobile apps, S3= OSN management, S4= automation, S5= marketing, S6=news content, S0= all other], Source count, CDN content size	This feature set was first used in Gilani et al, 2017 and is denoted as *F*
retweet (RT) count, tweet favourite count, reply and user mention. The engagement intensities/levels are defined as follows:	Engagement intensity features are denoted as *E* and are categorised based on intensity: low engagement, moderate engagement and high engagement.
lexical diversity, type-token ratio, usage of contractions and emoticons.	Lexical richness features are denoted as *L*. Thus, *FLE* refers to combination of all the three features sets.

Table 3. Datasets and respective description of features utilised in training the prediction model

Dataset	Description
Dataset2_F	Features in Table 2
Dataset2_FL	features from in Table 2 and our proposed lexical features (L)
Dataset2_L	Training Dataset2 on proposed lexical features (L)
Dataset3_L	Proposed lexical features on Dataset3
Dataset4_LE	Combination lexical and engagement features on Dataset4

5. RESULTS AND DISCUSSION

The following section presents the main findings of the study. Figure 1 shows empirical evidence that the proposed lexical features are among the important features for the identification of automated accounts. Similarly, Figures 2, 3 and 4 show how lexical features manifest in humans and bot accounts.

Lexical Diversity: Figure 2 shows the results of computing lexical diversity in human and bot accounts. *Lexical diversity* is expected to be higher in humans, since humans have been shown to generate better and novel content on Twitter (Gilani et al., 2017). However, this is not entirely true, especially in some automated accounts by prominent organisations, as shown in Figure 2. Accounts with a higher number of followers under the bot category are shown to have higher lexical diversities than the corresponding human counterparts. This is probably because such accounts are managed to update a large number of users on various topics on regular basis. Accounts in this category include organisational accounts, such as the BBC, politicians or popular celebrities. However, in *Dataset3,* humans have higher lexical densities which can be linked to the approach that was employed for the data collection. The dataset is a representative of an average user on Twitter. This confirms our earlier intuition that a typical human user account is expected to have a higher lexical diversity.

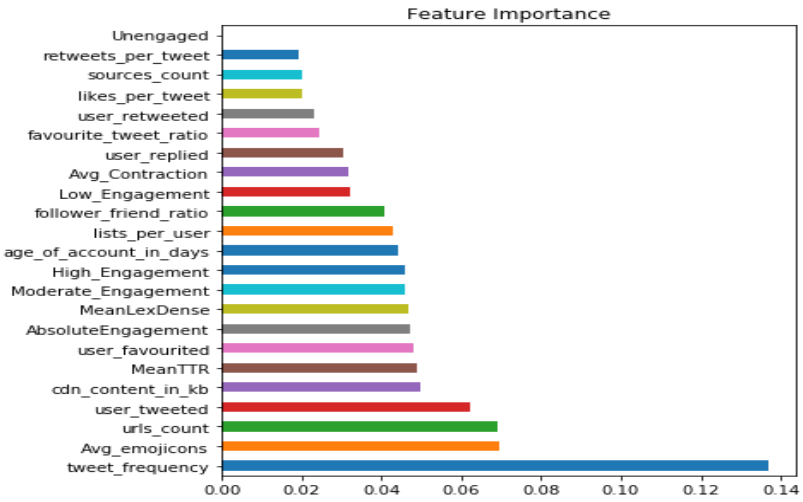

Figure 1. A comparison of importance of the proposed lexical features and features utilised in a related study

Usage of Contractions: Figure 3 shows how the usage of contraction varies across the datasets. With the exception of *Dataset3*, the difference in the use of contracted words between humans and bots is not very significant. This can be due to the fact that users with many followers on Twitter, such as organisational accounts or politicians, tend to use contracted words in tweets. With the exception of *Dataset3*, the difference in the use of contracted words between humans and bots is not very significant. This can be attributed to the fact that users with a high number of followership on Twitter will tend to use contracted words, e.g. (*Dataset2 1M and Dataset2 10M*), which mainly contain tweets of organisations, celebrities or politicians.

Usage of Emoticons: We found that the usage of *emoticons* is higher in bot accounts than in human accounts across all the different datasets as shown in Figure 4. Surprisingly, the results suggest that bot accounts utilise more *emoticons* in their tweets than humans. This is contrary to our prior intuition that humans are more likely to use more emoticons. The primary goal is to improve detection of bot accounts by adding lexical features into the detection system. *Emoticons* happen to be the most distinctive features between humans and bots. In Figure 4 we observe an agreement in the usage of *emoticons* in all the datasets, i.e. bots use *emoticons* more frequently than humans. It is evident from the classification result (Table 4) that lexical features significantly improve the performance of the detection system. The distinguishing power of *lexical features* appears to be less effective in *Dataset1* and *Datatset2* in relation to *Dataset3*. With extra measures during data collection and annotation, this effect can be mitigated. This is evident from *Dataset3* which was manually inspected, and emphasis should be placed in the collection of more robust and representative datasets for effective detection. Despite the variations, which we attribute to many false positives in *Dataset1* and Dataset2, *lexical features* prove to be strong indicators as shown in Figure 1. The figure shows that *emoticons (*captioned as *avg_emojicons)* is the second most important feature among the 18 features. The lowest performing of our proposed lexical features (*avg_contraction*) outperforms 6 features utilised in a related study.

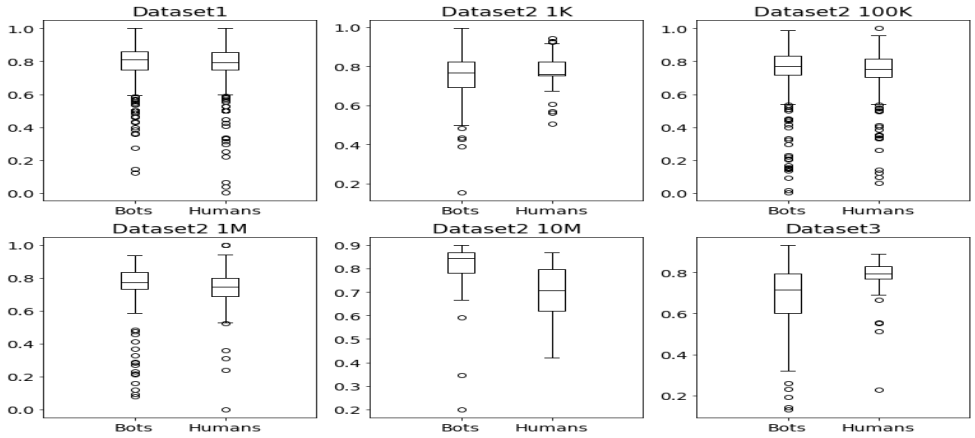

Figure 2. Average *lexical diversities* of *human* and *social bot* accounts

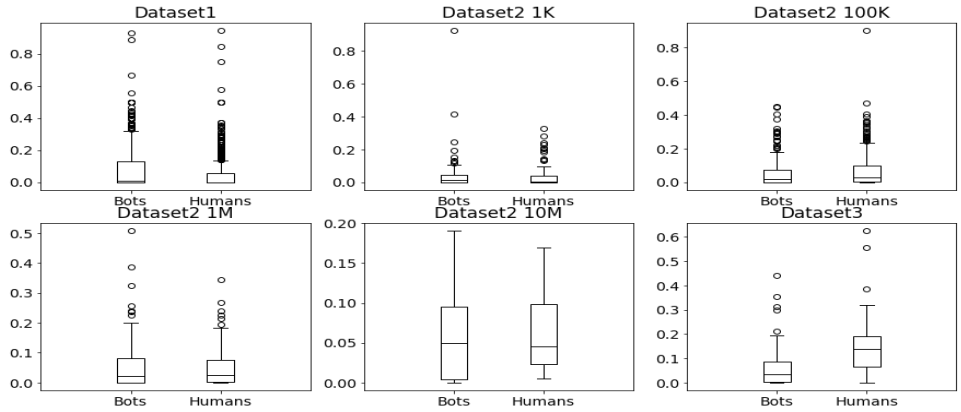

Figure 3. Average Contractions across datasets utilised in this study

Effect of engagement and lexical features on classification of bot and human accounts: We use machine learning algorithms to measure the extent at which our proposed lexical features aid detection of bot accounts. Table 4 shows the results of a trained random forest classifier across the datasets. Using the lexical features only we achieve an accuracy of 86% and AUC score of 87% in Dataset3. In Dataset2_FL we achieve an AUC score of 95% which is a significant improvement over 71% achieved using only the features utilised in (Gilani et al., 2017).

Table 4. Datasets and respective prediction performances

Dataset	***Accuracy***	***Precision***	***Recall***	***AUC Score***
Dataset1_L	0.65	0.65	0.65	0.65
Dataset2_F	0.71	0.72	0.72	0.71
Dataset2_L	0.66	0.67	0.67	0.66

Dataset2_FL	**0.95**	**0.96**	**0.96**	**0.95**
Dataset3_L	**0.86**	**0.87**	**0.87**	**0.87**
Dataset4_EL	**0.95**	**0.96**	**0.96**	**0.95**

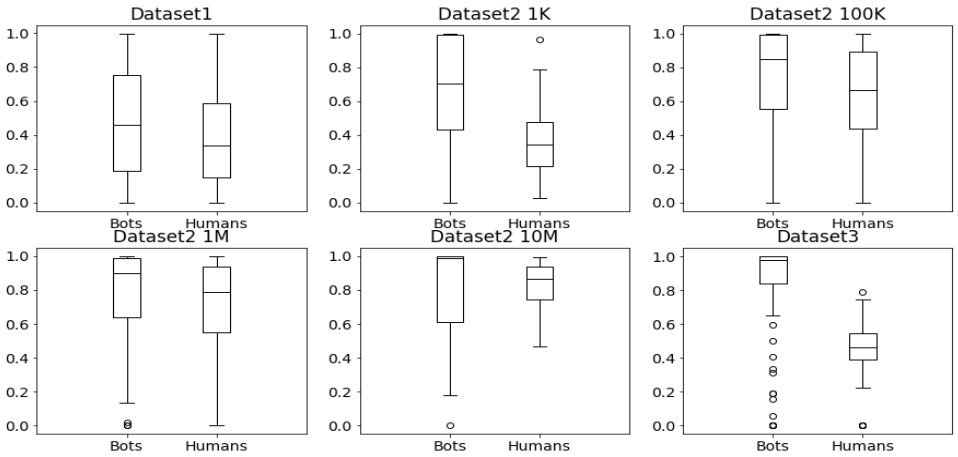

Figure 4. Average *emoticons* in the different datasets

The introduction of the engagement features can be seen to be influential. Table 4 shows the significant improvement in performance (see also Figure A1 and Figure A2 in appendix section). Figure 5 and Figure 6 shows the proportions of engagement features: low, moderate and high intensities from sampled accounts.

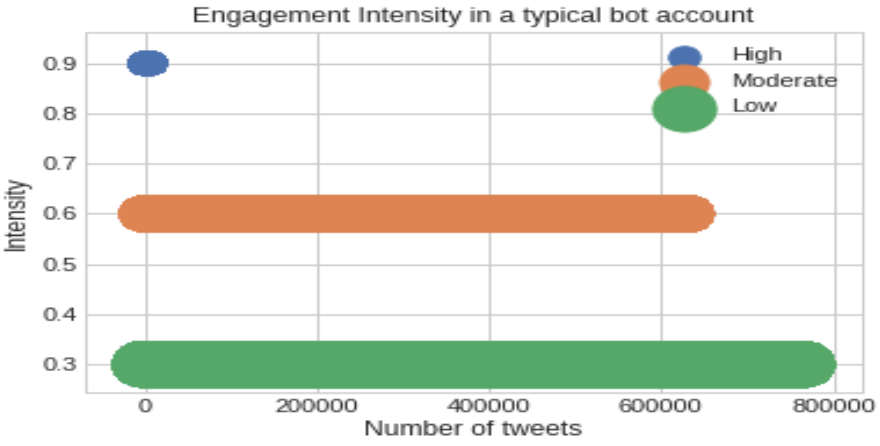

Figure 5. This figure shows the proportions per engagement intensity for bot accounts. There are a small number of bots displaying meaningful engagements. These accounts correspond to credible organisations that use automated accounts to respond to basic queries

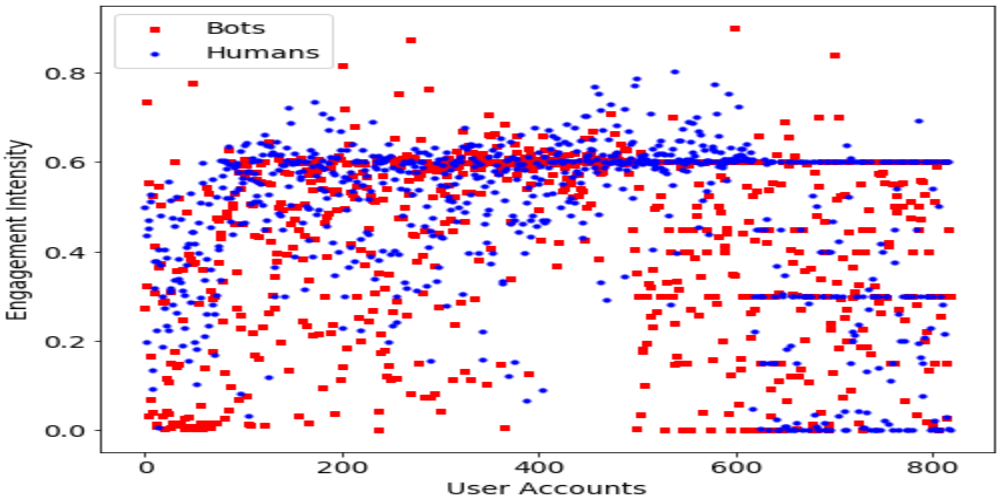

Figure 6. Proportions of engagement intensities in corresponding bot and human accounts. Both group of users show high number of low engagement intensity. The proportions of moderate and high engagements are more prominent in human users

6. CONCLUSION

Modern day social platforms have become part of our lives and effective social policing is required to ensure data credibility and civilised ways of interaction. However, with the growing sophistication of social bots, it is proving difficult to sanitise social platforms. The continuous increase in real-time streaming of tweets makes it practically ineffective to rely on many account features for detection. To effectively distinguish between a bot and a human user, an analysis of lexical richness of tweets posted by both users provides additional distinctive features. We train diverse classifiers to evaluate the role of lexical features in the detection of bot accounts. The newly proposed features significantly improve detection accuracy.

Our proposed an effective detection technique that utilises a set of *lexical* and *engagement* features to distinguish between *human* and *social bot* accounts on Twitter. Our approach is motivated by the premise that meaningful engagement will be difficult for *bot* account to sustain. We have shown the difference between humans and bots in terms of lexical diversity, usage of contraction and emoticons, based on three different datasets. Furthermore, we show how various forms of *engagement* manifest in both *human* and *bot account*. Findings from the study suggest that a combination of *lexical* and *engagement features* provide a powerful tool to improve detection systems. As the first study to introduce combination of *lexical* and *engagement features* for detection task, further research to investigate deeper interactions on Twitter will further widen the disparity between human and bot account independent of language usage. A recent study by Bello et al. (2018) focuses on reverse engineering the underlying mechanism of interaction in automated accounts. This new dimension will potentially be useful in informing how detection system can be improved further.

ACKNOWLEDGEMENT

The authors wish to thank the anonymous reviewers of our work. The third author has participated in this research work as part of the CROSSMINER Project, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 732223.

REFERENCES

Akimoto, A. (2011). *Japan the Twitter nation. The Japan Times. Available from https://www.japantimes.co.jp/life/2011/05/18/digital/japan-the-twitter-nation/#.XAHY4BCnzg5 [accessed 01/12/2018]*

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010, July). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)* (Vol. 6, No. 2010, p. 12).

Bello, B.S., Heckel, R. and Minku, L., 2018, October. Reverse Engineering the Behaviour of Twitter Bots. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 27-34). IEEE.

Cai, C., Li, L., & Zengi, D. (2017, July). Behavior enhanced deep bot detection in social media. In *Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on* (pp. 128-130). IEEE.

Chavoshi, N., Hamooni, H., & Mueen, A. (2017, April). Temporal patterns in bot activities. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1601-1606). International World Wide Web Conferences Steering Committee.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg?. *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811-824.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016, April). 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.

Dickerson, J. P., Kagan, V., & Subrahmanian, V. S. (2014, August). Using sentiment to detect bots on twitter: Are humans more opinionated than bots?. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 620-627). IEEE Press.

Inuwa-Dutse, I., 2018, April. Modelling Formation of Online Temporal Communities. In *Companion of the The Web Conference 2018 on The Web Conference 2018* (pp. 867-871). International World Wide Web Conferences Steering Committee.

Dutse, I.I, Bello, B. S., & Korkontzelos, I. (2018). Lexical analysis of automated accounts on Twitter. *In Proceedings of 17th International Conference on WWW/Internet (pp. 75-82). IADIS.*

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96-104.

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., & Crowcroft, J. (2017, July). Of Bots and Humans (on Twitter). In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*(pp. 349-354). ACM.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated "bot" accounts on T witter. *Journal of the Association for Information Science and Technology*, *67*(1), 232-238.

Howard, P. N., & Kollanyi, B. (2016). Bots,# StrongerIn, and# Brexit: computational propaganda during the UK-EU referendum.

Inuwa-Dutse, I., Liptrott, M., & Korkontzelos, I. (2018). Detection of spam-posting accounts on Twitter. *Neurocomputing*.

Lee, K., Eoff, B. D., & Caverlee, J. (2011, July). Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *ICWSM* (pp. 185-192).

Lee, S., & Kim, J. (2012, February). WarningBird: Detecting Suspicious URLs in Twitter Stream. In *NDSS* (Vol. 12, pp. 1-13).

Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M., & Liu, H. (2016, August). A new approach to bot detection: striking the balance between precision and recall. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 533-540). IEEE Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Sutton, J. N., Palen, L., & Shklovski, I. (2008). *Backchannels on the front lines: Emergency uses of social media in the 2007 Southern California Wildfires* (pp. 624-632). University of Colorado.

Šišková, Z. (2012). Lexical richness in EFL students' narratives. *Language Studies Working Papers*, *4*, 26-36.

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., ... & Stevens, A. (2016). The DARPA Twitter bot challenge. *arXiv preprint arXiv:1601.05140*.

Templin, M. C. (1957). Certain language skills in children; their development and interrelationships.

Thomas, K., Grier, C., Ma, J., Paxson, V., & Song, D. (2011, May). Design and evaluation of a real-time url spam filtering service. In *Security and Privacy (SP), 2011 IEEE Symposium on* (pp. 447-462). IEEE.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323-352.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*.

Wang, A. H. (2010, July). Don't follow me: Spam detection in twitter. In *Security and cryptography (SECRYPT), proceedings of the 2010 international conference on* (pp. 1-10). IEEE.

Wang, A. H. (2010, June). Detecting spam bots in online social networking sites: a machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy* (pp. 335-342). Springer, Berlin, Heidelberg.

Wilson, C., Sala, A., Puttaswamy, K. P., & Zhao, B. Y. (2012). Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web (TWEB)*, *6*(4), 17.

APPENDIX

Table A1. Examples of most frequent languages in the non-English dataset (Dataset4)

Language	***size***	***Language***	***size***
Arabic (ar)	443949	Portuguese(pt)	837
Espaniols (es)	17263	Haitian (ht)	634
Turkish(tr)	12160	Persian(fa)	451
Japanese(ja)	3062	Estonian(et)	446
French (fr)	1877	Catalan(ca)	444
Tagalog(tl)	1107	German(de)	351

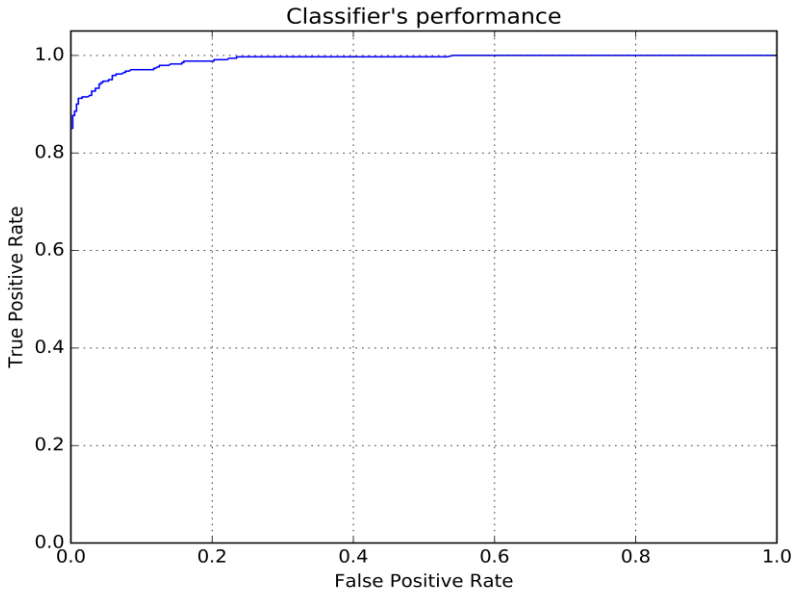

Figure A1. Performance of a classification model (random forest) on the non-English dataset (Dataset4)

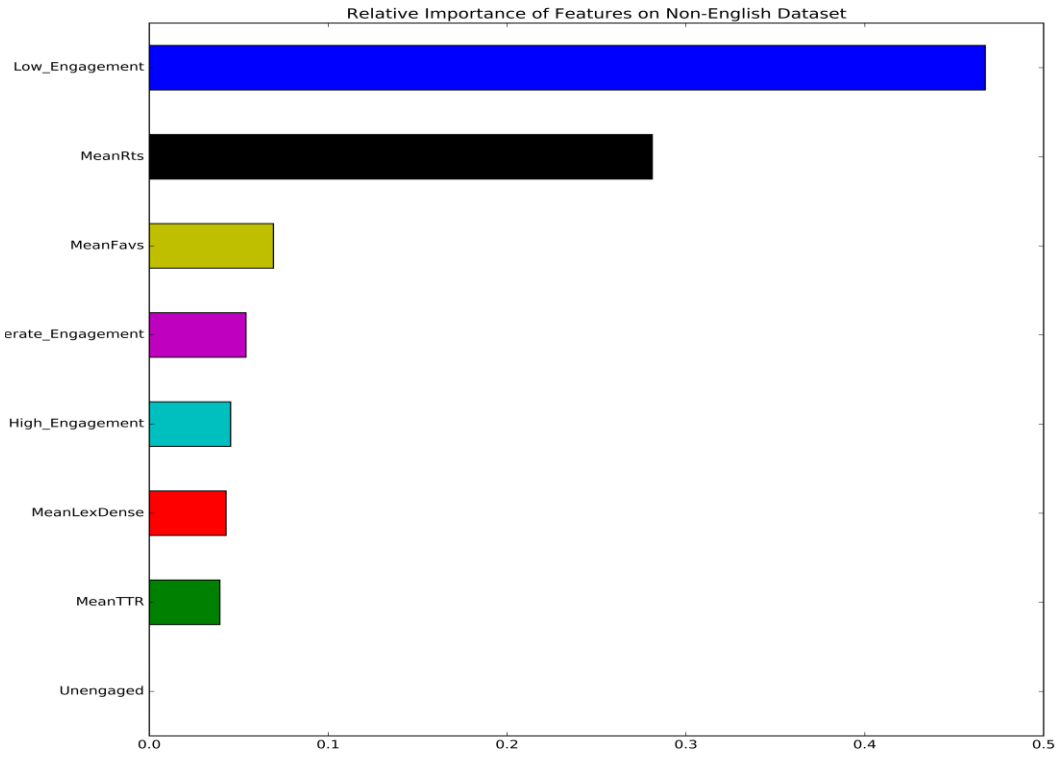

Figure A2. Feature importance of engagement and lexical features on Dataset4. Importance of each feature is assessed in the 0 – 1 continuum and the cumulative sum of all values for features sum to one. The low engagement feature is shown to be a strong distinguishing feature (about 0.45 value)