Assessing temporal and spatial features in detecting disruptive users on Reddit

***Abstract*—Trolling, echo chambers and general suspicious behaviour online are a serious cause of concern due to their potential disruptive effects beyond social media. This motivates a better understanding of the characteristics of disruptive behaviour on the internet and methods of detection. In this work we focus on Reddit which provides a rich social media platform for community focused interactions. Using network representations of user activity alongside temporal statistics and other features we assess the behaviour of a sample of potentially disruptive users, based on their assigned comment karma (an aggregate of a user's comment up-votes), relative to the wider population. We explore how these signals contribute to the accurate prediction of disruptive users, and note that this is achieved without requiring any semantic analysis. Our results show that it is possible to detect signs of disruptive behaviour with good accuracy using limited inputs that are primarily based on the reply patterns that users generate. This is of potential value for large-scale detection problems and operation across different languages.**

***Index Terms*—Reddit, Behaviour-Driven, Disruptive, Motifs, Complex Networks, Temporal**

I. Introduction

Internet trolling is the practice of disrupting user engagement with the intention to provoke a strong response (often negative) towards a discussion or user. This behaviour can be automated by social bots that masquerade as genuine users but often function at a faster pace [11], [14], [25]. The misuse of social media is a growing cause of public and academic concern. This encompasses issues including trolling [2], [13], [20], echo chambers [3], [5], [16], fake news [7] and suspicious social bot behaviour [10], [17]. While trolling is an issue across multiple communities [20], it is common to find troll-like behaviour within a political setting [1], [2], [13]. To combat this issue, methods have been applied on Twitter with bot detection [10] and fake account detection for generic online social networks [7].

In our research, we focus on the social news aggregation site Reddit specifically as the platform has been the subject of recent foreign interference from social bots manipulating users through political propaganda[1]. This is of significant concern given that it has the potential to disrupt the functioning of democracy.

As a platform, Reddit allows users to conveniently comment and share content within communities of interest. Participants can establish a reputation within a community of peers and engage with others having similar interests. This is achieved through so-called *subreddits* where users publicly post content of relevance to a particular topic. For example, the subreddit `r/funny` is a community dedicated posts that are amusing and humorous in content and `r/politics` is used to discuss the latest political affairs around the world.

Unlike other social media platforms, users on Reddit can express both positive or negative sentiment towards the comments and replies of other users. More specifically, users within a community can up-vote or down-vote a post depending on how well it is received within the community, in turn producing what is known as *karma*. Furthermore, users can produce comments and leave replies forming a nested discussion tree. As a result, Reddit serves as a rich platform to support direct user-to-user interaction through discussion threads. However this functionality means that users can be exposed to sophisticated echo chambers [3], [16] and filter bubbles [5] as well as more serious problematic behaviour such as *trolling*.

In this paper we contribute to the detection of disruptive behaviour on Reddit based upon a user's comment karma score. Users were labelled as 'disruptive' if the karma was negative ($k <= 0$) or 'normal' if positive ($k > 0$). We investigate new approaches to classifying troll-like activity in the presence of "normal" or non-disruptive user activity. Our approach is behavioural and network-based, where we consider the patterns of interactions that users establish, both in their structure and temporally. This involves defining egocentric reply networks as a component to frame our analysis and classify potential internet trolls in Reddit. We observe that relatively little research has addressed this important problem.

II. Related Work

The methods we present in this paper are similar to techniques used for processing bulletins board by modelling a directed network of users replying to others within nested discussion threads [19], [28]. These networks exhibit valuable metrics which provide the basis for analysing user behaviour such as leadership within discussions [6], assessing topic discovery with hierarchical quality [26] and predicting user interactions [15]. As a result, these network structures facilitate the discovery of interaction patterns and behaviour present within basic discussions.

Furthermore, temporal features also provide valuable insights towards better understanding user behaviour on social media. This is achieved by assessing the timings of various activities using spike train analysis to observe events as an ordered sequence. Evidence suggests that distinct temporal features contribute to the detection of social media bots,

[1]Reddit Transparency Report 2017: https://www.redditinc.com/policies/transparency-report

spammers as well under-performing students online [8], [12], [21].

In addition to this, research has been performed to understand the value of significant patterns in behavioural networks with examples such as situational understanding [4], information fusion [18], conspiracy theories [24], user influence [22], [23] and quality [9], [27]. However, little research has been performed to study the impact of user behaviour from the perspective of a single user within the discussion thread.

We observe that there is an opportunity to combine network-based and temporal features in detecting disruptive behaviour from bot-like activity. This has the potential to improve the prediction accuracy for classification of social bots and supports situational awareness for social media users and the platform itself.

III. Methods

We used Reddit's API to extract a sample of disruptive and non-disruptive normal users ($N = 794$ and $N = 850$ respectively). Firstly, we selected 100 random subreddits and sampled from 10 random posts within each subreddit with the intention to get a uniform sample of users across the platform. From this, we used the 'comment karma' score k assigned by Reddit for each user in our dataset. This derived from a calculation based upon the ratio of positive 'upvotes' and negative 'downvotes' given by other users. Overall, we processed a total of $N = 469,606$ comments with approximately 454 comments per non-disruptive user and 75.6 comments per disruptive user. We use karma as it represents the receptivity of the user's contribution to a discussion in a community, where something that is negatively received can be seen as being disruptive to the norms of that community.

We investigate whether network and temporal-based features of a user's behaviour can be used to predict whether they are considered disruptive or not. To begin, we generate an egocentric reply network for each user X by forming a directed edge between user X and the other users that X has either replied to or has received a reply from. For each such network, we count the frequency of induced 4-node subgraphs in a star formation, where X is central to the induced star. We choose 4-node subgraphs as this easier to process (in terms of computational overhead) and still preserves a suitable level of detail compared to that of 3-node subgraphs. There are 10 possible alternative configurations (see Figure 3). These capture the alternative edge configurations surrounding a target user, and the interactions taking place in terms of direction of communication. For each user X, we normalise the counts of each type of 4-node subgraph that are induced. This expresses the proportions of 4-node subgraphs in which X is the centre of the star, and allows us to present a profile for each user (e.g., Figure 3).

We supplement this with further frequency statistics relating to user activity on Reddit, including account age, number of historical comments, mean comments per week, and mean duration between comments. Together these features form a basis to characterise user activity and to predict disruptive behaviour based on karma.

TABLE I: Spearman correlation coefficients of temporal features with overall comment karma. All correlations are statistically significant $p<0.05$.

	# Comments	**Age**	**μ Comment rate**	**μ Duration**
Normal	0.459	0.441	0.559	-0.682
Disruptive	-0.173	-0.099	-0.08	0.114

IV. Results

A. Temporal features

To begin, we first consider attributes associated with user activity and comment timings. Figures 1a, 1b, 1c and 1d show histograms of comment count, account age, mean comment rate and mean comment duration, split between the two user groups.

The histograms in Figure 1 reveal distinct behaviour between normal and disruptive users. It is apparent that normal users produce a uniformly distributed number of comments, whereas disruptive users are less likely to have a large comment history. This "long-tail" log-normal type distribution is consistent, to varying degrees, across each histogram, where disruptive users are far more likely to less active between commenting (see Figure 1d) and much less mature in age (see Figure 1c).

To understand the relationship each variable has with respect to comment karma, we assess the correlation between each variable using Spearman tests (see Table I). Contrary to disruptive users, variables such as age and comment rate reveal a partial correlation suggesting that more-mature accounts are less likely to be used for disruptive behaviour on Reddit.

B. Egocentric reply networks

Figure 2 provides two examples of user interactions, where the centre node is our target user and edges going out indicate a reply and edges being received represent receiving a reply. These are featured around individual users rather than groups. The examples provided here are used to illustrate that normal users are more active that disruptive users and are more likely to participate in a two-way conversation.

We observe the user interactions by counting and enumerating over all 4-node subgraphs resembling a tree-like structure where the root node serves as our target user of interest (see Figure 3). This allows us to examine the occurrence of edge configurations surrounding an individual user's interactions, when we consider the induced 4-node star configurations. The proportion of all induced 4-node star configurations for each user is presented in Figure 3.

Each of the subgraph profiles presented in Figure 3 provide the basis for inferring the structure of social interactions. The frequency plot reveals how the third (one-in, two-out), fifth (one-in-out, one-in, one-out) and seventh (all-out) subgraphs from Figure fig:subgraphs stand out compared with normal user subgraphs. Normal users produce more consistent activity

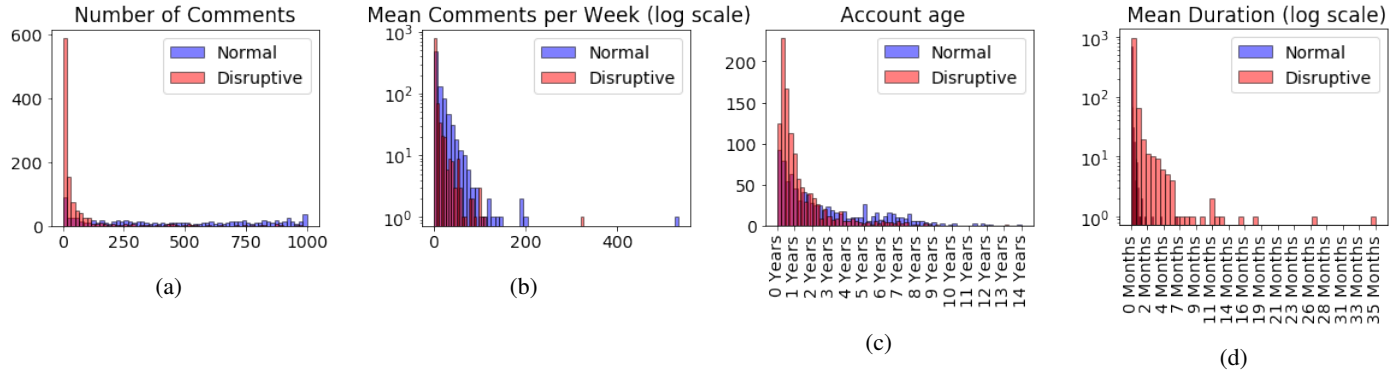
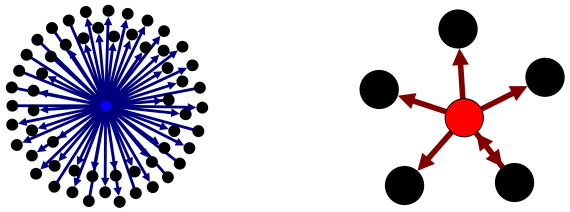


Fig. 1: Histogram comparison of normal and disruptive users with respect to user activity



(a) (b)

Fig. 2: Examples of typical normal (left) and disruptive (right) egocentric user reply networks

on the second (two-in, one-out), third (one-in, two-out) and fifth (one-in-out, one-in, one-out) subgraphs in Figure 3.

C. Prediction

Combining subgraph frequencies and temporal features together we train three binary classifiers to classify user behaviour as either non-disruptive "normal" (ND) or disruptive (D). Firstly, we use binary logistic regression (BLR) using the LIBLINEAR solver. Secondly, a support vector machine (SVM) using an (RBF) kernel. Finally, a random forest classifier (RFC) with $N = 100$ trees at a max depth of $D = 2$. We combine features with the intention to understand if the accuracy of these models can be improved. These result can be found in Table II where we compute the precision (P), recall (R), F1-score (F1) and accuracy (A) using a train-test split ratio of 75:25.

From the prediction results, it is clear that subgraph features perform better in comparison to temporal features. The temporal features appear to lack support for prediction in the case of BLR and SVM however, RFC consistently outperforms BLR and SVM on every data set. Overall, the accuracy of the model is significantly improved when both subgraph counts and temporal features are combined.

V. Discussion

The results help to reveal the value that egocentric subgraphs and temporal features have towards assessing and classifying user behaviour. The temporal features we collected provide early insights into the subtle differences between normal and disruptive activity. Simple attributes such as comment count and age provide initial indications as to whether behaviour is suspicious.

Counting the 4-node star subgraphs that are induced by users' communication provide a means to discover the relationship a user has with other users irrespective of the temporal domain. Our results show distinct differences between the group of disruptive users and the "normal" users, noting that some subgraphs are much more likely to be present for disruptive users and vice versa.

In our first example, we observe that the subgraphs featured in Figures 4a and 4c indicate that normal users are more likely to receive a reply and form two-way conversation (4c). In particular, we found that disruptive users vary much in the subgraphs featured in Figure 4b and 4d providing evidence that disruptive users are likely to initiate a reply to a user and are less likely to receive a reply or two-way response.

Furthermore, we report that disruptive users are less likely to preserve at least one symmetric edge, hence the subgraph in Figure 4c is significantly lacking in appearance. In contrast, normal users are more active and consistent across the nearly all subgraphs featured in Figure 4 where normal users are likely to receive at least one reply during the discussion potentially leading to a two-way conversion at some point (Figures 4b and 4c).

Our prediction analysis for user classification provides strong evidence for basic temporal statistics complementing subgraph frequencies. While classification performs reasonably well using each feature set in isolation, the combination of the two produces a significant improvement overall. The three classifiers we deploy in this paper serve as a basis to demonstrate the potential for behaviour classification and proof-of-concept, however we note that alternative classifiers can be used, and may further improve performance.

VI. Conclusion and Future Work

The methods and results featured in this paper indicate that the behaviour of users in Reddit, independent of the content and semantics, is sufficient for relatively accurate categorisation of disruptive users. This is based on extracting subgraph

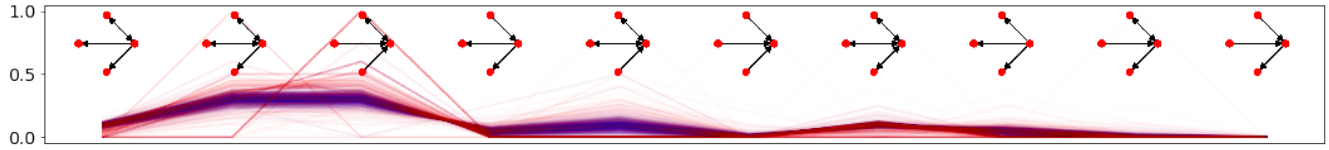


Fig. 3: Frequency plot of all featured subgraphs represented as a ratio of disruptive (red) and normal (blue) users overlapping

TABLE II: Prediction results of three leading classifiers for temporal features, subgraph features and the two combined improving the overall accuracy.

		Temporal				**Subgraphs**				**Both**			
		P	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**
BLR	**ND**	0.51	0.87	0.64	0.55	0.66	0.98	0.79	0.7	0.64	0.95	0.76	0.7
	D	0.72	0.28	0.41		0.94	0.35	0.51		0.88	0.41	0.56	
SVM	**ND**	0.71	0.38	0.5	0.67	0.74	0.96	0.83	0.78	0.72	0.92	0.81	0.77
	D	0.63	0.87	0.73		0.91	0.56	0.7		0.88	0.61	0.72	
RFC	**ND**	0.82	0.75	0.78	0.81	0.79	0.87	0.83	0.8	0.85	0.89	0.87	0.86
	D	0.8	0.86	0.83		0.81	0.71	0.76		0.88	0.83	0.85	

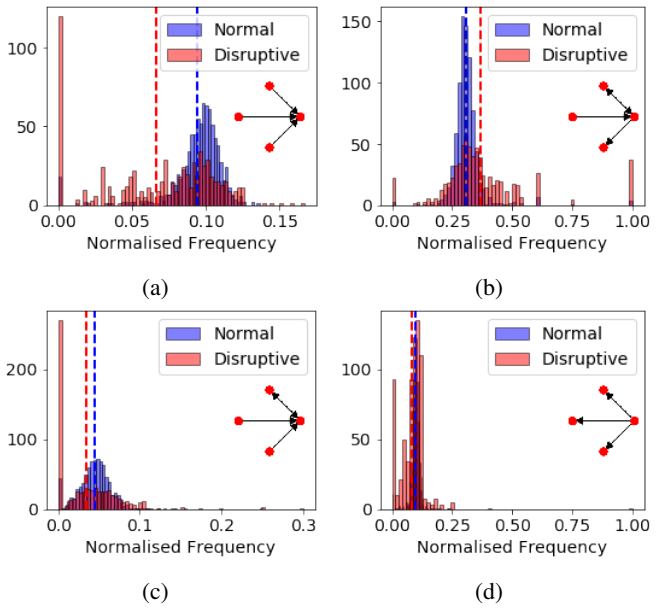


Fig. 4: Collection of motifs discovered comparing disruptive against normal user activity displayed as frequency histograms shown with mean marked by dashed lines.

features from user reply networks, where we can begin to describe the fundamental features surrounding user-to-user interaction. As a result, we determine that reply networks are a valuable tool for analysing discussion threads between users. Additionally, temporal features can be combined with this approach to improve classification accuracy. However, we note that our method doesn't consider repeated discussions and the depth of two-way debates between users. Future work can involve considering weighted edges (or timestamped edges) for repeated replies and more-advanced classifiers can further improve overall performance and classification accuracy. In view of these results, these methods provide key insights into communication patterns of disruptive users such that moderators can reduce this behaviour by minimising and discouraging certain interactions.

REFERENCES

[1] An, J., Kwak, H., Posegga, O., Jungherr, A.: Political discussions in homogeneous and cross-cutting communication spaces. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 13, pp. 68–79 (2019)

[2] Bergstrom, K.: "don't feed the troll": Shutting down debate about community expectations on reddit. com. First Monday **16**(8) (2011)

[3] Boutyline, A., Willer, R.: The social structure of political echo chambers: Variation in ideological homophily in online networks. Political Psychology **38**(3), 551–569 (2017)

[4] Braines, D., Felmlee, D., Towsley, D., Tu, K., Whitaker, R.M., Turner, L.D.: The role of motifs in understanding behavior in social and engineered networks. In: Next-Generation Analyst VI. vol. 10653, p. 106530W. International Society for Optics and Photonics (2018)

[5] Bruns, A.: Echo chamber? what echo chamber? reviewing the evidence (2017)

[6] Buntain, C., Golbeck, J.: Identifying social roles in reddit using network structure. In: Proceedings of the 23rd international conference on world wide web. pp. 615–620 (2014)

[7] Cao, Q., Sirivianos, M., Yang, X., Pregueiro, T.: Aiding the detection of fake accounts in large scale social online services. In: Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12). pp. 197–210 (2012)

[8] Costa, A.F., Yamaguchi, Y., Traina, A.J.M., Jr, C.T., Faloutsos, C.: Modeling temporal activity to detect anomalous behavior in social media. ACM Transactions on Knowledge Discovery from Data (TKDD) **11**(4), 1–23 (2017)

[9] Cunningham, P., Harrigan, M., Wu, G., O'CALLAGHAN, D.: Characterizing ego-networks using motifs. Network Science **1**(2), 170–190 (2013)

[10] Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botornot: A system to evaluate social bots. In: Proceedings of the 25th international conference companion on world wide web. pp. 273–274 (2016)

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

[12] Ferraz Costa, A., Yamaguchi, Y., Juci Machado Traina, A., Traina Jr, C., Faloutsos, C.: Rsc: Mining and modeling temporal activity in social media. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 269–278 (2015)

[13] Flores-Saviaga, C.I., Keegan, B.C., Savage, S.: Mobilizing the trump train: Understanding collective action in a political trolling community. In: Twelfth International AAAI Conference on Web and Social Media (2018)

[14] Gehl, R.W., Bakardjieva, M.: Socialbots and their friends: Digital media and the automation of sociality. Taylor & Francis (2016)

[15] Glenski, M., Weninger, T.: Predicting user-interactions on reddit. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 609–612 (2017)

[16] Guest, E.: (anti-) echo chamber participation: Examining contributor activity beyond the chamber. In: Proceedings of the 9th International Conference on Social Media and Society. pp. 301–304 (2018)

[17] Hurtado, S., Ray, P., Marculescu, R.: Bot detection in reddit political discussion. In: Proceedings of the Fourth International Workshop on Social Sensing. pp. 30–35 (2019)

[18] Iribarren, J.L., Moro, E.: Impact of human activity patterns on the dynamics of information diffusion. Physical review letters **103**(3), 038702 (2009)

[19] Medvedev, A.N., Delvenne, J.C., Lambiotte, R.: Modelling structure and predicting dynamics of discussion threads in online boards. Journal of Complex Networks **7**(1), 67–82 (2019)

[20] Merritt, E.: An analysis of the discourse of Internet trolling: A case study of Reddit. com. Ph.D. thesis (2012)

[21] Młynarska, E., Greene, D., Cunningham, P.: Time series clustering of moodle activity data. In: 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), University College Dublin, Dublin, Ireland, 20-21 September 2016 (2016)

[22] Müngen, A.A., Kaya, M.: Influence analysis of posts in social networks by using quad-motifs. In: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). pp. 1–5. IEEE (2017)

[23] O'Callaghan, D., Harrigan, M., Carthy, J., Cunningham, P.: Network analysis of recurring youtube spam campaigns. In: Sixth International AAAI Conference on Weblogs and Social Media (2012)

[24] Samory, M., Mitra, T.: Conspiracies online: User discussions in a conspiracy community following dramatic events. In: Twelfth International AAAI Conference on Web and Social Media (2018)

[25] Subrahmanian, V., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F.: The darpa twitter bot challenge. Computer **49**(6), 38–46 (2016)

[26] Weninger, T., Zhu, X.A., Han, J.: An exploration of discussion threads in social news sites: A case study of the reddit community. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013). pp. 579–583. IEEE (2013)

[27] Wu, G., Harrigan, M., Cunningham, P.: Classifying wikipedia articles using network motif counts and ratios. In: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration. pp. 1–10 (2012)

[28] Zhongbao, K., Changshui, Z.: Reply networks on a bulletin board system. Phys. Rev. E **67**, 036117 (Mar 2003). https://doi.org/10.1103/PhysRevE.67.036117, https://link.aps.org/doi/10.1103/PhysRevE.67.036117