This Must Be the Place: Predicting Engagement of Online Communities in a Large-scale Distributed Campaign

Abraham Israeli, Alexander Kremiansky, Oren Tsur
{isabrah,kremians}@post.bgu.ac.il,orentsur@bgu.ac.il
Department of Software and Information System Engineering
Ben-Gurion University of the Negev
Israel

ABSTRACT

Understanding collective decision making at a large-scale, and elucidating how community organization and community dynamics shape collective behavior are at the heart of social science research. In this work we study the behavior of thousands of communities with millions of active members. We define a novel task: predicting which community will undertake an unexpected, large-scale, distributed campaign. To this end, we develop a hybrid model, combining textual cues, community meta-data, and structural properties. We show how this multi-faceted model can accurately predict large-scale collective decision-making in a distributed environment. We demonstrate the applicability of our model through Reddit's `r/place` – a large-scale online experiment in which millions of users, self-organized in thousands of communities, clashed and collaborated in an effort to realize their agenda.

Our hybrid model achieves a high F1 prediction score of 0.826. We find that coarse meta-features are as important for prediction accuracy as fine-grained textual cues, while explicit structural features play a smaller role. Interpreting our model, we provide and support various social insights about the unique characteristics of the communities that participated in the `r/place` experiment.

Our results and analysis shed light on the complex social dynamics that drive collective behavior, and on the factors that propel user coordination. The scale and the unique conditions of the `r/place` experiment suggest that our findings may apply in broader contexts, such as online activism, (countering) the spread of hate speech and reducing political polarization. The broader applicability of the model is demonstrated through an extensive analysis of the WallStreetBets community, their role in `r/place` and four years later, in the GameStop short squeeze campaign of 2021.

CCS CONCEPTS

• **Computing methodologies** → Neural networks; • **Information systems** → **Social networks**; • **Applied computing** → **Ethnography**; **Sociology**; • **Social and professional topics** → *Cultural characteristics*;

KEYWORDS

Online Communities, Natural Language Processing, Social Networks, Computational Social Science, Reddit, rPlace, wallStreetBets, GameStop

ACM Reference Format:
Abraham Israeli, Alexander Kremiansky, Oren Tsur. 2022. This Must Be the Place: Predicting Engagement of Online Communities in a Large-scale Distributed Campaign. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3485447.3512238

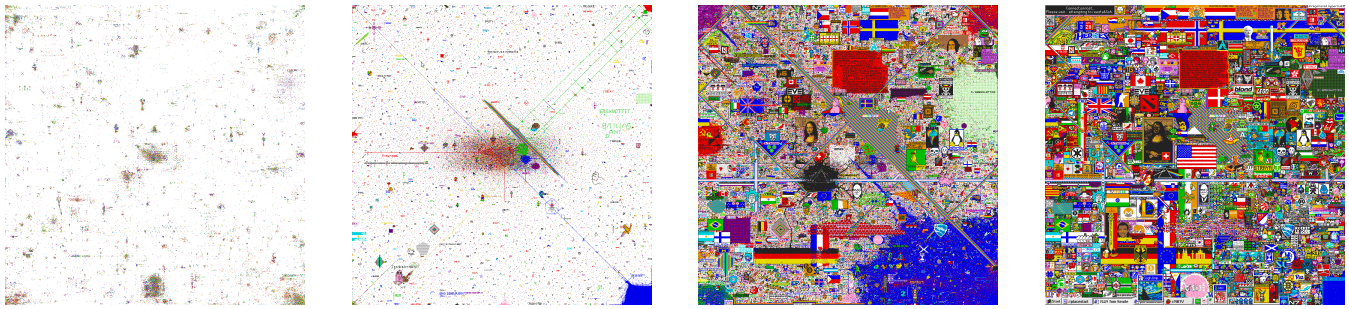
WWW '22, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512238

1 INTRODUCTION

Group dynamics, organization, norms, structure, and collective action are at the core of social sciences research, e.g., [22, 27, 46, 61, 62, 81] to mention just a few works. The surge of online activity provides an unprecedented opportunity to study these patterns organically and at a large scale [45, 53]. Coordinated online activity was found to fuel street protests [22, 34, 37], stimulate political conversation [3], model the support for a social change [31], and change traditional financial behaviours [49]. The susceptibility of millions of users to emotional manipulation [9, 42] and the echo-chamber effect [4, 17], were utilized to disseminate misinformation, discredit democratic institutions and to interfere with political processes [8, 28, 30]. Notorious examples are the activity of Russian trolls [7, 35, 57] and micro-targeting practices used by firms like Cambridge Analytica [33, 78]. Understanding the factors that impact the behavior of online communities is a crucial step toward increased resilience of online communities to manipulation efforts [5]. Despite its significance, a large-scale study of the ways decentralized communities operate, internally and with respect to other communities, is scarce (see survey in Section 2).

In order to model collective actions, we develop a hybrid algorithm to predict the collective action of thousands of communities, integrating multiple signals, ranging from the language-use to the community structure. Specifically, we use the activity of Reddit communities before the `r/place` "experiment" – a massive online game, which we view as an external shock to the platform – predicting the community reaction to the shock. We refer to `r/place` as a naturally occurring-large scale controlled experiment. The simple design of the experiment (explained below) provides a unique opportunity to study how decentralized online communities act in response to the external shock. Our analysis illuminates the factors that facilitate the undertaking of a collective effort.



(a) t_0 + 2 hours (b) t_0 + 7 hours (c) t_0 + 25 hours (d) t_0 + 72 hours (final state)

Figure 1: The evolution (from left to right) of the `r/place` experiment virtual canvas.

Reddit. Reddit[1] is one of the most popular websites worldwide, falling short behind Google and YouTube, with an average of over 430 million active users per month [67]. Reddit evolved into an active system of dedicated forums (3.1M forums, as of September 2021). Forums' URLs are marked with a 'r/' followed by the forum's name, and are therefore called *subreddits*. Subreddits are usually topical, although the specificity of topics range widely, from the very broad *r/music* or *r/politics* subreddits to the more specific *r/talkingheads* or *r/brexit*. Each subreddit forms a community in which members (called *redditors*) can start a new discussion thread, add comments to a thread, up/down-vote posts, etc. Each subreddit is maintained by a few moderators and has its own rules specifying community code of conduct (or lack of). In the current study, we view each subreddit as a community, having its own internal language, style, interaction dynamics, and norms.

r/place – a large-scale online "experiment". The `r/place`[2] subreddit was conceived by Reddit as a social experiment gag for the 2017 April Fools' Day. A shared white canvas of one million pixels (1000 x 1000) appeared in a new subreddit called r/place. Redditors could change the color of any pixel, one pixel at a time. Every change of a pixel was reflected on the *shared* canvas, thus viewed by all users. Viewing it as a weird online game, users were not aware of any clear purpose of the experiment. To the users' surprise, the experiment abruptly terminated after 72 hours. During its 72 hours of operation, the canvas attracted 16.1 million pixel changes performed by 1.2 million redditors. Figure 1 presents four snapshots attesting to the progression of the canvas' state, from its early chaotic state to its final form – an intricate collage of complex logos and artworks.

The social experiment of `r/place` was controlled, to some degree, as Reddit's developers served as moderators and enforced certain rules – albeit users were oblivious to those rules. Two examples of such rules are: (i) Only accounts created prior to the unexpected beginning of the experiment could manipulate the color of a pixel, and (ii) Once a redditor manipulated the color of a pixel, he or she was blocked by the system for a random time (5–20 minutes), thus effectively preventing any single redditor from having a significant influence on the canvas.

A careful look at the final state of the canvas (Figure 1d) shows[3] that many of the symbols reflect some form of group identity, including national flags, university and sports teams logos, developer's communities (e.g., Linux), and online gaming communities. Other artworks include popular memes and reproductions of iconic art works (e.g., Da Vinci's Mona Lisa and Van Gogh's Starry Night). Some communities pursued a more abstract presence (e.g., 'the blue corner', bottom right in Figures 1b-1d) or the 'rainbow road' which crosses the canvas diagonally. Few communities tried to drew hate-related symbols (e.g., swastikas and Pepe the Frog), or vandalized other symbols. The most successful vandalizing effort was 'The-Black-Void' (TBV) – an ever expanding black fractal-like shape in the middle of the canvas, noticeable in the center of Figure 1c. It is important to note that while TBV's declared purpose was vandalizing other symbols, clashes erupted between many other communities competing for "real estate" for their logo. In this work we use the complex signals *preceding* the experiment in order predict community engagement in the game. The rich community dynamics observed *during* the experiment will be addressed in a subsequent work.

The remainder of the paper is organized as follows: in Section 2 we provide a brief review of the relevant literature. In Section 3 we describe the data in detail and specify the prediction task. In Section 4 we present our computational approach, the various algorithms we use and the experimental setting. In Section 5 we present the results, provide a social interpretation to the results, present an error analysis, and dive into one exemplar of the *r/WallStreetBets* community. Lastly, in Section 6 we summarize our work and suggest future research directions.

2 RELATED WORK

Community behaviour has been studied for decades. Lewin [46] established the modern field of group dynamics, organization, norms and collective action. Naturally, recent research puts increased emphasis on online communities [45, 54, 83].

Many works study collaborative patterns in crowd sourced projects like Q&A sites, Wikipedia or open-source projects [38, 39, 68]. These works differ from ours in that such works model individuals forming a community dedicated to a project (whether a wiki page, a

[1] https://www.reddit.com
[2] https://www.reddit.com/r/place/

[3] A high resolution image can be found at https://bit.ly/39e1E9a

QA forum or an open source project), whereas we are interested in existing organic communities that are mobilized to participate in a new campaign.

A general overview of the uses of Reddit data to study its communities is presented by Medvedev et al. [52]. Reddit communities, patterns of information sharing, and evolving community norms are studied in a series of works [13, 14, 16, 21, 54, 60, 63, 69, 75, 79], among others. These works address various aspects in community organization as an interest group, the dealings with heated topics and the inherent tension between anonymity and identity. Recent works study the structure and other characteristics of Reddit communities [29, 44, 51, 83, 84]. Zhang et al. [83] suggests a new representation of communities through two complex dimensions ('distinctiveness' and 'dynamicity') and prove that these representation successfully reflects different user engagement measures (e.g., retention rate). Kumar et al. [44] suggests a novel way to model conflict between communities on the web. Their approach combines textual features with communities meta features (all in the form of numeric vectors). This methodology is similar to the way we combine different representations of online communities. Datta and Adar [15] expand this work and study the landscape of conflicts among subreddits. Other works focus on a single community, presenting its uniqueness and norms [2, 10, 36]

Works using the rich `r/place` data are still scarce. The dependency between logo size and canvas density over time is studied by Müller and Winters [58], while latent patterns of collaboration between individuals are modeled by Rappaz et al. [69] and Armstrong [1]. Conflicts between communities during the `r/place` campaign are studied by [76]. These works differ from ours in four fundamental ways: (i) These works focus on collaborations between individual redditors and the ways they correlate with other individuals, using the canvas as the main shared focal point. We, on the other hand, are interested in coordination on the *community* level. Thus, we model the ways a community is organized and operates reflect on its interest and ability to undertake a competitive, large scale community effort. Consequently, (ii) We define a *novel prediction task* – which community will engage in the experiment. (iii) We only use data *preceding* the experiment while prior works use the data generated during the experiment, and (iv) We combine *multiple types of features* (language, community structure, user dynamics, etc.) while others use only the `r/place` pixel allocations data.

To the best of our knowledge, this is the first work to make use of historical Reddit data, language use and community meta features in the context of organizational development at large and in the context of the `r/place` experiment in particular.

3 DATA

For the purpose of this research, we collected two datasets. The first, denoted *DS*1, is composed of all data posted on reddit in the six month *prior* to the experiment (10/1/16–3/31/17). The second dataset, *DS*2, contains all data posted *during* the 72 hours of `r/place` (3/31–4/3/17) and is used to validate and improve the gold labels, as we describe below. We also used Reddit's API[4] to retrieve meta data per each subreddit (e.g., creation time, number of subscribers).

[4]https://www.reddit.com/dev/api/

Predicting that an inactive community (i.e., by which no post was added to the platform in the months before the experiment) will not engage in the game is trivial. Hence, we created a labeled dataset, balancing the positive samples (participating communities) coupled a negative set with relatively similar attributes. The careful process of creating the balanced gold standard is described below.

Participating Communities (S^+). The positive set is composed of the communities that took part in the experiment. While some of these communities are easy to pinpoint, others require some effort. Hence, we took two approaches toward identifying the participating communities: (i) Utilizing the place-atlas[5] resource, and (ii) A machine learning approach, based on the *DS*2 dataset.

The place-atlas is a crowd-sourced effort created by a group of redditors after the termination of the experiment. This platform allowed other redditors to "claim their victory" by posting their community name, the location of their final artwork, and some extra details about it (e.g., some logos were the result of joint efforts by multiple communities). Manually analyzing the atlas data we identified 802 communities participating in `r/place`. However, most of the artworks indexed in the place-atlas are attributed to 'winners' – communities whose effort is recognized on the canvas at the termination of the experiment. Many other communities did participate, but failed to leave a mark, as their logo was vandalized or over-ridden by a competing community. In order to identify these, as well as other communities that were not indexed in the place-atlas, we trained a simple classification model based on regular expressions.

We used the data that were generated *during* the 72 hours of the experiment (*DS*2) matching specific regular expressions that could link communities to the experiment (e.g., *draw*, *rplace*, *pixel*, *canvas*). All matched subreddits are *candidates* for the positive set. This pool of candidates needs to be filtered further since many communities only discuss the sudden `r/place` hype but do not end up participating. We thus manually labeled ~100 communities as positive (drawing) and another ~100 communities as negative (not drawing). This way we had a small labeled seed to be used in a bootstrapping manner, discovering more participating communities with each iteration. Following Kozareva and Hovy [41], only candidates with a very high confidence score were added to the seed in each iteration. After three iterations we discovered a few hundreds of new positive examples (i.e., communities that were not indexed in the place-atlas). We manually validated[6] a sample of these newly found positive examples, and found it highly accurate (94.9% accuracy over 156 communities). Adding the discovered positive communities to those that were identified by the place-atlas, resulted in a set of 1231 communities. We refer to this set of communities as S^+.

Non-participating communities (S^-). Most sureddits did not take an organized part in the `r/place` experiment. Using all subreddits not in S^+ as the negative set is inappropriate for several reasons, ranging from the relative simplicity of classifying an extremely unbalanced datasets (~1.2K positive vs. ~1.2M), to the fact

[5]https://draemm.li/various/place-atlas

[6]By reading the posts that were written in the subreddit during the time of the experiment. We looked for explicit statements by the subreddit members describing a joint effort of participation in `r/place`.

Table 1: Subreddits statistics for S^+ (drawing), S^- (not drawing) and S (all the subreddits active in the 6-month period prior to the `r/place` experiment). Mean values are computed over all submissions in a subreddit. *Age* denotes the number of days since the subreddit was established. Inactivity denotes the number of days from the most recent post to the launch of `r/place` (effectively bound by 180 – the span of $DS1$).

	S^+(1231)				S^-(1289)				S (258.8K)			
	Total	Mean	Median	STD	Total	Mean	Median	STD	Total	Mean	Median	STD
Subscribers	219.7M	178.5K	24.1K	1.13M	409.2M	317.5K	15.9K	1.9M	1.7B	6.6K	30.0K	234.6K
Active Users	6.9M	5.6K	1.47K	15.9K	4.4M	3.47K	0.7K	12K	54.7M	211.4	2.0	5.9K
Age	–	2.15K	2.34K	0.89K	–	1.7K	1.7K	1.01K	–	0.98K	0.87K	0.8K
Inactive	–	0.9	0.0	6.7	–	2.23	0.0	11.1	–	55.2	39.0	52.74
Submissions	12.8M	10.4K	2.3K	55.6K	12.3M	9.6K	2.4K	51.7K	53.3M	209.1	3.0	6.1K
Comments	–	12.0	8.6	25.3	–	6.1	3.9	7.5	–	1.4	0.25	23.9

that many subreddits are old, inactive, very new or very small (in terms of active members). In order to create a balanced and challenging dataset we opted to subsampling – trying to have the meta features of communities in the negative set similar to those in the positive set. The negative training set, denoted S^-, was therefore created in the following heuristic manner: we first matched each drawing community in S^+ to a non-drawing community of similar size where size is measured by: (i) Number of subscribers, or (ii) Number of submissions posted in the community page. Both measures have pros and cons, so we experimented with both. Both heuristics ended up showing very similar results. In this paper we only report results obtained from the second heuristic.

General statistics describing various aspects of the datasets are presented in Table 1. Calculations are based on the $DS1$ (historical) dataset. Many subreddits(~950K) did not show any activity in the six month span of $DS1$, thus are not accounted for in the table. For example, the table shows that communities in S^+ and S^- are bigger and more active compared to a random Reddit community.

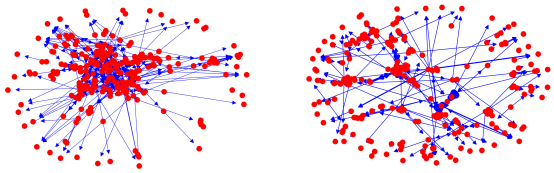
We wish to reiterate that $DS2$ is only used for the creation of the gold standard (participating/non-participating) feature. In the remainder of the paper we use *only* the "historical" data ($DS1$) in representing and predicting community participation.

3.1 Community Representation

Communities are multifaceted and could be characterized from different perspectives. To this end, we represent Reddit communities by features of three general types: (i) Textual features, (ii) Meta-features, and (iii) Network features.

Textual features. We normalized the data by lower casing, tokenization, removing punctuation and conversion of full URL addresses to the domain name (e.g., www.youtube.com/XYZ → www.youtube.com). We used both classic Bag-of-Words (BOW) representation for the GBT model (see Section 4) and word embeddings representation for the neural models. In the BOW models, we used the *TF-IDF* score [71]. Using bigrams/trigrams did not yield any improvement so all BOW results are reported for the unigram setting, using the 300 most frequent tokens.

Meta-features. Reddit was originally conceived as an interest-based message board, therefore a subreddit can be represented by meta features like the number of users subscribed to a community, the average number of posts per day, the average number of votes per post, the "age" of the community (days since its creation), etc. We used a total of 25 meta-features for each subreddit.



(a) *r/srilanka* subreddit (b) *r/insults* subreddit

Figure 2: Network structure of two communities with a similar number of nodes. Both networks were rendered using the Fruchterman-Reingold layout.

Network features. A community can be characterized by the patterns of communication between its members. These interaction patterns could be thought of as a social network in which a direct reply by user u to a post by user v constitutes a directed edge $u \rightarrow v$. These networks provide another perspective on the organizational principles of a community and the dynamics between its members. The intuition behind the use of the network perspective is illustrated through Figure 2. The figure shows the network structure of two subreddit communities with a similar number of nodes: *r/srilanka* ($\in S^+$) and *r/insults* ($\in S^-$). It is visually apparent that these communities are organized in a very different way with *r/srilanka* presenting a tighter community structure. In total, we used 32 network statistics as features (e.g., #nodes, #edges, avg. and std. of various centrality measures, #triangles)[7]

4 EXPERIMENTAL SETTING

We cast the likelihood of a community to participate in `r/place` as a classification problem. We experiment with an array of algorithms ranging from a logistic regression based on bag of words to a neural architectures, taking a collection of complex features as input. The algorithms we experiment with could be divided into two broad

[7]The list of all meta and network features, together with a short description of each, is provided in Table 4 in Appendix B.

categories: sequential and non-sequential models. Sequential models take the sequential nature of the data (text, discussion threads) into account while non-sequential models ignore this structure. We consider a number of classifiers of each type. However, due to space constraints, we report only on the following five: Gradient Boosted Decision Trees (GBT), Multi-Layer Perceptron (MLP), Zero-Shot BERT with an MLP layer, Parallel LSTM with an MLP layer, and a Max-Pooling CNN with an MLP layer. A deviance loss function is used for the GBT algorithm and a binary log-loss is used for the MLP and the parallel LSTM models. We provide a broader explanation for each classifier in Appendix A. Each subreddit is represented by an array of features extracted from the "history" of the community (*DS*1) – meta features, textual features, and network features (see Section 3).

We execute all algorithms in an ablation manner, in order to learn the importance of the different feature types. Accuracy, Precision, Recall, and F1-score are reported for each setting. The different algorithms are optimizing the F1-score as precision and recall are equally important, given the task definition. We evaluate all algorithms and settings using a stratified 5-fold cross validation. Neural architectures are restricted to a maximum of ten epochs with an early stopping.

For the MLP implementation, we apply a simple neural model – a single hidden layer of 150 nodes. For the parallel LSTM deep neural model (see Figure 5a in Appendix A), we setup 150 hidden units in each LSTM hidden state.

We use the sklearn [65] implementation of the GBT algorithm. We use the DyNet [59] and PyTorch packages [64] for building the neural architectures, due to the dynamic construction of the computation graph which is highly efficient given the high variance in number and length of textual data. We make the code developed as part of the research and all data that were collected public in our project's GitHub repository.[8]

Table 2: Prediction results. L/M/N denote *linguistic*, *meta* and *network* features, respectively. The Zero-shot BERT, parallel LSTM, and parallel CNN models require linguistic features in all settings, hence only three options were evaluated for these models.

Model	Features	Precision	Recall	F1-Score
GBT Models	L	0.776±0.027	0.758±0.033	0.766±0.022
	M	0.766±0.027	0.8±0.039	0.782±0.017
	N	0.67 ±0.044	0.75±0.031	0.707±0.027
	M + N	0.765±0.062	0.814±0.031	0.788±0.011
	L + M	0.8±0.024	0.82±0.027	0.81±0.016
	L + M + N	0.814±0.03	**0.84±0.018**	**0.826±0.011**
MLP Models	L	0.76±0.049	0.74±0.063	0.74±0.017
	M	0.737±0.033	0.632±0.031	0.679±0.013
	N	0.68±0.054	0.501±0.134	0.57±0.083
	M + N	0.73±0.028	0.76±0.067	0.77±0.02
	L + M	0.78±0.039	0.82±0.064	0.784±0.021
	L + M + N	0.755±0.056	0.82±0.006	0.78±0.021
Zero-Shot BERT Models	L	0.582±0.006	0.825±0.085	0.681±0.028
	L + M	0.697±0.036	0.806±0.108	0.744±0.047
	L + M + N	0.761±0.062	0.686±0.128	0.713±0.051
Parallel LSTM Models	L	0.741±0.06	0.752±0.085	0.741±0.03
	L + M	0.743±0.024	0.799±0.07	0.767±0.011
	L + M + N	0.779±0.04	0.756±0.089	0.763±0.032
Parallel CNN Models	L	0.761±0.019	0.722±0.08	0.737±0.037
	L + M	0.795±0.012	0.726±0.062	0.756±0.021
	L + M + N	**0.825±0.02**	0.753±0.042	0.786±0.021

5 RESULTS AND ANALYSIS

5.1 Prediction Accuracy

The results obtained by each model in the various ablation settings are detailed in Table 2. The parallel CNN neural model, using all feature types, achieved the best precision score (0.825). However, the best F1-score (0.826) was obtained by combining all feature types in the GBT model. Using each feature type independently performs well, yielding a significant improvement compared to a random baseline. The explicit network features play a lesser role compared to other feature types.

Using a single feature type, the highest F-score is obtained by the community meta features (GBT), beating the language features by a small but a significant margin. However, this holds only for the GBT model, probably due to the power of the word embeddings used in the neural models. We note that while trailing behind, network features alone achieved a decent (∼ 20%) improvement over a random baseline. These results show that all feature types capture a signal inherent to the community's participation in the experiment, whether linguistically or structurally.

Looking at ablation tests of the best performing algorithm, combining all three feature types achieves superior results over all subsets of features. Surprisingly though, three of the neural models (i.e., the MLP, the Zero-Shot BERT, and the Parallel LSTM) performed better without the network features. This does not hold for the CNN – the best performing neural architecture we tried. These results suggest that while the network features are instrumental in improving the classification, this feature type could benefit from further tuning.

As noted, the GBT model performs better than the neural models across all settings. This result demonstrates the often overlooked limitations of neural networks and language models [55]. Specifically, we attribute this result to the discrepancy between the relatively small number of instances in each class and the richness of the feature types. We elaborate on that in Section 5.5.

5.2 Error Analysis

We further conducted an error analysis, focusing on the false positive/negative predictions of the best performing model.

Table 3 contains the most striking errors of the GBT model. Interestingly, many of the false negative cases are 'ImagesOf<geolocation>' subreddits. While these communities have a strong and defined identity, related to the specific location in focus, they present patterns of general picture-posting communities rather than those

[8] https://github.com/NasLabBgu/rplace-engagement-prediction

Table 3: The ten false positive/negative errors with the highest deviation from the expected class score. Communities in each type are ordered according to the likelihood that the GBT model assigns the community to participate in `r/place` (values in brackets).

False Positives	False Negatives
r/MemeEconomy (0.97)	r/ImagesOfTennessee (0.03)
r/mega64 (0.94)	r/ImagesOfNewZealand (0.04)
r/KillLaKill (0.93)	r/RealGirls (0.05)
r/SubredditDrama (0.93)	r/ImagesOfColorado (0.05)
r/nomanshigh (0.93)	r/Pigifs (0.06)
r/MST3K (0.93)	r/thetreesnetwork (0.06)
r/Dirtybomb (0.92)	r/ImagesOfTexas (0.07)
r/totalwar (0.92)	r/srpska (0.07)
r/Battlefield (0.92)	r/BDSMcommunity (0.09)
r/ericprydz (0.92)	r/the_donald_discuss (0.1)

of an active community that promote engagement between its members. For example, while participating in `r/place`, joining efforts with the *r/texas* community, the *r/ImagesOfTexas* subreddit presents dynamics closer to those of *r/catPics*. In that sense, *r/ImagesOfTexas* can be seen as a sub-subreddit of *r/texas*, a fact that our models fail to account for.

On the other hand, the false positive list contains some communities that are, indeed, expected (albeit naively) to participate in `r/place`. Gaming communities that were correctly predicted to participate (e.g., *r/PuzzleAndDragons*, *r/ScottPilgrim*, and *r/ClashRoyale*) are very active, presenting high degree of inter-user engagement. It is therefore understandable why other communities that revolve around video games (e.g., *r/mega64*, *r/Dirtybomb*), bearing strong similarity to gaming communities, would be falsely predicted to participate in `r/place`.

5.3 Social Interpretation

We use the SHAP framework [50] to quantify the impact of specific features on the prediction and derive social interpretation and insights. A SHAP value is calculated for each explanatory variable (feature) and input instance (i.e., a community in our case) for a specific model. High (low) SHAP value indicates the positive (negative) impact of the feature on the prediction of the specific instance. Looking at the aggregate values for a specific feature provides a way to interpret a given model.

The SHAP values of the most prominent features (of each feature type) are presented in Figure 3. The SHAP value (X-axis) indicates the contribution of the feature to the prediction – high value means positive contribution to the model's prediction. The Y-axis simply indicates the density of a specific SHAP value. The color corresponds to the actual value of the feature for each instance – high values in red and low values in blue.

5.3.1 Community Meta Features.

Feedback. Feedback, particularly a positive feedback, is known to have a positive influence on the cohesiveness of communities and to increases engagement and retention [12, 25]. Reddit provides its users with a feedback mechanism – users can up/down-vote submissions made by other users. We find the 'comment average score' to be the most powerful feature among the meta features (see Figure 3a). The *comment average score* per community for $s \in S^+$ is 4.39, and only 3.05 for $s \in S^-$ – suggesting that positive feedback (more up-votes than down-votes) supports a positive atmosphere, increases solidarity, and contributes to the community preparedness to engage in a campaign.

Age. The SHAP values of the *age* meta-feature are clearly distributed between two distinct clusters (Figure 3a). These clusters appear to be highly correlated with the actual value of this feature, as indicated by color: high SHAP values are red, and low SHAP values are blue. This supports our intuition that the longer history an active community shares, the more it will be inclined to join a distributed campaign, while a shorter shared history (blue) has a negative (low SHAP values) impact on the community's inclination to join the cause.

5.3.2 Network Features.

Closeness Centrality. The median nodes' *closeness* is significantly higher in participating communities (note that higher values of closeness indicate higher centrality, Figure 3b). This is an indication that participating communities are denser, maintaining tighter relations between members.

Triads. Communities participating in `r/place` have a significantly higher number of triads. On average, participating communities have 148.46K triads compared with only 26.8K triads in non-participating ones. While this could be a attributed to the fact that S^+ communities tend to have a larger number active users (Table 1), it is important to note that of 23% of S^- communities had *no* triads at all, compared with only 4.95% in S^+. Indeed, community cohesion is known to impact the inclination of its members to undertake a collective task [20, 24].

Connected Components. The number of connected components is negatively correlated with the SHAP values, suggesting that fractured communities are less likely to engage in a collective action. On average, participating communities have 46.79 connected components, while non-participating ones have 69.32. While this result seems trivial (group cohesiveness is a major factor in acting toward a common goal), the relative cohesiveness of the participating communities deviates from the expected number of components given community size. Considering only the size of the community (nodes amount), the number of connected components is expected to positively correlate with the number of nodes in the network. However, in the `r/place` case, we observe the opposite – participating communities (S^+) have a significantly *higher* number of active users[9], compared to S^- communities (average and median, see Table 1 second row) but still present a much *higher* connectivity.

5.3.3 Textual Features.

Practicing Good Citizenship. Higher *tf-idf* values (purple and red) for words like *vote* and *discussion* correlate with higher SHAP values (see Figure 3c). This suggests that communities that opted

[9]Note that the number of subscribers (first row in Table 1) is less relevant when analysing connected components as we only consider users that took an active role in their community in the months preceding the experiment.

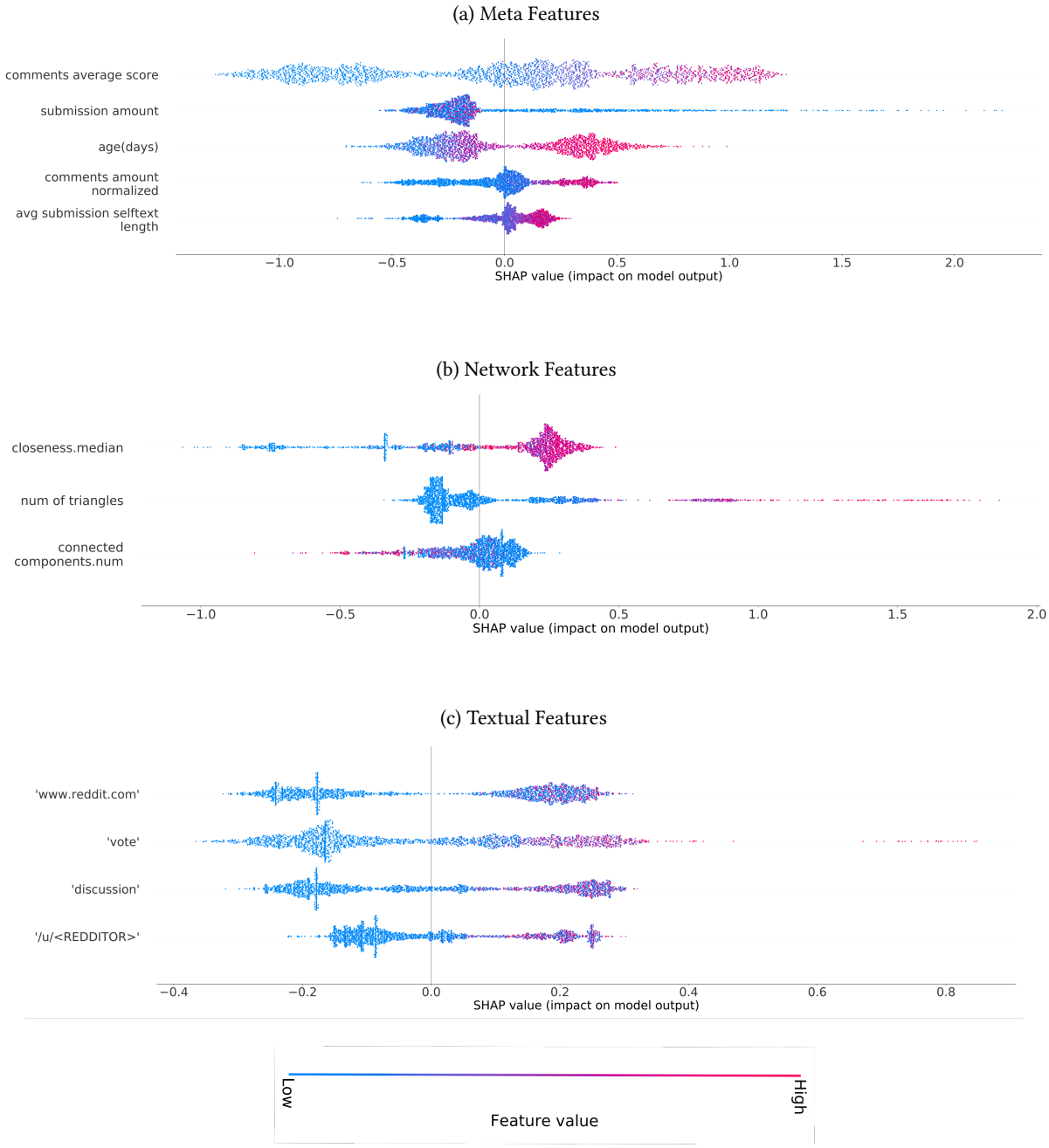


Figure 3: SHAP values of twelve prominent features. Each point represents the SHAP value of an instance for a specific feature (in the GBT model). All features in the figure are significantly important (p-value $< 10^{-4}$) in the prediction model.

to participate in `r/place` are those who tend to promote discussions among their members (rather than serve merely as message boards) and encourage their members to engage and practice "good citizenship", evident by a high frequency of tokens like *vote* and *discussion*.

Explicit Mentions. Members of participating communities tend to explicitly mention other users (using the '/u/' prefix), significantly more than members of non-participating communities. This is in line with well established social theory i.e., the strong correlation between the *sense of community* to social factors such as knowing your neighbor names [11].

Internal URLs. Interestingly, referencing Reddit, indicated by the *www.reddit.com*, also correlates with high SHAP values. We

hypothesize that frequent references to Reddit indicate strong engagement with the platform and the community.

5.4 Beyond r/place: WallStreetBets (WSB)

The GameStop short squeeze of early 2021, organized and promoted in the *WallStreetBets* (WSB) subreddit,[10] was argued, at the time, to have shifted the financial power balance. The unfolding of the events initiated a debate among economists and sociologists, trying to understand its causes and its impact on future trade [19, 48, 49]. This successful short squeeze campaign involved a tight, though decentralized, coordination toward the realization of a common goal. Similarly to the r/place sudden appearance, the GameStop short squeeze provides another unique, though anecdotal, opportunity to examine the way community characteristics (norms, textual, structural, etc.) correspond to a collective undertaking.

The *WSB* community took part in r/place. Its participation in the experiment is somewhat surprising. Most of the communities that participated, were naturally gathered around a national flag, a video game logo, a sport team symbol, etc. These are natural causes for participation – the community members share solidarity at a basic level and have a predefined logo (insignia) to place on the canvas. This solidarity is not expected in a community like *WSB* that is used for sharing stock-trading advice. [11] In spite of our early disposition regarding the nature of *WSB* and similar communities, we observe the community ability to collaborate as reflected in r/place, leaving a clear mark of a sit size on the canvas[12]. Indeed, keeping *WSB* out of the training set, the model predicts the participation of the community with a likelihood of 0.82.

We use a SHAP "waterfall" plot to further analysis of the community. The top features contributing to the model's prediction are presented in Figure 4. The 'comments average score' is the most dominant feature for this community. All three feature types (linguistic, meta, and network) appear among the most dominant features. We observe that most of the features in Figure 4 are also found to be central in the *aggregated* SHAP analysis (Figure 3). However, both 'submissions average score' and 'urls ratio' are ranked higher in *WSB* compared with the aggregated SHAP analysis . This observation supports our conclusions regarding the importance of a positive feedback by community members, and of link sharing as positive influences on the engagement and cohesiveness of online communities (see Section 5.3).

5.5 Performance of the Neural Models

Neural models are considered to be the state-of-the-art in many classification and prediction tasks. All the neural architectures we experimented with achieved decent results, significantly outperforming a naive baseline (Table 2). However, the neural models were outperformed by the GBT. We briefly consider the factors that may have affected the performance of the models, though a proper experiment is out of the scope of this paper:

- *Dataset size*: Although the size of the dataset ($S^+ \cup S^-$) is considerable (tens of millions of posts submitted by over ten million users), the total number of *instances* is relatively small (< 2500). Deep learning models perform well on large datasets, with high number of instances that allows a productive back-propagation of error terms to a large number of neurons.
- *Data complexity*: While the number of instances is relatively small, each instance is a complex unit, composed of thousands of words, threads and users, making feature definition and extraction less trivial.
- *Neural architectures*: The neural architectures we designed are aimed to process very different feature types – word tokens (at the post, thread and community level), meta features, and social features. Optimizing a neural architecture to handle multiple feature types is not a trivial task [74, 77].

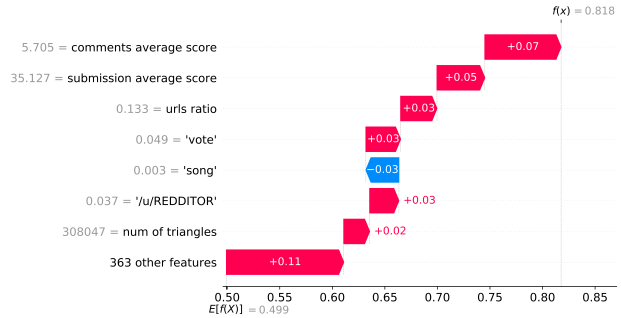


Figure 4: SHAP analysis - *WallStreetBets* subreddit. The top features that contribute to push the model output from the base likelihood value (0.499) to the its final value (0.818). Features pushing the prediction higher (lower) are shown in the right-red (left-blue) arrows. Values on the arrows are the marginal contribution of each feature. Values on the y-axis are the actual values of each feature.

We wish to note that while the neural architectures could be optimized to achieve higher accuracy, our primary motivation in this work is to model and understand the social factors that drive the community behaviour toward a campaign, rather than optimizing the network.

6 CONCLUSIONS AND FUTURE WORK

In this work we study how community structure, norms and language can be used to predict the community's engagement in a large scale distributed campaign. Specifically, we predict which Reddit community is likely to take place in the r/place experiment. We use Gradient Boosted Trees and complex neural models, experimenting with various representations of community (language, network, meta). We demonstrated that all type of features contribute to the classification, that feature types enhance each other and that meta features are as important as linguistic features.

Future work takes two trajectories: (i) Analysing the rich community dynamics observed *during* the experiment and model the communities' success level, and (ii) Improving the ways complex meta and network features are defined, extracted and introduced into the learning models.

[10] A short squeeze campaign that incurred loses of billions to a number of hedge-funds in just a few days during January and February of 2021.
[11] Indeed, other trading communities like *r/investing* and *r/stocks* did not participate in r/place.
[12] See the right side of the final canvas, Figure 1d – above the GNU/Linux penguin.

REFERENCES

[1] Ben Armstrong. 2018. *Coordination in a Peer Production Platform: A study of Reddit's/r/Place experiment.* Master's thesis. University of Waterloo.

[2] Tal August, Dallas Card, Gary Hsieh, Noah A Smith, and Katharina Reinecke. 2020. Explain like I am a Scientist: The Linguistic Barriers of Entry to r/science. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[3] Christopher Andrew Bail. 2016. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences* 113, 42 (2016), 11823–11828.

[4] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[5] Melisa Basol, Jon Roozenbeek, and Sander van der Linden. 2020. Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *Journal of Cognition* 3, 1 (2020).

[6] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[7] Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

[8] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.

[9] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* 114, 28 (2017), 7313–7318.

[10] Brian C Britt, Rebecca K Britt, Jameson L Hayes, Elliot T Panek, Jessica Maddox, and Aibek Musaev. 2021. Oral healthcare implications of dedicated online communities: A computational content analysis of the r/Dentistry subreddit. *Health communication* 36, 5 (2021), 572–584.

[11] David M Chavis and Abraham Wandersman. 2002. Sense of community in the urban environment: A catalyst for participation and community development. In *A Quarter Century of Community Psychology*. Springer, 265–292.

[12] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[13] Daejin Choi, Jinyoung Han, Taejoong Chung, Yong-Yeol Ahn, Byung-Gon Chun, and Ted Taekyoung Kwon. 2015. Characterizing conversation patterns in Reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*. ACM, 233–243.

[14] Tiago Oliveira Cunha, Ingmar Weber, Hamed Haddadi, and Gisele L Pappa. 2016. The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*. ACM, 99–103.

[15] Srayan Datta and Eytan Adar. 2019. Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, Vol. 13. 146–157.

[16] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity.. In *ICWSM*.

[17] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports* 6 (2016), 37825.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[19] Tim Di Muzio. 2021. GameStop Capitalism. Wall Street vs. The Reddit Rally (Part I). (2021).

[20] Justin Fagnan, Osmar Zaïane, and Denilson Barbosa. 2014. Using triads to identify local community structure in social networks. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 108–112.

[21] Casey Fiesler, Jialun" Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance.. In *ICWSM*. 72–81.

[22] Dana R Fisher, Kenneth T Andrews, Neal Caren, Erica Chenoweth, Michael T Heaney, Tommy Leung, L Nathan Perkins, and Jeremy Pressman. 2019. The science of contemporary street protest: New efforts in the United States. *Science advances* 5, 10 (2019), eaaw5461.

[23] Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 4 (2002), 367–378.

[24] Adrien Friggeri, Guillaume Chelius, and Eric Fleury. 2011. Triangles to capture social cohesion. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 258–265.

[25] Maria Glenski and Tim Weninger. 2017. Predicting user-interactions on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 609–612.

[26] Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies* 10, 1 (2017), 1–309.

[27] Mark S Granovetter. 1973. The strength of weak ties. In *Social networks*. Elsevier, 347–367.

[28] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.

[29] William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Eleventh International AAAI Conference on Web and Social Media*.

[30] Simo Hanouna, Omer Neu, Sharon Pardo, Oren Tsur, and Hila Zahavi. 2019. Sharp power in social media: Patterns from datasets across electoral campaigns. *Australian and New Zealand Journal of European Studies* 11, 3 (2019).

[31] T Hässler, Johannes Ullrich, Michelle Bernardino, Nurit Shnabel, D Valdenegro, C Van Laar, S Sebben, E Visintin, L Tropp, R González, et al. 2020. A large-scale test of the link between intergroup contact and support for social change. *Nature Human Behaviour* (2020).

[32] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[33] Margaret Hu. 2020. Cambridge Analytica's black box. *Big Data & Society* 7, 2 (2020), 2053951720938091.

[34] Sarah J Jackson and Brooke Foucault Welles. 2016. # Ferguson is everywhere: initiators in emerging counterpublic networks. *Information, Communication & Society* 19, 3 (2016), 397–418.

[35] Kathleen Hall Jamieson. 2018. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. Oxford University Press.

[36] Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. r/science: Challenges and opportunities in online science communication. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[37] John T Jost, Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A Tucker. 2018. How social media facilitates political protest: Information, motivation, and social networks. *Political psychology* 39 (2018), 85–118.

[38] Brian Keegan, Darren Gergle, and Noshir Contractor. 2011. Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tohoku catastrophes. In *Proceedings of the 7th international symposium on Wikis and open collaboration*. ACM, 105–113.

[39] Brian Keegan, Darren Gergle, and Noshir Contractor. 2012. Do editors or articles drive collaboration?: multilevel statistical network analysis of wikipedia coauthorship. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, 427–436.

[40] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[41] Zornitsa Kozareva and Eduard Hovy. 2010. Not all seeds are equal: Measuring the quality of text mining seeds. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 618–626.

[42] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[44] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 933–943.

[45] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)* 323, 5915 (2009), 721.

[46] Kurt Lewin. 1947. Frontiers in Group Dynamics: Concept, Method and Reality in Social Science; Social Equilibria and Social Change. *Human Relations* 1, 1 (1947), 5–41. https://doi.org/10.1177/001872674700100103 arXiv:https://doi.org/10.1177/001872674700100103

[47] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[48] Cheng Long, Brian M Lucey, and Larisa Yarovaya. 2021. "I Just Like the Stock" versus "Fear and Loathing on Main Street": The Role of Reddit Sentiment in the GameStop Short Squeeze. (2021).

[49] Lorenzo Lucchini, Luca Maria Aiello, Laura Alessandretti, Gianmarco De Francisci Morales, Michele Starnini, and Andrea Baronchelli. 2021. From Reddit to Wall Street: The role of committed minorities in financial collective action. *arXiv preprint arXiv:2107.07361* (2021).

[50] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[51] Joan Massachs, Corrado Monti, Gianmarco De Francisci Morales, and Francesco Bonchi. 2020. Roots of trumpism: Homophily and social feedback in donald trump support on reddit. In *12th ACM Conference on Web Science*. 49–58.

[52] Alexey N Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. 2017. The anatomy of Reddit: An overview of academic research. In *Dynamics on and of Complex Networks*. Springer, 183–204.

[53] Alberto Melucci. 1996. *Challenging codes: Collective action in the information age*. Cambridge University Press.

[54] Humphrey Mensah, Lu Xiao, and Sucheta Soundarajan. 2020. Characterizing the Evolution of Communities on Reddit. In *International Conference on Social Media and Society*. 58–64.

[55] William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A Smith. 2021. Provable Limitations of Acquiring Meaning from Ungrounded Form: What will Future Language Models Understand? *arXiv preprint arXiv:2104.10809* (2021).

[56] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[57] Robert S Mueller. 2019. Report on the investigation into Russian interference in the 2016 presidential election. *US Dept. of Justice. Washington, DC* (2019).

[58] Thomas F Müller and James Winters. 2018. Compression in cultural evolution: Homogeneity and structure in the emergence and evolution of a large-scale online collaborative art project. *PloS one* 13, 9 (2018), e0202019.

[59] Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, et al. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980* (2017).

[60] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest.. In *ICWSM*. 279–288.

[61] Mancur Olson. 2009. *The Logic of Collective Action: Public Goods and the Theory of Groups, Second printing with new preface and appendix*. Vol. 124. Harvard University Press.

[62] Elinor Ostrom. 2000. Collective action and the evolution of social norms. *Journal of economic perspectives* 14, 3 (2000), 137–158.

[63] Elliot Panek, Connor Hollenbach, Jinjie Yang, and Tyler Rhodes. 2018. The Effects of Group Size and Time on the Formation of Online Communities: Evidence From Reddit. *Social Media+ Society* 4, 4 (2018), 2056305118815908.

[64] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019), 8026–8037.

[65] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[66] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[67] Sarah Perez. 2019. Reddit's monthly active user base grew 30% to reach 430M in 2019. *TechCrunch (Date accessed: 2/2/2020)* (2019). https://techcrunch.com/2019/12/04/reddits-monthly-active-user-base-grew-30-to-reach-430m-in-2019

[68] Sam Ransbotham and Gerald C Kane. 2011. Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *Mis Quarterly* (2011), 613–627.

[69] Jérémie Rappaz, Michele Catasta, Robert West, and Karl Aberer. 2018. Latent structure in collaboration: the case of Reddit R/place. In *Twelfth International AAAI Conference on Web and Social Media*.

[70] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. *Learning internal representations by error propagation*. Technical Report. California Univ San Diego La Jolla Inst for Cognitive Science.

[71] Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval. (1986).

[72] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[73] Robert E Schapire. 1990. The strength of weak learnability. *Machine learning* 5, 2 (1990), 197–227.

[74] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. SemEval-2020 Task 8: Memotion Analysis–The Visuo-Lingual Metaphor! *arXiv preprint arXiv:2008.03781* (2020).

[75] Greg Stoddard. 2015. Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News.. In *ICWSM*. 416–425.

[76] Prateek Vachher, Zachary Levonian, Hao-Fei Cheng, and Svetlana Yarosh. 2020. Understanding Community-Level Conflicts Through Reddit r/place. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 401–405.

[77] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. 2020. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems* 33 (2020).

[78] Ken Ward. 2018. Social networks, the 2016 US presidential election, and Kantian ethics: applying the categorical imperative to Cambridge Analytica's behavioral microtargeting. *Journal of media ethics* 33, 3 (2018), 133–148.

[79] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An exploration of discussion threads in social news sites: A case study of the reddit community. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*. IEEE, 579–583.

[80] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161* (2019).

[81] Wayne W Zachary. 1977. An information flow model for conflict and fission in small groups. *Journal of anthropological research* 33, 4 (1977), 452–473.

[82] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for Longer Sequences.. In *NeurIPS*.

[83] Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Eleventh International AAAI Conference on Web and Social Media*.

[84] Naitian Zhou and David Jurgens. 2020. Condolences and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 609–626.

A ALGORITHMIC APPROACHES

In this work we use comments data only through meta-features creation and *not* as part of the textual input data for each model. We assume that submissions texts well represent communities' properties and characteristics. We limit to the number of submission texts per community that are used by the model to 10K. This limitation is due to the high complexity of the models when handling thousands of potential long text submissions.[13]

As mentioned in Section 3.1, we use embedding vectors for training the neural models. We considered two embeddings alternatives: an off-the-shelf pretrained model [56] and a dedicated GloVe model [66] trained on *DS*1. All reported results were obtained with the latter model which achieved better performance.

In the rest of this section, we elaborate about the classification models we experiment with.

Gradient Boosted Trees (GBT). We experiment with a various non-sequential classifiers (classifiers that ignore the sequential structure of the data) – logistic regression, decision tree, SVM, naive Bayes, ada-boost, random-forest, and gradient boosted trees (BGT). However, we report only on the performance of the GBT model, the best performing non-sequential classifier.

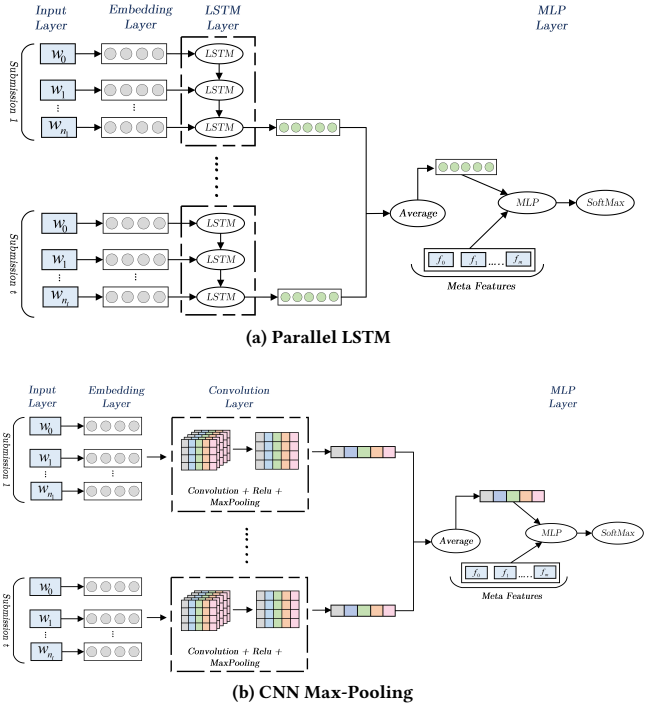
GBT is a class of information-theoretical discriminative classifiers. A series of weak learners (decision trees) is constructed, boosting classification accuracy by combining the respective learners, trained on modified and over/under-sampled data [23, 73]. GBT classifiers tend to work well on relatively small datasets and with a combination of different and unnormalized feature types. Hence, the GBT classifier seem to be well suited to the task at hand. We limited the number of trees to 100 and the depth of each tree to 3.

Multulayer Perceptron (MLP). MLPs are a class of simple feed-forward artificial neural network classifiers, with at least one hidden layer of neurons, successfully learn nonlinear functions [70].

BERT Zero-Shot. Bidirectional Encoder Representations from Transformers (BERT) [18] is a transformer-based technique that achieves state-of-the-art results in many NLP domains. Most of the BERT models are applicable for short snippets pf text [47, 72]. Lately, new BERT variants were suggested to handle longer texts [6, 82]. However, none of them is able to handle the high number of submissions and comments that is generated by each community. We use BERT representation of sentences in a zero-shot manner [80]. We use a pre-trained BERT model to get a vector representation of each submission in the corpus. We use the last layer of the BERT model, of size 768, for this purpose. This series of *submission embeddings* is averaged to represent the whole subreddit. The averaged *submission embeddings* is fed to the an *MLP component* along with a vector of the network and the meta features. Its dimension depends on the experimental setting (see Section 4). Finally, a binary *softmax* function is applied in order to predict the class.

Parallel LSTM with an MLP Layer. Long-Short-Term Memory (LSTM) [32] are special type of recurrent neural network (RNN). The network contains internal loops, allowing an adaptive memory effect. LSTM networks are proven to perform well on many language-related classification tasks as the sequential nature of language and its dependencies are being captured accurately by a series of LSTM cells. Beyond the sequential nature of language, the sequential nature of discussion threads promotes the use of LSTMs.

We design the network to take different types of features and combine them in different layers. An illustration of the neural architecture is presented in Figure 5a. Each input iteration operates on a subreddit. A subreddit is composed of multiple submissions.[14] The input layer is dynamic, depending on the number of submissions in the current subreddit. Each tokenized submission (i.e., $w_0, w_1 \ldots$) is translated to a sequence of *embedding vectors* that are fed into a respective sequence of *LSTM* cells. The LSTM layers yield a series of *submission embeddings* (note – each submission is a thread in the subreddit) which is averaged to represent the whole subreddit (dark green circles in Figure 5a). The *submission embedding* is fed to the an *MLP component* along with a vector of the network and the meta features. Its dimension depending on the experimental setting (see Section 4). Finally, a binary *softmax* function is applied in order to predict the class.



(a) Parallel LSTM

(b) CNN Max-Pooling

Figure 5: Parallel LSTM and CNN Max-Pooling architectures. Each input submission is splitted into tokens, and represented as an embedding vector. All Submissions are aggregated into one vector at the end of the LSTM/CNN layer, and then concatenated with the meta features input vector. Dimension of all inputs (number of submissions, length of each submission, and number of meta features) are dynamic. Last layer includes a standard MLP and a softmax part. Blue shaded variables indicate model's input.

[13] Each submission may consist of a very high number of sentences and tokens (unlimited number in Reddit).

[14] Each submission is the root of a discussion thread, see Section 1

Table 4: Meta and social network features. We calculate each feature per community (subreddit). For the last four features in the table (marked as <feature-name>*), we use average, maximum, minimum, median, and standard deviation measures.

	Feature Name	Type	Explanation
Meta Features	submission amount	Int	Number of submissions
	submission amount normalized	Float	Number submissions % number of users
	submission average score	Float	Submission up-vote average score
	submission median score	Float	Submission up-vote median score
	comments average score	Float	Comments up-vote average score
	comments median score	Float	Comments up-vote median score
	comments submission ratio	Float	Number of comments % number of submissions
	deleted removed submission ratio	Float	Number of deleted or removed submission % number of submissions
	distinct comments to submission ratio	Float	Percentage of submissions which were commented (at least once)
	distinct comments to comments ratio	Float	Number of submissions commented % number of comments
	users amount	Int	Number of users registered (not necessarily wrote anything)
	submission distinct users	Int	Distinct number of users who wrote a submission (i.e., started a thread)
	average submission per user	Float	Average number of submissions per user (out of those who wrote something)
	median submission per user	Float	Median number of submissions per user (out of those who wrote something)
	submission to comments users ratio	Float	Distinct number of commented % distinct number of users who wrote a submission
	submission users std	Float	Users standard deviation, out of the users who wrote a submission
	comments users std	Float	Users standard deviation, out of the users who wrote a comment
	users deleted normalized	Float	Number of submission or comments deleted user % number of total users
	submission title length	Float	Average length of a submission title
	median submission title length	Float	Median length of a submission title
	submission selftext length	Float	Average length of a submission selftext (content of the submission)
	median submission selftext length	Float	Median length of a submission selftext (content of the submission)
	empty selftext ratio	Float	Percentage of submissions with an empty selftext
	submissions2comments words used	Float	Number of comments distinct words % number of submission distinct words
	age(days)	Int	Number of days which the community exists (counting back from 31/3/2017)
Social Network Features	num of nodes	Int	The overall number of active users)
	num of triangles	Int	Number of triads
	num of edges	Int	The total number of edges in the graph
	is biconnected	Binary	Is the graph biconnected
	num of nodes to cut	Int	The min-cut value of the graph's biggest component
	density	Float	Number of nodes % number of edges
	num connected components	Int	The number of connected components in the graph
	num connected components > 2	Int	The number of connected components that contain more than two nodes
	max group in a connected components	Int	The largest connected component size
	num s.connected components	Int	The number of strongly connected components in the graph
	num strongly connected components > 2	Int	The number of strongly connected components that contain more than two nodes
	max group in a s.connected components	Int	The largest strongly connected component group size
	betweenness*	Float	Each node's betweenness
	centrality*	Float	Each node's centrality
	closeness*	Float	Each node's closeness value
	in degree*	Int/Float	Each node's in-degree value

Note that meta-features are related to a subreddit as a whole, hence the vector of the meta-features is added just before the final MLP component, as illustrated in Figure 5a.

Max pooling CNN with an MLP Layer. Convolutional neural network (CNN) [43] are a special type of multilayer perceptron, which apply different convolutions to the input data. Recently, CNNs proved useful in NLP related problems [26, 40]. The architecture we designed is presented in Figure 5b and contrasted with the LSTM architecture. The main difference is our usage of a convolutional layer instead of a sequence of LSTM cells. Such convolution step, allows us to capture relations between words sequences (in the embedding form). Empirically, a long width and a short length mask yielded the best results. We use a mask of (100, 2) and a 300 dimensions embedding vector.

B META AND NETWORK FEATURES

As mentioned in Section 3.1, we use meta features (e.g., comments amount) as well as social network features per community. These network and meta features are commonly used in social network analysis (SNA) research. We preferred *not to* represent the network and meta features as embedding vectors, in order to gain useful insights from the model's interpretation.

In Table 4 we provide the full list of 25 meta features followed by the 32 social network features, with a short description of each.