Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1885–1889
Marseille, 20-25 June 2022

The Engage Corpus: A Social Media Dataset for Text-Based Recommender Systems

Daniel Cheng,[1], Kyle Yan,[1] Phillip Keung,[1] Noah A. Smith[1,2]
[1]University of Washington, [2]Allen Institute for Artificial Intelligence
d0@uw.edu, kyleyan@uw.edu, pkeung@uw.edu, nasmith@cs.washington.edu

Abstract

Social media platforms play an increasingly important role as forums for public discourse. Many platforms use recommendation algorithms that funnel users to online groups with the goal of maximizing user engagement, which many commentators have pointed to as a source of polarization and misinformation. Understanding the role of NLP in recommender systems is an interesting research area, given the role that social media has played in world events. However, there are few standardized resources which researchers can use to build models that predict engagement with online groups on social media; each research group constructs datasets from scratch without releasing their version for reuse. In this work, we present a dataset drawn from posts and comments on the online message board Reddit. We develop baseline models for recommending subreddits to users, given the user's post and comment history. We also study the behavior of our recommender models on subreddits that were banned in June 2020 as part of Reddit's efforts to stop the dissemination of hate speech.

Keywords: Reddit, social media, recommender systems

1. Introduction

User engagement with online communities has come under significant scrutiny in light of the increasing polarization in global politics. Many researchers have examined social media's role in recent headlines, like the spread of coronavirus-related vaccine misinformation and the dissemination of conspiracy theories in political discourse (Puri et al., 2020; Douglas et al., 2019, *inter alia*). Since social media platforms typically use recommender systems to help users discover new content and online groups, a natural question that arises is what role such algorithms have played in changing user behavior. Previous studies and investigations (Horwitz, 2021; Bakshy et al., 2015) have discussed the impact of recommendation algorithms on how users discover and join online groups, and researchers are also interested in understanding how NLP is used in generating recommendations. However, we are not aware of any standardized datasets specifically designed for constructing text-based recommender systems in the social media setting; research groups tend to scrape their own datasets from various web resources without releasing their corpora, which hinders the reproducibility of reported results.

We introduce the Engage corpus, a large-scale, realistic dataset derived from Reddit to address this resource gap. Reddit is a rich source of social media data on a variety of subjects, and there are public APIs that allow researchers to download years of posts and comments for each user. Reddit is divided into 2.5 million communities called 'subreddits', and users interact within subreddits by writing 'posts' or submitting 'comments' on existing posts. We say that a user *engages* with a subreddit if they post or comment in that subreddit.

In this work, we focus on the subreddit recommendation task: predict whether a user will engage with a subreddit within the next three months, given the user's post and comment history thus far.

In the sections that follow, we discuss the corpus construction process and present the key statistics for various dimensions of the Engage corpus. We create baselines for this task with neural collaborative filtering models (He et al., 2017) and we demonstrate that incorporating information from a user's post and comment history significantly improves recommendation performance (versus using user-subreddit interaction data alone). We also study the behavior of text-based recommender systems on certain controversial subreddits (such as r/The_Donald and r/Incels), which were banned by Reddit's administrators in June 2020. The instructions for getting the processed dataset and our code can be found at `https://github.com/engage-corpus/dataset`.

2. Corpus Processing and Statistics

The primary portion of our dataset comes from all Reddit posts and comments from 01–12/2019. The Reddit corpus is very large (250 GB when compressed), and we downsample the dataset to retain a random 6% sample of users (131,544 users after downsampling). We split the dataset into two distinct periods: 01–09/2019 and 10–12/2019. We extract training data from the first period and we extract evaluation data from the second period. Furthermore, we only include posts and comments from the 5,000 largest subreddits (as measured by active users at the time of dataset construction).

The training data consists of all the posts and comments for 131,544 users in 01–09/2019. The development set contains the first post or comment from each user in 10–12/2019, and the test set contains the second post or comment from each user in 10–12/2019. Therefore, a

# Users	# Subreddits	# Interactions	Sparsity
131,544	5000	2,206,986	99.7%

Table 1: User-subreddit interactions in the corpus.

	50th	90th	99th
Posts & comments per user	46	418	2,278
Tokens per post or comment	13	64	221
Users per subreddit	209	1,070	6,382
Posts & comments per subreddit	1,472	8,816	55,787

Table 2: Percentiles for various dataset characteristics.

recommender model that's tuned on the development set maximizes the prediction performance on the users' first subreddit interaction in the evaluation period, and the performance on the test set shows how well the model generalizes to the next (i.e., second) interaction in the evaluation period.

2.1. Data Fields

In our dataset, each user is represented with a JSON object with an anonymized username field (a string), a posts field (a list of post objects, sorted by date), and a comments field (a list of comment objects, sorted by date).

A post object has the following fields: a randomly generated post ID, title, date, the subreddit that the post was submitted to, total number of upvotes, and text in the body of the post. A comment object has all of the fields of a post object (except for the title) plus the ID of the post the comment belongs to.

2.2. Corpus Statistics

Table 1 provides some basic statistics for the dataset. Most users only engage with a handful of subreddits in any given year. As is typical of recommendation datasets, the user-subreddit interactions are sparse, and less than 0.3% of all possible interactions actually occur: 655M = 131K × 5K user-subreddit interactions are possible, but only 2.2M interactions actually occurred. Table 2 shows text-related statistics for the dataset. We observe the usual 'super-user' behavior in our dataset, where a small percentage of users engage especially frequently within each subreddit. We also see long-tailed distributions in the number of posts/comments in each subreddit, where the 99th percentile is ~50× larger than the median.

3. Experiments

We use two metrics to evaluate the quality of our recommendations: the hit rate (HR) at 10 and normalized discounted cumulative gain (NDCG) at 10 (Järvelin and Kekäläinen, 2002). HR@10 (which is also known as precision@10) is the rate at which a system's top 10 recommendations for each user contain the correct subreddit. NDCG@10 is a rank-based metric whose value increases as the rank of the correct subreddit increases among the top 10 recommendations. In our setting, NDCG@10 is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\log_2(r_i+1)}$$

where n is the number of users and r_i is the 1-indexed rank of the correct subreddit among the top 10 ranked interactions for the ith user (e.g., $r_i = 3$ if the correct subreddit is the third recommendation in the top 10), where $r_i = \infty$ when the correct subreddit is not in the top 10 recommendations. Both of these metrics are commonly used in the recommender systems literature.

3.1. Baseline Model

We adopt the neural collaborative filtering (NCF) model (He et al., 2017) as our baseline system for generating user-subreddit recommendations. NCF models are standard baseline systems for recommender systems; we will not describe the model architecture in detail here (Figure 1). In brief, NCF models rely on *collaborative filtering* (Goldberg et al., 1992), which assumes that the subreddit preferences of a user can be inferred from those of other users that engaged with similar subreddits. The model learns user-specific and subreddit-specific embeddings and outputs a score between 0 and 1 for each user-subreddit pair. The model is trained on positive and negative user-subreddit pairs, where positive pairs are drawn from observed user-subreddit interactions and negative pairs are constructed by sampling at random from the universe of 5,000 subreddits. To perform inference, we rank the list of 5,000 subreddits by their scores for each user. We implement NCF in PyTorch using the default hyperparameter settings described by He et al. (2017) with a factor size of 64.

3.2. NCF with User Text

The NCF model uses the presence or absence of user-subreddit interactions to predict future subreddit interactions for each user; the sequential order of the subreddits that each user engaged with and the user-generated text from posts and comments are ignored. Of course, incorporating user-generated text is computationally expensive, since a user in the 90th percentile produced 10,927 tokens across all posts and comments in 2019. Encoding very long document contexts with transformers remains an active area of research; see Tay et al. (2020) for a survey. We take a two-stage approach, where we first train a linear classifier[1] that uses trigrams from a user's post and comment history to predict the probability that he will interact with each of the 5,000 subreddits. These 5,000 scores from the linear classifier are then provided to the NCF model as a user-specific, text-based feature vector (Figure 1). The trigram features are extracted from a user's posts and comments in

[1] https://github.com/VowpalWabbit/vowpal_wabbit/wiki

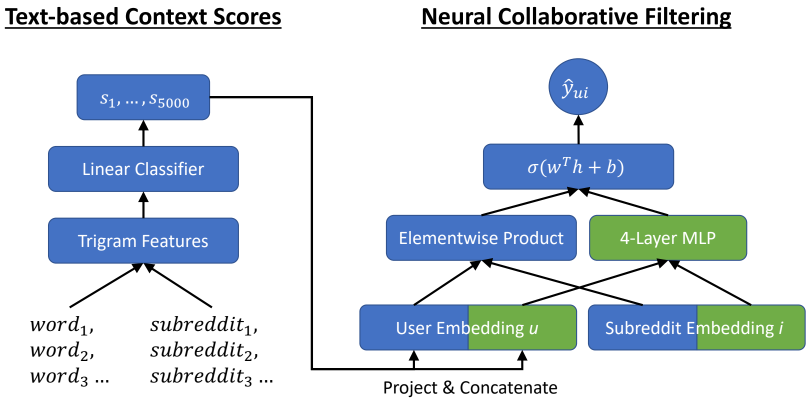

Figure 1: Our baseline neural collaborative filtering model (right) and text-based linear classifier (left). The neural collaborative filtering (NCF) model learns user- and subreddit-specific embeddings, which are used to calculate the score for each user-subreddit pair. The NCF model is trained on positive and negative pairs of user-subreddit interactions. We also use a linear classifier to improve the baseline NCF model by modeling 6 months of post and comment text and sequential subreddit history for each user (see Sec. 3.2).

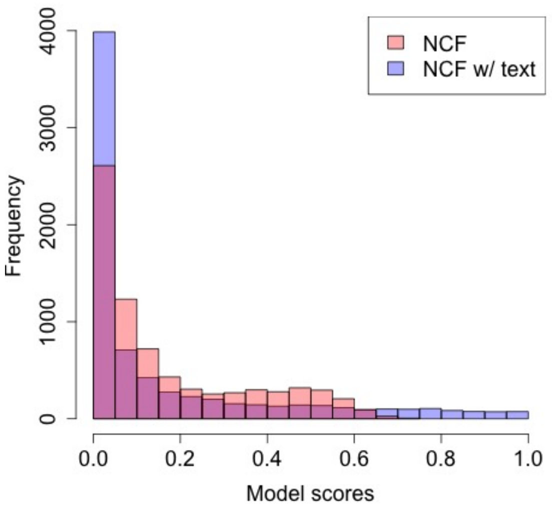

Figure 2: Histogram of model scores for the 8 largest banned subreddits. The NCF model that uses post and comment text tends to generate more extreme scores (i.e., close to 0 and 1) compared to the baseline model.

Metric@10	HR	NDCG
NCF baseline	53.5	33.5
+ text-based context scores	60.3	38.5

Table 3: Hit rate and NDCG at 10 for our baseline NCF model and NCF with text-based context scores. Higher scores are better.

a 6-month window, and the classifier is trained to predict the subreddits they engaged with in the subsequent 3-month window. For example, we take the trigrams from the user's text and sequence of subreddits between 01/2019–06/2019, and we predict which of the 5,000 subreddits they engaged with between 07–09/2019. We treat this as a multi-label classification problem over the 5,000 subreddits included in the corpus.

After training the linear classifier, we use it as a feature extractor for each user's post and comment text. We then project the 5,000 scores to a 128-dimensional vector using a learned linear projection, which we concatenate with the user embeddings in the NCF model. For a given user-subreddit interaction during training of the NCF model, we use the 6 months of the user's post and comment history prior to that interaction as text input for the linear classifier. The development set is used to select hyperparameters and model checkpoints. We report results on the test set in our tables.

3.3. Baseline Results

In Table 3, we provide the HR and NDCG at 10 for the NCF model and the NCF model with the text-based context scores from the trigram classifier. The improvements to both HR and NDCG at 10 are substantial after adding the text-based scores, which demonstrates the utility of modeling user context with NLP in the recommendation setting.

3.4. Analysis on Controversial Subreddits

In June 2020, the administrators of Reddit banned hundreds of subreddits in an effort to combat hate speech on the platform (Reddit, 2020). Political subreddits like r/The_Donald (a community for supporters of Donald Trump) and r/ChapoTrapHouse (a community for listeners of a left-wing political podcast) were among the subreddits that were banned as part of this initiative. Recommender systems may play a significant role in driving user traffic to controversial online groups. We investigate the effect of incorporating textual context into NCF models by analyzing the behavior of our recommender systems on the 8 largest banned (8-LB) sub-

	Top 1	**Top 10**
NCF	0.5%	11.4%
NCF + text-based scores	2.4%	13.2%
Δ	4.8×	1.2×

Table 4: Percentage of 'vulnerable' users who receive a recommendation to a banned subreddit in their top 1 and top 10 recommendations. Using textual context in the model increases the likelihood of recommending a controversial subreddit by as much as 4.8×.

reddits. We evaluate our models on a subset of 918 'vulnerable' users, who are the users that did not engage with any of the 8-LB subreddits between 01–09/2019, but did engage with at least one of them between 10–12/2019. We take this group as a proxy for the users who would be willing to engage with a controversial subreddit if the recommender system exposed them to one, even if they've never done so before.[2]

In Figure 2, we observe that using the user-generated text from posts and comments makes the recommender model much more confident in its recommendations for the 8-LB subreddits, in the sense that the NCF model scores for those subreddits become much closer to 0 and 1.

In Table 4, we show that adding text-based context scores increases the risk of recommending an 8-LB subreddit by as much as 4.8×. In other words, when the model is provided with the last 6 months of a user's post and comment history, it can target vulnerable users more effectively and give them recommendations to controversial subreddits. We do not believe that the use of text is inherently tied to pushing people towards controversy; our study only suggests that extra attention needs to be given to the issue, and systems should be evaluated carefully for such tendencies and designed to avoid them.

4. Related Work

Reddit data has been used in the past for community recommendation, but prior work on Reddit has generally made limited use of the semantic content of the posts. Jamonnak et al. (2015) uses association rule mining on the user and subreddit IDs (i.e., based on user-subreddit co-occurence statistics) to build a Reddit recommendation service. However, this approach does not leverage the text in the posts and comments of each subreddit. Tuomchomtam and Soonthornphisaj (2019) create a recommender system for subreddits based on statistical features like the average length of posts in the subreddit and the proportion of posts that only contain links. Subreddits are then clustered together with DBSCAN based on these features. This approach does not use the text content to generate recommendations, whereas our approach does.

The textual content of posts has been used in other Reddit applications unrelated to community recommendation. For example, Chandrasekharan et al. (2019) construct a model that helps Reddit moderators detect posts and comments that violate community guidelines (e.g., hate speech, pornography, etc.) Tan and Lee (2015) study the patterns in the trajectories of Reddit user behavior across multiple subreddits.

Furthermore, while Twitter and Reddit have been used as sources for social media text in other publications (Pak and Paroubek, 2010; Petrović et al., 2010; Weninger et al., 2013), the work that we've surveyed did not release their processed datasets for other researchers to use. To the best of our knowledge, no previous work has organized Reddit data into a form suitable for creating text-based recommender systems or provided baselines for such systems. Our corpus provides a way to generate reproducible results in this area.

Other social media datasets have been used for text-based recommender systems. Lan et al. (2018) propose a topic model based on point processes to recommend relevant discussion threads to students on the Coursera website to facilitate social learning. Risch and Krestel (2020) predict whether users will engage with other user-generated comments on the Guardian newspaper website. Recommender systems that use text as part of their input have been examined in the context of books (Mooney and Roy, 2000), news articles (Lu et al., 2015), scientific articles (Popescul et al., 2001), etc.

There is a substantial body of work that studies hate speech, political polarization, disinformation, and other related phenomena on social media. Chandrasekharan et al. (2017) study the reduction in hate speech on Reddit after a 2015 ban on selected subreddits. Tucker et al. (2018) provide a comprehensive survey of the literature on disinformation and polarization on social media. However, previous studies of Reddit data did not publish the corpora that the researchers used, which causes difficulties for other researchers who are attempting to build upon prior work or reproduce those findings.

5. Discussion

The Engage corpus provides a realistic, large-scale dataset for constructing text-based recommender systems. We expect that this resource will help researchers build and better understand the systems that affect user interactions with online groups on social media. We also show that leveraging user-generated text can result in recommendations that drive more user traffic to controversial subreddits. We hope that future work will explore techniques to avoid this phenomenon.

6. Acknowledgements

We gratefully acknowledge the helpful discussions and pointers to related work from our colleagues and anonymous reviewers.

[2] This proxy isn't perfect; for example, it's possible that the users engaged with these subreddits prior to 2019.

7. Bibliographical References

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), December.

Chandrasekharan, E., Gandhi, C., Mustelier, M. W., and Gilbert, E. (2019). Crossmod: A cross-community learning-based system to assist Reddit moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30.

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., and Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, 40:3–35.

Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. (2017). Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Horwitz, J. (2021). Facebook knew calls for violence plagued groups, now plans overhaul. *Wall Street Journal*, Jan.

Jamonnak, S., Kilgallin, J., Chan, C.-C., and Cheng, E. (2015). Recommenddit: a recommendation service for Reddit communities. In *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 374–379. IEEE.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Lan, A. S., Spencer, J. C., Chen, Z., Brinton, C. G., and Chiang, M. (2018). Personalized thread recommendation for MOOC discussion forums. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 725–740. Springer.

Lu, Z., Dou, Z., Lian, J., Xie, X., and Yang, Q. (2015). Content-based collaborative filtering for news topic recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media*, pages 25–26.

Popescul, A., Ungar, L. H., Pennock, D. M., and Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, page 437–444, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Puri, N., Coomes, E. A., Haghbayan, H., and Gunaratne, K. (2020). Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Human vaccines & immunotherapeutics*, 16(11):2586–2593.

Reddit. (2020). Update to our content policy.

Risch, J. and Krestel, R. (2020). Top comment or flop comment? Predicting and explaining user engagement in online news discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 579–589.

Tan, C. and Lee, L. (2015). All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1056–1066.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2020). Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Tuomchomtam, S. and Soonthornphisaj, N. (2019). Community recommendation for text post in social media: A case study on Reddit. *Intelligent Data Analysis*, 23(2):407–424.

Weninger, T., Zhu, X. A., and Han, J. (2013). 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)*, pages 579–583. IEEE.