Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 1131–1136
DOI: 10.15439/2023F6275
ISSN 2300-5963 ACSIS, Vol. 35

Text embeddings and clustering for characterizing online communities on Reddit

Jan Sawicki
0000-0002-8930-7564
Warsaw University of Technology
Email: jan.sawicki2.dokt@pw.edu.pl

***Abstract*—This work analyses Reddit, the largest public, topic-centered social forum. In the experiments, contextualized text embeddings, obtained using DistilBERT, represented subreddit content. Next, clustering was performed, using an unsupervised K-means algorithm and evaluated with multiple clustering metrics. The obtained clusters were analyzed. Moreover, changes of cluster structure, between 2019 and 2022 have been examined.**

I. Introduction

REDDIT is the largest, public, topically-separated forum [1]. Its unique structure allows users with common interest to find their place for discussion, i.e. a subreddit; a subforum dedicated to a particular topic. The platform policy, and Reddit administration, put no restrictions on the topic of subreddits (except for the rules regarding illegal content). Moreover, any user with at least 30-day old account and a non-negative "karma score" (reputation metric) can create a subreddit. This allows "communities" to blossom, with little-to-no supervision. There are subreddits, which are very close thematically, e.g. r/worldnews or r/news (news and information), or r/leagueoflegends and r/Overwatch (video games). There are also subreddits with distant, or even opposite, topics, e.g. r/Conservative and r/Libertarian.

The freedom and scale of subreddits raise multiple research questions, e.g. what are the most popular topics? Are there topically similar subreddits? Can subreddits be reasonably grouped into clusters? Are there, and if so what are, migrations of subreddits between clusters? This contribution explores these questions, for a Reddit dataset spanning 2019-2022, using natural language processing and data clustering.

II. Related works

Reddit's communities have been analyzed with different methods and from different perspectives. The main inspiration for this work are the results of a 2015 study [2] clustering 15,000 subreddits, from the first half of 2013 using scale-free backbone graph networks. The subreddits were grouped into 57 clusters and further, manually, annotated into 10 metaclusters (categories) such as: Electronic Music, Fitness, Sports, Soccer, Video Games, my Little Pony, LGBT, Pornography, Programming, Guns. Captured relations were based on interactions of over 800,000 users. However, the actual content of the posts, or comments, were not analyzed.

Two years later, "community2vec" [3], was introduced. This study also focused on users, by encoding post authors and user co-occurrences and applying PCA . Additionally, post content was encoded with static GloVe embeddings [4]. The main result was showing that vector representations of communities can encode meaningful analogies and semantic relationships, similarly to what has been previously seen for words.

A 2020 study of Reddit and Twitter [5] focused exclusively on texts of 54.5 million Reddit comments and 23,684 tweets. Its goal was to compare text embedding methods: TF-IDF Word2Vec [6] and Doc2Vec [7] applied to topic modelling, with document clustering using k-means, k-medoids, hierarchical agglomerative clustering and non-negative matrix factorization (NMF). For these, different settings and hyperparameters have been tested. It was established that combining Doc2Vec and K-means achieved the best results.

Finally, in [8], instead of static clustering, community evolution over time was analyzed. Here, active users and textual content, processed with LIWC analysis , has been applied. The results represent patterns of user engagement at different stages of community lifespan. However, they do not show how the subreddit topic clusters evolve over time.

To summarize, over time, a shift from user-based to content-based embeddings can be observed. Additionally, since 2018, an influence of NLP advancements [9] is noticeable; from basic text processing (e.g. LIWC, TFIDF, PCA) and static embeddings (e.g. GloVe, Word2Vec and Doc2Vec), to contextualized text embeddings (e.g. BERT [10] and BERT-like models, e.g. DistilBERT [11]). Here, note that older techniques underperform, against BERT-like models (e.g. LIWC [12]).

Moreover, Reddit continues to grow, since its launch in 2005, with past studies completed in 2015, 2017 and 2020. Therefore, a Reddit structure study needs to be revisited, applying modern approaches, to understand what the subreddits communities look today like and how they change in time. Therefore, this contribution presents results of explorations based on a dataset spanning four years (2019-2022), while applying contextual text embeddings with a BERT-like model, with the goals of analysis of subreddit community structure and its evolution over time.

III. Methodology

Let us now briefly discuss (1) dataset, (2) text embedding method, (3) clustering approaches, and (4) cluster quality assessment methods used in this contribution.

A. Dataset

Reddit consists of over 3.5 million communities (and over 1.5 billion monthly visitors). The most popular Reddit data source is the Pushshift database [13], [14]. The subreddit data was extracted from Pushshift subreddit dumps. Furthermore, Pushshift REST API could not have been used due to an outage that happened in early 2023. Overall, content of 3090 "largest" subreddits, i.e. subreddits with at least 100,000 subscribers, has been extracted.

Overall, Reddit is a subject to the 1% rule that appears in the majority of social networks [15]. The majority of posts gain little-to-no attention ("upvotes"), while a small fraction "goes viral" and appears on the frontage of Reddit (the main Reddit forum r/all). Hence, to reduce the computational cost, while capturing subreddit structure, 1000 posts with the highest *scores*, have been extracted from each subreddit. The *score* is the Reddit's measure of "appraisal by a community of Reddit subscribers of an item" [16]. Finally, the dataset spans 4 years: 2019-2022, to allow the analysis of subreddits cluster evolution over time. The resulting dataset consisted of over 12 million unique user posts.

B. Text embedding

After gathering, text embeddings has been applied to the posts. Since the introduction of BERT (in 2018), multiple models, for different NLP goals, were introduced [17], [18]. This work needs a general feature extraction models that deliver multipurpose text embeddings. The model should be "general" and multipurpose, because input data originates from over 3000 communities, and covers topics from politics and news (r/politics, r/news), through video games (r/DOTA, r/gaming), memes (r/hmmm, r/me_irl), drug usage (r/LSD, r/shrooms), to plants (r/Bonsai, r/gardening), fishkeeping (r/Aquariums, r/PlantedTank) or military (r/military, r/guns).

Moreover, the NLP part takes the longest processing time (over 50% of total runtime). Therefore, a general multipurpose and fast, but efficient model is required. In 2019 a "smaller, faster, cheaper and lighter" version of the BERT model has been introduced, the DistilBERT. It retains 97% of the original BERT performance on downstream tasks, while being 40% smaller and 60% faster [11]. Therefore, to reduce computation time, DistilBERT has been selected.

Here, it should be noted that different Reddit communities have different posts "styles". For example, r/politics consists mostly of links to news websites, while r/AbruptChaos contains mostly GIFs or short videos. There is, however, one part of posts that is forced by Reddit – the post's title, which has to be present on every posts regardless of subreddit. While there are ways to overcome this (Reddit post, over 99% of posts in the dataset have a textual title. Overall, DistilBERT embedded posts titles to 768 dimensional vectors, which were clustered.

C. Clustering

Use of K-means for clustering followed results found in [19], [20]. However, the biggest downside of K-means is that it requires specification of the number of clusters. This problem can be overcome by using unsupervised clustering metrics [21], [22] to find "best" clustering. In this context, Silhouette Score (previously used on Reddit [23]), Davis-Bouldin score , Caliński-Harabasz score and k-means inertia (sum of squared distances of samples to their closest cluster center) have been tried. The most suitable cluster size has been sought by evaluating clustering results for cluster sizes: 10, 20, 30, ... 1530, 1540, where 1540 is half o the number of subreddits. Davis-Bouldin metrics is the only metric where lower values are better (for others, higher is better). For easier interpretability of the results, presented in Figure 1, Davis-Bouldin metrics is presented with a minus sign. Interestingly, the metrics were practically identical for considered time periods (annually for 2019-2022). Hence, it can be stated that the number of subreddit clusters, and hence the topical dispersion, does not change much over time (see, also, Section IV).

Fig. 1. Cluster evaluation with different metrics (Davis-Bouldin score is actually negative score, to keep up with "higher is better" interpretation.

The choice of the cluster number was a challenge since not all metrics have been consistent (see, Figure 1). The Silhouette Score differed a lot. This is related to the fact that the Silhouette Score ranges from -1 to 1, where the best value is 1 (all points assigned to the right cluster) and the worst value is -1 (all points assigned to the wrong cluster), while scores near 0 indicate cluster overlap. All Silhouette Scores were close to 0, indicating existence of overlaps. To find a compromise between Davis-Bouldin, Caliński-Harabasz and inertia metrics, the Elbow Method was applied, as previously used in similar settings [24], [25] (also on Reddit [26]). Overall, any number between 180 and 450 represented a "good fit". However, to achieve interpretability of results, 200 clusters were selected. Here, note that smaller numbers of clusters were checked (e.g. 100), but they produced clusters with "not-fitting" topics. Larger numbers (e.g. 300), on the other hand, resulted in fragmented topics.

After clustering, the subreddit groups were manually assigned to meta groups (categories). For example, subreddits related to games (r/LeagueOfLegends, r/DOTA, r/Overwatch, r/gtaonline) or subreddits related to politics (r/Conservative, r/politics, r/Libertarian, r/Political_Revolution, r/geopolitics). These groups are further described in Section IV-A.

D. Cluster similarity and time-evolution

To automatically detect cluster dynamics, they were compared annually using the Jaccard Index [27]. Each cluster from a period was compared to each cluster from the next chronological period. The pair of sets with the highest Jaccard Index is considered a transition, from the predecessor to the successor. Note that the predecessor and the successor may be the same, i.e. the cluster did not change from period to period. This way, ordered lists of cluster transitions were created. Then for each list a generalized multi-set Jaccard index was calculated for all sets (2019, 2020, 2021, 2022). The results are discussed in Section IV.

IV. Results and their analysis

Let us now discuss the key experimental findings.

A. Subreddit clusters characterization

Let us first look into clusters of subreddits established for year 2022. Table I presents number of subreddit by manually annotated categories. Most categories are obvious, but some require some explaining.

"**Pictures**" aggregates subreddits dedicated to posting pictures, GIFs and videos. The themes range from wallpapers (r/wallpapers), to content that is supposed to amaze on-lookers (r/nextfuckinglevel, r/woahdude) or disquiet/scare them (r/cursedcomments, r/cursedimages, r/cursedvideos).

Some subreddits do not have a "theme", and are extremely broad, such as r/gif, r/gifs, r/pics. Interesting is a group of "X_Porn" subreddits, where X is some subject. Here, the term "porn" is a synonym to "amazing", "beautiful", "wonderful" (not pornography). Subreddits in this group (r/CabinPorn, r/CityPorn, r/EarthPorn, r/InfrastructurePorn) showcase pictures of things, places or phenomena that are to be perceived as "porn", i.e. most spectacular of its kind. There are also "meta-themes", such as r/BetterEveryLoop, where the author of a post claims that the more times a GIF/video is watched, the better it gets. The actual content is discretionary. Finally, there are also subreddits with random pictures, e.g. r/nocontextpics.

In the "**states**" category, there are subreddits related to individual US states and cities, e.g. r/Atlanta, r/Austin, r/Calgary, r/California, r/Dallas, r/Denver, r/LosAngeles.

Interestingly, while the applied NLP model is meant for English, it clustered subreddits in other languages into the category "**language specific**". There are also separate clusters for: German subreddits (r/de, r/de_IAmA), the Polish subreddit r/Polska, the Netherlands, containing r/thenetherlands and a cluster related to Scandinavian subreddits, containing r/norge, r/svenskpolitik, r/swedishproblems.

The "ask" category contains subreddits with questions. Here, questions can be general (r/AskReddit), topic specific (r/morbidquestions, r/AskRedditAfterDark), or answerer specific (r/AskMen, r/AskMenOver30, r/AskWomen, r/AskEurope, r/AskUK).

There was a group that was separated from "pictures" were "**animals**". This group contains clusters of subreddits about (mostly) dogs and cats and other small animals. It appears that the similarity between some "animals" subreddits and some "pictures" is in the feelings that the pictures are supposed to invoke, i.e. happiness, or cuteness. For example, subreddits: r/aww (described as: "Things that make you go AWW! Like puppies, bunnies, babies, and so on... A place for really cute pictures and videos!" and r/MadeMeSmile ("A place to share things that made you smile or brightened up your day. A generally uplifting subreddit." . On the "other side", one can find "animals" in subreddits dedicated to brutality and violence in animal kingdom (r/natureismetal, r/Natureisbrutal).

Next, there is the "**irl**" group, standing for "in real life". It contains subreddits, such as: r/meirl, r/2meirl4meirl, r/bi_irl, r/discord_irl, r/egg_irl, r/anime_irl, r/gay_irl, r/me_irlgbt, r/woof_irl, r/ich_iel . All of them contain pictures with strict post title policy. Depending on the subreddit, the titles are always "meirl" (r/meirl), "2meirl42meirl4meirl" (r/2meirl42meirl4meirl), etc. It was clear how to characterize the content of these subreddits, other than it being memes. Interestingly, even though r/ich_iel is a German subreddit, it got clustered with other English "irl" subreddits.

Let us now consider "**social chatting**" group. Here, clusters include both general (r/CasualConversation, r/MakeNewFriendsHere) and specialized chatting topics (r/BreakUps, r/LongDistance, r/Marriage). There are also subreddits where users explicitly ask for an advice: r/Advice, r/askwomenadvice, r/dating_advice or seek approval/disapproval of their actions: r/AmItheAsshole (the latter with a dedicated study, from 2023 [28]).

Moving to smaller subreddits, there is the "**NSFW**" (Not Safe For Work) group. Here, confirmation that the user is an adult is required. However, these are different from "pornography", since they discuss adult topics, such as fetishes, fantasies and other sex-related issues. The examples are: r/BDSMAdvice, r/BDSMcommunity, r/Swingers, r/polyamory, r/DeadBedrooms, r/NoFap, r/bigdickproblems. There are also subreddits devoted to looking for other people with similar interests. These often use the acronym "r4r" (Redditors for Redditors): r/DirtySnapchat, r/Kikpals, r/dirtykikpals, r/dirtyr4r, r/exxxchange, r/r4r, r/snapchat, r/swingersr4r.

The "**Reddit meta**" category contains Reddit administration (r/announcement) and technical support subreddits (r/help). Next, "Deals" category contains subreddits about free goods, or goods on sale, e.g. r/GameDeals, r/NintendoSwitchDeals, r/PS4Deals, r/deals, r/eFreebies, r/freebies, r/googleplaydeals. Finally, "**Help me find**" is the group of subreddits where users ask others to help them find something, or find what something is, e.g. r/HelpMeFind, r/RBI (Reddit Bureau of Investigation), r/Whatisthis. Here, subreddits for identifying pornographic performers or scenes r/pornID, r/sources4porn, r/tipofmypenis are included.

What is clearly visible in Table I, is that the most of the clusters are related to pornography, pictures with vague themes, video games, memes and technology. These themes also aggregate the biggest number of subreddits.

TABLE I
CLUSTERING EVALUATION

category	subreddits count	cluster count
pornography	692	33
pictures	253	22
games	215	6
memes	201	8
mixed	195	13
tech	135	9
social chatting	119	6
tv series	99	3
animals	95	5
politics	92	5
music	64	6
sports	64	3
finance	58	4
NSFW	52	4
hate	48	1
cooking	39	3
drugs	38	2
popculture	36	1
science	31	3
states	29	1
education	27	3
fashion	27	2
game consoles	27	1
language specific	26	12
military	24	3
ask	23	2
fitness	21	2
science fiction	21	1
art	20	2
camping	20	1
irl	19	7
movies	19	2
cars	17	2
plants	17	1
Reddit meta	16	3
craftsmanship	16	1
food	14	1
horror	14	1
mental health	12	1
deals	11	3
writing	11	2
crime	11	1
religion	11	1
anime	10	1
trading	9	3
help me find	7	1
hiring	5	1
surveys	1	1

B. Subreddit clusters findings

Let us now report key findings regarding the clustering.

1) Subreddit naming: There are naming patterns and conventions of subreddits. Reddit's users employ multiple acronyms, e.g. "IRL" (In Real Life), "AMA" ("Ask Me Anything") or "NSFW" ("Not Safe For Work"). These 3 alone materialize in 42 subreddits. There are also subreddits acronym names, e.g. r/ATBGE ("Awful Taste But Great Execution"), including the longest name: r/UNBGBBIIVCHIDCTIICBG ("Upvoted Not Because Girl, But Because It Is Very Cool; However, I Do Concede That I Initially Clicked Because Girl."). As noted, common is using the word "porn" to name content that is supposed to be beautiful, aesthetically pleasing, interesting, well-made, etc. Moreover, only 10 out of 71 "X_porn" subreddits contain actual pornography.

Finally, while many subreddits are descriptive of the topic (e.g. movie or TV series title, music genre, area of science or name of a video game), multiple subreddits focus on describing a general phenomenon/feeling (r/aww, r/INEEEEDIT, r/iwanttobeher). This shows how important it is to analyze the content of the subreddits instead of just the names.

2) Country subreddits: There is a group of subreddits, in English, dedicated to countries, e.g. r/UnitedKingdom, r/russia, r/China, r/canada. Interestingly, they appear in *different* clusters, but all in the category "politics". Interestingly, r/russia and r/China appear in a single cluster, consisting of: {r/ANormalDayInRussia, r/China, r/MapPorn, r/PropagandaPosters, r/imaginarymaps, r/russia, r/vexillology, r/vexillologycirclejerk}. This suggests that text embeddings of map-related subreddits and a propaganda poster subreddit, are similar to the content of r/russia and r/China. Subreddits r/canada, r/europe, r/unitedkingdom are in the same cluster with political and general news subreddits: {r/CanadaPolitics, r/Conservative, r/DemocraticSocialism, r/canada, r/Economics, r/Libertarian, r/Political_Revolution, r/europe, r/Republican, r/news, r/The_Mueller, r/democrats, r/geopolitics, r/politics, r/ukpolitics, r/unitedkingdom, r/worldnews}.

Note that even though the US is the third-largest country by population, and has the highest number of users on Reddit, it does not have a dedicated subreddit. However, as noted (in Section IV-A), there exist subreddits dedicated to US states and cities, and they form a separate cluster.

3) Technology + Finance = Cryptocurrencies: There is an interesting overlap between subreddits in "finance" and "tech". There is a cluster containing both investment subreddits (r/algotrading, r/pennystock, r/RobinHoodPennyStocks), cryptocurrencies subreddits (r/BitcoinMarkets, r/btc, r/ethereum, r/ethtrader) and technology subreddits (r/tech, r/technews, r/technology). This captures that fact that cryptocurrencies became a "middle-ground" conversation joining finances and technology.

4) Real life and gaming: In the cluster dedicated to crafts, there is a subreddit with "digital craftsmanship". It is r/Minecraftbuilds, containing building concepts created in the game Minecraft. It appears in the cluster with subreddits such as: {r/HomeImprovement,

r/Justrolledintotheshop, r/Tools, r/electricians, r/longboarding, r/redneckengineering, r/whatisthisthing, r/woodworking}. Similarly, among "fashion" subreddits there is one related to the Animal Crossing video game fashion designs (r/ACQR): {r/AsianBeauty, r/Embroidery, r/Makeup, r/MakeupAddiction, r/RedditLaqueristas, r/crafts, r/crochet, r/femalefashionadvice, r/knitting, r/malefashion, r/malefashionadvice}. This illustrates interfusion of real life craftsmanship and fashion with in-game craftsmanship and fashion.

5) Are you eating? Watch a documentary.: There exists a relatively small cluster: {r/Documentaries, r/mealtimevideos}, where the first subreddit is about a documentary movie and the second contains video suggestions for watching during lunch or dinner. It appears that both of them have similar content, meaning that documentary videos would be a good suggestion for watching during mealtime.

6) Pornography mix: As visible previously in Table I, most of the clusters are dedicated to pornography. There are subreddits dedicated to fetishes, body parts, looks activities performers, sexual preference (e.g. heterosexual, homosexual etc.), amateur vs professional or performers. However, all of their content seems alike, as there are no particular patterns in pornography clusters, except for one. The subreddits dedicated to particular performers have similar content (often about praising a particular performer), e.g. {r/AdrianaChechik, r/AlexisTexas, r/AngelaWhite, r/DaniDaniels, r/KimmyGranger, r/Miakhalifa, r/RileyReid, r/abelladanger, r/leahgotti}.

The lack of any other patterns when it comes to clustering pornography subreddits shows that their content is extremely overlapping and similar regardless of the subreddit.

C. r/worldpolitics in NSFW subreddits

There is an interesting anomaly in one of the clusters. Subreddit r/worldpolitics appears in a cluster nearly exclusive to NSFW content, e.g. {r/BDSMAdvice, r/Rapekink, r/SexWorkers, r/Swingers, r/mbti, r/bigdickproblems, r/bisexual, r/lgbt, r/polyamory, r/sexover30, r/BDSMcommunity}. At first, this looks like a clustering error, since r/worldpolitics should be in a political cluster with subreddits such as r/politics. Due to permissive rules of this subreddit it is full of all kinds of posts. Its description states "reddit's anything goes subreddit, no topic imposed or opposed by the mods" . Additionally, when posts are sorted by Reddit's "top of all time", the first 100 are marked NSFW, even though they do not include adult content. Overall, r/worldpolitics seen from the perspective of text-embedding, is very close to other adult-content subreddits, i.e. it is either very chaotic, or contains adult content.

1) Current clustering vs. previous studies: As mentioned, the study from 2015 [2] performed similar clustering and manual annotation into "meta clusters". Let us compare meta-clusters from 2015 with these from 2022.

First, the 2015 groups: "Fitness", "Sports", "Video Games", "Pornography" all map one-to-one to cluster groups established in 2022. Second, "Electronic Music", "Programming", "Soccer" and "Guns" map to wider/similar cluster categories, which are "music", "tech", "sports", and "military", respectively. Third, there are 2 groups of clusters with no one-to-one correspondence: "my Little Pony" and "LGBT". Subreddits marked as "LGBT" in 2015 appear in clusters of "NSFW" (e.g. {r/BDSMAdvice, r/Rapekink, r/SexWorkers, r/Swingers, r/mbti, r/bigdickproblems, r/bisexual, r/lgbt, r/polyamory, r/sexover30, r/BDSMcommunity}). This can be a question of the naming convention. Moreover, it seems that LGBT issues are close to NSFW, which is logical, since many of them involve sexuality and sex. "My Little Pony" subreddits were completely absent in this analysis. Even though they are large enough (over 100,000 subscribers), they did not appear in the Pushshift dumps, probably due to inconsistencies in the Pushshift database. Hence, this cluster has no mapping to 2022 clusters.

Similarly to the original work, a dimensionality reduction of the embeddings has been performed with the t-SNE method, to create a two-dimensional visualization. Figure 2 shows the clusters in the top 20 categories, by subreddit count in cluster. T-SNE reduced the dimensions of vectors 768 to 2, Even in two dimensions the clusters such as "pornography", "sports", "music" or "tech" appear close within the group and far between the groups. This is consistent with the 2015 study. Hence, it supports quality of embeddings and category annotations reported in current contribution.

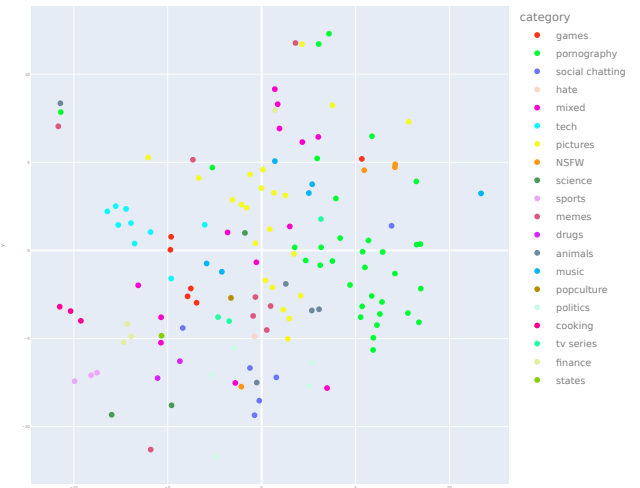

Fig. 2. The clusters in the top 20 categories by subreddit count in cluster. The X and Y axis are insignificant due to dimensionality reduction with t-SNE

D. Subreddit cluster transitions

The second group of results focuses on cluster evolution between 2019 and 2022. Due to space limitations, only selected key findings are presented.

1) Gardening, hair, writing and vehicles stay unchanged: The highest Jaccard index (0.78) between subreddit clusters is achieved for clusters about plants, gardening and fish tanks.

Here, almost no change has been observed for 4 years. The cluster lost one subreddit: r/shrooms, and gained 3 new ones: r/snakes, r/thingsforants, r/whatsthisbug. Interestingly, in 2021, the r/shrooms subreddits migrated to a drug-related cluster. Similarly, between 2019 and 2022, the hair-related cluster (Jaccard index of 0.7) went through a couple changes, but finally lost one subreddit (r/FancyFollicles) and gained one (r/beauty).

With Jaccard index of 0.61 there is also the writing cluster, which moved closer to it's writing theme by dropping r/MovieSuggestions, r/TrueFilm and gaining r/stephenking.

Another barely changed cluster concerns vehicles. In 2019, it was mostly related to cars, but in 2020 it gained and retained r/MTB (mountain bike), r/bicycling and r/cycling subreddits. Interestingly, the r/Cartalk and r/MechanicAdvice subreddits, in 2022 formed a completely new cluster. The main difference between these two and other clustered subreddits is that they focus on discussions rather than showcasing vehicle models.

2) COVID-19: As expected, COVID-19 pandemic is noticeable in subreddit evolution. A cluster that, in 2019, contained general science, health and chemistry subreddits ({EverythingScience, r/Health, r/physicsgifs, r/science}) was extended, in 2020, with r/COVID19, r/China_Flu, r/Coronavirus. It remained unchanged until 2022.

3) Pornographic subreddits migrations: Over time, significant migrations between pornographic clusters have been observed. "Pornography" category has the second-lowest mean, 4-year, Jaccard index of about 0.02. There even was one cluster in 2019 of 30 subreddits which finally got reduced to a single-subreddit cluster (containing only r/sarah_xxx). However, similarly as described in Section IV-B6 these cluster migrations are chaotic and random, and no pattern was detected.

V. Concluding remarks

This work is devoted to study of structure of and time evolution of the Reddit platform. Current text-embedding methods have been applied to the dataset covering 2019-2022 period. Overall, Reddit is a place containing content and discussion on various topics, with both very wide and very narrow scopes. Majority of the most popular subreddits are dedicated to pornography, pictures and videos about "anything and everything". Furthermore, popular are video games, memes and technology subreddits. While some of the topical clusters stay unchanged over the years, there are subreddit migrations between most of the clusters. Future studies will focus on more particular groups of subreddits and researching new methods for inter-subreddit topical modelling, such as crossposts.

References

[1] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, "Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics," *Social Media+ Society*, vol. 7, no. 2.

[2] R. S. Olson and Z. P. Neal, "Navigating the massive world of reddit: Using backbone networks to map user interests in social media," *PeerJ Computer Science*, vol. 1, p. e4, 2015.

[3] T. Martin, "community2vec: Vector representations of online communities encode semantic relationships," in *Proceedings of the Second Workshop on NLP and Computational Social Science*.

[4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing*.

[5] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit," *Information Processing & Management*, vol. 57, no. 2, p. 102034, 2020.

[6] K. W. Church, "Word2vec," *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017.

[7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[8] H. Mensah, L. Xiao, and S. Soundarajan, "Characterizing the evolution of communities on reddit," in *International Conference on Social Media and Society*, 2020, pp. 58–64.

[9] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," *arXiv preprint arXiv:2003.07278*, 2020.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[12] J. Biggiogera, G. Boateng, P. Hilpert, M. Vowels, G. Bodenmann, M. Neysari, F. Nussbeck, and T. Kowatsch, "Bert meets liwc: Exploring state-of-the-art language models for predicting communication behavior in couples' conflict interactions," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 385–389.

[13] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 830–839.

[14] J. Sawicki, M. Ganzha, M. Paprzycki, and A. Bădică, "Exploring usability of reddit in data science and knowledge processing," *Scalable Comput. Pract. Exp.*, vol. 23, pp. 9–22, 2021.

[15] E. Hargittai and G. Walejko, "The participation divide: Content creation and sharing in the digital age," *Information, Community and Society*, vol. 11, no. 2, pp. 239–256, 2008.

[16] P. Van Mieghem, "Human psychology of common appraisal: The reddit score," *IEEE Transactions on Multimedia*, vol. 13.

[17] P. Xia, S. Wu, and B. Van Durme, "Which* bert? a survey organizing contextualized encoders," *arXiv preprint arXiv:2010.00854*, 2020.

[18] M. Koroteev, "Bert: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021.

[19] G. Ahalya and H. M. Pandey, "Data clustering approaches survey and analysis," in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management*.

[20] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*.

[21] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE access*.

[22] T. Sai Krishna, A. Yesu Babu, and R. Kiran Kumar, "Determination of optimal clusters for a non-hierarchical clustering paradigm k-means algorithm," in *Proceedings of International Conference on Computational Intelligence and Data Engineering: ICCIDE 2017*.

[23] J. Sirait, "Investigating news source characterizations using reddit audiencebased metrics," 2022.

[24] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, vol. 336. IOP Publishing, 2018, p. 012017.

[25] M. Cui *et al.*, "Introduction to the k-means clustering algorithm based on the elbow method," *Accounting, Auditing and Finance*, vol. 1.

[26] V. Veselovsky, I. Waller, and A. Anderson, "Imagine all the people: Characterizing social music sharing on reddit," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15.

[27] L. d. F. Costa, "Further generalizations of the jaccard index," *arXiv preprint arXiv:2110.09619*, 2021.

[28] S. Giorgi, K. Zhao, A. H. Feng, and L. J. Martin, "Author as character and narrator: Deconstructing personal narratives from the r/amitheasshole reddit community."