applied sciences

MDPI

Article

Applying Named Entity Recognition and Graph Networks to Extract Common Interests from Thematic Subfora on Reddit

Jan Sawicki [1,*], Maria Ganzha [1], Marcin Paprzycki [2] and Yutaka Watanobe [3]

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland; maria.ganzha@pw.edu.pl

[2] Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland; marcin.paprzycki@ibspan.waw.pl

[3] Department of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu 965-8580, Japan; yutaka@u-aizu.ac.jp

* Correspondence: jan.sawicki2.dokt@pw.edu.pl

Abstract: Reddit is the largest topically structured social network. Existing literature, reporting results of Reddit-related research, considers different phenomena, from social and political studies to recommender systems. The most common techniques used in these works, include natural language processing, e.g., named entity recognition, as well as graph networks representing online social networks. However, large-scale studies that take into account Reddit's unique structure are scarce. In this contribution, similarity between subreddits is explored. Specifically, subreddit posts (from 3189 subreddits, spanning the year 2022) are processed using NER to build graph networks which are further mined for relations between subreddits. The evaluation of obtained results follows the state-of-the-art approaches used for a similar problem, i.e., recommender system metrics, and applies recall and AUC. Overall, the use of Reddit crossposts discloses previously unknown relations between subreddits. Interestingly, the proposed approach may allow for researchers to better connect their study topics with particular subreddits and shows promise for subreddit similarity mining.

Keywords: data mining; social networks; Reddit; natural language processing; graph networks

check for updates

Citation: Sawicki, J.; Ganzha, M.; Paprzycki, M.; Watanobe, Y. Applying Named Entity Recognition and Graph Networks to Extract Common Interests from Thematic Subfora on Reddit. *Appl. Sci.* **2024**, *14*, 1696. https://doi.org/10.3390/app14051696

Academic Editor: Andrea Prati

Received: 2 January 2024
Revised: 28 January 2024
Accepted: 14 February 2024
Published: 20 February 2024

1. Introduction

Reddit is a very large social network, with information divided into topical fora called subreddits. In March 2023, Reddit had about 52 million active users (https://backlinko.com/reddit-users, accessed on 13 February 2024) and over 3.4 million subreddits (https://www.businessdit.com/how-many-subreddits-are-there, accessed on 13 February 2024). Analysis of pertinent literature shows that there are two main approaches to using Reddit-extracted data in research. The first focuses on a *single subreddit* and analyzes various phenomena within it. Here, existing studies have been focused, among other topics, on politics [1], gaming [2], online harassment [3], mental health [4,5], suicide prevention [6], dermatology [7], parenting [8] or teaching [9], and others [10]. On the other hand, Reddit-based data, *treated as a single dataset*, have been used in natural language [11] or image processing [12]. Here, the Reddit structure(s) was (were) ignored, and all considered posts were treated as a single large collection of texts. However, recent researchresearch ([13]) suggested that additional big scale studies are needed, to properly capture Reddit information structure.

In this context, to the best of our knowledge, no research used a large Reddit dataset, while taking into account the existence and caveats (such as crossposts) of thematic subfora. This is somewhat surprising, taking into account the fact that it is the existence of the topical subfora that distinguishes Reddit from the other social networks. Therefore, the aim of this contribution is to deliver a more comprehensive understanding of the information structure of Reddit. In particular, the focus of reported work is on automatically establishing common topics of interest shared by readers of individual subreddits.

Here, we note that there exists a Reddit-only phenomenon that is completely absent from the Reddit research. These are crossposts, i.e., posts that were posted to one subreddit and later linked in another one. In other words, crossposts represent the situation when individual users of a specific subreddit believe that selected posts could be of interest to readers of another subreddit. As such, crossposts explicitly capture how individual users conceptualize similarities between subreddits. However, it has to be acknowledged that crossposts are relatively infrequent (vis-à-vis the volume of data posted regularly on Reddit). This means that their appearance can be used only as a corroboration of the existence of common interests between readers of different subreddits (i.e., it cannot be treated as a ground truth for subreddit similarity). Nevertheless, it is our belief that they are worthy of a closer examination.

In this context, the aim of this contribution was to establish (1) ways to automatically uncover topical similarities between subreddits and (2) the role that crossposts can play in this process. The proposed approach to reach these goals is based on natural language processing (NLP), Named Entity Recognition (NER), and graph networks. Here, NER is used to find entities which are then used to build graph networks. The goal of the process is to capture and elaborate on similarities between found named entities.

We note that this work can have practical application. Let us assume that researchers are interested in topics related to mental health (as they are represented in the Reddit posts). In this case, the most natural subreddit for their work would be r/mentalhealth. However, if it would be possible to establish which other subreddits involve topics similar to these discussed in r/mentalhealth, then such subreddits could be also included in the research. Similarly, researchers interested in former President Donald Trump could consider all subreddits where Donald Trump is the "common topic", instead of focusing only on the r/The_Donald.

The remaining parts of this contribution are organized as follows. Section 2 presents the state of the art pertinent to the area of interest of this contribution, i.e., research related to use of natural language processing and graph networks as applied to Reddit. Next, Section 3 describes the collected dataset. The proposed approach is introduced in Section 4 and followed by analysis of experimental results presented in Section 5. Finally, the research is summarized in Section 6, where future research directions are also elaborated. In addition, this work is accompanied by supplementary information in Appendix A.

2. Pertinent State of the Art of Reddit-Related Research

Two recent overviews of Reddit-related research [14,15] suggested that natural language processing and graph networks are the most popular techniques used to analyze various aspects of Reddit-derived datasets. The following sections discuss the state of the art of the NLP methods, graph networks, and performance evaluation found in Reddit-related work.

2.1. Reddit and Natural Language Processing

One of the main NLP-anchored research directions found in Reddit-related literature concerns extracting the main point(s) of a text [16]. The umbrella term for these methods is *topic modeling*. In this research, typically, one attempts to extract "main topics" from the textual data. Next, the output is further modeled, e.g., with graph networks. In this context, we note that graph network model relations are easier to find for a small space of unambiguous features, which clearly point to specific things, phenomena, people, etc. Hence, of particular value are methods with a reasonably small output space. Such methods also deliver results that are easier to comprehend by humans. Therefore, let us now briefly discuss the selected most popular and recent topic modeling methods [17–19] and justify the choice of the method used in this work.

The prime example of a “classical” methods is Non-negative Matrix Factorization (NMF). This method utilizes algebraic tools for matrix factorization to turn the word-document matrix into matrices of document-topic and word-topic dimensions. Algebraically speaking, given a set of text documents (D) and words (W), a matrix is built where rows are word embeddings with dimensions $W \times D$ (i.e., a word-frequency matrix). The resulting matrix is always non-negative since word frequency has to be always positive (a word can only appear a positive number of times in a document). Next, using one of the methods for matrix factorization, two matrices are calculated. Their product would return the original $W \times D$ matrix. The two matrices have dimensions $W \times T$ and $T \times D$ and provide information regarding which word belongs to which topics (T) and which topic belongs to which document, respectively. Hence, the output (the topics) are a combination of the words from the input documents.

The second approach is Latent Dirichlet Allocation [20] (LDA). LDA is a generative model that uses Dirichlet distribution to assign documents their topics based on the words within them. This method is outlined in Algorithm 1.

Algorithm 1: LDA pseudocode (simplified).

```
input: k—number of topics to assign
for document in documents do
    for word in document do
        randomly assign each word a topic from k topics
$N \leftarrow$ number of iterations;
while $N \neq 0$ do
    for document in documents do
        for word in document do
            compute $p(w|t)$—the proportion of all documents that are assigned to
             topic k (for a given word) compute $p(t|d)$—the proportion of words in
             documents that are assigned to topic t
    Recalculate probability $p(w|t,d) = p(w|t) \times p(t|d)$ $N \leftarrow N - 1$;
```

In LDA, there are two hyperparameters not mentioned in the simplified version of Algorithm 1. These are:

- α, a document density factor (i.e., weight of topic in a document) controlling the number of topics expected in the document (the higher the value, the more topics are envisioned to exist in a document),
- β, a topic word density factor (i.e., weight of a word in a topic) controlling the distribution of words per topic in the document (the higher the value, the larger number of words may belong to each topic).

Similarly to NMF, the output of this approach consists of a probability that a given word is a part of a specific topic in a document. The output space consists, again, of different combinations of words from the input documents. We note that the two hyperparameters provide some form of “manual control” of the expected size of the output space.

Other methods that can be used for statistical topic modeling are, for instance, the correlated topic model (CTM) [21] and the Pachinko Allocation Model (PAM) [22]. However, these two (and many others) share the same problem, namely the size of the output space. Since the output topic model is built from the word within the documents, the number of possible topics is enormous. Even though the number of topics and the number of words in a topic can be manipulated (as described in the case of LDA), the final range of topics remains gigantic as, theoretically, any word from any document can be included in a topic. Therefore, methods from this area are not easily applicable to the analysis of large-scale Reddit-derived dataset(s).

Moving to more recent approaches, it is well known that the transformer [23] architecture revolutionized the Natural Language Processing field [24]. It can also be used for topic modeling. Here, BERT-like [24] models use a pre-trained model which is capable of capturing the general semantic meaning of text in low-dimensional vectors (most commonly reported size is 768 or 1024). These models are then fine-tuned on a particular task, e.g., topic modeling. Among many fine-tuned BERT-like models, in the context of this contribution, the most popular would be BERTopic [25], which has shown its applicability to Reddit-derived data in previous studies [26]. BERTopic's process can be explained in the following simplified steps:

1. Conversion of text to vectors using a pre-trained BERT model.
2. Dimensionality reduction with the UMAP model [27].
3. Clustering using the HDBSCAN method [28].
4. Conversion from vectors to topics using TF-IDF (i.e., extraction of most meaningful words for each cluster).

Even with its highly accurate results in various applications [29–31], this approach, again, is problematic due to the fact that the topics are built from the whole space of words in the documents. Moreover, there are no simple mechanisms to control the size of the output space.

Depending on the definition, one may also consider text summarization [32] as part of text modeling methods. There are two main categories of text summarizing models [33]: extractive (which extract particular sentences or sub-text) and abstractive (which generate the summary not necessarily using the words from the input text).

An example of an extractive method is Latent Semantic Analysis (LSA) [34]. As in many other extractive methods, instead of selecting words and assigning them to topics, LSA looks for the most meaningful sentences (or other "larger units" of text) and uses them to construct a summary. This, however, produces an even larger space of possible topics extracted from the text because each extracted topic is basically a set of sentences.

Now, let us consider exemplary methods from the abstractive text summarization category. These methods generate a new text (the summary) based on the input text. The output does not necessarily need to include any of the sentences (or even words) from the original text. Here, the most recent research used deep neural networks, with an encoder–decoder architecture as the backbone of summarization models [35]. As representatives of this approach, the most popular models are BERTSUM [36], Pegasus [37] (further challenged by SimCLS [38]) or Prophetnet [39].

Apart from the neural networks approach, there are also probability-based methods such as BRIO [40]. BRIO challenges the deterministic approach to modeling and uses a novel training paradigm assuming a non-deterministic distribution. We note that the abstractive approach is still a subject to the gigantic topic (summary) space, the vocabulary of which can easily extend to any word similar to the input text in terms of text embeddings.

Finally, feature extraction has been approached using Named Entity Recognition (NER) [41–43]. Although NER is not strictly a topic modeling method, it allows extracting crucial features—named entities. A named entity is a phrase that clearly identifies a person (PER), location/place (LOC), organization (ORG), or others (MISC). The most recent NER models are based on BERT-like [24] transformers with the attention mechanism. The most popular NER model found in the literature is dslim/bert-base-NER (https://huggingface.co/dslim/bert-base-NER, accessed 1 February 2024). This approach is based on a pre-trained model for general language modeling.

It has been further fine-tuned in the context of named entity recognition (which is a subset of feature extraction). Specifically, dslim/bert-base-NER has been trained on a single NVIDIA V100 GPU with recommended hyperparameters from the original BERT paper [24] using CoNLL-2003 [44]. Here, performance of a 91.3% F1 score, a 90.7% precision and a 91.9% recall on the test dataset was reported. Although it is not a necessity for this contribution, the model also returns classes of extracted named entities.

The main advantage of the NER models over the previously covered topic modeling methods is their output. They return results from the domain of named entities, which is far smaller than previously mentioned techniques. This reduction in the size of the output set is very beneficial, considering the next step of the method—building graph networks. Furthermore, building graph networks and finding similarities between them is much easier and unambiguous if the nodes (named entities) refer to particular real-life entities instead of sets of words (like in the previous topic modeling methods). Obviously, it is possible to build networks from topics [45,46] or even key phrases [47]. However, in what follows, the aim is to build simpler and easily cross-referenceable graphs.

In this context, we consider an example of two pairs of graphs (two graphs made with named entities and two graphs made from keywords/topics). A graph built from named entities *NASA*, *Stephen Hawking* and *The Sun* can be easily matched with the second graph, with nodes *NASA*, *USA* and *United States Congress*. Here, *NASA* is unambiguously the common ground topic (entity). If one graph had topics built from keywords, e.g., *national-astronomy-organization*, *famous-science-people*, *astronomy-planets* and the other captured concepts *american-organizations*, *countries-national* and *government-organization*, finding the correspondence would require use of additional techniques such as semantic metrics, text embedding similarity, or others, which would effectively hinder the methods' accuracy and readability. Obviously, this is by no means the best metaphor for comparing topic modeling and NER, rather a visualization of an important issue that has to be considered when developing data processing pipelines.

For completeness, let us note that, in the case of disambiguity, there are also models for Named Entity Linking (NEL) and/or Named Entity Disambiguation (NED) [48] which determine whether two or more named entities refer to the same entity. However, at this stage, these approaches are out of scope of the reported results. Nevertheless, their application may be worth investigating.

2.2. Graph Networks

The second part of the literature review concerns graph networks and their construction and utilization in social media research. Specifically, in the context of this work, after topics are extracted, graph networks can be formed and analyzed from the perspectives of selected aspects of the data [49]. Since in this contribution the relations that are captured are bidirectional, and since their significance varies (as described in Section 4), weighted undirected graphs are of particular interest.

First, let us discuss the selected most common metrics used in the majority of graph analysis. Here, let us assume the node of the graph is denoted as v.

- Degree—the number of edges connected to a node v. For weighted graphs, the sum of weights of all edges connected to a node v may also be used.
- Degree centrality (normalized)—the fraction of all nodes that a given node v is connected to.
- Average neighbor degree [50]—the average of degrees of all nodes that node v is connected to: $\frac{\sum degree(u_i)}{degree(v)}$, where u_i is the neighbors of v.
- Clustering—the fraction of all possible triangles (a K3 graph, a complete graph of Size 3) that could pass through a node v. A triangle is a set of three nodes, which are all connected to each other (i.e., a complete graph of Size 3). The derived formula is $\frac{2T(v)}{degree(v)(degree(v)-1}$, where $T(v)$ is the actual number of triangles that pass through node v [51].

Let us now shortly outline the relevant graph network methods. There are many applications of building graphs from textual features: from general ontologies [52], modeling the real world with graph relations and knowledge graphs [53] enabling reasoning methods, to particular applications in biology [54] and social graph networks [55]. The last one is of particular interest, provided that the dataset under study is derived from Reddit, which is a social network.

In the literature, graph networks modeling social media most commonly apply the user-as-node [56–58], and community-as-node [59–62] representations. There, connections (edges) are built on the basis of interactions (users commenting/talking) or common belonging (e.g., users subscribing to the same groups/subreddits). However, in this work, a different approach is undertaken. Specifically, the features (mined from the posts) are used to find similarities between communities. In this context, the closest domain covered in the literature is the work related to broadly understood recommender systems. Therefore, while appreciating the difference, works related to such system were reviewed to find general methods, performance evaluation methods, etc.

First, there are works concerned with user recommendation. For example, there is a user-to-user recommendation studied in a work about X (formerly Twitter) [63]. Interestingly, this work is based on application of LDA. In particular, the topics extracted using LDA are used to rank users as potential recommendation targets and to build the final recommendation set. The results were tested for a 2010 dataset of tweets. Depending on the subset of the dataset, the recommending systems achieved 20–50% recall. Since the specific *recall* metric was not defined, the general definition, $\frac{True\_Positive}{Positive}$ (also known as hit rate, hit ratio, sensitivity, or power), was assumed. As it is clear from multiple related works, recall is the main method for evaluating rankings of recommendation systems. Additionally, the 2010 contribution claims that "graph-based methods have high precision, since graph information is known to be a reliable estimator of social influence, but also have low recall due to possible low connectivity". This claim is further addressed when presenting obtained results in Section 5.

Moving closer to the core topic of this work, there are hashtag recommendation systems. Hashtags are conceptually similar to named entities, since they point to a specific phenomenon, person, event, organization, etc. One of the works [64] builds the graphs using hashtags, mentions, following information, and topics. Next, with a graph community detection algorithm (the Clique percolation method [65], the Louvain algorithm [66], and the label propagation algorithm [67]), the existing communities are determined. Finally, the original algorithm produces top N hashtags to be used in the recommendation. The method achieved better results in terms of hit rate than previous similar research (43% vs. state-of-the-art 37%).

Another work [68] notices the sparsity of hashtags in tweets and focuses on utilizing external sources to build the tweet–tweet similarity. This method employs disambiguation based on Wikipedia and The Guardian searches, word embeddings, and translation to extract the most significant hashtags for recommendation. This work provided results in terms of *precision@k* and *recall@k* for $k = 2, 3, 5$, where k is the number of top results returned by the model. The state-of-the-art results provided in the work reach recall in the range of 30–40% and precision in the range of 20–35% (exact numbers were not provided). The proposed method performed better, and reached recall in the range of 40–50% and precision of about 40–50% (again, exact numbers were not provided).

Moving away from hashtags and towards Reddit, there were studies where Reddit was used together with other sources. For example, a study used Reddit and Pinterest for ownership recommendation [69]. Since this topic is far from the research reported here, the specific methodology is skipped. However, an interesting part is performance evaluation. This recommender system uses hit rate (recall) similarly to the previous study. In addition, it also employs *area under the receiver operating characteristic curve* (ROC AUC). The use of the ROC AUC metric is particularly important because it prevents the algorithm from maximizing recall by recommending all potential choices.

Another approach, dealing with two data sources, is a study based on Reddit and Twitter [70]. There, the goal is to cross-reference the user's tweets with potentially interesting Reddit threads (this term was not explicitly explained and is assumed to be equivalent to a post or a post + comments). Applying natural language processing with WordNet to tweets, the authors built models (Naive Bayes, Random Forest and SVM) to generate an interest profile for the user and then to recommend the matching genres. Here, again, the important aspect is the evaluation methodology which accounts for top five results and evaluates them using the *accuracy*, *precision* and *recall* metrics.

Finally, the closest to the topic of recommendation on Reddit is a work from 2019 [71]. Here, the authors built two networks: a User–subreddit–User (UsU) network, where nodes are users, and edges exist if users have at least seven subreddits in common; and a Subreddit–user–Subreddit (SuS) network, where nodes are subreddits, and edges exist if subreddits have at least one user in common. The UsU network contained 2751 nodes and 845,128 edges, while the SuS network contained 847 nodes and 22,940 edges. Hence, the analysis ultimately focused on 2751 users and 847 subreddits. The features of each node were standard graph metrics: node degree (weighted and unweighted), degree centrality, closeness centrality (weighted and unweighted), betweenness centrality (weighted and unweighted), clustering coefficient (weighted and unweighted), HITS hub score [72,73] and PageRank score (weighted and unweighted) [73,74]. Additionally, in the node-to-node interaction, two more metrics were included: hop count and weighted distances between the corresponding nodes. Then, modularity-based community detection was used to discover 1965 communities in the UsU network and 292 communities in the SuS network. The community label was also added to the node feature set. To further augment the feature set, based on the network architecture, a node embedding method was employed. Here, Node2Vec [75] was used. Briefly speaking, Node2Vec is a graph network node embedding model based on the Word2Vec [76,77] concept of building embeddings of nodes (words) based on their context. The needed context was generated using random walks. Further, the paper proposed using content-based analysis to extend the features. It used keywords extracted from users' posts with the TF-IDF method [78].

Using all these features, a vector representing a user–subreddit relation was built. The negative examples (cases where a recommendation of a subreddit to a user was incorrect) were sampled from the set of users and subreddits which they do not belong to. This way, the dataset was augmented with negative cases. The evaluated models were logistic regression, a neural network, and a random forest classifier. The best results were achieved using all features with a random forest classifier with a 93% accuracy, a 93% precision and a 93% recall, a 93% F1-score and a 99% ROC AUC. The research concluded by highlighting the relevance of both network and content linguistic features when fusing different sources of information. The 2019 research is a particular inspiration for the contribution of this paper, where extending the graph network information with meta-data of posts (see Section 4) is explored.

Additionally, there appeared a meta study for the recommender systems, which raised an important issue of bias in the domain of recommendation data [79]. It stated that biases based on gender, ethnicity, race, etc., need to be avoided for ethical reasons. Therefore, in the design of this contribution, manual labeling and any external interference (such as dataset manipulation, subjective subreddit choice) that could introduce bias in the dataset was eliminated. Moreover, it can be claimed that none of the applied methods are known to introduce bias of the type listed above to the resulting models.

To summarize, while there are many works on hashtag, user, topic, and subreddit recommendation, the literature has very few examples of content-based small output space discovery methods. In this context, what follows uses a small output space method, named entity recognition, to build graph networks for the large-scale Reddit dataset. Moreover, the proposed evaluation methods are based on methods that have been used in similar works (reported above).

2.3. Use of Crossposts in Reddit Information Structure and Content Analysis

As noted earlier, one of Reddit's features that is almost non-existent in the literature are the so-called crossposts. Crossposts were introduced in 2017 (https://www.reddit.com/r/modnews/comments/7a5ubn/crossposting_coming_soon_to_your_subreddit, accessed on 13 February 2024). These are posts which appeared in one subreddit and were manually linked (crossposted) to a different one by a user. Since crossposts reference the original subreddit, they (1) show which posts (and named entities found in them) are seen by the readers of one subreddit as being of likely interest to the readers of another subreddit; and (2) establish a directional link between subreddits. We note that crossposts are the only user-generated content that explicitly indicates potential existence of shared interests between separate communities (the originating subreddit and the subreddit the post was crossposted to). In this context, authors of [80] suggested that crossposts may help to better understand the information structure of Reddit [14].

Obviously, knowledge brought by crossposts needs to be taken with caution. The fact that a user A believes that a post X from subreddit $S1$ would be of interest to the readers of subreddit $S2$ does not represent universal truth. It is possible that the user A is mistaken (or their crosspost is a result of a deliberate misinformation campaign). However, after manual inspection of a randomly picked sample of 200 crossposts, it was established that only less than 5% can be seen as potentially malicious or misguided. Hence, application of crossposts in analysis of subreddits is further explored in this contribution.

Nevertheless, it is important to note that since crossposts are scarce and may be accidental (rather than a result of in-depth analysis performed by the user), they should be seen more as a hint than hard evidence. So, their use in analysis of data should not apply standard effectiveness metrics, e.g., accuracy, precision, etc. In the literature, classification problems with an uncompleted response variable are often addressed with the *positive-unlabeled* (PU) approach [81], which offers two solutions: (1) assumption that that unlabeled cases are negative, or (2) estimation of the distribution of Y. However, in the case of crosspost-based analysis, missing are features that the PU needs. First, there are no negative samples (i.e., topics that are definitely not interesting to the two subreddit communities). Second, the positives (i.e., topics interesting to the two groups) are captured too rarely. Moreover, estimators of Y are not known to exist (and cannot be expected to be established in the future). Therefore, instead of PU-evaluation, the use of *recall* and *AUC* was selected. Here, we notice that recall rewards capturing the entities that actually appeared in crossposts, while AUC counters the bias in using recall alone. Here, let us recall that this approach was also used for ranking evaluation in similar problems described in Section 2.2, e.g., subreddit recommendation [82,83], hashtag recommendation [64,68,84–86], or in the case of user recommendation systems [63]. There, quality of rankings was evaluated using hit-rate (i.e., recall@5, or recall@10) and ROC AUC. Taking into account the size of the available dataset, it was decided to apply *recall@10* and *ROC AUC* to estimate how well the entities detected with the proposed method as being common to two subreddits fit with the entities found in the crossposts.

3. Dataset Preparation

Let us now describe the dataset that was used in the performed experiments. Here, we recall that Reddit is separated into topical subfora (subreddits). Each subreddit consists of posts that, in most cases, are moderated. Posts include a title (always) and a body, i.e., text, and/or multimedia. Text-to-media ratio varies between subreddits, but each Reddit post must have at least a text title.

3.1. Collection and Filtering of the Dataset

The initial dataset was collected via Pushshift API [87], with custom scripts realizing the preprocessing pipeline. It contained all posts from 3000 most popular subreddits by subscriber count, spanning the whole year 2022. However, preliminary data check indicated that very few crossposts exist between these 3000 subreddits. Therefore, the dataset was extended with posts from additional 250 subreddits, which contained most crossposts from/to the initially selected 3000.

Next, the following categories of inadequate subreddits were filtered. (1) Administrative subreddits (e.g., r/announcements—with posts by Reddit administrators, or r/help—a technical support subreddit). (2) Image-only subreddits (e.g., r/aww) that cannot be analyzed directly using NLP. (3) Subreddits that impose a specific post structure (e.g., r/copypasta, or r/hmm), which require specialized tools for entity recognition. (4) Subreddits, in which NER run into technical issues (e.g., r/meirl, r/meirlgbt and similar). For example, all posts in r/meirl are forced to have "meirl" as the post title, negatively influencing results of NER analysis, because the NER model is only fed strings saying "meirl" and mistakenly considers it as a named entity (the only named entity in this subreddit).

In addition to subreddit-level cleaning, pruning was applied to individual posts. Here, it was noticed that multiple posts did not capture the attention of other users. These posts ere filtered out based on their *score*. The *score* is the Reddit's appreciation mechanism. A user can give a post an upvote (+1) or a downvote (-1). Here, we note that all posts are upvoted by default (the starting *score* is 1). The resulting *score* is the sum of upvotes and downvotes. Interestingly, it very rarely happens that a post accumulates a negative *score*. As a matter of fact, such cases were absent in the collected dataset. However, the number of upvotes and downvotes of a given post is not known. Hence, it is possible that a highly controversial post with a lot of upvotes and downvotes ended up as being *score*-neutral. Here, it was assumed that posts with a low *score* can be omitted. In view of the above, this decision may seem as somewhat controversial, but it was based on the observation that the probability that the number of upvotes and downvotes is almost equal is relatively low. We note that, obviously, there is no universal threshold of when a post becomes significant. However, typically, post popularity fits the Internet "1% rule" [88]. Therefore, in reported work, only the top 20% of most popular posts were retained. This reduced the size of the dataset, allowing faster processing. Using all (more) posts is one of potential future research directions. This is appealing also since it takes care of the potential of upvotes and downvotes cancelling out. Nevertheless, preliminary explorations suggested that the use of all posts does not bring significantly different results while substantially increasing processing time.

Overall, after filtering, the dataset contained 32,203,763 posts from 3189 subreddits (median of 341,243 subscribers per subreddit), with a median of 3875 posts per subreddit. The posts had medians of 192 *score*, 14.5 comments, and 0.5 named entities per posts. We note that, to the best of our knowledge, no research focused on various subreddits used such a large Reddit dataset (see also [82]). Finally, about 650,000 posts (2.3% of all posts) were crossposted.

3.2. Named Entity Recognition

As noted, NER is prominent in data preprocessing when similarity between texts is to be established. Here, the content of posts was processed with two NER models: dslim/bert-base-NER (https://huggingface.co/dslim/bert-base-NER, accessed on 13 February 2024) (pre-trained BERT large models fine-tuned on the CoNLL-2003 dataset) and flair/ner-english-large (https://huggingface.co/flair/ner-english-large, accessed on 13 February 2024) (an XLM-RoBERTa [89] pre-trained on a cleaned Common Crawl dataset). Named entities extracted by both models were deduplicated. As a result, 15,308,754 named entities were identified. This set was used in this study.

4. The Proposed Approach

Let us now describe the method, which is the core contribution of the reported research. For easier understanding, Figure 1 describes the process step by step, starting with already described data acquisition up to the similarity calculation.

To offer a better understanding of the approach, the following simplified pseudoalgorithm summarizes it as well:

1. Acquire posts from two subreddits.
2. Preprocess the posts.
    2.1. Clean deleted/removed posts;
    2.2. Filter meaningful posts based on their *score*;

   2.3. Extract and store the required metadata for each post (*score*, *number of comments*; *total awards received*, does *it contain an image*).

3. Extract named entities from posts using NLP models and deduplicate them, if needed.
4. Create a graph network using named entities for each of the subreddits; here,
   4.1. Named entities become nodes;
   4.2. An edge joins two named entities if they appear in a post together. Multi-edges (when the entities appear jointly in multiple posts) are merged and (for each metrics) their weight becomes the sum of *scores* originating from posts where the two named entities appear jointly.
5. Calculate graph network characteristics for the two subreddit graphs separately and assign to each node (degree, clustering, average neighbor degree).
6. Filter out named entities (nodes and their edges) that appear only in a single subreddit from the pair (from here on, only subgraphs with named entities common to both subreddits are considered).
7. For each measure (metadata-based and network characteristics) independently, apply the following:
   7.1. For each (common) named entity,
      7.1.1. Calculate the similarity between the measures for the named entities in the two subgraphs.
   7.2. Create a ranking of named entities based on their similarity;
   7.3. Obtain the top $N = 10$ entries (named entities).

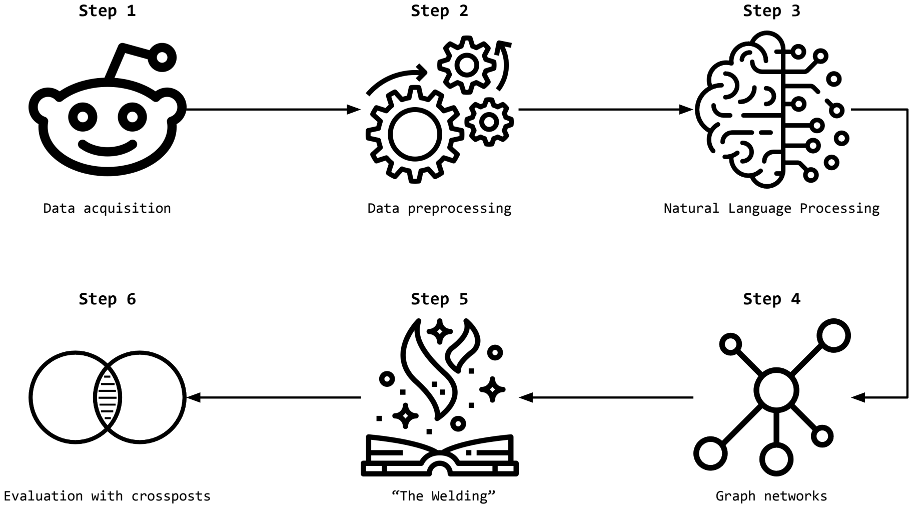

Figure 1. A scheme of the proposed method.

The final result for two subgraphs representing two subreddits is a set of rankings (consisting of 10 named entities each) obtained independently for each of the metadata-based and network characteristics-based similarities. For better understanding, a step-by-step example is provided in Appendix A to this work.

Let us now provide more details of each step and describe the method fully.

The dataset described above was used to create a graph network for each subreddit. As noted, thus far, graph networks have been created with users or subreddits as nodes. As one of the contributions of this work, graph networks were built for each subreddit, with named entities as nodes. Hence, rather than focusing on relations between users and/or communities (as considered in the past), *the spotlight was shifted to the content of individual subreddits*.

The first step of the approach consists of extracting and preprocessing all posts, as described in Section 3.

In the second step, a post is assigned aggregated metadata. Specifically, the used metadata were (i) *score*—the sum of the number of upvotes reduced by number of downvotes, for each post containing a given entity; (ii) *number of comments*; (iii) *number of awards received*—awards are a special, paid badges that users can assign to other user posts (https://www.reddit.com/r/awards, accessed on 13 February 2024); and (iv) *does a post contain an images*—number of posts that, in addition to the entity under consideration, contained also an image (in any format).

It should be noted that the following metadata were also tried: *is_meta* (mean) (https://reddit.com/r/NoStupidQuestions/comments/b9xe4d/what_exactly_is_a_meta_post, accessed on 13 February 2024), *is_original_content* (mean) (https://www.reddit.com/r/OutOfTheLoop/comments/1vlegj/what_does_oc_mean, accessed on 13 February 2024), *is_video* (mean), *over_18* (mean), *spoiler* (mean), *upvote_ratio* (mean). However, they did not visibly influence the results, most likely due to them being represented as boolean values (except the upvoe_ratio). We interpret this fact as follows. The boolean information was not enough for any tried similarity measure to influence the similarity. The zero or one values can either be equal or not, there is no more nuanced similarity that could be represented, as it is in case of, e.g., floating point numbers. However, joint use of all the metadata, together with all available posts, may be explored in the future.

The next step is named entity extractions with dslim/bert-base-NER (https://huggingface.co/dslim/bert-base-NER, accessed on 13 February 2024) (pre-trained BERT large models fine-tuned on the CoNLL-2003 [44] dataset), which was already covered in Sections 2.1. Extracted entities are deduplicated. The result of this step is a set of named entities extracted from posts.

The following step is the creation of graphs representing each subreddit. Here, an edge connects a pair of nodes (recognized named entities if they appear in at least one post together. Multi-edges are combined into single edges and their *edge weight* equals to the sum of *scores* of posts in which the connected entities appear. The result is a graph representing the structure of the information content of a given subreddit. It includes all recognized named entities and their features encoded in nodes, edges and edge weights.

All individual subreddit graphs are available in a Zenodo repository (https://zenodo.org/record/8037573, accessed on 13 February 2024). Table 1 summarizes the basic characteristics of obtained graphs. It also illustrates their scale. What is important to note is the large heterogeneity of obtained graphs. This is clearly visible by the differences between the means and the medians of, for instance, the node count (1919 vs. 918) and the edge count (2397 vs. 495).

Furthermore, the graphs have nonuniform distribution of periphery sizes [90]. In 47% of graphs, 10% of the nodes contribute to a periphery. On the other hand, in 25% of graphs, over 50% of nodes belong to a periphery. A similar situation emerges for isolates (nodes with zero degree). A total of 40% of graphs have over 50% of isolated nodes. This is consistent with graph density [91] being, on average, barely 0.002. To further highlight the "1% rule" of the graph structure, it is worth mentioning that in 74% of graphs, the average shortest path is less than three, meaning that, on average, any node is connected with any other node with a distance of two.

Table 1. Network statistics aggregated.

Statistic	Mean	Median
named entity count	5113.14	1727
unique named entity count	1888.92	918.5
node count	1919.44	940.5
edge count	2397.38	492
degree	1.32	1.08
degree centrality	0.004	0.001

Extracting Common Entities from Graphs

After the network graphs are created for each subreddit, the key stage of the process ensues. Its goal is to find the most similar named entities for each subreddit *pair*. Let us now present, in some detail, how the proposed method works for a pair of subreddits A and B.

First, only the named entities, common to A and B, are considered, but their characteristics (e.g., degree in a graph) are calculated in the context of all nodes. Named entities that appear in either A or B but not in both are discarded. All common entities from A and B already have their metadata (e.g., the aggregated *score* of posts where each entity appeared and their network characteristics (e.g., the aggregated degree of the node representing an entity in the graph of subreddit A and a separate value for subreddit B) calculated. The aggregation methods for each similarity measure are described and justified in Section 2.3.

The metadata and network characteristics are used to calculate the similarity measures of an entity, independently for each of the considered similarity measures (for the subreddit A and B pair). We note that since no single similarity measure was determined to be best (see Section 5), the comparison process involved each considered similarity measure (i.e., separate ranking for *score*, separate ranking for degree, etc.). Obviously, an attempt to create a combined similarity measure could have been undertaken (e.g., following the discussion found in [92]). However, this is outside the scope of this contribution.

What is important is to note that subreddits and their graphs differ in size. As specified above, a subreddit graph may contain hundreds of times more nodes than another one. As a result, the same named entity may have, for instance, a very large degree in the graph of subreddit A and a small one in B. To deal with this issue, in all comparisons, all metadata metrics and all network characteristics were normalized.

To select the best metadata and network characteristics as well as the similarity measures, the rankings were evaluated with appropriate data science metrics (recall and AUC as further described in Section 2.3).

The final result for subreddit pair A and B is a set of lists of their common named entities, ranked according to each considered similarity measure. Since five metadata-based and three network characteristics-based measures were considered, a total of eight separate rankings was obtained for each pair of subreddits. Although the resulting rankings were often unanimous, they do not have to be. Therefore, instead of choosing only a single similarity measure (metadata- or network characteristics-based), multiple metrics were calculated, as each of them may provide insightful information.

5. Experimental Results and Their Analysis

The outcomes of applying the proposed method to the collected dataset are now discussed.

Let us start from describing the evaluation process. For each pair of subreddits, named entities from posts and named entities from crossposts are considered separately. We note that named entities that were extracted from crossposts from subreddit A to subreddit B and from B to A are combined into one set (regardless of the direction of a crosspost). In other words, named entities from crossposts are treated as being present in both subreddits. This can be understood in the following way. If a post with entities $NE1$, $NE2$ and $NE3$ was crossposted from subreddit A to subreddit B, then these three named entities materialized in the set of named entities in crossposts in subreddit B.

We recall that for each individual similarity measure, named entities (from posts appearing in pairs of subreddits) are compared and ordered according to a given similarity measure (rankings resulting from the application of the proposed approach). Next, the set of named entities from posts and the set of named entities from crossposts (only) are compared and evaluated using metrics applied in similar problems (as described in Section 2.2)—recall@10 and AUC. Specifically, recall@10 and AUC metrics are calculated between the ranking of similarity of named entities originating from posts and crossposts.

Recall is calculated using the following formula: $|X \cap Y|/|X|$, where X is the set of named entities found in crossposts between two considered subreddits, Y is the set of named entities found in the two considered subreddits using the proposed method, and $|X|$ is the size of the set X. AUC is calculated with *roc_auc_score* from the sklearn library [93] based on the overlap of named entities found in crossposts and those selected from all posts in the two considered subreddits.

The best results, in terms of the recall and AUC, are achieved for summing the metadata (e.g., summing the aggregated *score* of an entity in subreddits A with B) and calculating the negative absolute value of network entities (e.g., calculating $-|degree\ of\ entity\ in\ graph\ of\ subreddit\ A - degree\ of\ entity\ in\ graph\ of\ subreddit\ B|$).

The aggregated results for all pairs of subreddits are presented in Table 2. Individual results for each subreddit pair can be found in the Zenodo repository (https://zenodo.org/record/8037573, accessed on 13 February 2024). The main observation is that metadata metrics achieved better recall, while network characteristic provided better AUC results.

Table 2. Recall@10 and AUC results for different metadata and network characteristics. The "Isolates" column shows results for graphs with many isolate nodes (over 50% of the network nodes are isolates, i.e., have no neighbors). Periphery refers to results where the periphery was large (over 50%).

	Recall	Recall	Recall	AUC	AUC	AUC
	All	Isolates	Periphery	All	Isolates	Periphery
average_neighbor_degree	0.20	0.31	0.67	0.57	0.52	0.56
clustering	0.13	0.25	0.62	0.50	0.50	0.51
degree	0.17	0.30	0.67	0.52	0.52	0.56
is_original_content	0.18	0.26	0.60	0.46	0.49	0.48
has_image	0.28	0.31	0.62	0.20	0.19	0.32
num_comments	0.35	0.43	0.73	0.24	0.28	0.46
score	0.38	0.47	0.76	0.26	0.31	0.50
total_awards_received	0.35	0.41	0.66	0.24	0.29	0.43

To evaluate the results, a reference point is needed. Although there is no direct comparison, previous works about recommendation systems use similar metrics (recall@10 and AUC) [63,64,68–70,79,82,85,94,95]. These works are discussed in Section 2.2 and consider the problems such as hashtag recommendation, subreddit recommendation, and user recommendation. While these problems are different to the task studied in this contribution, the top results, from studies of similar scale datasets, range from 25 to 60% for the recall and 60 to 90% for the AUC. While concerning different problems, they provide some indication as to what could be reasonably expected performance measures when dealing with Reddit data. In Table 2, the best recall results are achieved for the metadata: *score* (38%), *number of comments* (35%) and *total awards received* (35%). Moreover, the best AUC results achieved for the network characteristics are as follows: degree centrality (60%), degree (51%) and clustering (50%). Therefore, the best recall and AUC results are in the middle of the range found in other, similar studies. However, we note also that none of the cited works dealt with a dataset with tens of thousands of pairs of compared subreddits. Separately, the question as to what should be reasonably expected when measuring recall for the node-based metadata and the AUC for the network-based metadata requires further investigation. However, this is out of scope of this contribution.

Since there was no single best similarity metric, additional attempts at combining the rankings to achieve better results were completed. Taking the top results from two separate rankings (for example, top five *score*-based entities and top five degree-based entities) did not yield better results. In most cases, the obtained results were simply an average of two rankings, which makes this worse than accepting the better ranking of the two.

In the next steps, the obtained results were further explored. First, a working hypothesis was posed that topically close subreddits, like subreddits about gaming (e.g., like r/gaming, r/esports, r/Games, r/pcgaming), or politics and news (e.g., r/politics, r/news, r/worldnews), have better recall and AUC scores. Although there are solitary cases described in Section Exploration of Particular Subreddit Pairs, no major noticeable correlation between the quality of results and subreddit topical areas was observed.

Second, it was postulated that subreddits with similar named entity network structures (measured with network characteristics) may deliver better quality of results. This assumption was correct. Networks with larger peripheries (measured as the set of nodes with eccentricity [96] equal to the diameter) achieved recall values that were better by approximately 30–60 percentage points. Similarly, networks with a large number of isolates (nodes with a degree equal to zero) achieved results approximately eight percentage points higher than the mean. Upon further reflection, this phenomenon is easy to explain. In most cases, networks with a large periphery or with many isolates happen to also have a low number of high-degree nodes. These high-degree nodes are the focal points of the network (i.e., of discussion). Hence, they achieve a higher *score*, *greater numbers of comments* and higher values of other metadata. Moreover, they are very likely to appear in similar network configurations in subreddit named entity networks. This means that they are more often chosen as the results by both the metadata and the network characteristics rankings. Moreover, they are also likely to be crossposted due to their general popularity. Finally, this means that the proposed method works more reliably on non-complex networks, i.e., networks with the high periphery and/or a large number of isolates. As a by-product, this presents an interesting case of the "1% rule" for the Internet networks previously shown in other studies [88,97].

Many other approaches to combining metadata and network characteristics were tested, e.g., using the percent difference with the *score* or the sum with the degree of centrality, etc. (we note that there are almost infinitely many ways to calculate a similarity between two characteristics (see, for instance, [98])). Moreover, multiple popular graph metrics from the previous study [71] were also attempted, i.e., pagerank [99], voterank, closeness centrality [100], betweenness centrality [101], current flow closeness centrality [102], current flow betweenness centrality [102,103]. Node2Vec embeddings with different hyperparameters (p: 1, 2; q: 1, 2; walk length: 10, 100; number of walks: 10, 100, vector dimensions: 16, 32, 64) were also checked. However, all of them failed to obtain results of at least 10% recall@10 or at least 20% AUC, so they were omitted from this discussion. Overall, the final best results in terms of recall and AUC were achieved for formulas introduced in Section 5. These are *degree*, *degree centrality*, *clustering* [51] and *average neighbor degree* [50]. Nevertheless, this issue is not resolved, and further investigation is planned.

Exploration of Particular Subreddit Pairs

Let us now briefly share the most interesting results obtained in the cases where the algorithm reached the highest recall and AUC. These results are summarized in Table 3. There, when applicable, metrics values are given in parentheses, with the following demarcation: $r.$ = *recall*@10 and $a.$ = AUC.

Table 3. Selected similarities between subreddits with the highest recall and AUC.

Subreddit Pair	Entities	Recall [%]	AUC [%]	Metadata/Network Characteristic
r/SteamDeck, r/totalwar	Steam Deck, Total War	100	76	score
r/SteamDeck, r/totalwar	Steam Deck, Total War	100	80	# comments
r/manga, r/Meika	Meika	81	78	degree
r/manga, r/Meika	Meika	82	100	avg. neighbor degree
r/PornStars, r/KristyBlack	Kristy Black	89	80	total awards received
r/MovieDetails, r/UnexpectedMulaney	John Mulaney	63	98	degree
r/hqcelebvideos, r/EmmaWatson	Emma Watson	86	82	deg. centrality
r/nvidia, r/AyyMD	AMD, EVGA, NVIDIA GeForce, RTX	83	73	score, # comments, has image
r/carporn, r/NASCAR	Circuit Zolder, Death Valley, NASCAR Camaro, Suzuki, XB Falcon	67	67	deg. centrality
r/German, r/germany	German, Deutsch	100	100	score, # comments, total awards received, has image
r/Borderlands2, r/Borderands3	Borderlands	100	100	score, # comments, total awards received, has image
r/graphic_design, r/technology	Adobe	100	100	# comments
r/onions, r/ethereum	Tor	100	100	score, # comments, total awards received
r/frugalmalefashion, r/eagles	Nike	100	100	score, # comments, total awards received,
r/Piracy, r/vpnnetwork	Blackfriday VPN, Disney Plus, Internet Service Provider	71	92	deg. centrality
r/vpnnetwork, r/fantasybball	NBA, Firefox, Android	100	100	score, degree, deg. centrality, avg. neighbor degree
r/Kanye, r/IAmA	Auschwitz, Holocaust Survivor	100	63	score
r/Kanye, r/IAmA	Auschwitz, Holocaust Survivor	100	69	# comments
r/AITA, r/justdependathings	AITA	100	100	score, # comments, avg. neighbor degree
r/AITA, r/weddingshaming	AITA	63	98	avg. neighbor degree
r/AITA, r/houseplants	AITA	100	100	score, # comments, avg. neighbor degree

Table 3. *Cont.*

Subreddit Pair	Entities	Recall	AUC	Metadata/Network Characteristic
r/crappyoffbrands, r/AwesomeOffBrands	China	86	88	deg. centrality
R/UNBGBBIIVC-HIDCTIICBG, r/lingling40hrs	Electric Harp, Kiki Bello, Van Halen	66	91	avg. neighbor degree
r/freefromwork, r/wholesomememes	Japan	100	100	score, # comments, total awards received

Let us now summarize the most interesting findings. First, many similarities occur when one of the subreddits has a strictly narrower or wider topic, and it is this topic (named entity) that is the common one for the two subreddits. For example, r/SteamDeck (gaming console) and r/totalwar (video game) share entities Steam Deck and Total War (*score*: r. 100%, a. 76%, *number of comments*: r. 100%, a. 80%). Between r/manga and r/Meika (a Manga character) it is Meika that is the most common (degree: r. 81%, a. 78%, average neighbor degree,: r.82%, a. 100%). Between r/PornStars and r/KristyBlack (a porn actress) the entity is Kristy Black (*total awards received*: r. 89%, a. 80%); for r/MovieDetails and r/UnexpectedMulaney (subreddit about references to John Mulaney), John Mulaney is the most common topic (degree, r.63%, a. 98%); r/hqcelebvideos (subreddit about celebrities) and r/EmmaWatson share Emma Watson as the main similarity (degree centrality r. 86% a. 82%).

Second, there exist groups of subreddits that intuitively could have been expected to be similar or have common grounds. For example, r/nvidia and r/AyyMD are both about digital computing companies and the subreddits share many models of GPUs, such as AMD, EVGA, NVIDIA GeForce, RTX (*score*, *number of comments* and has image all have recall over 83% and AUC over 73%). Another example is r/carporn (related to beautiful cars) and r/NASCAR, where the common grounds are names of races and cars (e.g., Circuit Zolder, Death Valley, NASCAR Camaro, Suzuki, XB Falcon) (degree centrality r. 67%, a. 67%). German (language) subreddit r/German and Germany (country) r/germany share German and Deutsch (*score*, number of comment, *total awards received*, has image—all r. 100% and a. 100%). R/Borderlands2 and r/Borderands3 (both related to the video game) share the name of the game as the most common topic (*score*, *number of comments*, *total awards received*, has image—all have r. 100%, a. 100%). r/graphic_design and r/technology share Adobe (the company creating software for a.o. graphic design) (*number of comments*: r. 100%, a. 100%). R/onions (subreddit about anonymous access to the Internet) and r/ethereum (cryptocurrency) share Tor, software for anonymous Internet browsing (*score*, *number of comments*, *total awards received*—all r. 100% and a. 100%). R/frugalmalefashion (fashion subreddit) and r/eagles (sports team) share Nike (the sport fashion company) as the main common interest (*score*, *number of comments*, *total awards received*—all r. 100% and a. 100%).

R/Piracy and r/vpnnetwork are both mostly interested in Blackfriday VPN, Disney Plus, Internet Service Provider (ISP), which are the intuitive common topics between the two (degree centrality r.71% a.92%). Further, r/vpnnetwork and r/fantasybball (Fantasy Basketball) share the NBA league, Firefox and Android (*score*, *number of comments*, degree, degree centrality, average neighbor degree—all r. 100% and a. 100%).

There are also similarities based on Internet scandals (also called dramas). For example, r/Kanye and r/IAmA ("where the mundane becomes fascinating and the outrageous suddenly seems normal."). Here, the middle ground is, among others, Auschwitz and Holocaust Survivor, which reflects an Internet scandal from 2022 (https://www.washingtonpost.com/history/2022/12/02/hitler-kanye-west-black-germans-holocaust/, https://www.

ajc.org/news/5-of-kanye-wests-antisemitic-remarks-explained, accessed on 13 February 2024).

Looking from yet another angle, some rather unintuitive similarities between subreddits were found. For example, r/crappyoffbrands and r/AwesomeOffBrands, which discuss bad and great offbrands, both share China as the main similarity (degree centrality r. 86%, a. 88%). R/UNBGBBIIVCHIDCTIICBG (subreddit about engaging videos) and r/lingling40hrs (subreddit about string instruments) share Electric Harp, Kiki Bello (harpist) and Van Halen (guitarist). This seems intuitive that these are the similar-named entities, but it is not intuitive that these subreddits share any topic whatsoever.

Separately, it is relatively easy to realize that this work has potential to help other research to expand its scope. For instance, there were studies [104,105] on social norms based on the r/AITA subreddit. What is interesting is that the acronym AITA ("Am I The Asshole?") is the main similarity between r/AITA and r/justdependathings (*score*, *number of comments*, average neighbor degree—all r. 100% and a. 100%) and also between r/AITA and r/weddingshaming (average neighbor degree: r. 63%, a. 98%), and r/AITA and r/houseplants (*score*, *number of comments*, average neighbor degree—all r. 100% and a. 100%). In this way, two additional communities could have been worth looking into when studying social norms. Moreover, there was a master thesis concerning The Legend of Korra [106] which analyzed subreddit r/TheLastAirBender. Using the introduced tool, benefit could be obtained from using r/ecchi and r/AndavaArt which are two subreddits where The Legend of Korra is one of the main similarities (average neighbor degree—all r. 66% and a. 80%). Another example is a study on subreddit r/NBA [107]. As it appears, NBA is the main similarity between subreddits r/vpnnetwork and r/fantasybball, the latter being dedicated to fantasy basketball, which also aggregates fans of the league. Finally, various studies on the 4chan Reddit community (r/4chan) [108–110] could also look into the subreddit r/justneckbeardthings, because r/4chan and r/justneckbeardthings share 4chan and Anon as the main similarities (*score*, *number of comments*, *total awards received*—all r. 100% and a. 100%).

Finally, a general observation visible in the results is that, most often, metadata metrics are consistent with each other, i.e., when *score* achieves high recall, *number of comments* and *total awards received*, achieve high results, too. For example, this is the case for r/German and r/germany, r/Borderlands2 and r/Borderands3, r/onions and r/ethereum, r/freefromwork and r/wholesomememes (see results in Table 2). A similar situation occurs with network characteristics, e.g., for r/vpnnetwork and r/fantasybball, r/AITA and r/justdependathings, r/manga and r/Meika.

On the other hand, there is no correlation between the node and network characteristics. As a matter of fact, it is possible that one group of metrics may report high similarity measures, while the other very low ones. This observation will be investigated in future research.

Other potentially interesting findings are available in the Zenodo repository (https://zenodo.org/record/8037573, accessed on 13 February 2024).

6. Concluding Remarks

The aim of this contribution was twofold. First, our goal was to apply named entity recognition and graph networks to propose a method for studying the structure of information content of subreddits. The proposed method treats subreddits as individual structures and establishes similarities between them. Second, we aimed to explore the potential of crossposts as (1) natural indicators of topics shared by subreddits and (2) as a supporting measure of success of the proposed method. Application of the proposed approach allowed for to capture a number of expected similarities between subreddits. Moreover, some unexpected relations were also found. Finally, when using crossposts, it was shown that the proposed method achieved performance metrics that match those reported for similar problems. These results combined deliver a basic level of confidence in the proposed approach.

Since the obtained results are promising, further explorations are planned. Here, a few potential examples based on use of the already developed tool set are presented. (1) In the text, some suggestions are made as to the ways to make the results more comprehensive. This can include, among other methods, extending the dataset by one more year, e.g., 2021 (or two more years, including 2020). We note also that use of 2023 data has to wait till after the end of the year for the posting and crossposting to subside. Adding data would also allow following evolution of the information content structure for the time frame of 2020–2022. (2) Taking into account that the strength of connections is already measured, it would be possible to start from considering almost all connections and then systematically increase the threshold of keeping connections. In this way, connection strength-based evolution of the network graphs can be explored. (3) Reported work involved only textual data. Since a large portion of Reddit content is multimedia (mostly as images and videos or short clips), there is space for extension of the proposed approach with multi-modal image processing. This, however, requires substantially more work.

Finally, future work may involve the generation of crosspostable content. Though this is just a weak hypothesis (outside the scope of this work), the proposed approach could help to generate crosspostable content, i.e., posts of interest to readers of multiple precisely identified subreddits. Here, our solution could help to generate traffic/coverage desired by content creators and bridge gaps between information bubbles [111] and/or reach out to echo chambers [112].

Author Contributions: Conceptualization, J.S.; methodology, J.S.; software, J.S.; validation, M.P. and M.G.; formal analysis, J.S.; investigation, J.S.; resources, M.G.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S., M.P., M.G. and Y.W.; visualization, J.S.; supervision, M.P., M.G. and Y.W.; project administration, M.P.; funding acquisition, M.G. and M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: This research was carried out with the support of the Laboratory of Bioinformatics and Computational Genomics and the High Performance Computing Center of the Faculty of Mathematics and Information Science Warsaw University of Technology.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Step-by-Step Example

To better understand the proposed approach, let us consider a detailed example. This example uses artificial data to highlight key aspects of the proposed method. Moreover, to simplify the description, only *score* and the *node degree* are considered.

Let us assume that there are two subreddits: r/technology (T) and r/programming (P) with posts represented in Tables A1 and A2. Moreover, the only available metadata are the post's *score*. For readability, the named entities are marked in italic.

Table A1. Artificial posts from subreddits r/technology used in a step-by-step example.

ID	Title	Text	Score
T1	**Microsoft** releases a new library in **Python**!	The tech company is said to release an open source library in **Python** solely for **AWS** and **Azure** integration.	7
T2	Social media—what do people use?	I have been having problems with **Facebook** recently. What other social media (apart from **Reddit**) do you use? Is **Instagram** a good alternative?	11
T3	I have been waiting for 3 months! FOR A WATCH?!	Apparently, the **COVID-19** caused a freeze in microchips production. All smartwatch shipment to **Canada** have been put on hold. I guess I'll never see my new watch. . .	31
T4	What are your thoughts on the new **Python Reddit** API?	Good/bad?	3
T5	Have you tried using **Raspberry Pi Python** library?	Anyone has docs for that?	1

Table A2. Artificial posts from subreddits r/programming used in a step-by-step example.

ID	Title	Text	Score
P1	**Python** or **R**?	Let's settle the debate once and for all: What language do you prefer for data science applications: **Python** or **R?**	13
P2	AMA: Ex **Twitter** engineer	I worked for **Twitter** for over 10 years. I have worked both with front-end **JavaScript**, but also a bit of backend with **Python** and **MySQL**. Ask me anything!	5
P3	I will store everything in my **AWS** from this point	The president of the **US** is going to institute access to high speed internet as a part of human rights. **Canada**, but also **UK** and 20 other European countries, already declare support of this idea. What do you guys think? Are we migrating to the cloud? Is **AWS** the right choice?	23
P4	[deleted]	[deleted]	2

During preprocessing, post *T*5 is deleted, since its *score* is lower than the threshold (2). The "[deleted]" tag means that post *P*4 was deleted by the user. Hence, it is also removed.

Next, named entities are extracted from the posts, with the results represented in Table A3.

Table A3. Named entities extracted from example posts.

Post	Named Entities
T1	Microsoft, Python, Azure, AWS
T2	Facebook, Instagram, Reddit
T3	COVID-19, Canada
T4	Python, Reddit
P1	Python, R
P2	Twitter, JavaScript, Python, MySQL
P3	AWS, US, Canada, UK

Then, using entities and their metadata, two graph networks are created. The entities are mapped into nodes, and the sum of *scores* of posts they appeared in is calculated.

Naturally, the node network characteristics are added to the node properties after the edges are instantiated. The network for subreddit r/technology is presented in Figure A1, while that for r/programming is in Figure A2.

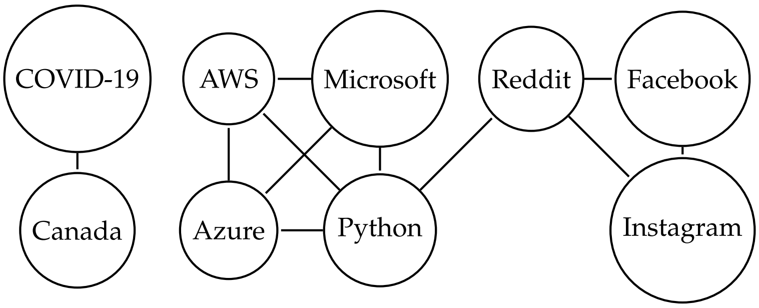

Figure A1. Network for r/technology.

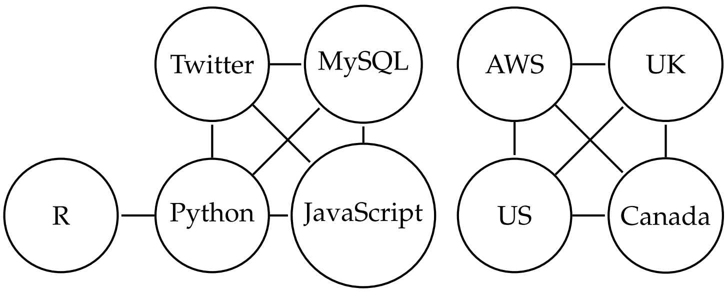

Figure A2. Network for r/programming.

Finally, graph commonality is established. Entities appearing in networks *Python*, *AWS* and *Canada* are retained, while the remaining ones are ignored. In this example, the only two properties are *node score* and *node degree*. All nodes have their (two) similarity metrics calculated. The results are min–max normalized. The final values of node *scores* and node degrees are shown in Table A4 for r/technology and Table A5 for r/programming.

Table A4. Named entities (node) metadata and network characteristics for r/technology.

Node	Python	AWS	Canada
Summed score	10	7	31
Normalized score	0.125	0	1
Degree	4	3	1
Normalized degree	1	0.(6)	0

Table A5. Named entities (nodes) metadata and network characteristics for r/programming.

Node	Python	AWS	Canada
Summed score	18	23	23
Normalized score	0.7(2)	1	0.58
Degree	4	3	3
Normalized degree	1	0	0

Next, the normalized values from both subreddits are compared by calculating their similarity (here, for the reasons outlined above, as a sum of *score* and a negative absolute difference of the *node degree*), with the results presented in Table A6. The final outcome are two rankings of common entities according to *score* and degree.

Table A6. Example similarities of metadata and network characteristics.

Node	Python	AWS	Canada
Score similarity	0.847(2)	1	1.58
Degree similarity	0.0	-0.(6)	0.0

While in the general case 10 top entities are used (see the next section), here, there are only three entities. Let us assume that only the top two results are of interest. These are Canada and AWS for *score* and Python and Canada for *degree measure*.

Let us now illustrate how this information can be confronted with that brought by the crossposts. Let us assume that the following post from r/technology was crossposted to r/programming:

Post TP1

Title: Python crowned as the most versatile language!

Text: You can do anything in *Python*! From backend development, through machine learning to embedded systems. See more in this report: (. . .).

Score: 13

Here, the only named entity that matches these captured in both subreddits is *Python*. However, degree-based ranking and *score*-based rankings may not agree. The node degree-based approach pointed to *Python* and *Canada* as the two top ranked common entities (as this includes *Python*). On the other hand, the *score*-based top two are *Canada* and *AWS*. This means that, from this perspective, *Python* is unimportant.

Obviously, if the top three (or more, in the general case) common entities were selected, the two approaches would be consistent in terms of results.

References

1. Soliman, A.; Hafer, J.; Lemmerich, F. A characterization of political communities on reddit. In Proceedings of the 30th ACM conference on hypertext and Social Media, Bavaria, Germany, 17–20 September 2019; pp. 259–263.
2. Bergstrom, K.; Poor, N. Reddit gaming communities during times of transition. *Soc. Media+ Soc.* **2021**, *7*, 20563051211010167. [CrossRef]
3. Marwick, A.E. Morally motivated networked harassment as normative reinforcement. *Soc. Media+ Soc.* **2021**, *7*, 20563051211021378. [CrossRef]
4. Park, A.; Conway, M.; Chen, A.T. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: A text mining and visualization approach. *Comput. Hum. Behav.* **2018**, *78*, 98–112. [CrossRef]
5. Yoo, M.; Lee, S.; Ha, T. Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit. *Inf. Process. Manag.* **2019**, *56*, 1565–1575. [CrossRef]
6. Yeskuatov, E.; Chua, S.L.; Foo, L.K. Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10347. [CrossRef] [PubMed]
7. Buntinx-Krieg, T.; Caravaglio, J.; Domozych, R.; Dellavalle, R.P. Dermatology on Reddit: Elucidating trends in dermatologic communications on the world wide web. *Dermatol. Online J.* **2017**, *23*, 7. [CrossRef]
8. Ammari, T.; Schoenebeck, S.; Romero, D.M. Pseudonymous parents: Comparing parenting roles and identities on the Mommit and Daddit subreddits. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QB, Canada, 21–27 April 2018; pp. 1–13.
9. Staudt Willet, K.B.; Carpenter, J.P. Teachers on Reddit? Exploring contributions and interactions in four teaching-related subreddits. *J. Res. Technol. Educ.* **2020**, *52*, 216–233. [CrossRef]
10. Bergstrom, K.; Poor, N. Signaling the Intent to Change Online Communities: A Case From a Reddit Gaming Community. *Soc. Media+ Soc.* **2022**, *8*, 20563051221096817. [CrossRef]
11. Botzer, N.; Ding, Y.; Weninger, T. Reddit entity linking dataset. *Inf. Process. Manag.* **2021**, *58*, 102479. [CrossRef]
12. Chevrier, N. Automating Hate: Exploring Toxic Reddit Norms with Google Perspective. Ph.D. Thesis, Université d'Ottawa/University of Ottawa, Ottawa, ON, Canada, 2022.
13. Dey, J. Topic Mining and Categorization in Online Discussion Forums. Ph.D. Thesis, University of Illinois, Urbana, IL, USA, 2020.
14. Proferes, N.; Jones, N.; Gilbert, S.; Fiesler, C.; Zimmer, M. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Soc. Media+ Soc.* **2021**, *7*, 20563051211019004. [CrossRef]
15. Sawicki, J.; Ganzha, M.; Paprzycki, M.; Badica, A. Exploring Usability of Reddit in Data Science and Knowledge Processing. *Scalable Comput. Pract. Exp.* **2022**, *23*, 9–22. [CrossRef]
16. Torfi, A.; Shirvani, R.A.; Keneshloo, Y.; Tavaf, N.; Fox, E.A. Natural language processing advancements by deep learning: A survey. *arXiv* **2020**, arXiv:2003.01200.

17. Zhao, H.; Phung, D.; Huynh, V.; Jin, Y.; Du, L.; Buntine, W. Topic modelling meets deep neural networks: A survey. *arXiv* **2021**, arXiv:2103.00498.
18. Alghamdi, R.; Alfalqi, K. A Survey of Topic Modeling in Text Mining. *Int. J. Adv. Comput. Sci. Appl.* **2015**, *6*, 1. [CrossRef]
19. Vayansky, I.; Kumar, S.A. A review of topic modeling methods. *Inf. Syst.* **2020**, *94*, 101582. [CrossRef]
20. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
21. Blei, D.M.; Lafferty, J.D. A correlated topic model of science. *Ann. Appl. Stat.* **2007**, *1*, 17–35. [CrossRef]
22. Li, W.; McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 577–584.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
25. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
26. Kedzierska, M.; Spytek, M.; Kurek, M.; Sawicki, J.; Ganzha, M.; Papryzcki, M. Topic modeling applied to the Reddit posts. In *Proceedings of the Big Data Analytics in Astronomy, Science, and Engineering*; University of Aizu: Aizuwakamatsu, Japan; National Institute of Technology Delhi and IIT: Delhi, India, 2023.
27. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
28. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]
29. Egger, R.; Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef] [PubMed]
30. Uncovska, M.; Freitag, B.; Meister, S.; Fehring, L. Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany. *NPJ Digit. Med.* **2023**, *6*, 115. [CrossRef]
31. Jeon, E.; Yoon, N.; Sohn, S.Y. Exploring new digital therapeutics technologies for psychiatric disorders using BERTopic and PatentSBERTa. *Technol. Forecast. Soc. Chang.* **2023**, *186*, 122130. [CrossRef]
32. Giarelis, N.; Mastrokostas, C.; Karacapilidis, N. Abstractive vs. Extractive Summarization: An Experimental Review. *Appl. Sci.* **2023**, *13*, 7620. [CrossRef]
33. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.; Gutiérrez, J.; Kochut, K. Text Summarization Techniques: A Brief Survey. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **2017**, *8*, 397–405. [CrossRef]
34. Dumais, S.T. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* **2004**, *38*, 189–230. [CrossRef]
35. Alomari, A.; Idris, N.; Sabri, A.Q.M.; Alsmadi, I. Deep reinforcement and transfer learning for abstractive text summarization: A review. *Comput. Speech Lang.* **2022**, *71*, 101276. [CrossRef]
36. Ma, T.; Pan, Q.; Rong, H.; Qian, Y.; Tian, Y.; Al-Nabhan, N. T-bertsum: Topic-aware text summarization based on bert. *IEEE Trans. Comput. Soc. Syst.* **2021**, *9*, 879–890. [CrossRef]
37. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 13–18 July 2020; pp. 11328–11339.
38. Liu, Y.; Liu, P. SimCLS: A simple framework for contrastive learning of abstractive summarization. *arXiv* **2021**, arXiv:2106.01890.
39. Qi, W.; Yan, Y.; Gong, Y.; Liu, D.; Duan, N.; Chen, J.; Zhang, R.; Zhou, M. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv* **2020**, arXiv:2001.04063.
40. Liu, Y.; Liu, P.; Radev, D.; Neubig, G. BRIO: Bringing Order to Abstractive Summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Volume 1: Long Papers; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 2890–2903. [CrossRef]
41. Grishman, R.; Sundheim, B. Message Understanding Conference- 6: A Brief History. In Proceedings of the COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, Vienna, Virginia, 6–8 May 1996.
42. Albared, M.; Ocaña, M.G.; Ghareb, A.; Al-Moslmi, T. Recent progress of named entity recognition over the most popular datasets. In Proceedings of the 2019 First International Conference of Intelligent Computing and Engineering (ICOICE), Hadhramout, Yemen, 15–16 December 2019; IEEE: New York, NY, USA, 2019; pp. 1–9.
43. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
44. Tjong Kim Sang, E.F.; De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Edmonton, AB, Canada, 31 May–1 June 2003; pp. 142–147.
45. Xuan, J.; Lu, J.; Zhang, G.; Luo, X. Topic model for graph mining. *IEEE Trans. Cybern.* **2015**, *45*, 2792–2803. [CrossRef] [PubMed]
46. Chen, H.; Luo, X. An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. *Adv. Eng. Informatics* **2019**, *42*, 100959. [CrossRef]
47. Bougouin, A.; Boudin, F.; Daille, B. Topicrank: Graph-based topic ranking for keyphrase extraction. In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Taipei, Taiwan, 20–23 November 2013; pp. 543–551.
48. Oliveira, I.L.; Fileto, R.; Speck, R.; Garcia, L.P.; Moussallem, D.; Lehmann, J. Towards holistic Entity Linking: Survey and directions. *Inf. Syst.* **2021**, *95*, 101624. [CrossRef]

49. Chakraborty, A.; Dutta, T.; Mondal, S.; Nath, A. Application of graph theory in social media. *Int. J. Comput. Sci. Eng.* **2018**, *6*, 722–729. [CrossRef]
50. Barrat, A.; Barthelemy, M.; Pastor-Satorras, R.; Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3747–3752. [CrossRef]
51. Costantini, G.; Perugini, M. Generalization of clustering coefficients to signed correlation networks. *PLoS ONE* **2014**, *9*, e88669. [CrossRef]
52. Thiéblin, E.; Haemmerlé, O.; Hernandez, N.; Trojahn, C. Survey on complex ontology matching. *Semant. Web* **2020**, *11*, 689–727. [CrossRef]
53. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Philip, S.Y. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Networks Learn. Syst.* **2021**, *33*, 494–514. [CrossRef]
54. Frisoni, G.; Moro, G.; Balzani, L. Text-to-Text Extraction and Verbalization of Biomedical Event Graphs. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 2692–2710.
55. He, Q.; Yang, J.; Shi, B. Constructing knowledge graph for social networks in a deep and holistic way. In Proceedings of the Companion Proceedings of the Web Conference, Singapore, 13–17 May 2020; pp. 307–308.
56. De Pril, R. *User Classification Based on Public Reddit Data*; Ghent University: Ghent, Belgium, 2019.
57. Garibay, I.; Oghaz, T.A.; Yousefi, N.; Mutlu, E.Ç.; Schiappa, M.; Scheinert, S.; Anagnostopoulos, G.C.; Bouwens, C.; Fiore, S.M.; Mantzaris, A.; et al. Deep agent: Studying the dynamics of information spread and evolution in social networks. In Proceedings of the Conference of the Computational Social Science Society of the Americas, Virtual Event, 8–11 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 153–169.
58. Kolomeets, M.; Chechulin, A.; Kotenko, I.V. Bot detection by friends graph in social networks. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.* **2021**, *12*, 141–159.
59. Datta, S.; Phelan, C.; Adar, E. Identifying Misaligned Inter-Group Links and Communities. *Proc. ACM Hum. Comput. Interact.* **2017**, *1*, 37. [CrossRef]
60. Aggarwal, A.; Gola, B.; Sankla, T. Data mining and analysis of reddit user data. In Proceedings of the Cybernetics, Cognition and Machine Learning Applications: Proceedings of ICCCMLA 2020, Goa, India, August 2020; pp. 211–219.
61. Cai, B.; Decker, S.; Zheng, C. The Migrants of Reddit: An Analysis of User Migration Effects of Subreddit Bans; Preprint. 2019. Available online: https://snap.stanford.edu/class/cs224w-2019/project/26424942.pdf (accessed on 1 February 2024).
62. Nadiri, A.; Takes, F.W. A large-scale temporal analysis of user lifespan durability on the Reddit social media platform. In Proceedings of the Companion Proceedings of the Web Conference, Lyon, France, 25–29 April 2022; pp. 677–685.
63. Pennacchiotti, M.; Gurumurthy, S. Investigating topic models for social media user recommendation. In Proceedings of the 20th international Conference Companion on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 101–102.
64. Alsini, A.; Datta, A.; Huynh, D.Q. On utilizing communities detected from social networks in hashtag recommendation. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 971–982. [CrossRef]
65. Palla, G.; Derényi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, *435*, 814–818. [CrossRef] [PubMed]
66. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
67. Raghavan, U.N.; Albert, R.; Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* **2007**, *76*, 036106. [CrossRef] [PubMed]
68. Kumar, N.; Baskaran, E.; Konjengbam, A.; Singh, M. Hashtag recommendation for short social media texts using word-embeddings and external knowledge. *Knowl. Inf. Syst.* **2021**, *63*, 175–198. [CrossRef]
69. Wu, Z.; Paul, A.; Cao, J.; Fang, L. Directional Adversarial Training for Robust Ownership-Based Recommendation System. *IEEE Access* **2022**, *10*, 2880–2894. [CrossRef]
70. Nguyen, H.; Richards, R.; Chan, C.C.; Liszka, K.J. RedTweet: Recommendation engine for reddit. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France, 25–28 August 2015; pp. 1381–1388.
71. Janchevski, A.; Gievska, S. A Study of Different Models for Subreddit Recommendation Based on User-Community Interaction. In Proceedings of the ICT Innovations 2019. Big Data Processing and Mining: 11th International Conference, ICT Innovations 2019, Ohrid, North Macedonia, 17–19 October 2019; Gievska, S., Madjarov, G., Eds.; Springer: Cham, Switzerland, 2019; pp. 96–108.
72. Kleinberg, J.M. Authoritative sources in a hyperlinked environment. *J. ACM* **1999**, *46*, 604–632. [CrossRef]
73. Langville, A.N.; Meyer, C.D. A survey of eigenvector methods for web information retrieval. *SIAM Rev.* **2005**, *47*, 135–161. [CrossRef]
74. Ma, N.; Guan, J.; Zhao, Y. Bringing PageRank to the citation analysis. *Inf. Process. Manag.* **2008**, *44*, 800–810. [CrossRef]
75. Grover, A.; Leskovec, J. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 855–864.
76. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
77. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.

78. Rajaraman, A.; Ullman, J.D. Data Mining. In *Mining of Massive Datasets*; Cambridge University Press: Cambridge, UK, 2011; pp. 1–17. [CrossRef]
79. Edizel, B.; Bonchi, F.; Hajian, S.; Panisson, A.; Tassa, T. FaiRecSys: Mitigating algorithmic bias in recommender systems. *Int. J. Data Sci. Anal.* **2020**, *9*, 197–213. [CrossRef]
80. Krohn, R.; Weninger, T. Subreddit Links Drive Community Creation and User Engagement on Reddit. In Proceedings of the International AAAI Conference on Web and Social Media, Virtual Event, 6–9 June 2022; Volume 16, pp. 536–547.
81. Elkan, C.; Noto, K. Learning classifiers from only positive and unlabeled data. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 213–220.
82. Das, A.K.; Bhat, N.; Guha, S.; Palan, J. A Personalized Subreddit Recommendation Engine. *arXiv* **2019**, arXiv:1905.01263.
83. Chandrasekharan, E.; Gandhi, C.; Mustelier, M.W.; Gilbert, E. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum. Comput. Interact.* **2019**, *3*, 1–30. [CrossRef]
84. Li, M.; Gan, T.; Liu, M.; Cheng, Z.; Yin, J.; Nie, L. Long-tail hashtag recommendation for micro-videos with graph convolutional network. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 509–518.
85. Li, Y.; Liu, T.; Hu, J.; Jiang, J. Topical co-attention networks for hashtag recommendation on microblogs. *Neurocomputing* **2019**, *331*, 356–365. [CrossRef]
86. Belhadi, A.; Djenouri, Y.; Lin, J.C.W.; Cano, A. A data-driven approach for Twitter hashtag recommendation. *IEEE Access* **2020**, *8*, 79182–79191. [CrossRef]
87. Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; Blackburn, J. The pushshift reddit dataset. In Proceedings of the International AAAI Conference on Web and Social Media, Virtual Event, 8–11 June 2020; Volume 14, pp. 830–839.
88. Carron-Arthur, B.; Cunningham, J.A.; Griffiths, K.M. Describing the distribution of engagement in an Internet support group by post frequency: A comparison of the 90-9-1 Principle and Zipf's Law. *Internet Interv.* **2014**, *1*, 165–168. [CrossRef]
89. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv* **2019**, arXiv:1911.02116.
90. Oellermann, O.; Swart, H. On the Steiner Periphery and Steiner Eccentricity of a Graph. In *Topics in Combinatorics and Graph Theory: Essays in Honour of Gerhard Ringel*; Springer: Berlin/Heidelberg, Germany, 1990; pp. 541–547.
91. Black, P.E. Dictionary of Algorithms and Data Structures. 1998. Available online: http://www.nist.gov/dads (accessed on 1 February 2024).
92. Szmeja, P.; Ganzha, M.; Paprzycki, M.; Pawłowski, W. Dimensions of semantic similarity. In *Advances in Data Analysis with Computational Intelligence Methods: Dedicated to Professor Jacek Żurada*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 87–125.
93. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
94. Eberhard, L.; Walk, S.; Posch, L.; Helic, D. Evaluating narrative-driven movie recommendations on Reddit. In Proceedings of the 24th International Conference on Intelligent User Interfaces, Los Angeles, CA, USA, 17–20 March 2019; pp. 1–11.
95. Zhang, M.; Wu, S.; Gao, M.; Jiang, X.; Xu, K.; Wang, L. Personalized graph neural networks with attention mechanism for session-aware recommendation. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 3946–3957. [CrossRef]
96. Hage, P.; Harary, F. Eccentricity and centrality in networks. *Soc. Netw.* **1995**, *17*, 57–63. [CrossRef]
97. Hargittai, E.; Walejko, G. The participation divide: Content creation and sharing in the digital age. *Inf. Community Soc.* **2008**, *11*, 239–256. [CrossRef]
98. Szmeja, P.; Ganzha, M.; Paprzycki, M.; Pawlowski, W. Dimensions of Semantic Similarity. In *Advances in Data Analysis with Computational Intelligence Methods*; Gaweda, A.E., Kacprzyk, J., Rutkowski, L., Yen, G.G., Eds.; Studies in Computational Intelligence; Springer: Berlin/Heidelberg, Germany, 2018; Volume 738, pp. 87–125.
99. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*; Technical Report; Stanford InfoLab: Road Stanford, CA, USA, 1999.
100. Freeman, L. Centrality in networks: I. conceptual clarifications. *Soc. Netw.* **1979**, *1*, 215–239. [CrossRef]
101. Freeman, L.C. A set of measures of centrality based on betweenness. *Sociometry* **1977**, *40*, 35–41. [CrossRef]
102. Brandes, U.; Fleischer, D. Centrality measures based on current flow. In Proceedings of the Annual Symposium on Theoretical Aspects of Computer Science, Stuttgart, Germany, 24–26 February 2005; Springer: Berlin/Heidelberg, Germany, 2005, pp. 533–544.
103. Zhang, J.X.; Chen, D.B.; Dong, Q.; Zhao, Z.D. Identifying a set of influential spreaders in complex networks. *Sci. Rep.* **2016**, *6*, 27823. [CrossRef]
104. Botzer, N.; Gu, S.; Weninger, T. Analysis of moral judgment on reddit. *IEEE Trans. Comput. Soc. Syst.* **2022**, *10*, 947–957. [CrossRef]
105. De Candia, S.; De Francisci Morales, G.; Monti, C.; Bonchi, F. Social Norms on Reddit: A Demographic Analysis. In Proceedings of the 14th ACM Web Science Conference 2022, Barcelona, Spain, 26–29 June 2022; pp. 139–147.
106. Caza, K. "The World Has Always Been Like a Comic Book World to Me": Examining Representations of Queer Stories in Comics and Other Media. Master's Thesis, University of Waterloo, Waterloo, ON, Canada, 2022.
107. Zhang, J.S.; Tan, C.; Lv, Q. "This is why we play" Characterizing Online Fan Communities of the NBA Teams. *Proc. ACM Hum. Comput. Interact.* **2018**, *2*, 1–25.

108. Horne, B.D.; Adali, S.; Sikdar, S. Identifying the social signals that drive online discussions: A case study of reddit communities. In Proceedings of the 2017 26th International Conference on Computer Communication and Networks (ICCCN), Vancouver, BC, Canada, 31 July–3 August 2017; IEEE: New York, NY, USA, 2017; pp. 1–9.
109. Mittos, A.; Zannettou, S.; Blackburn, J.; De Cristofaro, E. "And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan. In Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 8–11 June 2020; Volume 14, pp. 452–463.
110. Rieger, D.; Kümpel, A.S.; Wich, M.; Kiening, T.; Groh, G. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Soc. Media+ Soc.* **2021**, *7*, 20563051211052906.
111. Linder, R.; Stacy, A.M.; Lupfer, N.; Kerne, A.; Ragan, E.D. Pop the feed filter bubble: Making Reddit social media a VR cityscape. In Proceedings of the 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Tuebingen/Reutlingen, Germany, 18–22 March 2018; IEEE: New York, NY, USA, 2018; pp. 619–620.
112. Cinelli, M.; De Francisci Morales, G.; Galeazzi, A.; Quattrociocchi, W.; Starnini, M. The echo chamber effect on social media. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2023301118. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.