DSM160 Coursework 1: Social Media Dataset Essay

Submitted for partial fulfilment for the Social Media and Network Science course.

By

Hendrik Matthys van Rooyen

University of London

December 2023

CONTENTS

CHAPTER 1 ESSAY

1.1 INTRODUCTION

The "Social Network: Reddit Hyperlink Network" dataset, available on the Stanford University SNAP platform, encompasses a detailed, directed network defined by hyperlinks connecting various subreddits. This paper delves into the dataset's complexities, its origins, and the ethical aspects concerning its use. It also examines the potential impact of automated bots and misinformation on the dataset's integrity.

Additionally, the paper considers the dataset in the context of the "Four V's" of big data: volume, variety, velocity, and veracity, and discusses their implications. The essay also analyses the applications of network science in relation to this dataset. It investigates the possible interpretations of the networks, as well as what network structures might be beneficial to construct or examine in the context of the dataset.

This exploration not only enhances our understanding of the dataset but also opens avenues for further research, especially in the areas of data veracity and the influence of external factors like bots on data reliability. By dissecting the network structures, the paper aims to contribute to the broader field of network science, especially as it applies to complex, online social networks.

1.2 CHOOSING THE DATASET

In deciding upon a dataset to discuss, multiple data sources were explored, ranging from those made available on the popular Data Science platform Kaggle, the Humanitarian Data Exchange datasets, as well as some others. Ultimately the "Social Network: Reddit Hyperlink Network" dataset was chosen for it being readily available without requiring signup or access requests and being concisely documented. The chosen dataset also provides a wide range of properties which can be utilized for various analytical purposes.(*SNAP: Social network: Reddit Hyperlink Network*, no date)

The dataset is sourced from the popular community and social news-oriented platform, Reddit. The dataset has been generated as part of a project exploring how different subreddits (communities) interact, specifically exploring inter-community conflicts.(*Community Interaction and Conflict on the Web*, no date)

When using the dataset it is important to consider that it has been compiled on data on a timespan of January 2014 to March 2017, Reddit has since implemented and enforced multiple significant policy decisions, including the decision to ban the “altright” subreddit and the revisions of its content policy in the wake of the George Floyd protests, which may mean that findings of the initial study, and trends within this dataset may no longer be applicable to the same extent on the current status of Reddit.(‘Timeline of Reddit’, 2023)

1.3 THE DATA

The dataset, presented in a ".tsv" format (tab-separated values), contains 858,490 entries detailing connections between various subreddits. Each entry provides the names of the source and target subreddits, a post ID, a timestamp, and a label classifying the post as either positive or negative towards the target subreddit. Additionally, the dataset includes numerous characteristics related to the post's content, such as text length, word type distribution, sentiment scores, and thematic elements. A supplementary dataset links these entries to the titles of the respective posts.

While the primary dataset maintains a degree of anonymity by excluding personal information and searchable text, the presence of post IDs and their linkage to a dataset with post titles could potentially enable identification of the posters.

In terms of legal and ethical considerations, the Reddit Data API Terms, Content Policy, Privacy Policy, and User Agreement do not explicitly prohibit the use of its content for academic or research purposes. All posts on Reddit are public and searchable. However, the original posters maintain ownership of their content. As a result, the usage of this data is governed by California State law, where Reddit is headquartered, and the data use laws of the poster's country of residence. (*Data API Terms - Reddit*, no date) (*Reddit Privacy Policy*, no date)

Since the outbreak of the Covid-19 pandemic, platforms such as YouTube, Facebook, X (subsequently Twitter), and Reddit have begun to confront, or at least address, the issue of misinformation. Although the quality of content has been monitored since then, this specific dataset, collected from January 2014 to March 2017, remains unaffected by these measures. Consequently, it is reasonable to infer that a higher proportion of the dataset may be compromised by misinformation. (*How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19*, no date)

Bot activity is also a significant concern on Reddit, with many bots either spamming messages or duplicating entire threads to drive engagement. Over time, Reddit has implemented some preventive and reporting mechanisms in this context. However, reports suggest that Reddit's actions have been somewhat antagonistic towards moderators. This situation suggests that the dataset might contain a considerable amount of bot-generated content. (Harding, 2023)

1.4 THE 4 V'S

Volume: The dataset contains a considerable amount of data, both in size and scale. It encompasses 55,863 nodes, representing subreddits, and 858,490 edges, which are hyperlinks between these subreddits. Each hyperlink is annotated with various properties, such as timestamp, sentiment, and a text property vector, contributing to the overall data volume.

Variety: This dataset demonstrates a broad range of data types and formats. It includes structured data, such as subreddit names and timestamps, as well as unstructured data, like text property vectors and sentiment analysis results. Additionally, it features subreddit embeddings, offering a rich and diverse dataset for analysis.

Velocity: This characteristic refers to the rate at which data is generated and updated. The dataset spans a period exceeding 2.5 years, from January 2014 to April 2017. While the dataset itself represents a static snapshot of this timeframe, the underlying social network, Reddit, is characterized by a high velocity of data generation, with continuous creation of new posts, comments, and hyperlinks.

Veracity: Veracity relates to the reliability and accuracy of the data. The dataset's credibility is enhanced by its comprehensive annotation, including sentiment analysis and text property vectors. The sentiment labels are generated through a text-based classifier, reflecting a systematic method for ensuring data quality. Nevertheless, as with any social media-derived data, there are potential issues concerning the accuracy and representativeness of the information.

1.5 DATASET POTENTIAL

In this dataset's context, the exploration and exploitation of various network structures are feasible. The dataset allows for the creation of a directed network, wherein subreddits serve as nodes and hyperlinks function as directed edges, thereby indicating the directional flow of content. An alternative approach involves constructing a weighted network, in which the edges are weighted according to the frequency of hyperlinks, emphasizing the strength of the connections. Furthermore, the development of a temporal network that integrates timestamps can indicate the evolution of subreddit interactions. Additionally, constructing a sentiment-weighted network, where edges are weighted based on sentiment scores, provides insights into the nature of these interactions.

The analysis of these networks necessitates the application of diverse network science techniques. Degree Centrality could be employed to identify highly interconnected subreddits (Jilbert, no date), while PageRank can be used to pinpoint those with significant influence (*PageRank - Neo4j Graph Data Science*, no date). Community Detection can discover clusters within Reddit, shedding light on common interests or themes. Time Series Analysis tracks changes in network characteristics over time (Holme and Saramäki, 2012). In contrast, Sentiment Analysis offers an in-depth understanding of the tone of subreddit interactions.

To effectively understand the insights leant from the analysis, a variety of visualization techniques can be employed. These include graphical representations, such as visual maps of the network with nodes and edges symbolizing subreddits and hyperlinks. Interactive timelines would provide a dynamic perspective on the network's evolution, and heatmaps can be used to illustrate the intensity of connections or sentiments between subreddit clusters.

1.6 CONCLUSION

In conclusion, the examination of the "Social Network: Reddit Hyperlink Network" dataset from Stanford's SNAP platform reveals its value for network science research. The dataset's rich content offers insights into subreddit interactions. However, it also presents ethical challenges, including data veracity and the impact of bots. Despite these limitations, this analysis can enhance our understanding of online social networks and provide information for future studies in data and network science. The Reddit Hyperlink Network dataset thus stands as a significant tool for comprehending the complexities of online communities.

Wordcount: 1349

CHAPTER 2 REFERENCES

Community Interaction and Conflict on the Web (no date). Available at: http://snap.stanford.edu/conflict/ (Accessed: 9 December 2023).

Data API Terms - Reddit (no date). Available at: https://www.redditinc.com/policies/data-api-terms-2 (Accessed: 10 December 2023).

Harding, S. (2023) *Reddit mods fear spam overload as BotDefense leaves "antagonistic" Reddit, Ars Technica*. Available at: https://arstechnica.com/gadgets/2023/07/reddit-mods-fear-spam-overload-as-botdefense-leaves-antagonistic-reddit/ (Accessed: 10 December 2023).

Holme, P. and Saramäki, J. (2012) 'Temporal networks', *Physics Reports*, 519(3), pp. 97–125. Available at: https://doi.org/10.1016/j.physrep.2012.03.001.

How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19 (no date) *New America*. Available at: http://newamerica.org/oti/reports/how-internet-platforms-are-combating-disinformation-and-misinformation-age-covid-19/ (Accessed: 10 December 2023).

Jilbert, O.L. and I. (no date) *4.2 Degree Centrality | Social Networks: An Introduction*. Available at: https://bookdown.org/omarlizardo/_main/4-2-degree-centrality.html (Accessed: 10 December 2023).

PageRank - Neo4j Graph Data Science (no date) *Neo4j Graph Data Platform*. Available at: https://neo4j.com/docs/graph-data-science/2.5/algorithms/page-rank/ (Accessed: 10 December 2023).

Reddit Privacy Policy (no date). Available at: https://www.reddit.com/policies/privacy-policy (Accessed: 10 December 2023).

Silva, V.F. *et al.* (2021) 'Time series analysis via network science: Concepts and algorithms', *WIREs Data Mining and Knowledge Discovery*, 11(3), p. e1404. Available at: https://doi.org/10.1002/widm.1404.

SNAP: Social network: Reddit Hyperlink Network (no date). Available at: http://snap.stanford.edu/data/soc-RedditHyperlinks.html (Accessed: 9 December 2023).

'Timeline of Reddit' (2023) *Wikipedia*. Available at: https://en.wikipedia.org/w/index.php?title=Timeline_of_Reddit&oldid=1178481396 (Accessed: 9 December 2023).

User Agreement - September 25, 2023 - Reddit (no date). Available at: https://www.redditinc.com/policies/user-agreement (Accessed: 10 December 2023).