Investigating the True Crime Phenomenon on Reddit: A Semantic Network Analysis Perspective

Kelly Ortega

Department of Computational Linguistics
Montclair State University
ortegak2@montclair.edu

Abstract

In recent years, the growing interest in true crime content across social media has given rise to various communities focusing on discussing and dissecting real-life crime. Among these platforms, Reddit, in particular, has become a popular hub for true crime enthusiasts, providing a variety of subreddits on the matter. This study aims to investigate these subreddits through semantic network analysis. This methodology exposes hidden discussion patterns and identifies key themes and relationships between different topics. Semantic links are examined within these subreddits using TF-IDF values, co-occurrence matrix, and graph-based methods to determine interconnected and prevalent concepts. The findings of this research project improve our insight into true crime fan interests and online true crime communities while highlighting the usefulness of semantic network analysis when analyzing unstructured and complex textual data.

1 Introduction

Reddit is among the largest social media platforms today (Proferes and Zimmer, 2021). There are thousands of different subreddits to browse, each representing a community focusing on niche aspects of information. Redditors actively participate in subreddits to share opinions, facts, or storytelling. Previous research on Reddit discourse has observed communities relating to politics or medical conditions such as Covid-19 vaccination discussion (Melton, 2023) and depression and bipolar disorders (Minjoo Yoo, 2019). Given the surging popularity of the true crime genre, particularly on Reddit, a unique opportunity is presented to explore public discourse relating to crime and criminal justice. Just as past research has leveraged this platform to track the flow of political discussion or gain insight into mental health conditions, the study of online true crime communities could enrich our understanding of how the public engages with and is affected by crime narratives.

Millions of cases in the United States range from property crime, which includes offenses such as burglary and larceny-theft, to violent crime, defined in the Uniform Crime Reports (UCR) program as those offenses that involve force or threat of force (DOJ, 2020). To the best of the author's knowledge, there is currently little to no research on the social media discourse about true crime and how this information spreads across platforms.

Certain criminal cases are famous on social media, sparking widespread conversation and speculation. These cases often contain elements that capture the public's interest for years, such as the involvement of high-profile individuals or shocking and unexpected circumstances. To this day, people debate the outcome of the 1995 OJ Simpson Trial: Does Simpson's not guilty verdict prove his innocence (Hudson, 2021)? And more than 20 years following the murder of a child beauty pageant winner, true crime enthusiasts continue to wonder: Who killed Jon-Benet Ramsay? 'Fame' in this context encompasses the intensity of public discourse and the variety of posts or comments. While it is apparent that numerous criminal cases have achieved notoriety, the degree and nature of this recognition warrant further explanation.

This research will use various natural language processing techniques to investigate the discourse on multiple true crime subreddits. By gathering data, preprocessing the text, creating word matrices, and performing semantic network analysis, it is possible to gain insight into public perception and correspondence. Therefore, this study aims to explore true crime content and its influence on public perception of criminal cases by analyzing these communities and the conversations taking place within them.

This paper's organization is as follows: it will first introduce key concepts and methodologies, then present the data collection and preprocessing stages. Next is a detailed walkthrough of the semantic network creation. It will conclude by discussing results, limitations, and potential implications.

2 Semantic Network Analysis

Semantic network analysis is a method to discover semantic structures in texts (Minjoo Yoo, 2019). Semantic networks are knowledge structures that provide graphical representations of words or concepts as nodes and their interrelationships as edges. The graphs highlight relationships by illustrating how terms or ideas connect within a specific context.

Unlike topic modeling, a technique used to classify documents based on common words, semantic network analysis classifies words within a document reliant on their proximity or co-occurrences (Segev, 2021). The semantic network approach is helpful because it allows researchers to view extensive textual data from a bird's eye perspective. Combining content and network analysis allows the resulting graph to capture sentiment, opinion, and topic categorizations visually.

Past research has used semantic network analysis to explore various topics, ranging from scientific collaborations to political discourses. The work of Yoo, Lee, and Ha (2019) is particularly relevant to this study, as it analyzed how people perceive both bipolar and depressive disorders among r/Bipolar and r/Depression subreddits. The authors pinpointed clusters in semantic networks using subreddit data and observed the overlapping and differentiating semantic traits. Despite this existing research, there is a lack of comprehensive semantic network analysis within true crime communities on Reddit, which this study seeks to address.

3 Methodology

3.1 Data Collection

The initial phase involved the extraction of user-generated posts and comments from r/Truecrime, r/Unresolvedmysteries, and r/Serialkillers. These subreddits are relevant to this study as the discussion focuses on various crime types: all crime, cold and unsolved cases, and any information related to serial killers, respectively. The author used Python Reddit API Wrapper (PRAW) to gather data. They created a RedditScraper class, which allows multiple parameters such as the subreddit's name, specification of data extraction from the "New" tab, a default post limit of 1000, and an optional keyword search - though this process did not implement keyword search.

Furthermore, the posts collected in this study were from each subreddit's "New" section, meaning that posts and subsequent comments were organized by the most recent when downloading data. This decision avoided the analysis of the "Hot" and popular posts or "Top" posts of all time. Focusing on these popular posts would provide an inaccurate representation of true crime discourse or everyday discussion for this study. Rather, it would alter this analysis to one of virality on Reddit.

The RedditScraper class also has a default limit of up to 1000 posts to avoid hitting the API limit of Reddit. However, the "New" posts thread returned fewer than 1000 results for each subreddit. The RedditScraper class saved all data to separate JSON files, structured so that each post is a dictionary object and comments are a list associated with each post. A function titled count_posts_and_comments calculated the total number of posts and comments collected. Table 2 displays these results.

Subreddit	Total Posts	Total Comments
r/Truecrime	895	50191
r/UnresolvedMysteries	965	78137
r/Serialkillers	900	57340

Table 1: Total data collected

3.2 Data Preprocessing

Once the data was collected, the next stage involved preprocessing it to prepare it for analysis. The text preprocessing stage comprises two primary tasks: text cleaning and Term Frequency-Inverse Document Frequency (TF-IDF) calculation.

In order to clean and prepare the corpora for semantic network creation, the author followed a five-step procedure. The first step, slang conversion, consisted of converting standard internet abbreviations such as "LOL" and "TL;DR" to their extended form, "laughing out loud," and "too long, didn't read." Slang dictionaries, specifically the comprehensive list of slang for text preprocessing (Mbaye, 2020), helped identify known abbreviated terms. Converting these phrases and more helped reduce noise in the analysis, as most slang did not add semantic importance.

The next step, bot deletion, was necessary because this study focuses on user-generated content. On Reddit, bots such as "RemindMeBot" are created to comment under posts to remind users of updates made by original posters. There are thousands of more bots, and each has its purpose. Identifying and removing bot content ensures the semantic network focuses on human interactions alone.

The third step, lowercase conversion, was carried out to eliminate data redundancy and enhance consistency. Converting all text to lowercase guarantees that words are considered separate entities due to case differences. For example, 'Murder' and 'murder' were viewed as one word or node.

Subsequently, special character removal involved eliminating non-alphanumeric characters, such as emojis, special symbols, and punctuation marks. This measure was crucial in data refining, making processing easier for machine learning algorithms.

The next step was to export cleaned datasets to separate JSON files. This export provided an ready to use dataset for semantic network creation. It ensured that the following analysis utilized clean and reliable data.

The final task in the data preprocessing stage was to perform TF-IDF calculation. Term frequency-inverse document frequency is one of the most popular techniques in natural language processing (Amir Jalilifard, 2020). It is a numerical statistic that indicates the importance of a document in a collection or corpus by scoring words in a text. It increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus - this helps adjust for the general frequency of specific terms. TF-IDF calculation was performed on the cleaned subreddit data.

3.3 Semantic Network Construction

The construction of the semantic network began immediately following data preprocessing. This step entails creating a network structure to capture relationships between words in the corpora. More specifically, this involves creating a TF-IDF weighted co-occurrence matrix. The matrix quantifies how often pairs of words appear together in a text window, weighted by their TF-IDF scores. This mathematical structure enfolds the co-occurrence and relative importance of word pairs in the corpora.

The created TF-IDF weighted co-occurrence matrix allowed the construction of the semantic network. Each node in the network signifies a unique word, and an edge represents the co-occurrence relationship between two words. The weight of the edge mirrors the TF-IDF weighted co-occurrence of the words. The author utilized the NetworkX python library to accomplish this semantic network construction and convert the co-occurrence matrix into a network graph. NetworkX is a library for creating, manipulating, and studying complex networks' structure, dynamics, and functions (Hagberg, 2008). It enables the visualization of semantic networks in this study by providing many network analysis functions.

The final step was visualizing the semantic network. The network was exported to GEXF format to be compatible with the network visualization tool, Gephi. Gephi allowed the exploration of many network properties, including node degree and edge weight, and a unique view of thematic structures across subreddits r/Truecrime, r/UnresolvedMysteries, and r/Serial Killers. Figures 1, 2, and 3 show the overall semantic networks.

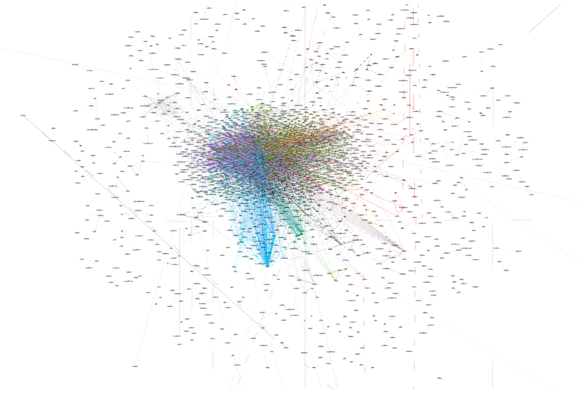

Figure 1: r/Truecrime semantic network

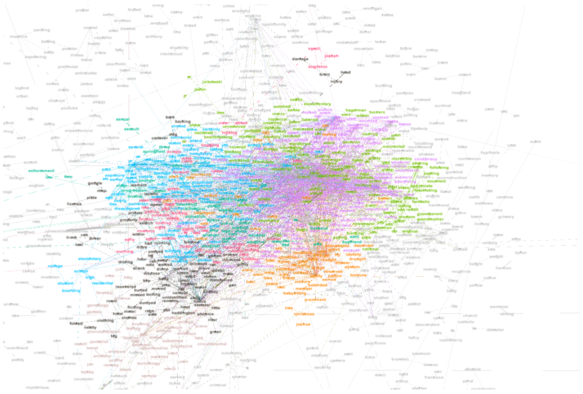

Figure 2: r/Unresolvedmysteries semantic network

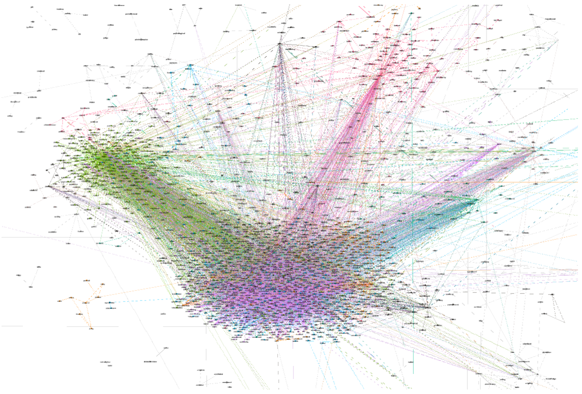

Figure 3: r/Serialkillers semantic network

4 Results

The general findings and observations of the three semantic networks are limited. The most prominent finding of this semantic network analysis is that the r/Serialkillers subreddit contains highly graphic material. The conversation of users in this subreddit often leads to serial killers' psychological background or state, and introduces more explicit content than in r/Truecrime or r/Unresolvedmysteries. To view this observation, let us first understand the visualization methods in Gephi.

As previously mentioned, Gephi is used for network visualization and analysis. For this study, several statistical calculations influenced the overall insight and semantic representation of subreddit discourse. First, the modularity calculation helped identify communities in the network, and the different color edges and nodes represent these clustered communities. The communities are partitioned based on the statistical modularity value.

Modularity Class	Partition
12	15.75%
45	11.6%
2	9.58%
1	7.46%
5	5.93%
4	5.68%

Table 2: r/Truecrime modularity Statistics

However, the labeling of each community relies on personal interpretation. The sizes of nodes are determined using eigenvector centrality, which is a calculation that represents the influence of a node within a network. For the semantic networks shown in this study, the larger the node, the higher its eigenvector centrality - thus indicating that the word represented significantly impacts the overall discussion in each subreddit. Lastly, the layout algorithm chosen for these visualizations is titled OpenOrd. This layout algorithm is for large-scale network visualization. In this case, the strength of connections, or edge weights, determine the spatial relationships between nodes.

Figure 4 previews the explicit content from the r/Serialkillers semantic network. Figure 5 shows a cluster of words likely used in the subreddit to diagnose or discuss the diagnoses of serial killers and victims.

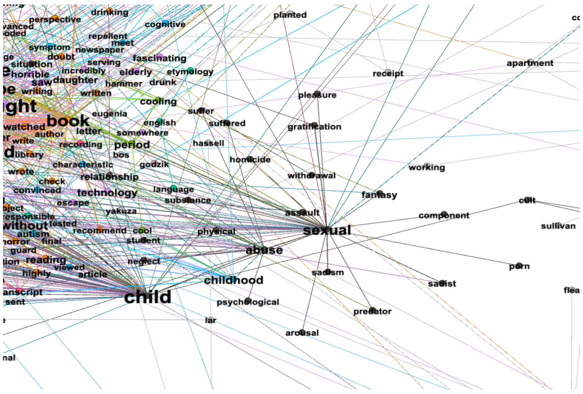

Figure 4: r/Serialkillers explicit cluster

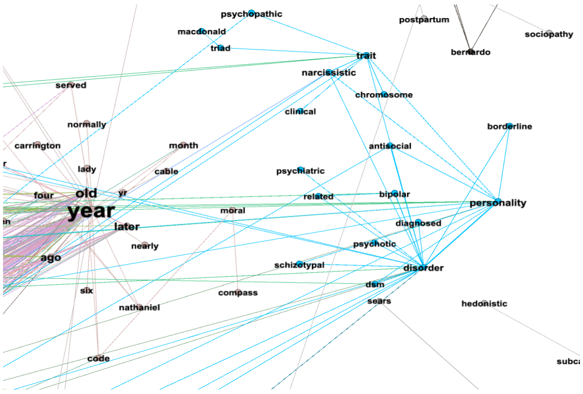

Figure 5: r/Serialkillers psychology cluster

In the other subreddit networks, such clusters do not exist. Instead, as expected, r/Truecrime sees clusters related to other social media platforms and discusses various criminal cases. r/Unresolvedmysteries displays the most emotion conveyed among the three subreddits. Figures 6 and 7 reinforce these observations, respectively.

5 Conclusion

In conclusion, this research has successfully employed semantic network analysis to visualize

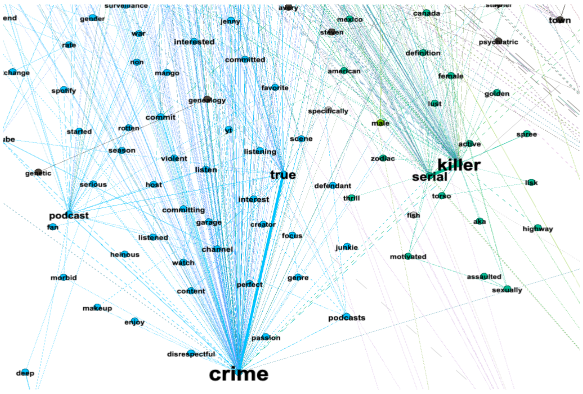

Figure 6: r/Truecrime media cluster

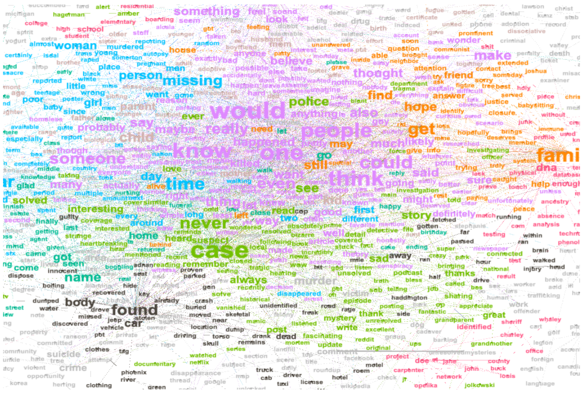

Figure 7: r/Unresolvedmysteries emotion cluster

and explore the co-occurrence relationships between words in the subreddits of r/Truecrime, r/Unresolvedmysteries, and r/Serialkillers. The process began with data collection and preprocessing, followed by constructing a semantic network based on a TF-IDF weighted co-occurrence matrix. Finally, we visualized the network using the Gephi software and OpenOrd layout algorithm. The visualizations allowed for an exploration of thematic clusters. It also revealed patterns and associations that may have stayed hidden in the raw text data. However, this study showed that while semantic network analysis provides a unique and visually compelling representation of word relationships, it remains a surface level view of true crime trends and discussions.

Future Research

For future research, the author proposes using Word2Vec embeddings to create the semantic network. Word2Vec embeddings capture the similarity between words. Unlike TF-IDF and co-occurrence matrices, which focus on word frequencies, Word2Vec considers the context in which words appear. This approach may give a more robust look into the discussions of true crime subreddits, relating unexpected topics and concepts.

Acknowledgments

I want to thank Dr. Larissa Goulart da Silva for her guidance and contribution to this research. I also thank the faculty at Montclair State University and my fellow students in the Computational Linguistics department.

References

Alex F. Mansano Rogers S. Cristo-Felipe Penhorate C. da Fonseca Amir Jalilifard, Vinicius F. Caridá. 2020. Semantic sensitive tf-idf to determine word relevance in documents.

U.S. DOJ. 2020. Offenses known to law enforcement.

Swart Pieter S Chult Hagberg, Aric. 2008. Exploring network structure, dynamics, and function using networkx.

Carol Hudson. 2021. From the archive: Opinion: O.j. simpson: Morally guilty, but legally not.

M Mbaye. 2020. Up-to-date list of slangs for text preprocessing.

Bae J. Olusanya O.A. Brenas J.H. Shin E.K. Shaban-Nejad A. Melton, C.A. 2023. Semantic network analysis of covid-19 vaccine related text from reddit. *Studies in Computational Intelligence*, 1060.

Taehyun Ha Minjoo Yoo, Sangwon Lee. 2019. Semantic network analysis for understanding user experiences of bipolar and depressive disorders on reddit. *Information Processing Management*, 56.

Naijyan; Gilbert Sarah; Fiesler Casey; Proferes, Nicholas; Jones and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. 56.

E Segev. 2021. *Semantic network analysis in social sciences*. Routledge.