A Network-Analysis-Based Approach on Extracting Main Idea and Internal Word Relations of Applicational Texts

Chuqiao Wang, Hanquan Zhong, Yuqi Yan

Washington University in St. Louis, St. Louis

May, 2024

Abstract

Inspired by a two-way classification approach (Celardo, 2020), we take a new network analysis method for text analysis. We use mental-health topic articles extracted from website *Healthline* as the base of our text analysis. We use NLP and network analysis tools such as *nltk* and *networkx* packages in Python to extract the topics of the text, and build network(s) that represent the occurrences and semantic similarity of the vocabulary. WordNet (Princeton, 2024) is a database that provides the semantic distances of vocabulary. We measured the degree distribution, clustering, and betweenness of the vocabulary network.

***Keywords*— Similarity-Based Networks**, **Text Analysis**, **Clustering**

Github Repo $https://github.com/cse416a-sp24/final-project-chuqiaowang_hanquanzhong_yuqiyan$

1 Introduction

1.1 Problem statement

In recent years, both qualitative and quantitative analysis on natural languages are gaining more attention. With the introduction of large language models such as ChatGPT, Gemini, and more recently, Llama, there has been an increasing demand on the efficiency and accuracy of the tools. Inspired by this global

trend, we would like to conduct an experimental research on extracting key words from sample texts, construct networks, and see if the result incorporates any useful information.

Our project mainly consists of two research questions:

- Could our proposed network analysis methods provide a direct visualization that concludes the main idea and key words of the target text?
- What other linguistic information can be obtained from the network, with modifications on nodes and edges?

1.2 Significance of the project

Though this project provides only a superficial analysis and attempt on constructing a network to process natural language, it provides an innovative idea of using network instead of text flow to present information. It potentially has advantages of visual directness and explicitness, and there may be other strong points that can be discovered in future research.

2 Methodology

2.1 sample selection

All samples in this project are selected from articles on health on a website called *Healthline*. The six articles are:

1. ***Understanding Financial Stress and Tools to Help You Cope***
2. ***17 Strategies for Coping with Stress in 30 Minutes or Less***
3. ***15 Ways to Soothe Your Mind and Body During Times of Distress***
4. ***16 Simple Ways to Relieve Stress***
5. ***The Mental Load: Managing a Burden You Can't Actually See***
6. ***What is a Worry Journal for Stress?***

As it may also be seen through the titles, they are all passages that offer suggestions on stress, while the source of stress may vary from financial difficulties, personal pressure, or other sources. They are selected because these can be good samples to start with, since the main idea is straightforward and easy to attest if the network analysis actually gives the accurate information. The internal connected nature also makes it reasonable for other experiments, such as identification and differentiation between problems (sources of stress) and solutions.

2.2 Data cleaning and processing

The sample texts are cleaned and processed so that morphemes and propositional phrases are excluded from the focus of analysis since they are thought o contribute little to understanding the key problems or solutions stated in the original texts. Starting from raw texts, all special symbols are firstly removed. Then all stopwords are removed using XXX. After that, the words are converted to all lowercase and are gone through a process of lemmatization made possible by the ntlk package in Python.

2.3 Network construction

To set up a network that can be analyzed on, the nodes an edges of the network should be well-defined. Considering the mechanism of networks and examing the sample texts. We proposed the following definition for the constituents of the target network.

- **Nodes** refers to word cleaned after data processing (see section 2.2).
- **Edges** are added based on specific function of the network. Generally, if two words are in neighbouring positions, an edge is added between the two nodes that represent the two words. Based on analysis requirement, for each network, a directed and an undirect version is provided. In the directed graph, the edge will point from the word that is in precedent position to the word that follows. For instance, for the segment "network analysis", we will have an edge $network \rightarrow analysis$ if the graph is directed.

Based on the following criteria, an overall network is initialized. The network has 1459 nodes and 3521 edges, indicating the number of potential keywords and connections.

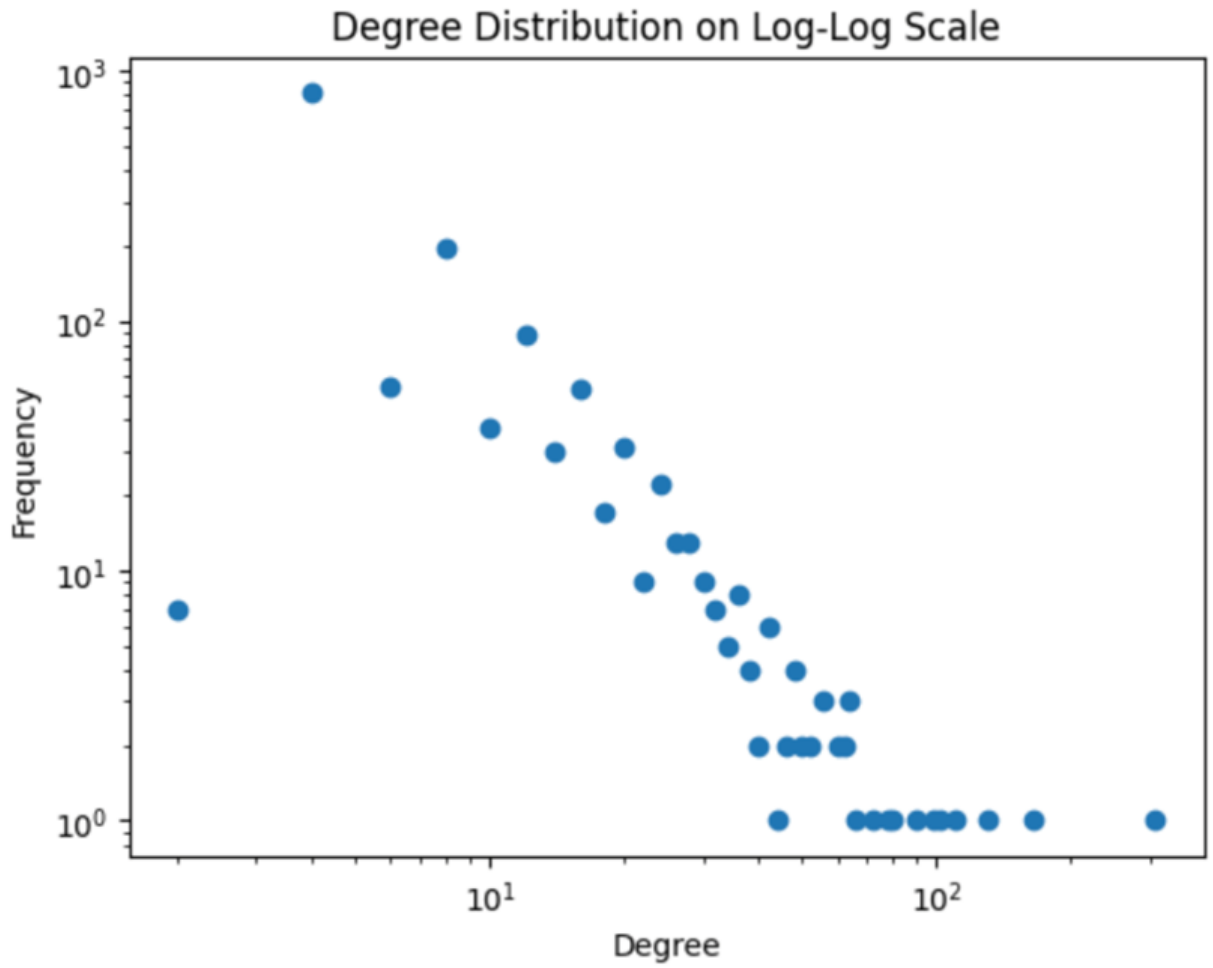

Figure 1: Log-log plot of the degree distribution of the overall network

Figure 1 shows the degree distribution of the overall network on a log-log scale. It can be observed that the network has an iconic ”heavy tail” distribution, which means most nodes have low degrees while a few nodes have very high degrees. In the context of this project, this indicates that there are a few words that are frequently used (presumably words related to stress and stress management), and most words are used only a few times.

2.4 Hypothesizes on communities

Immediately after the overall network is set up, communities are detected. When specific subgraphs are extracted, such characteristics become more explicit. However, it's hard to directly conclude the denotation of such communities in reality, as natural languages are complex. Thus, we have three hypothesizes of these communities, which will be tested in section 3.

- **H1: Document Community** - Each community represents a specific article or document.

- **H2: Thematic Word Community** - Each community contains vocabulary centered around specific themes (such as ”actions”, ”reason”, etc.)
- **H3: Part of Speech Community** - Each community corresponds to a part of speech (such as nouns or verbs).

Aside from the main research questions, this project will also attempt to provide an answer to these hypothesizes and predict whether it is common on other similar applicational texts.

2.5 Two sub-graphs based on different node selection criteria

The overall graph is hard to extract any useful information, thus extracting meaningful sub-graph will be crucial. With the progress of this project, we find it necessary to have nodes selected so that they can be representative in various ways. Thus, we finally decide to have two smaller networks for direct analysis, with node selected with different criteria. Note that despite the node is different in each network, the edges are still added based on the definition in section 2.3.

The first network, **abbreviated as H in following texts**, selects nodes based on ranking of degrees. With definition in section 2.2, it can be inferred that to be selected in H as a node with high degree, the word has to be both significant itself and neighbours other significant words (since for an edge to exist in H, both nodes should be selected as nodes of H, which indicates they both have high degrees). A total of 37 words are selected as nodes, and 74 edges are found between them.

The second network, **abbreviated as F in following texts**, selects nodes based on its frequency in the original text. Any word that appear at least 10 times are selected, because the fact that they are constantly stressed suggests their significance. A total of 37 nodes are selected, and 109 edges are found between them.

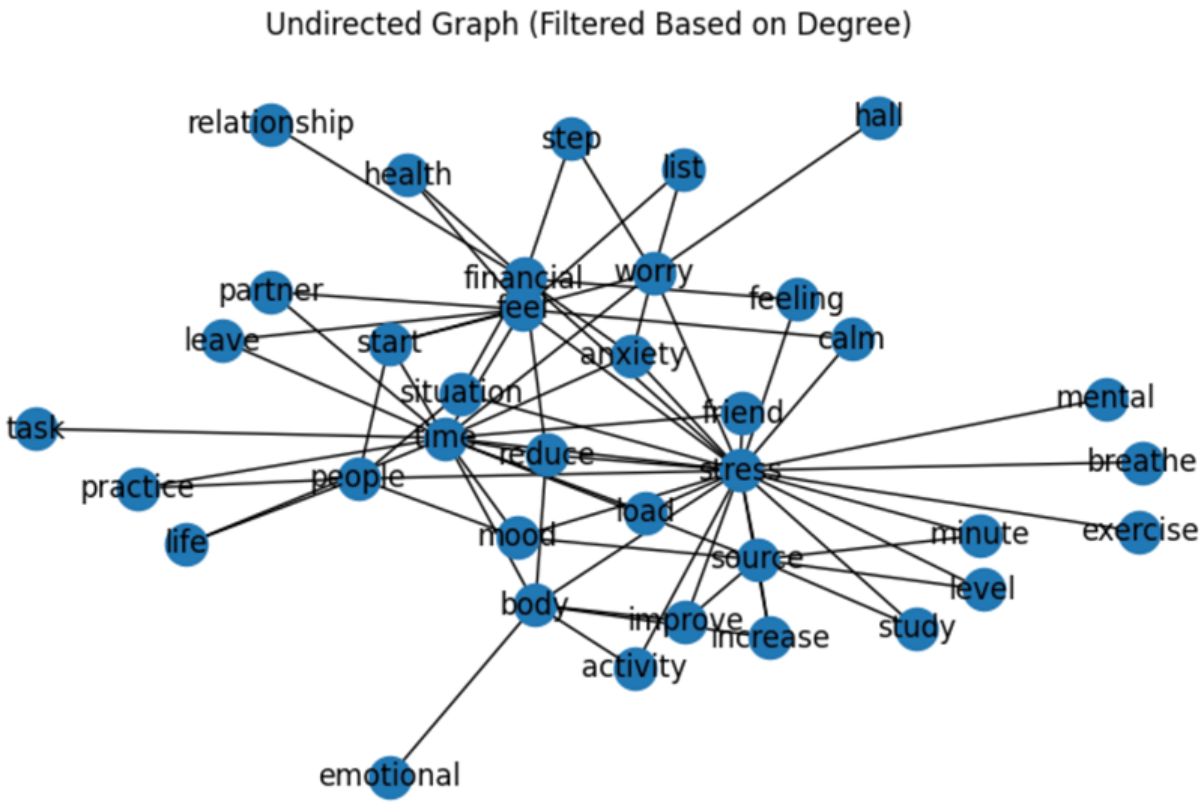

Figure 2: Sub-graph H (nodes selected based on degree ranking)

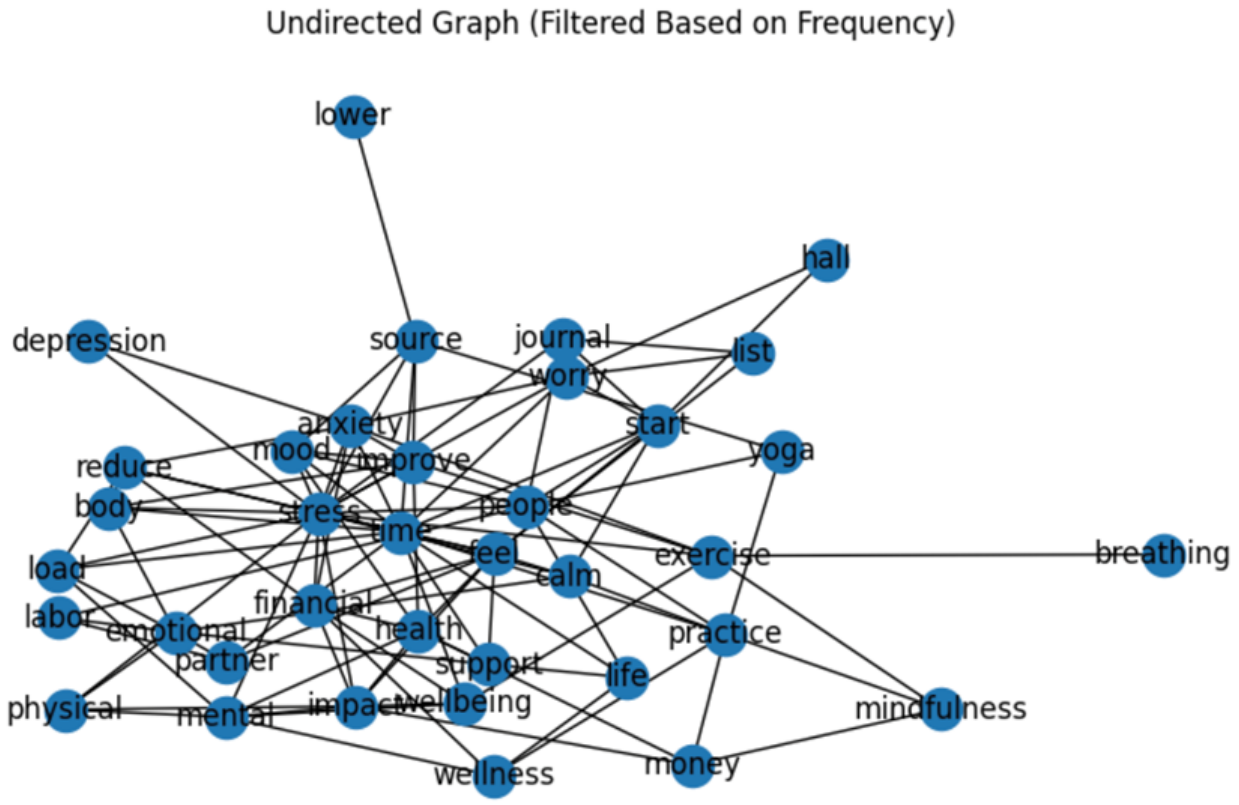

Figure 3: Sub-graph F (nodes selected based on frequency ranking)

Figure 2 and Figure 3 show the visualizations of two sub-graphs generated based on the above principles. Such separation reflects the intuition of texts as the consecutive flow of languages. While F values the individual significance of single words, H implies 'important phrases' - two or three consecutive significant words. While this presents some problems for final interpretation (see section 4), differences are indeed detected during the analysis.

3 Findings

3.1 Centrality Analysis

As a useful and necessary investigation process, centrality measures are calculated for both sub-graphs H and F. Among the four, betweenness centrality and Eigenvector Centrality are selected because in both graphs, the significance of a single word is somewhat represented through the selection of the word in the respective sub-graph, and Close centrality does not have a good real-word implication in this case, since key words don't have to connected to most other key words. Additionally, high betweenness and eigenvector centrality reflects signifcance in information flow and influence within the network, which exactly matches out demand - find the main idea and internal relations of words.

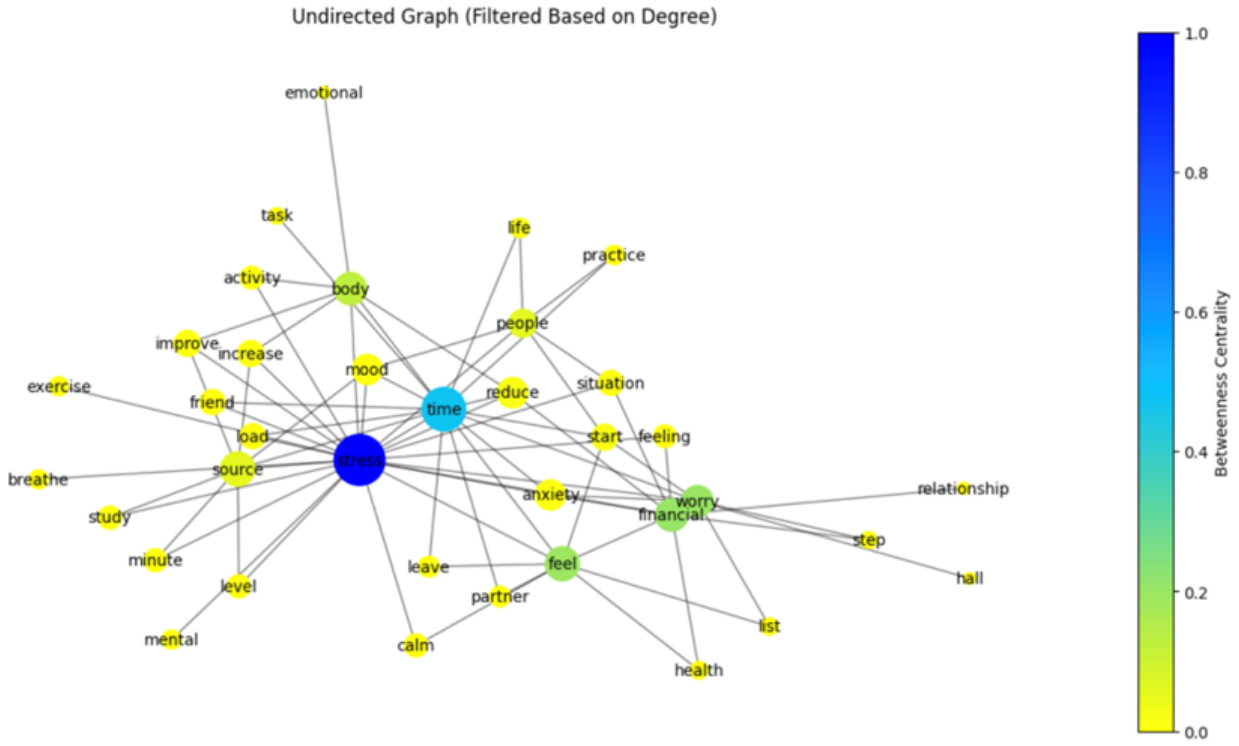

Figure 4: Centrality Analysis on Graph H (degree)

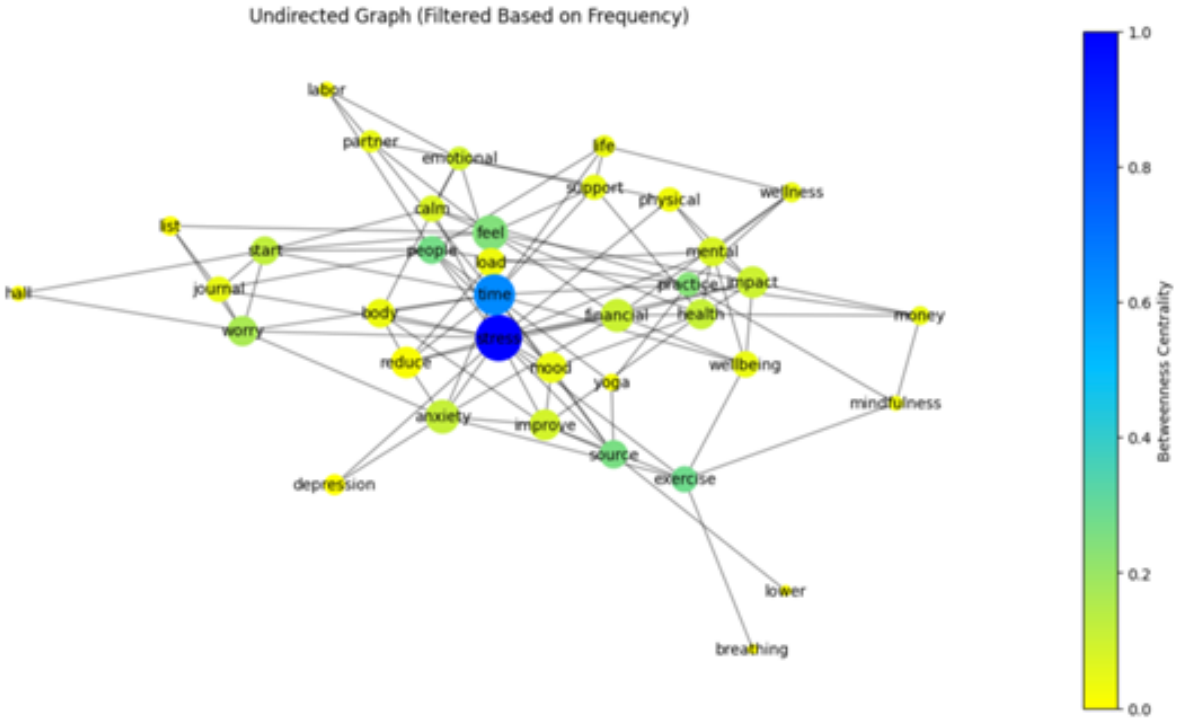

Figure 5: Centrality Analysis on Graph F (frequency)

Figure 4 and Figure 5 are visual representations of the results of centrality analysis. The node size reflects eigenvector centrality, while the color reflects betweenness centrality. The bigger a node is, the higher eigenvector centrality it has, and vice versa. More 'dark' (blue) a node is, the higher betweenness centrality it has; more 'shallow' (light yellow) a node is, the lower betweenness centrality it has.

	Node	Betweenness	Node	Eigenvector
0	stress	0.550967	stress	0.516852
1	time	0.258859	time	0.381328
2	financial	0.110461	source	0.235413
3	worry	0.107778	feel	0.231979
4	feel	0.106107	financial	0.210700

Figure 6: Top 5 key words in H based on centrality measures

	Node	Betweenness	Node	Eigenvector
0	stress	0.278389	stress	0.432386
1	time	0.178184	time	0.340327
2	exercise	0.077981	feel	0.249786
3	people	0.074108	anxiety	0.226834
4	source	0.072606	financial	0.216958

Figure 7: Top 5 key words in F based on centrality measures

Figure 6 and Figure 7 provides tables of Top 5 nodes with the highest centrality measures in both H and F. There are 4 rankings in total (2 centrality measures * 2 graphs), but the results looks similar. "Stress" comes in the top in all four columns, which well indicates that the main topic is found and it is accurate based on the original intention of the sample text. Below "stress", some common problems such as "anxiety" and useful solutions such as "exercise" can be spotted, but tehre are also some verbs such as "feel" that are frequent simply because people use it when express emotions. The problem of verbs will be covered in the discussion (see section 4).

3.2 Community Detection

Besides rankings of nodes based on centrality measures, it is also important to see if certain communities can be identified based some community detection algorithms. In this project, Girvan-Newmann algorithm is applied, and the result is shown as follow.

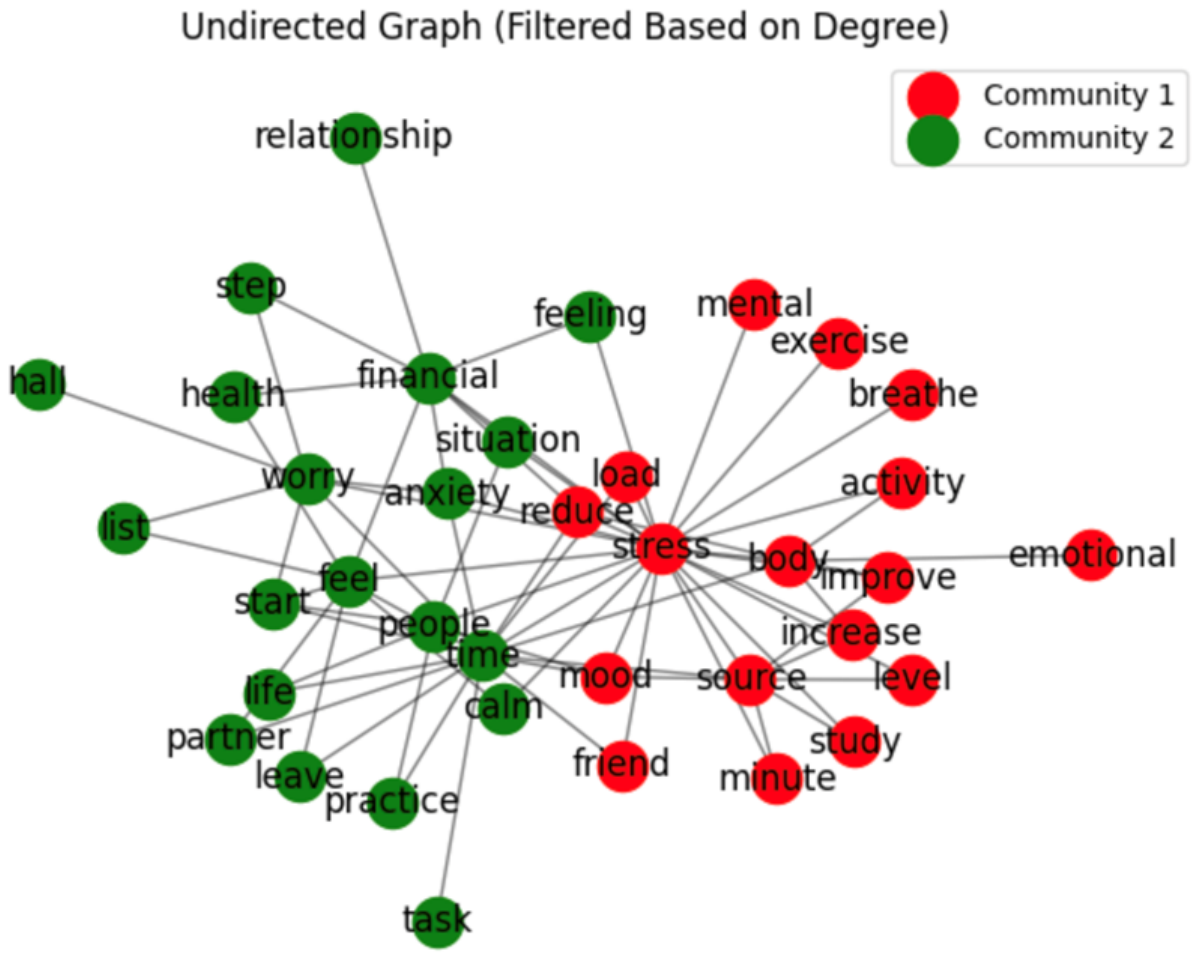

Figure 8: Community Detection on Graph H (degree)

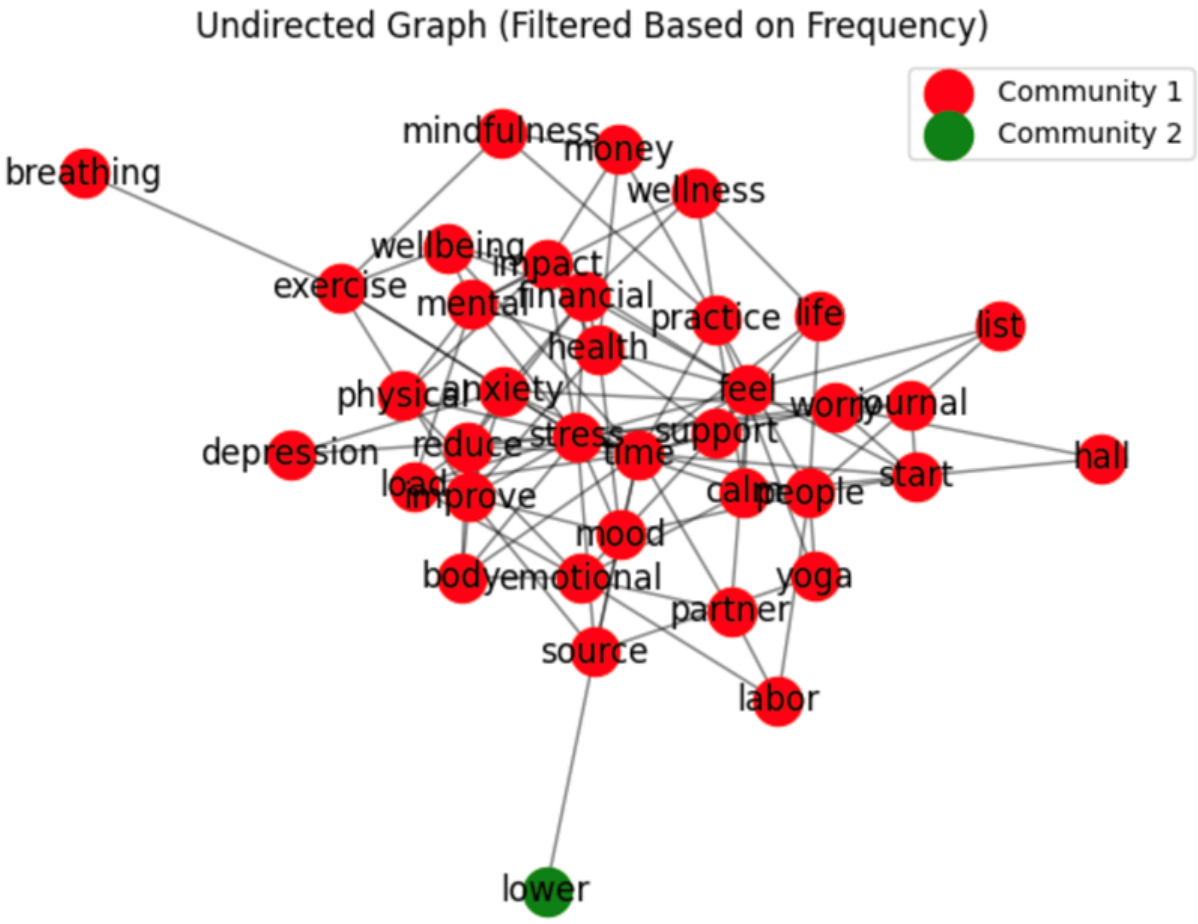

Figure 9: Community Detection on Graph F (Frequency)

Figure 8 and Figure 9 show the result visualization after running the algorithm. Two colors of nodes represent two communities. Figure 8 gives a nice balanced community distribution, while the result in Figure 9 behaves not as good as the one in Figure 8. Despite the fact that the number of nodes are nearly equal in Figure 8, however, it's hard to directly interpret the meaning in each community. There might be two reasons for this: 1. interference of verbs and adjectives that don't actually address anything meaningful; 2. a mix of problems and solutions, which are two natural categories in the logic of application texts. Regardless of the reason, community detection using Girvan-Newman algorithm does not work well with the two given sub-graph in this project. There may still be meaningful information enclosed, but further processing is required.

3.3 Alternative method: node labeling through sentiment analyzer

As seen in section 3.2, community detection based on an algorithm don't always give ideal results. Especially for the result on sub-graph F (Figure 9), only 'lower' being recognized in one community does not help with explaining the main idea of the texts. Thus, an alternative classification methods based on a

package called sentiment analyzer under nltk in Python is applied to see if better results are shown.

Sentiment analyzer gives a score for an input based on the 'sentiment' the word carries. If the word is mostly used positively, it receives a positive score between 0 and 1, else it receives a negative score between -1 and 0. The grading criteria is shown in the following table:

Score range	Meaning (mood)
$(0, 1]$	Positive
0	neutral
$[-1, 0)$	negative

Table 1: Scoring criteria for sentiment analyzer

The tool is also used for its strong power in detecting (or having a great database of) sentiments, even with non-text combinations, certain sentiments can be detected.

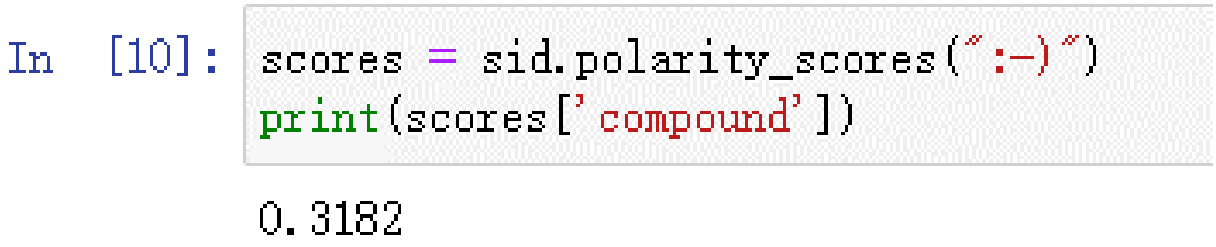

Figure 10: Computation of score on the smiley face ":-)"

Figure 10 is a smiley face in horizontal position, and the sentiment analyzer 'detects' it and gives it a positive score based on the scoring criteria in Table 1. This serves as a piece of evidence for the reliability of the sentiment analyzer, which can be further attested on the results. After applying it to the nodes in both graph H and F, following results are retrieved.

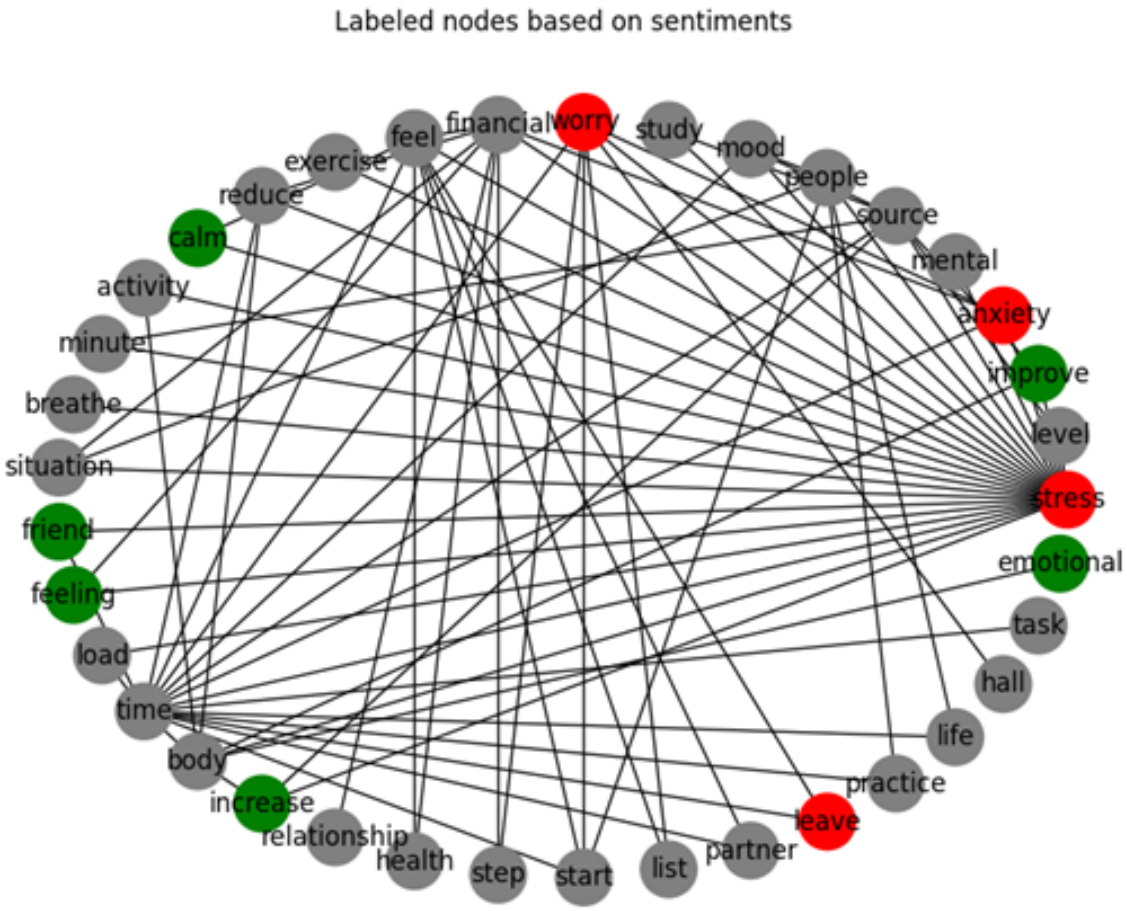

Figure 11: Labeled Sub-graph H based on Sentiment Analyzer

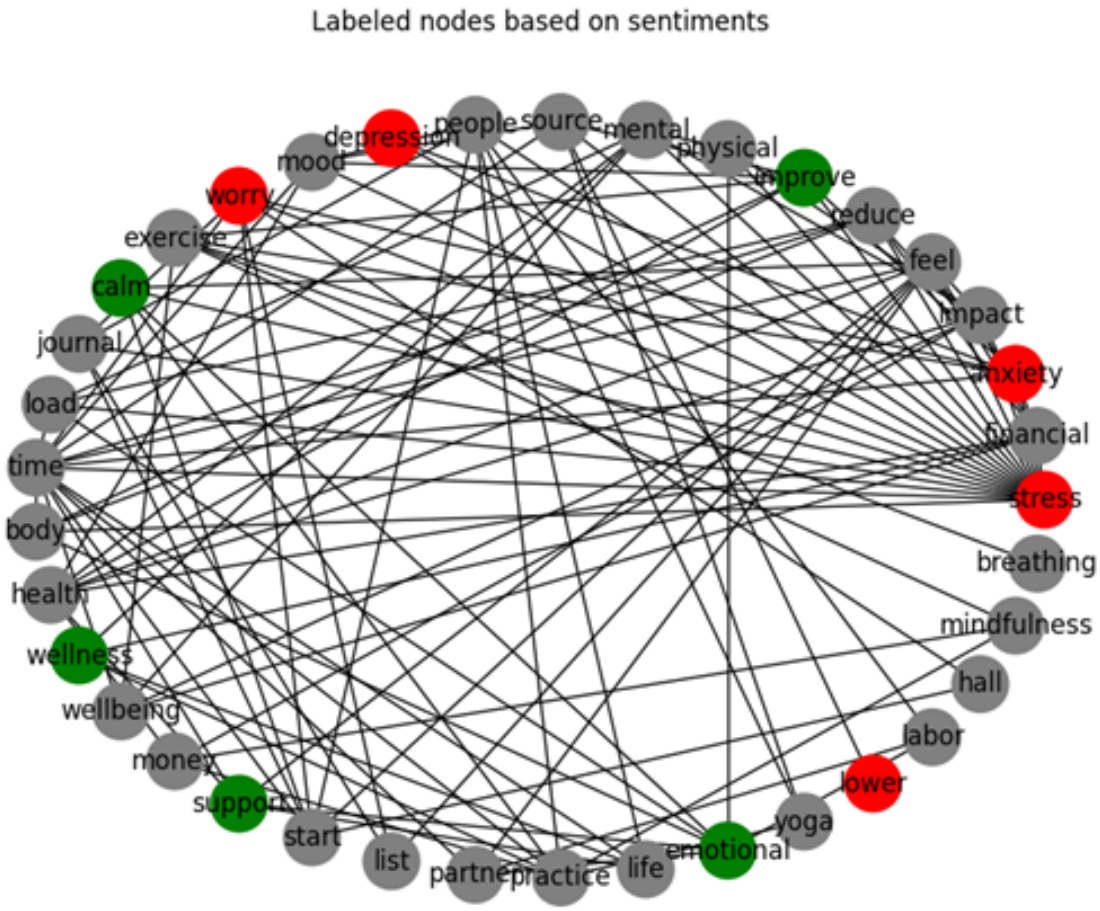

Figure 12: Labeled Sub-graph F based on Sentiment Analyzer

As seen in Figure 11 and 12, Figure 11 is the results after node labeling on graph H, with nodes selected based on degrees, and Figure 12 is that of F, with nodes selected based on frequencies. The green nodes represents ones with positive score, the gray represents neutral ones, and the red represents ones with negative scores. Some really good classifications can be observed, especially on Figure 12, where stress-related problems such as ”anxiety””stress””depression” and solutions such as ”wellness””support” are properly addressed. In Figure 11, some key words such as ”friend” and ”stress” are also recognized, but the outcome is not as clear as in Figure 12, probably because consecutive words in texts usually indicates different part of speech, and certain parts of speech create problems for differentiating problems and solutions (see section 4).

4 Discussion

Our results support H2 that the communities are centered around vocabulary with specific meanings (themes). However, as results shown in 3.2, a naive community detection based on node degree centrality (i.e. the occurrence of the vocab in the text) does not represent the key topic and information in the source text well. This is likely because during the data procession, some context information are lost when converting words into network nodes. For example the position of the vocabulary in the source text. Normally text appeared in title, first, and last paragraph. Using a sentiment analyzer yields a somewhat better result by including more contextual information. Transition and filler words such as ”a” ”the” etc have high counts and would skew the result representation.

Potential work includes obtaining more data and using the semantic distance of the words as the weight of the edges (e.g. words with close meanings such as depression and sadness should be indicated as close distance.) We would also strip the transitional words and only keep the nouns and verbs.

5 Conclusion

In this project, we built a network of mental health keywords by processing the article's text. We performed analysis tasks including community detection, node labeling, and visualization. We show that while our network is helpful for revealing the information with top-frequency in the source text, more NLP-related processing need to be done for the network to represent community relations.

Reference

Celardo, "Network text analysis: A two-way classification approach," *International Journal of Information Management*, 2020.

Princeton, *WordNet*. https://wordnet.princeton.edu/, 2024.

Healthline, *Mind & Body*. https://www.healthline.com/mental-health/mind-and-body, 2015-2024.