SMU Data Science Review

Volume 2 | Number 3 Article 5

2019

Personalized Detection of Anxiety Provoking News Events using Semantic Network Analysis

Jacquelyn Cheun PhD
Southern Methodist University, jcheun@smu.edu

Luay Dajani
Southern Methodist University, ldajani@smu.edu

Quentin B. Thomas
Southern Methodist University, qmanthomas@gmail.com

Recommended Citation

Cheun, Jacquelyn PhD; Dajani, Luay; and Thomas, Quentin B. (2019) "Personalized Detection of Anxiety Provoking News Events using Semantic Network Analysis," *SMU Data Science Review*: Vol. 2 : No. 3 , Article 5.
Available at: https://scholar.smu.edu/datasciencereview/vol2/iss3/5

Personalized Detection of Anxiety Provoking News Events using Semantic Network Analysis

Jacquelyn Cheun PhD.[1], Luay Dajani[1], Quentin B. Thomas[1]

Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {jcheun,ldajani,qthomas}@smu.edu

Abstract. In the age of hyper-connectivity, 24/7 news cycles, and instant news alerts via social media, mental health researchers don't have a way to automatically detect news content which is associated with triggering anxiety or depression in mental health patients. Using the Associated Press news wire, a semantic network was built with 1,056 news articles containing over 500,000 connections across multiple topics to provide a personalized algorithm which detects problematic news content for a given reader. We make use of Semantic Network Analysis to surface the relationship between news article text and anxiety in readers who struggle with mental health disorders. Based on a reader's anxiety profile collected by the network, a personalized dataset can be established to better understand the type of news that impacts a reader's mental health in a negative way. This study can benefit two groups. The first group is the mental health community who can use our approach to understand the impact of news content on those combating anxiety disorders and depression. The second group is for readers of news content in general who might not be aware of the type of news topics they are sensitive to. The insight from the Semantic Network should provide more information about specific triggers of anxiety that were previously unknown.

Keywords: Semantic Network Analysis · Network Science · Graph Analysis · Generalized Anxiety Disorder · Depression

1 Introduction

Currently, readers have little control over what they see while visiting news websites. The decisions regarding what a reader of news consumes on a given day are under the control of the news media organizations. For increased engagement, news organizations might elect to publish embellished content to elicit an emotional response from the reader. The more invested a reader is in the content, the more likely he or she is to be susceptible to clickbait titles, and sensationalized stories. News content can have behavioral effects on readers, mainly when the underlining event results in an uncontrollable tragedy[10, 11]. While this may not be a problem for most people, others with a predisposition to anxiety, may be affected. Unfortunately, anxiety sufferers do not have many options to control what the news organizations present. Additionally, current news recommendation algorithms tend to supply readers with popular trending news items of the day and do not take into account potential mental health issues of individuals in the society. This may extenuate the anxiety inducing effect.

In this paper, using our method of detecting individual anxiety triggers, we hypothesize that there is a relationship between anxiety and the language used in news articles. Words and their use carry with them psychological meaning [26]. We examine language in news stories and how it might contribute to Generalized Anxiety Disorder (GAD) with the help of techniques and tools from the emerging field of Network Science [6, 19, 25, 3]. We examine the structure in a language network based on statistical mechanics and principles highlighted by Albert and Barabasi [1]. We present a way to detect this hidden structure in English written news content in order to identify the text that has an association to GAD or Depression.

2 Background

In the United States, an estimated 6.8 million people are suffering from Generalized Anxiety Disorder, and according to the Anxiety and Depression Association (ADAA), 43% of those individuals are not receiving treatment [20]. The ADAA recognizes Generalized Anxiety Disorder as a legitimate problem in the United States. In its mildest form, GAD is characterised as excessive worrying. On a more severe level, researchers have linked worrying as a potential precursor to depression [23, 21]. We assert that the more connected the world becomes, the more difficult it will be for those suffering from GAD to be able to enjoy the comfort of not worrying about specific events. We come to this assertion based on evidence that news consumption and coverage can influence the way we interact with the web [27].

2.1 Understanding Generalized Anxiety Disorder

The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) defines Generalized Anxiety Disorder (GAD) as

> *"... Excessive anxiety and worry (apprehensive expectation) about a number of events or activities. The intensity, duration, or frequency of the anxiety and worry is out of proportion to the actual likelihood or impact of the anticipated event. The individual finds it difficult to control the worry and to keep worrisome thoughts from interfering with attention to tasks at hand."*

The DSM-5 also mentions that GAD in adults can become so severe that these episodes can extend beyond six months. For most individuals, worry comes and goes, but for those with GAD, worry is a much more prolonged condition. These prolonged worries can bring with them symptoms such as difficulty concentrating, irritability, and muscle tension. These emotions eventually force individuals into a state of restlessness, which inundates them with the feeling of constantly being on edge. Relaxation becomes an impossibility. These symptoms can lead to impairment in one's social environment, work, or other important functional areas of life [2].

Anxiety disorders have been researched since 1988 (DiNardo et.al) [9]. Anxiety can often be a precursor to depression, and there are various ways of detecting its early stages [14]. Hirsch and Mathews provide a road map to understanding the stages of GAD as pathological worry [16]. Everyone worries to a varying degree at one point or another. However, a GAD patient's experience of worry carries a much greater risk of leading to depression [15].

More importantly, very little research so far has gone into the relationship of news content and GAD in general. Most studies focus on an individual's encounters with emotionally charged events in the real world. GAD, however, is not limited to what we see in the real world but can be triggered by what is termed *"verbal and imagery based stimuli* [15]. Individuals with GAD usually focus their attention on the negative and the worst-case scenario. This focus on the negative becomes uncontrollable in many extreme cases and often leads to varying levels of depression, which can alter the way information is processed [4]. Some researchers coin this problem as *attentional bias* [16]. For our study, we focus on the verbal or linguistic triggers associated with GAD, and particularly the ones that might exist within web based news content.

We assume that negative linguistic stimuli coming from news is a problem for those suffering from GAD. There are a number of treatments for people with GAD, but neither news or mental health organizations currently have an automated way to detect triggers in web based news content.

2.2 Attempts to treat Generalized Anxiety Disorder

There is a growing body of research that covers the topic of GAD. We focused primarily on a study published in 2000 by The American Psychological Association (APA) by Ladouceur, Dugas, Freeston, and other researchers titled *Efficacy of a cognitive–behavioral treatment for generalized anxiety disorder: Evaluation in a controlled clinical trial* [18].

The study included a total of 26 individuals who were previously diagnosed with GAD. The process measure used in the study was the IUT or Intolerance of Uncertainty Scale questionnaire, which is useful because uncertainty is the cornerstone of GAD [13]. The questionnaire allows researchers to quantify the intensity of a worry based on the way patients answer the questions.

In the Ladouceur study, 14 individuals were randomly allocated into the treatment group, while 12 individuals were placed in what the study terms the *wait list control group condition.* According to the Ladouceur study [18], the questionnaire was composed of 27 items relating to the following areas: uncertainty, emotional and behavioral reactions to ambiguous situations, the consequences of uncertainty, and attempts to control future events. The 12 wait-list patients were told that treatment would begin 16 weeks after their first assessment. As they waited, they were summoned by highly trained therapists once a month to monitor their state and were provided a small amount of support. At the same time, the treatment group received weekly one hour sessions for sixteen weeks. This cognitive behavioral treatment consisted of five parts: presentation of treatment rationale, awareness training, correction of erroneous beliefs about worry, problem-orientation training, and cognitive exposure. In our approach, we pay close attention to the cognitive exposure component of this treatment because the calibration session in our model described in Figure 2 makes use of a similar concept. The researchers increased tolerance of uncertainty by changing the meaning given to future events that the patients found threatening. The study concluded that there was a significant post-treatment improvement in regards to handling uncertainty for those in the treatment group when compared to the waitlist group. A two way repeated measure MANOVA test revealed that there was a significant time effect on those patients in the control group versus those that got the treatment right away.

2.3 Why identify triggering news for those suffering from GAD?

News content on the web is published every single day, and many of these websites receive millions of visitors per day. Each one of these visitors has his or her unique uncertainties regarding real-world events. The research we have highlighted in section 2.2 mentions uncertainty as a significant precursor of worry. News content is presented to readers without any solutions to problems highlighted in the content of the news article, and many opinion-based stories run the risk of containing embellished viewpoints that might not be good for readers combating anxiety, or depression.

For example, articles that talk about the slowing economy might mention that a top company has cut the jobs of 15,000 people. An individual with GAD could read this story and possibly worry about the safety of his or her job. In this scenario, the reader of the article might not be aware that the probability of he or she losing his or her job at the time the article hit the web is extremely low. If the reader suffers from GAD, the probability of this isolated event happening to him or her might appear to be higher than it is.

In the case of news dispersion and coverage, it is not yet possible to compute the risk probability of a specific event highlighted in a news article happening to a specific reader. News organizations often display all top stories to all people at all times as they receive them. Today, any news organization from any website can post any news content they wish without any understanding of the mental health needs of the individual reader. However, this lack of understanding is not the fault of any of the news organizations. The research necessary for understanding the triggers of anxiety from news events has not been explored in depth. The data necessary to understand the impacts of news events on mental health patients have not been collected in an automated fashion either.

We provide a method to address this problem by developing a way to automatically collect data regarding anxiety triggers based on news content for those dealing with GAD or Depression in general. A concerned reader can benefit from our approach as they gain understanding of their personal sensitivities during their cognitive exposure to news. In this study, we will show that it is possible to automatically collect useful data for mental health researchers and mental health patients. Our approach can help the field further understand what news events have a negative impact on readers.

Up until now, there has not been a way to automatically detect the kinds of news stories that might trigger GAD. Anyone suffering from GAD might need to disconnect from the web entirely in order to maintain mental stability. Completely disconnecting from the web is a solution, although it is not always practical for modern-day audiences integrated into the digital world.

3 Methodology

Our approach works to automate the process of identifying triggers in news content for individuals with GAD based on what we call GAD profiles, which come from reader calibration sessions. GAD profiles consist of extracted article features that our system infers are a contributor to their anxiety based on the reader's critique of random headlines. News articles, along with the reader's GAD profile, can then be represented as a semantic graph that contains details of the relationship problematic news stories have with other news stories that might not be as triggering of GAD. These relationships between anxiety-triggering articles and articles that do not cause anxiety for readers exist within the same semantic network. The network detects problematic news content based on what the reader's personalized profile defines as triggering to read.

Our methodology includes a mixture of techniques. These techniques include an initial calibration session with the network. This session is essentially our process for locating articles that may trigger reader anxiety based on reader feedback or critiques. These dynamic critiquing methods are based on ideas from Zanker et al. in 2010 [28]. We make heavy use of an automated semantic graph extraction process, which serves as a platform for network analysis techniques on our knowledge graph of news articles. Utilizing these techniques together results in a reliable semantic graph that can automatically identify triggering news content for people suffering from GAD or Depression. This technique for the detection of anxiety triggers from news events can also result in useful data containing previously unseen relationships between news and anxiety, which can be leveraged in the mental health research community as well as the Network Data Science research community. Research has shown the usefulness of network topology as it applies to music recommendations [7]. We show that semantic networks can be used by mental health researchers to understand patterns and unseen relationships in news events that individuals with GAD or Depression find problematic.

3.1 Data Collection

In order to successfully define a recommendation process that detects anxiety-inducing articles, we chose a reliable source of constant news content to represent our entire population of news articles A. Associated Press represents our news source because AP is a non-profit organization, and its news wire data is publicly available. In addition to the full availability of AP's news content, the organization also covers a wide variety of news content across several categories. This wide variety of coverage allows for our population of A news articles to be representative of a diverse network. We represent the nodes in our network as articles A.

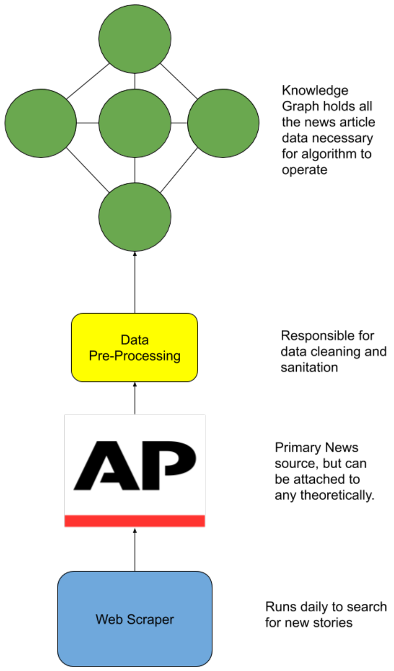
Knowledge Graph holds all the news article data necessary for algorithm to operate
Data Pre-Processing
Responsible for data cleaning and sanitation
AP
Primary News source, but can be attached to any theoretically.
Web Scraper
Runs daily to search for new stories

Fig. 1. Data Pipeline for AP News.

AP publishes news articles daily. To keep up with publishing, we developed an automated pipeline that scrapes the AP News website on a scheduled basis[1]. In Figure 1, we illustrate the step-by-step process for our data acquisition pipeline. The steps operate in sequence to allow the pipeline to keep the semantic graph up to date.

The semantic graph contains a total of 1056 nodes or news articles across 14 different topics. These topics, which AP lists on their news website, include anything from technology, sports, politics, to entertainment. Our pipeline executes daily within a specified period in order to get to a sufficient number of news

[1] https://www.apnews.com/

stories to consider for our experiment. At the end of our collection process, we have the data of interest for our problem described in Equation 1.

$$G_{ij} = \{\ a \in A\} \tag{1}$$

Where G_{ij} represents our entire graph and where i, j represent un-directed relationships between each article a. We will define the connections between the articles in the following sections. Any record a that is a duplicate is removed immediately by the data collection process.

Table 1. Features Collected from AP News

Field	Description
Article Headline	The title given to the article by AP News
Article Author	The author listed by AP News
Article Text	The main body of text within the article.
Article Post Date	The date the article was posted into the format mm/dd/yy

Each article has the following information extracted: article headline, article author, article text, article post date, which we show in Table 1. In order for the data to become useful, we designed pre-processing steps within the pipeline which takes the article data through a series of cleaning steps defined in Equation 2.

$$\{f(a) \forall a \in A\} \tag{2}$$

Where $f(a)$ represents our pre-processing function that runs for all a in articles A. Finally, we represent our data as a 1056 × 4 matrix of articles.

3.2 Cleaning Methodology

Next, we discuss the overall processing of $f(a)$ in order to get the article text from the news organization in the correct state. Each one of the fields listed in Table 1 has its series of operations that must complete. Special characters and stop words were removed from article titles and text as they are not useful for examining words and sentence meaning. We remove numerical values in the text unless they were attached to proper nouns or represented specific dates. In the cases where data met this criterion, we transformed the values into words representing the numbers.

We were interested in examining news article text, so we opted to keep words that fit into parts of speech, which are useful for detecting articles that could cause worry and anxiety for the readers. These parts of speech include nouns and proper nouns, adjectives, adverbs, and verbs. These selections are based on research as it applies to the types of objects that caused anxiousness and worries in the patients studied [15]. If some of these items were broken down into neuro-linguistic stimuli, the items could be best captured in language by their parts of speech. We experimented with these parts of speech by removing all words that do not belong to the parts of speech chosen. For example, some patients displayed signs of anxiety when they saw threatening text based on specific trigger words. These words can come in the form of any of the parts of speech we have outlined above.

All words were tokenized and lemmatized in order to get rid of repetitive representations of a specific word phrase. We did not want the algorithm to double count any of the words that we represent to the network so as to ensure we get the meaningful representation of the text of the news event. All punctuation was removed for this analysis, as well.

For word representation, we made use of term weighting [24] in order to demonstrate the importance of words to the semantic network. Finally, Latent Semantic Analysis [8] was used in order to get the final representation of our article into our network. Because of the high dimensional nature of text data, we followed this process so as to not overload the entire network with data that was not relevant for our calculations.

3.3 Collecting Reader Preferences Through Calibration Sessions

To account for the subjectivity involved in anxiety triggers, we designed what we call a calibration session with participation from the reader. The reader was responsible for responding to the system in a dialogue to establish an appropriate threshold that he or she deemed useful for anxiety trigger detection. Taking a recommendation approach based on user-guided critiquing is a well-researched field [12]. Figure 2 shows a detailed flow of how our calibration session works.

During the session, the network presented the reader with a specified random sample of article headlines. The reader's job was to answer whether or not the story headline caused them anxiety by responding yes or no to the network. The calibration session algorithm carried out the data collection process without the reader having to navigate to the site in question, and the feedback given from the reader created the construction of the GAD profile within the network. The network tracked the articles that the reader flagged as headlines that caused a feeling of anxiety. The network also tracked the articles where no such indication was made. The network used these responses during the execution of the network utility function.

3.4 Quantifying Article language associations by proximity metric

Every article in G is represented by a node A_{ij}. A link is established with each article $a \in A$ representing the language captured in both articles a_i, a_j. The links in our network are considered symmetric and un-directed. Let A_i represent one node, and A_j represents a different node, we consider the pair (i, j) as representing the links between the article one and article two. We describe the relationship between articles with the cosine distance metric defined in Equation 3.

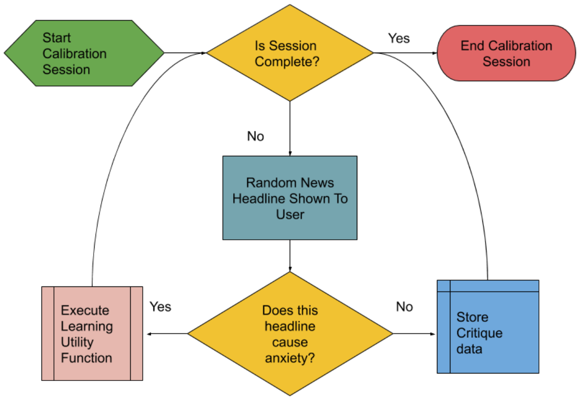
Start Calibration Session
Is Session Complete?
Yes
End Calibration Session
No
Random News Headline Shown To User
Execute Learning Utility Function
Yes
Does this headline cause anxiety?
No
Store Critique data

Fig. 2. Calibration Session

$$d_{ij} = \frac{1 - A_i A_j}{(\parallel A_i \times A_j \parallel)} \tag{3}$$

We found the cosine distance calculation to perform the most reasonably for the problem as compared to other metrics like the Pearson correlation coefficient [5] and Jiccard similarity [22]. We use d_{ij} as the length or distance between articles. Using this proximity metric allows the network to establish a representative relationship between each article as soon as they come in from the AP news wire. Therefore our network of article links can be represented by $A_{ij} = d_{ij}$.

3.5 The Network Utility Function

In order for the network to be able to detect triggers appropriately, a useful component is the network utility function. The network utility function's entire purpose is to locate articles or nodes which are close to or in the same neighborhood as the articles the reader identified were anxiety-inducing in the calibration session. The network search is designed to detect articles with the highest weighted degree. The weighted degree in the context of our semantic network is the number of words or phrases shared by a set of articles. During the session, the reader has identified a sampled set of the type of problematic news stories, and we capture those stories within the Semantic Network as they exist within their respective localities. These selected articles make up what we call the GAD profile, which we discuss in section 4.2.

Our knowledge Graph **G** contains the entire group of articles. These articles naturally form communities based on the relationship between the articles and their semantic network structure. These relationships appear as clusters in the network. We define our clustering coefficient in equation 4 as follows:

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{4}$$

Where C_i represents our local clustering coefficient for each of the connected articles, and L represents the number of links shared by the articles within the neighborhood k_i of articles, which are considered problematic. We leverage the local clustering coefficient as well as the global clustering coefficient of the entire network in order to understand what the natural boundaries are between the articles. These boundaries are partitioned into module classes. We describe these module classes in the analysis section 4.2. In Equation 5 we show our filtering process where a_ω is the weighted degree within the network for a specific article. The network utility function utilizes these weighted degrees to know the connectivity within each cluster C_i:

$$f \leftarrow \begin{cases} keep\ if\ a_\omega < \beta\ \forall a \in C_i \\ drop\ if\ a_w > \beta\ \forall a \in C_i \end{cases} \tag{5}$$

The threshold level maximum barrier for an article which is represented by β in equation 6 quantifies the level a article must not rise above if it is to remain outside the GAD profile. β is essentially our weighted degree threshold within the network and can be adjusted to find the optimal support level representation in the network. We summarize the Network utility function in the Algorithm 1 section.

Algorithm 1 Network Utility Function

```
procedure UPDATE($G, \beta, r$)                    ▷ G knowledge graph, r rounds
    $r \leftarrow set\,by\,user$                    ▷ Number of rounds in the session
    $f \leftarrow filtering\,function$              ▷ Removes Articles from G
    $\beta \leftarrow 115.434$                      ▷ Max. Threshold Representing Distribution
    while $i < r$ do                               ▷ Until Session Exit
        $\varphi \leftarrow calibsession(G, f(\beta))$    ▷ See Figure 2
        $i += 1$
    end while
    return $\varphi$                               ▷ The GAD Profile
end procedure
```

Understanding what the threshold level β is for the entire network gave us insight into the degree to which each article in the network is related to the other within their respective modularity class. This relationship is essential as it allows for a deeper understanding of how anxiety-provoking article language differs from non-anxiety-provoking articles in the network. We captured the relationship on an individual level giving readers the ability to understand what type of information is unhealthy for them to read. Analysis of the relationships between the two modalities of article types in the network can also help researchers discover the nuances in the troublesome events in the news and how they compare to one another in terms of triggering anxiety.

4 Analysis

Our network consists of 1,056 Nodes with 538,157 edges. We visualize our network in Figure 3 below. The semantic network is represented as a un-directed graph. The network links are represented by the distance between articles. Representing the news in this format allows us to understand the community of article clusters that naturally emerge based on their relationship to one another, which in turn is based on the language within the articles. In Figure 3, the article nodes are shown in a circular layout. The outer edges of the network show a large separation for those articles when compared to the articles closer to the center. In short, Figure 3 is what AP news looks like visualized as a Semantic Network.

Articles are green as all articles are assumed to be safe for the reader initially. However, after the calibration session begins, the network updates itself accordingly in order to keep current with the preferences of a specific reader.

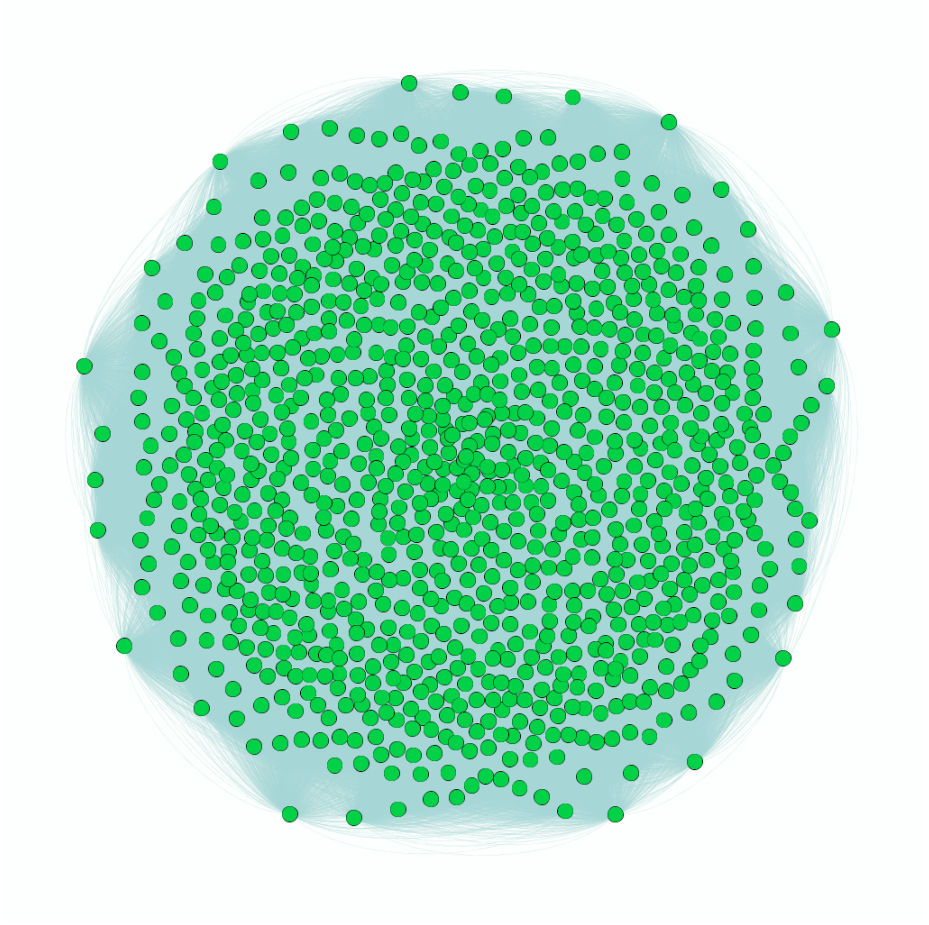

Fig. 3. AP News articles represented as a Semantic Network

In Table 2, we show a few network statistics for the AP News network. It is important to note that the network diameter will always be two, as each article will connect to at least one other article regardless of how far apart each article is from one another.

Table 2. AP News Network Statistics

Metric	Value	Description
Avg. Weighted Degree	115.434	Distribution of degree across Articles β
Communities	114	Number of groups or classes detected
Modularity	0.038	Suggests random connectivity

The average weighted degree, which we define in equation 6 demonstrates how many word phrase connections a specific article has if that article was selected. We see that any given article would have somewhere in the neighborhood of at least 115 connections. The weighted degree is significant, considering we are looking for a network that can detect triggers appropriately for individuals with GAD. If a reader selects one of the articles during his or her calibration session, we would expect the search space for neighboring articles to use this value as the threshold. In this case, we weight our network degree by the distance between the articles to reflect the genuine relationship between articles.

$$\beta = \frac{1}{A}\sum_{i=1}^{A}\omega_i d_{ij} \tag{6}$$

In Figure 4, we see our size distribution of each of our 114 communities detected by our network. A pattern emerges for the majority of the articles in the network. Many of the article node sizes are within the range of 5 to 10. The node size range suggests that many of the articles in the network will not affect many of the other articles when filtered during the calibration session with the reader. However, there are a few extreme values that carry substantial node sizes. If the reader selects these nodes, we expect to see a large number of neighboring nodes also get filtered due to their membership within the same modularity class. Having an evenly spread size distribution across the network allows for better filtering results from the network. In addition to an evenly sized spread, we also see based on our modularity value of 0.038, which is close to 0, the degree distribution in the network thoroughly explains the connections.

4.1 The GAD Profile and the AP News Semantic Network

As mentioned earlier, the semantic network must utilize individual feedback from the reader. With the feedback from the Calibration Session introduced in Figure 2, the network must update itself based on reader responses. The collection of reader responses is assembled into a GAD Profile. Figure 5 shows the updated

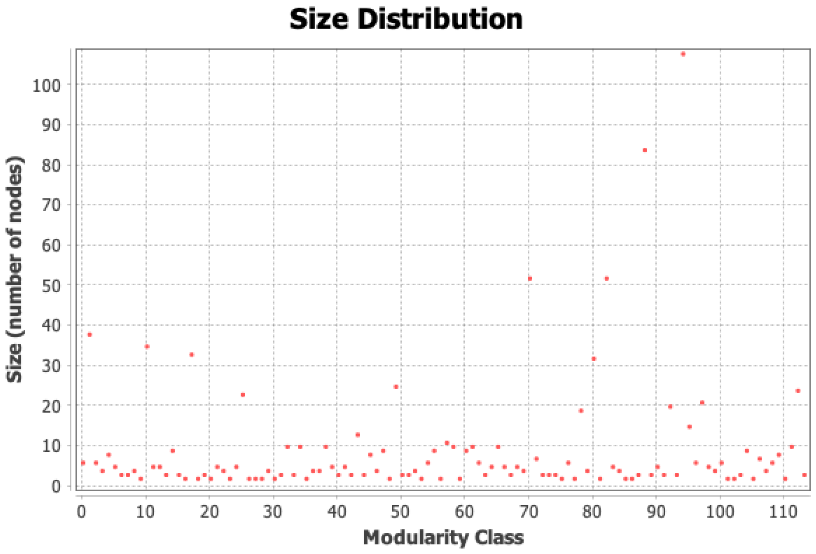

Fig. 4. Modularity and Communities within AP News Network

Semantic network after a reader enters their responses during the calibration session.

Figure 5 shows a user's interaction with the network during a calibration session. We can see the test user's selected articles in Figure 5 demonstrates that there is no clear pattern that describes the reader's anxiety when viewed in isolation. The user's selections are clearly in multiple areas throughout the network. However, these selected articles provide the network with a useful starting point in determining which articles are problematic based on the responses. It is also important to note that only 8 articles were selected during the user's calibration session as this represents a small subset of articles shown during the calibration session.

4.2 The Selection Process

GAD profile construction gathers together a list of articles that a specific individual believes is a contributing factor to his or her own anxiety. Anxiety is not a one size fits all phenomena. Any approach that seeks to help with the treatment of anxiety for any reader would need to account for the distinct nuances between individuals. Mathews and Hirsch [16] points to research which indicates worry itself differs between those who have personalized experiences of an event. We cannot identify each individual's edge case for anxiety without the user's input. Therefore the readers feedback to the network is essential for producing sensible results from the network. In order to factor in these nuances between individuals,

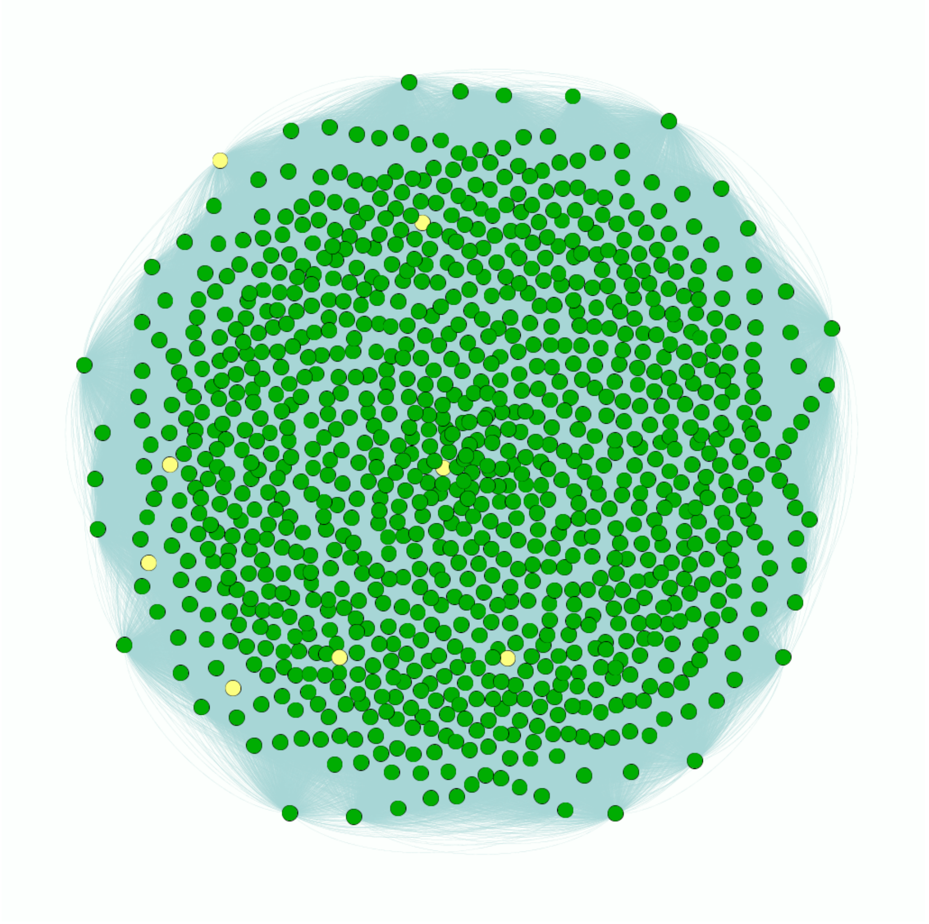

Fig. 5. Flagged Articles During a Test User's Calibration Session

the reader creates their critiques within the calibration session, as illustrated in Figure 5.

Network Modularity classes were a useful filtering mechanism for the accounting of personal anxiety indicators as articles start to form detectable communities within the network. We saw in Figure 4 that the modularity class sizes looked to be evenly distributed across node sizes. Therefore, we make use of Equation 7 shown below in order to determine our partitions in the network:

$$M_{class} = \sum_{class=1}^{n_{class}} \left[\frac{L_{class}}{L} - \left(\frac{k_{class}}{2L} \right)^2 \right] \tag{7}$$

Where k represents the number of articles within the same community, and L represents the links between the articles. Our classes represent the number of classes detected for each community, which we visualized in Figure 4. In short, modularity classes became essential in determining the boundaries for which stories should be selected for the profile, and which ones should be kept out of the profile. In Figure 6, the grey nodes are news events that should not be included in the profile, while the red nodes contain articles that should be a part of the profile based on the calibration session.

Figure 6 shows the problematic news events and their various connections. Also of interest is the number of articles that did not get selected by the network. Only 10.22% of the articles are selected for the GAD profile. We tracked the number of selected articles in the network in order to understand the proportion of articles that the network is not surfacing for a specific individual. Tracking this metric is of significant importance. If a mental health research organization decides on using this approach, it can determine how many stories are found to be problematic by tracking this ratio over time.

4.3 Understanding The GAD Profile Results

If we examine the articles selected by the network, a pattern between the articles which were selected emerges. Our assumption is that the network of selected articles serves as a representation of a reader's personal anxiety profile as it applies to news content. Figure 7, shows the articles which the network has surfaced. A close examination of these articles is shown in the study results in Table 3. We hypothesize that the headlines in Table 3 might contain triggers for those with depression or anxiety as the presented articles represent the most controversial news content which the network selected for a test user.

The results in Table 3 can and will vary according to the construction of the GAD profile and a reader's feedback. If we examine the article titles, we can see many of them have an associative pattern. For example, many of the articles deal with killings, sexual offenses, and assault, while others deal with issues of climate change and politics. The selected articles from the network can serve as useful data for the Network Data Science community as well as the mental health research community in order to better understand what news events are affecting each individual and why. We could then leverage this data to come up

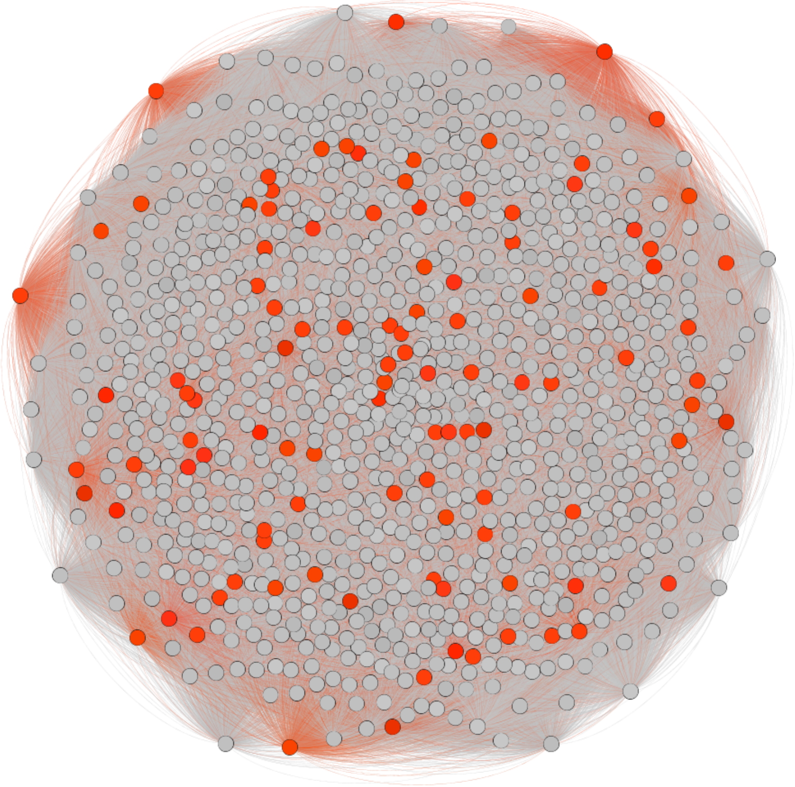

Fig. 6. Detected News Events (in RED) based on the Test User's Calibration Session

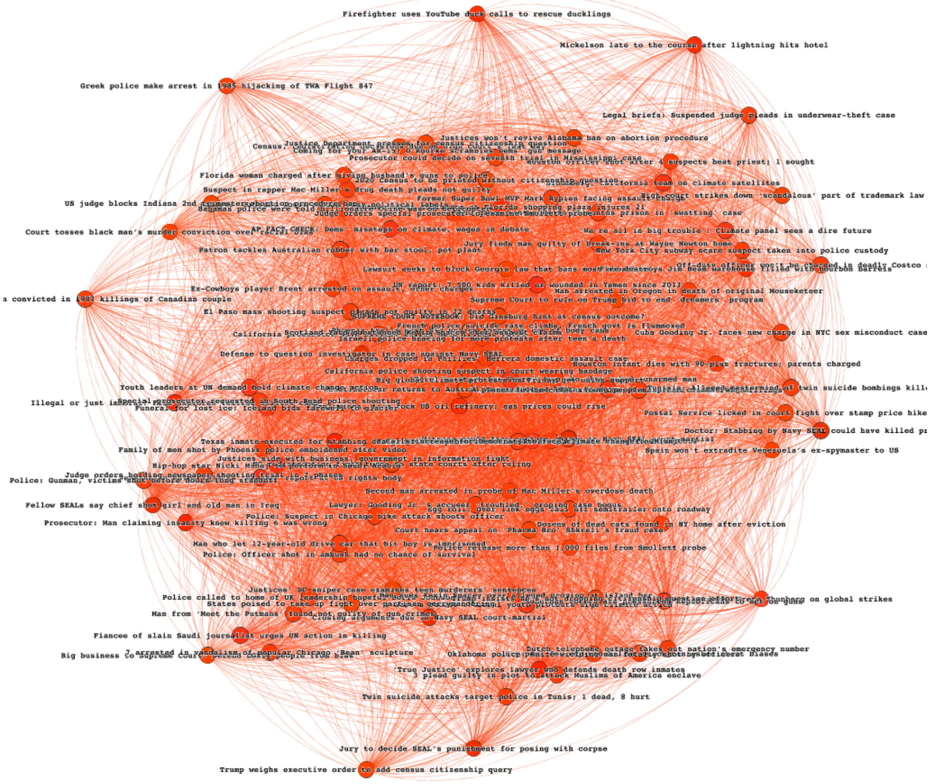
Firefighter uses YouTube duck calls to rescue ducklings
Mickelson late to the course after lightning hits hotel
Greek police make arrest in 1985 hijacking of TWA Flight 847
Legal briefs: Suspended judge pleads in underwear-theft case
Justices won't revive Alabama ban on abortion procedure
Justice Department presses for census citizenship question
Coming for your AR-15? O'Rourke scrambles Dems' gun message
Prosecutor could decide on seventh trial in Mississippi case
Houston officer shot after 4 suspects beat priest; 1 sought
Florida woman charged after giving husband's guns to police
2020 Census to be printed without citizenship question
Bloomberg, California team on climate satellites
Suspect in rapper Mac Miller's drug death pleads not guilty
High court strikes down 'scandalous' part of trademark law
US judge blocks Indiana 2nd trimester abortion procedure ban
Former Super Bowl MVP Mark Rypien facing assault charge
Bahamas police were told billionaire Cline was on helicopter
Judge orders special prosecutor to examine Smollett probe
Court tosses black man's murder conviction over racial bias
AP FACT CHECK: Dems' missteps on climate, wages in debate
'We're all in big trouble': Climate panel sees a dire future
Jury finds man guilty of break-ins at Wayne Newton home
Patron tackles Australian robber with bar stool, pot plant
New York City subway scare suspect taken into police custody
Off-duty officer won't be charged in deadly Costco
Lawsuit seeks to block Georgia law that bans most abortions
Fire destroys Jim Beam warehouse filled with bourbon barrels
UN report: 7,500 kids killed or wounded in Yemen since 2013
Ex-Cowboys player Brent arrested on assault, other charges
Man arrested in Oregon in death of original Mouseketeer
convicted in 1987 killings of Canadian couple
Supreme Court to rule on Trump bid to end 'dreamers' program
El Paso mass shooting suspect pleads not guilty in 22 deaths
SUPREME COURT NOTEBOOK: Did Ginsburg hint at census outcome?
French police suicide rate climbs, French govt is flummoxed
Cuba Gooding Jr. faces new charge in NYC sex misconduct case
Israeli police bracing for more protests after teen's death
Defense to question investigator in case against Navy SEAL
Charges dropped in Phillies' Herrera domestic assault case
Houston infant dies with 90-plus fractures; parents charged
California police shooting suspect in court wearing bandage
Youth leaders at UN demand bold climate change action
Tunisia: Alleged mastermind of twin suicide bombings kill
Special prosecutor requested in South Bend police shooting
Illegal or just
Funeral for lost ice: Iceland bids farewell to glacier
Postal Service licked in court fight over stamp price hike
Doctor: Stabbing by Navy SEAL could have killed pr
Texas inmate executed for stabbing death
Family of men shot by Phoenix police emboldened after video
Spain won't extradite Venezuela's ex-spymaster to US
Justices side with business, government in information fight
Hip-hop star Nicki Minaj to perform in Saudi Arabia
state courts after ruling
Judge orders holding newspaper shooting trial in 2 phases
Police: Gunman, victims shot before hours-long standoff
Second man arrested in probe of Mac Miller's overdose death
Fellow SEALs say chief shot girl and old man in Iraq
Lawyer: Gooding Jr.'s accuser 'troubled,' groping case bogus
Egg roll: Over 136K eggs fall off semitrailer onto roadway
Police: Suspect in Chicago bike attack shoots officer
Prosecutor: Man claiming insanity knew killing 6 was wrong
Dozens of dead cats found in NY home after eviction
Court hears appeal on 'Pharma Bro' Shkreli's fraud case
Man who let 12-year-old drive car that hit boy is imprisoned
Police release more than 1,000 files from Smollett probe
Police: Officer shot in ambush had no chance of survival
Justices' DC sniper case examines teen murderers' sentences
Police called to home of UK leadership hopeful
States poised to take up fight over partisan gerrymandering
Thunberg on global strikes
Man from 'Meet the Putmans' found not guilty of gun crimes
Closing arguments due in Navy SEAL court-martial
Fiancee of slain Saudi journalist urges UN action in killing
7 arrested in vandalism of popular Chicago 'Bean' sculpture
Big business to Supreme Court
Dutch telephone outage takes out nation's emergency number
Oklahoma police
'True Justice' explores lawyer who defends death row inmates
3 plead guilty in plot to attack Muslims of America enclave
Twin suicide attacks target police in Tunis; 1 dead, 8 hurt
Jury to decide SEAL's punishment for posing with corpse
Trump weighs executive order to add census citizenship query

Fig. 7. Detected News Events which contain anxiety triggers based on User's Calibration Session

Table 3. Anxiety provoking news events for test user with highest weighted degree.

Detected Article	Class in Fig. 4
UN report: 7,500 kids killed or wounded in Yemen since 2013	52
Venezuelan envoy rejects biased report at UN rights body	52
Fiancee of slain Saudi journalist urges UN action in killing	52
Former Super Bowl MVP Mark Rypien facing assault charge	70
Rocker Cliff Richard urges anonymity for sex-crimes suspects	70
Ex-Cowboys player Brent arrested on assault, other charges	70
Houston officer shot after 4 suspects beat priest; 1 sought	70
Ohio gamer sentenced to 15 months prison in swatting case	70
Jury pool for cop who killed black man asked about biases	70
Jury finds man guilty of break-ins at Wayne Newton home	70
Houston infant dies with 90-plus fractures; parents charged	70
Judge orders holding newspaper shooting trial in 2 phases	70
Florida woman charged after giving husbands guns to police	70
El Paso mass shooting suspect pleads not guilty in 22 deaths	70
Special prosecutor requested in South Bend police shooting	70
Police release more than 1,000 files from Smollett probe	70
Cuba Gooding Jr. faces new charge in NYC sex misconduct case	70
Hungary: Death toll in Danube River boat crash rises to 27	70
Twin suicide attacks target police in Tunis; 1 dead, 8 hurt	70
Dozens of dead cats found in NY home after eviction	70
Dems missteps on climate, wages in debate	32
Climate activist Greta Thunberg on global strikes	32
Youth leaders at UN demand bold climate change action	32
Funeral for lost ice: Iceland bids farewell to glacier	32
A planet full of ifs: Young people express climate angst	32
Bloomberg, California team on climate satellites	32
Big global climate protests on Friday get union support	32
Calls increase for Democrats to face climate change in Miami	32
Fellow SEALs say chief shot girl and old man in Iraq	2
Closing arguments due in Navy SEAL court-martial	2
Witness could face perjury charge in Navy SEAL court-martial	2
Jury to decide SEALs punishment for posing with corpse	2
Doctor: Stabbing by Navy SEAL could have killed prisoner	2
Defense to question investigator in case against Navy SEAL	2
Illegal or just immoral? Film explores texting suicide case	69
Texas inmate executed for stabbing deaths of 2 stepsons	69
Prosecutor: Man claiming insanity knew killing 6 was wrong	69
‘True Justice explores lawyer who defends death row inmates	69
Explosions, fire rock US oil refinery; gas prices could rise	11
Mickelson late to the course after lightning hits hotel	11
Authorities: Explosion at Florida shopping plaza injures 21	11
Firefighter uses YouTube duck calls to rescue ducklings	11
Fire destroys Jim Beam warehouse filled with bourbon barrels	11
Coming for your AR-15? ORourke scrambles Dems gun message	9
Suburban voters are pressuring Republicans to act on guns	9
Justices DC sniper case examines teen murderers sentences	25
Trump weighs executive order to add census citizenship query	25

with better techniques which can help readers deal with handling the uncertainty surrounding specific news events as they unfold.

4.4 Discussion

In summary, our approach to anxiety detection can produce useful data which can help us further understand the relationship between anxiety and news events in general. The preliminary results suggest that an algorithmic approach to detecting anxiety triggers in news event content is achievable by careful attention to the nuances of the language in that content. Our approach attempts to locate harmful topics for anxious or depressed individuals. The algorithm attempts to detect only those stories which might have the most impact on the negative psychological state of the reader. In theory, this approach can be extended beyond news feeds and into social media forums as well.

For news organizations, having a product designed for readers suffering from GAD or other mental health disorders can allow for a more personalized experiences for the millions of Americans who are currently diagnosed with GAD. In order to successfully offer such a feature, the news organization would need to find secure ways of protecting GAD profiles and privacy information of readers. For example, generic GAD profiles can be created on the reader's system, which is encrypted locally. The local GAD profile should become readable only by the network unless otherwise stated by the user so that the filtering operation can be successfully carried out in private. The reader could then be given the option to share the profile with organizations if the organization used the profile to help them reduce their anxiety. Many readers might not feel comfortable expressing issues that make them feel anxious with an outside organization that is not a mental health care center. As a result of this possible reluctance, a serious effort towards anonymizing GAD profiles and keeping them in control of the individual must be considered standard practice for the approach we outline as the algorithm can uncover triggers a reader finds inappropriate to share.

For this type of approach to anxiety trigger detection, there are a good number of ethical and legal considerations, many of which cannot fully be resolved but can be addressed nonetheless. Ethical and legal considerations apply to the usage and gaining a profit from content and materials published and copyrighted by third-party sources. Our method can be compared to a specialized search engine, in other words showing articles from websites can potentially bring in more patrons or readers. Rupert Murdoch, chairman of News Corporation at the time of this writing, has mentioned that with the production of journalism being expensive and a major investment, aggregation of the news is not fair use, and described it as theft [17]. To sustain any solution on the internet, there needs to be a business benefit. As its most basic function, a news aggregator pulls information from multiple sources into one site, pulling valuable internet traffic towards its site. A news aggregator can exploit this by means of lucrative online advertising. In 2005, a prominent wire service Agency France Presse (AFP) brought up a suit against Google citing that the headline, lead, and photo displayed was an infringement of their copyrights[17]. This led to a settlement and license agreement between the two companies. Though the implementation and interpretation of the US or other countries copyright laws can be arduous and convoluted, news aggregators use of third party content generally falls under fair use [17]. In the article: *4 guidelines for aggregating news content* from the website pointer.org, guidelines for using online articles are as follows: [2]

1. Publish just the content to identify the story, such as the headline or excerpts, but not the full story or text.
2. Identify the source of the information.
3. Clearly link to the original source of the information.
4. Clearly identify what is being provided.

Ultimately, a news aggregation service must be careful how it handles third party content or risk being sued.

In our approach we are using a random sample of the available news articles by reading data that has already existed on the AP news website. This opens up the inherent issue of suffering from bias, which as Data Scientists, we must try to be aware of and mitigate. The inherent biases include subjectivity bias, source bias, convenience sampling.

As detailed, the algorithm utilized must surface articles that are deemed as problematic according to an individual's GAD profile. Though there are general considerations by the algorithm in making this decision, ultimately, what can be deemed as anxiety-inducing is inherently subjective, and can differ amongst one person to another. To counter this type of bias, we used the calibration sessions to handle selection bias for filtering, which removes our own biases as much as possible.

Due to the sheer volume of articles and information, any news related algorithm can fall prey to source bias. Attempting to include all internet news sources from all regions or countries is arduous and out of the scope or capability of this effort. Due to this, there is an inherent source bias that may include or exclude specific news outlets or genres focusing on liberal, conservative, religious, secular, or other types. Associated Press (AP) wire news service, which is an award-winning not-for-profit and cooperative news association, proved to be most practical for our specific experiment.

Another form of bias can occur due to convenience sampling, where only articles collected during a set period are used. There may be instances where news from the morning or the night may be more or less favorable. As our approach is more of a proof of concept, this type of bias cannot be currently mitigated. If this approach is used in a production setting, articles would be collected over a more extended period.

5 Conclusions

As demonstrated in this study, we were able to provide data based on our detection methodology that can possibly aid in a deeper understanding of anxiety

[2] https://www.poynter.org/educators-students/2017/4-guidelines-for-aggregating-news-content/

as it applies to news readers, particularly ones with symptoms or having signs of Generalized Anxiety Disorder (GAD). To sum up, our approach succeeded in being able to collect data from the AP news and pre-process it accordingly. We summarized all articles by leveraging dimensionality reduction with the help of Latent Semantic Analysis. We represented all news content as a Un-directed Semantic Network, where the language relationship between the articles was represented as the links and the articles as nodes.

In conclusion, we present a novel approach that attempts to offer up a new way to collect data which helps a reader uncover their anxiety triggers contained in news events. Quantifying the relationship between anxiety triggers and news content would be a beneficial development for sufferers of Generalized Anxiety Disorder (GAD) and other mental health disorders. We mitigated the bias associated with pre-labeling or pre-determining what is positive or negative news by opting for user-defined profiles constructed from calibration sessions with the reader.

Possible future work for this approach may be to include enhancing reader profiles by optimizing the β parameter to find the ideal threshold based on various optimization algorithms as the network locates which articles are added to a reader's profile. Additionally, increased testing of the network by soliciting for many more volunteers that would give greater independent feedback on how well the network works for different GAD profiles. The ultimate production goal will be to determine if the selections during the detection process help readers understand their anxiety triggers as their profile is constructed over an extended period of time.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of modern physics **74**(1), 47 (2002)
2. Association, A.P., et al.: Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub (2013)
3. Bastian, M., Heymann, S., Jacomy, M.: Gephi: An open source software for exploring and manipulating networks (2009), http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154
4. Beck, A.T., Clark, D.A.: An information processing model of anxiety: Automatic and strategic processes. Behaviour research and therapy **35**(1), 49–58 (1997)
5. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)
6. Börner, K., Sanyal, S., Vespignani, A.: Network science. Annual review of information science and technology **41**(1), 537–607 (2007)
7. Cano, P., Celma, O., Koppenberger, M., Buldu, J.M.: Topology of music recommendation networks. Chaos: An Interdisciplinary Journal of Nonlinear Science **16**(1), 013107 (2006)
8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
9. Di Nardo, P.A., Barlow, D.H.: Anxiety disorders interview schedule–revised (ADIS-R). Phobia and Anxiety Disorders Clinic, Center for Stress and Anxiety Disorders (1988)
10. Dutta-Bergman, M.: Depression and news gathering after september 11: The interplay of affect and cognition. Communication Research Reports **22**(1), 7–14 (2005)
11. Essau, C.A., Ollendick, T.H.: The Wiley-Blackwell handbook of the treatment of childhood and adolescent anxiety. John Wiley & Sons (2012)
12. Felfernig, A., Burke, R.: Constraint-based recommender systems: Technologies and research issues. In: Proceedings of the 10th International Conference on Electronic Commerce. pp. 3:1–3:10. ICEC '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1409540.1409544, http://doi.acm.org/10.1145/1409540.1409544
13. Freeston, M.H., Rhéaume, J., Letarte, H., Dugas, M.J., Ladouceur, R.: Why do people worry? Personality and individual differences **17**(6), 791–802 (1994)
14. Goldberg, D., Bridges, K., Duncan-Jones, P., Grayson, D.: Detecting anxiety and depression in general medical settings. Bmj **297**(6653), 897–899 (1988)
15. Goodwin, H., Yiend, J., Hirsch, C.R.: Generalized anxiety disorder, worry and attention to threat: A systematic review. Clinical Psychology Review **54**, 107–122 (2017)
16. Hirsch, C.R., Mathews, A.: A cognitive model of pathological worry. Behaviour research and therapy **50**(10), 636–646 (2012)
17. Isbell, K.A.: The rise of the news aggregator: Legal implications and best practices. Berkman Center Research Publication **2010**(2010-10), 1,6 (2010)
18. Ladouceur, R., Dugas, M.J., Freeston, M.H., Léger, E., Gagnon, F., Thibodeau, N.: Efficacy of a cognitive–behavioral treatment for generalized anxiety disorder: Evaluation in a controlled clinical trial. Journal of consulting and clinical psychology **68**(6), 957 (2000)
19. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical review E **69**(2), 026113 (2004)

20. Newman, M.G., Llera, S.J., Erickson, T.M., Przeworski, A., Castonguay, L.G.: Worry and generalized anxiety disorder: a review and theoretical synthesis of evidence on nature, etiology, mechanisms, and treatment. Annual review of clinical psychology **9**, 275–297 (2013)
21. Portman, M.E.: Generalized anxiety disorder across the lifespan: An integrative approach. Springer Science & Business Media (2009)
22. Real, R.: Tables of significant values of jaccard's index of similarity. Miscel· lania Zoologica **22**(1), 29–40 (1999)
23. Regier, D.A., Rae, D.S., Narrow, W.E., Kaelber, C.T., Schatzberg, A.F.: Prevalence of anxiety disorders and their comorbidity with mood and addictive disorders. The British Journal of Psychiatry **173**(S34), 24–28 (1998)
24. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
25. Staudt, C., Sazonovs, A., Meyerhenke, H.: Networkit: An interactive tool suite for high-performance network analysis. CoRR **abs/1403.3005** (2014), http://arxiv.org/abs/1403.3005
26. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology **29**(1), 24–54 (2010)
27. Weeks, B., Southwell, B.: The symbiosis of news coverage and aggregate online search behavior: Obama, rumors, and presidential politics. Mass Communication and Society **13**(4), 341–360 (2010)
28. Zanker, M., Jessenitschnig, M., Schmid, W.: Preference reasoning with soft constraints in constraint-based recommender systems. Constraints **15**(4), 574–595 (2010)