Proposal

Team CYB3RL4NG

Class of 2026

Oliver D'Esposito, Tzipporah Harker, Terranova Oh, Akhilesh Puranam, Archana Sathiyamoorthy, Yash Thakur

Team CYB3RL4NG pledges on our honor that we have not given or received any unauthorized assistance on this paper.

Table of Contents

Defining Terms

The terms will be used later in the paper:

- Reddit: A forum style social media platform intended for asking questions, where users communicate through public threads. It is organized into topic-specific subreddits which form “communities” and have their own unique social rules and jargon.
- Semantic: Meaning of words and phrases
- Pragmatic: Meaning of a word or phrase in context; the intention of a statement.
- Syntax: Meaning by relationships between words; the structural component of meaning.
- Variable reinforcement: concept in behavioral psychology where positive reinforcement is provided after a varying number of performances of a certain behavior (i.e. slot machines–there is no way of knowing how many rounds of using the machine will result in a jackpot; therefore the schedule of reinforcement is variable)
- Polysemy: linguistic property where a word has multiple meanings or can be used in multiple contexts
- Polar Sensitivity Item (PSI): Parts of speech that exclusively reflect extremes of emotion. Can be negative or positive.
- Open coding: qualitative practice of making coding categories based on data, also referred to as ‘bottom-up’ coding
- Systematic coding: qualitative practice of coding data used predetermined categories, also referred to as ‘top-down’ coding

Abstract

Our team aims to investigate the relationship between the usage of mental health jargon and its evolving semantic context by using a cross modal analysis program that utilizes both word embeddings and comprehensive statistical analyses. We posit that the growth of informal communication on the internet due to the environment created by social media platforms has proliferated semantic change, encouraging broadening, particularly tonal shifts. With this change, jargon previously contained to niche online communities – notably socio-cultural topics such as mental health issues – have become increasingly mainstream.

A popular methodology to track semantic change involves generating word embeddings from data derived from a specific online community. These embeddings are then analyzed and compared to a predetermined baseline meaning through n-grams and frequency analysis.

Our research aims to build on prior work with word embeddings by applying novel methods of comprehensive statistical analysis to allow us to better understand the collacative and concordative elements of semantic change within the context of online communities and online communication.

Chapter 1: Introduction

Previous research on the evolution of language has focused predominantly on spoken language under the assumption that linguistic change occurs through speech and that written text is only a reflection of speech; a literate society theoretically does not interact with language any differently than a non-literate society [1]. This is not necessarily true – not to say that non-literate societies are less advanced or are "culturally behind" – but a way of saying that having two formats for communication causes speakers of a language to have two different approaches to how they communicate within the same framework. A seemingly reasonable comparison to this is fingerspelling in sign language, or spelling out letter-by-letter phrases in the language associated with one's speaker community. However, signed languages are not just codifications of a spoken language, they have their own complex syntactic structure, pragmatic schemes, semantic systems, etc. and the process of fingerspelling could better be compared to bilingualism [2]. What we are seeing with written communication is an entirely different phenomenon. On the internet, where messages can be transmitted almost instantaneously – but not quite – and mixed media elements such as stylistic spellings and emoticons can be incorporated into utterances, there is a pseudo-environment ultimately failing to replicate the conditions of spoken communication [1].

The informality and expansiveness of the internet has resulted in an increase of writing, reading, and sharing of informal text and with this influx of text we are seeing rapid acceleration in language [1]. Previous literature has highlighted the functional differences between online discourse and spoken communication. There are inherent time delays that occur while communicating via texting or emailing which are significant as they prevent real-time parsing – the process of reading a sentence is not equivalent to the process of listening to someone produce

speech because it is possible to skim [3]. While texting, emailing, and other digital methods of communication cannot be perfectly equated to speaking, we argue that it is a misconception that they are entirely irrelevant to the study of language and that online communication has many potential insights on patterns in spoken semantic and syntactic change.

With the growing phenomenon of internet social justice and the rise of cancel culture, as well as the use of online platforms to disseminate medical knowledge [4], the colloquial use of clinical terms has caused mental health jargon to be bastardized [5]. This appropriation extends to both the derogatory use of terms to describe mental illness and their application in other contexts unrelated to mental health.

This paper will focus primarily on horizontal and vertical semantic broadening, or the process of a word's meaning expanding to include related concepts or ideas from within the same general category and the process of a word's meaning expanding to encompass broader, categorically distant, or more abstract concepts, respectively [6]. We will be looking at semantic broadening applied specifically to mental-health jargon on the internet, observing how the expansion of meaning leads to a perceived desensitization. In an increasingly literate society where online communications are the norm, internet language simply cannot be ignored. Through this examination we hope to provide insight on the greater role of the internet in contemporary language usage.

Chapter 2: Literature Review

2.1. The Internet and the Norm of Overstimulation

The internet broadly refers to the social and informational network shared electronically through computers, phones, tablets, and other digital devices [7]. The interconnected and ever-present nature of the internet has created a vast network of fast, easily accessible information. Particularly among younger generations, social media has become the culturally preferred method of communication and acquiring information [8]. Social media provides a constant flow of information, and subsequently a constant flow of stimuli. Psychologists have documented the behavioral phenomenon of variable reinforcement on social media, sometimes referred to as the *slot machine effect*. Essentially, social media reinforces users with positive stimuli after varying amounts of time and use, prompting the user to spend more time on a given platform for the possibility of positive reinforcement [9].

The result of this is an environment characterized by overstimulation. Within the context of technology, overstimulation is commonly discussed by researchers as a physiological, sensory phenomenon [10], [11]. However, researchers like Bąk-Sosnowska and Holecki introduce the idea of information overstimulation in its relation to the physiological and psychological symptoms of stress and anxiety [12]. Some posit that to counter the effects of information overstimulation, users develop a tolerance for certain stimuli that results in a "blasé attitude" about their online experiences [13].

Yet the internet is still a space where users exchange views on exigent social or political topics. Health and medical topics are particularly common points of online discourse, especially in the wake of the COVID-19 pandemic [5], [14]. Mental health in particular tends to have a large, volatile presence in online discourse. Younger users frequently use the internet as a source

of mental health information and in many cases, social media serve as a destigmatizing platforms for users to discuss their own mental health issues [15], [16]. However, there are other, more negative manifestations of mental health discourse online. For example, the ‘depression’ or ‘anxiety’ hashtags have been used by content creators to boost post popularity, and in more extreme cases, online communities have been built around promoting psychological conditions and risk taking behavior, particularly eating disorders [17].

While there are resources online dedicated to providing informed, objective information that are maintained by those with mental health expertise, many unqualified internet users have a tendency to make statements regarding mental health without understanding the weight of them. Fang and Zhu observed this in a study they conducted regarding the use of stigmatizing mental health language on Twitter as a result of the highly publicized defamation trial between actors Johnny Depp and Amber Heard [4]. This trial included various allegations from both sides, including claims of one another exhibiting personality disorders, specifically borderline, histrionic, and narcissistic personality disorder. These accusations were heavily debated by online communities. Fang and Zhu found that during the course of the trial, as well as in the months after its conclusion, there was an increase in stigmatizing language particularly surrounding the aforementioned personality disorders, as well as a decrease in destigmatizing language. Their results highlight the possible effects that mainstream discussions of psychological disorders can have on stigmas surrounding mental illness, particularly when users do not truly understand the nature of the conditions that they discuss.

2.2. Semantic Broadening

The constant influx of new information and content, and the overstimulating nature of online discourse, create a ripe environment for the linguistic phenomenon of semantic broadening. Vertical semantic broadening, also known as semantic bleaching, is the process by which a word's meaning is expanded to encompass broader, categorically distant, or more abstract concepts. For example, Lou et al. explore this concept with adverbs such as 'awfully' and 'insanely' [18]. They posited that these adverbs, while originally having very distinct meanings, were semantically bleached until their primary use became as intensifiers, rather than as their own descriptive terms. In application to the discussion of mental health jargon, the adverb insanely originally meant 'in a way that suggests unsoundness or disorder of mind,' but through semantic bleaching, developed a secondary meaning of communicating extremity (i.e. insanely delicious, Lou et al.'s titular example) [19].

Horizontal semantic broadening is the process by which a word's meaning expands to include related concepts. For example, "traumatic" initially was used to communicate an emotionally disturbing or distress situation that causes emotional shock; now it is commonly used to describe an experience that is merely unenjoyable or marginally upsetting [20].

John Haiman proposed that the cause for language change is its repeated usage, which he termed "ritualization" [21]. He posited that words used repeatedly lose meaning through *habituation*, a phenomenon coined by behavioral psychologists to describe a decreased response to a stimulus due to repeated exposure. This theory of semantic broadening was expanded upon by Bybee and Thompson, who argued that repeated usage creates a "spiraling effect" where in turn words become less stimulating, thereby expanding their usage further [22]. With the heavily explored connection between behavioral reinforcement and social media use [9], it is likely that

an emphasis on stimulation in online spaces may heavily influence the ritualization and semantic evolution of terminologies accepted by users.

Other theories propose that broadening is caused by Polarity Sensitive Items (PSIs), words that are sensitive to the positive or negative implications of their context [23]. The following is an example in English [24]:

1)

(a) Glinda has not ever robbed a liquor store.

(b) * Glinda has ever robbed a liquor store.

Here, it's shown that "ever" exists freely in negative environments but is ungrammatical in affirmative statements; therefore, it is a negative polarity item. In Early Modern English, "ever" did not have a polarity distinction – referred to as Universal Polarity – but over time its usage has narrowed. In Contemporary Modern English there are a few applications of "ever" specifically as a positive polarity item, however these are heavily constricted by syntactic constructions.

2)

(a) Mr. Higgins, ever the connoisseur, was eager to try the wine.

(b) Even Mr. Hives, hardly (*ever) the connoisseur, was eager to try the wine.[1]

The preceding negative trigger "hardly" allows "ever" to carry a positive connotation.

Online, where people are able to participate in multiple niche communities, more novel, thematically varied utterances are generated and circulated than internet users would ever encounter in their isolated physical lives. When PSIs trigger alternative interpretations of established words and phrases they introduce the idea of a new meaning; this is particularly relevant to semantic bleaching where words become abstracted from their original meaning. This

[1] Though this construction is formally recognized as ungrammatical, some speakers may consider it to be acceptable.

combines Haiman's philosophy of ritualization and habituation with a consideration to how the polar qualities of semantic context can lead to broadening. Broadening is not just caused by how words are used in conversation and the frequency of exposure to terms but also the grammatical rules and patterns involved in language processing.

In an examination of concept creep, or the horizontal broadening of specifically harm-related concepts, Haslam proposed that the major factors in proliferating concept creep are cultural shifts in sensitivity to harm related concepts and societal changes in the presence of these phenomena; as a society becomes more aware and critical to different forms of harm, the conceptual boundaries of harm-related language expand to accommodate for this reanalysis [25]. Though not covered in this article, this idea of increased relevance to increased usage to the increased applicability of a lexical item could reasonably be applied to non-harm related concepts.

It is important here to note that while vertical and horizontal bleaching appear to be opposite applications of the same concept there is reasonable evidence that they are entirely different processes and are influenced by different linguistic factors. As it stands, vertical broadening is understood to be a result of variation in usage in a purely lexical sense and horizontal broadening from cultural outlooks; both are influenced by frequency of usage or perceived relativity.

Lauren Squires investigated an adjacent concept that she termed indexical bleaching, which refers to the bleaching of a term's contextual origins rather than its meaning [26]. She examined how popularization in mainstream media caused words that originally had contextual social implications to become "diffused" and broadened in their usage, even though their general

meaning remained the same. Her research suggests that mass media has the potential to influence the adoption and bleaching of words in the public sphere, and alter their implications in usage.

This concept is further explored by Andreea Calude et al., who investigated the broadening and semantic bleaching of the term “woke” though the examination of the hashtag “#WokeAF” on Twitter [27]. They begin by discussing the origins of the word “woke”, and how it began as a term used in African-American Vernacular English (AAVE) to refer to awareness about racial injustice. They identify key turning points that introduced the term to wider and wider populations: first, to a generally left-leaning population that used it in the context of any social injustice; then to a right-wing population who used it as a pejorative to criticize left-wing ideologies. They then mapped this process by sampling Twitter posts from between 2012 and 2022, graphing the frequency of the hashtag’s usage, and randomly sampling 50 posts from each year to examine manually. Calude et al. were able to identify the semantic bleaching of the “wokeAF" hashtag online by examining data from social media. Their research demonstrates that terms or phrases popularized on social media platforms have the potential to experience both vertical and horizontal semantic broadening.

2.3. Detecting Semantic Change

In recent years, most papers dealing with detecting semantic change in the meaning of words over time have done so using word embeddings, which are representations of words as vectors [28]. Word embeddings make it easy to quantify the semantic similarity between pairs of words, because one can treat the distance between two words’ embeddings as a measure of their semantic similarity.

The text we will be processing will contain words that are new and not included in any database containing information on words and their meanings. Therefore, we will use algorithms that generate word embeddings in an unsupervised manner. This can be done by assuming the distributional hypothesis, which is the idea that words that appear in similar contexts tend to have similar meanings [29]. Word embeddings can be made such that words that occur in similar contexts have vector representations that are close together. The goal is that words with similar meanings will have word embeddings that are close together. Here, "close" refers to cosine similarity, which is the cosine of the angle between two vectors. Once these embeddings are made, one can quantify the semantic similarity between two words by calculating the cosine similarity between their word embeddings.

It is important to note that words appearing in similar contexts don't always have similar meanings. An example of this is the words "hot" and "cold." Even though they are antonyms, both are used in similar contexts, so their word embeddings are close together [30]. Therefore, semantic similarity measures produced by comparing word embeddings do not correspond with similarity in their definitions. However, comparing word embeddings generated based on word distributions is a convenient way to find the semantic similarity between words in an unsupervised manner. For the purposes of this paper, we will treat word embeddings as representing the meanings of words rather than merely their distributions.

We considered three word embedding techniques: PMI, LSA, and skip-gram Word2Vec, also referred to as Skip-Gram with Negative Sampling (SGNS).

The Pointwise Mutual Information (PMI) of a pair of words is a measure of the probability of them occurring together in a sentence, and the Positive Pointwise Mutual Information (PPMI) is a modification of this measure. Hamilton et al., whose work will be

discussed later, construct word vectors by enumerating the PPMI of each word in the vocabulary with a large set of pre-specified ‘context’ words [31], [32]. We decided against this approach because picking context words may be time-consuming. The specific context words chosen may also have a non-negligible impact on the outputs of the model, although Hamilton et al. did not comment on this possibility.

Latent Semantic Analysis (LSA) is another commonly used word embedding technique. LSA builds a co-occurrence matrix recording how many times each word appears in each segment of text (in the context of social media, this would be Reddit comments or tweets) [33]. A dimensionality reduction technique called Singular Value Decomposition (SVD) is then applied to decrease the number of columns in the matrix while preserving the number of rows. These low-dimensional row vectors are then used as the final word embeddings [33].

Unlike PMI and LSA, which use statistics, Word2Vec creates word vectors using a shallow neural network with a single hidden layer. There are two kinds of Word2Vec models: Continuous Bag-of-Words (CBOW) and Continuous Skip-gram. CBOW models are trained to predict words given their contexts (neighboring words surrounding the target word), while skip-gram models are trained to predict the context of a word given the word [33]. Both models are trained iteratively, with the weights of the hidden layer used as the components of each target word vector. Skip-gram models are better able to capture meaning than CBOW models when the corpus contains many infrequently occurring words. Therefore, skip-gram is better for our task because our data will contain slang words that are not very commonly used [34].

Word2Vec embeddings are better able to capture the meanings of words than LSA embeddings when trained on corpuses as large as the one we plan to compile (10 million words or more) [35]. Training Word2Vec models also requires less memory than training LSA.

Word embeddings generated separately from two different time periods generally cannot be compared directly. This is because most word embedding techniques are either non-deterministic (Word2Vec) or are non-unique (LSA and other techniques using SVD), i.e., they will not create the same vector spaces even if run on the same dataset twice [28], [31], [34]. However, there are workarounds to this issue.

Kim et al. provide one such workaround [36]. Their method requires word embedding models that can be trained incrementally, such as Word2Vec (specifically, Kim et al. use skip-gram Word2Vec). This rules out embedding techniques such as GloVe, which cannot be trained incrementally [37]. To compare word vectors between two years A and B, they first generate embeddings for year A using a Word2Vec model, and this process updates the model's weights. They then use this same model with already-initialized weights to generate embeddings for year B. After this, the word embeddings from the two years can be compared the same way word embeddings from the same dataset can be: using cosine similarity.

Another way to track semantic change is using co-word analysis, which is the approach used by Hagen and de Zeeuw [38]. They chose specific words to study and then calculated the frequencies of words that occurred in the same sentences as their chosen words over time. Specifically, they found that the word "based" was initially most often used alongside the word "god" because of its association with a rapper nicknamed "the BasedGod"; later, "based" began to be used more frequently alongside alt-right terms such as "patriot." Thus, the co-words of "based" helped Hagen and de Zeeuw analyze how and why the word changed over time. Although co-word frequencies cannot directly be used to quantify semantic change, they are less computationally expensive to generate than word embeddings, and can be used to manually

analyze the trajectory of a word's meaning(s) over time and confirm that there was semantic change.

Chapter 3: Methodology

3.1. Data Collection

3.1.1. Overview

Our goal in data collection is to produce a dataset of relevant comments on Reddit with a set date range. In this case, the relevant text will be a randomized representative subsample of all comments in a date range within a subreddit. The dataset will include date, title, text content, author, subreddit, up-votes, and down-votes. This data is necessary to our current plans for data analysis, but will provide flexibility if our plans change.

3.1.2. Data source validity

We will collect data using the Reddit Application Programming Interface (API) with PRAW as a wrapper. Reddit has a rate limit of 100 queries per minute when using authentication, allowing us to collect data at a rapid rate [39]. Reddit is an ideal corpus for many forms of NLP analysis as shown by prior research [38] and due to its accessibility and large user base it is reflective of the contemporary usage of language on the internet. There are some possible pitfalls relevant to our project, however. Due to the structure of the Reddit API there isn't an ideal way of sampling the entirety of the corpus within a date range [40]. Our method currently relies on sampling comment IDs which are not evenly distributed across the sample space; this may lead to a non-representative sample which would skew the data. Additionally, most comments on Reddit contain text but they can be quite short and not carrying significant information – we need to take this into account when choosing methods of analysis. Existing research has determined that these methods of data collection are empirically useful and that word embeddings and skip-gram models, which aim to predict the contextual surroundings of a target word, can be used as a proxy for semantic change using the platform [41], [38], [42].

3.1.3. Proxying the Ability to Collect by Date Range

To navigate around the inability to collect by date range through the Reddit platform , while valid for use, different ID ranges will be identified and data will be collected specifically from that range. Since the Reddit IDs are not perfectly sequential, the starting post will be decided by choosing a post posted a day before the chosen date range. We will repeat the process to find an ending ID. It should be noted that the IDs are numbers in base 36.

Our data collection script can be run on multiple systems at once to speed up the collection. To allow for this, each system will run the script with a chosen subsection of the data which will not overlap with any other.

Within the subsection, a list of every number in the range is generated from between the starting and ending IDs, then randomly permuted and saved. This is just a list of every possibly relevant ID. The permutation of all numbers in the range prevents repeats, ensuring mechanical efficiency, and allows us to set an expected amount of samples per subsection, guaranteeing a minimum number of samples for each subsection of the date range. This serves to solve the issue of uneven distribution of data across the ID space. After initial data collection it will still be possible to return to subsections and resample them – if the program crashes or after data collection it is decided that a certain section needs more data, we can return to where we left off.

For each ID in the subsection we will make a call to the /info end point of the Reddit API. For each call we can provide up to 100 IDs, which we will saturate fully. For each returned comment considered a hit, we will store the text content, title (if present), date, author, subreddit, upvotes, downvotes, id, and permalink in a CSV. If what is returned is not a comment, we will

discard it and the ID will be considered a miss. We will maintain a count of hits and misses and once the number of hits has met the requirement for the subsection we will stop collecting IDs.

As noted we will get information on 100 IDs in one /info call, per minute we can make 100 calls per system giving 100,000 IDs per minute. Not all IDs are comments however we have found within the Reddit corpus over 90% are. Therefore we can expect approximately 90,000 comments per minute per system. This rate of collection will determine the number of comments we choose to collect.

It is also important to consider recovering the program from failure. Given that this program may need to run on each system for a prolonged period, it is important to ensure that recovery is possible from a failed state. When our permutation for each subsection is generated it will be saved to the disk along with our current position in the permeation. As we go through the permutation, every time we call the API we will move the counter forward. Even though we are going through the list sequentially the list is in random order so we still get a random sample. The comments we get will also be getting saved to disk at a set interval. In the case of a crash or failure, our program will reopen the last permutation and start from the last position of the permutation, then continue to collect random ids that are distinct from our previously collected nes. This will also allow us to re-define a larger amount of data from each permutation and be able to collect it quickly without issue. Overall the ability to recover or redefine the amount of data collected should make our data collection far smoother.

3.1.4. Pre-Processing

After collecting data we will process it with the goal of having the data in a format that is easily used for later analysis. This will include finding what text data is relevant from posts, combining useful data from different files, and cleaning data that is not in plain text form. In this

step, only pre-processing that is relevant to all analysis methods is done. Specific tasks such as tokenization are left until later steps. The outcome of this step will be a set of CSV files containing text from comments with their associated date.

Our program will go through each comment and determine relevant data from each post. Top level comments (posts) and comments have different types of useful text data. Top comments posts have titles, while comments do not. If the post or comment is not of a textual type, such as an image or video, and does not have another relevant text field, it is disregarded. However, if there is any relevant text, it will be stored in the CSV along with the metadata of that text.

Before text data is stored, it will be processed. If the text is not of a normal format (e.g., HTML) then it will be processed into plain text. Also often markdown is used in Reddit posts; we will also process this out into plain text. All of our conversion to plain text will be done by a parser rather than regex as a parser can catch more complex structures and regular expressions cannot fully describe. By the end of the process we will only have plain text.

By the end of this preprocessing step, we will have a CSV containing that plain text data from each post and comment with its publishing date. This is the final step in data collection, as we now have usable data for analysis. From herein to disambiguate the word entry will be used to reference a single line of the CSV which contains two pieces of data, text and the date that text was created by a user.

3.1.5. Selecting Terms

We must consider multiple factors when selecting terms to study. Firstly, we need to ensure that the terms we used were either coined by psychologists or professionals in psychology-adjacent fields, or are almost entirely used in the context of psychology or mental

health. The more exclusive to mental health a term is, the less likely it is to be polysemous. This will make later analyses of expanding contexts less complicated as it will limit the confounding factors that may be leading to the results we get from our embedding model [43].

We should also consider how much data we will actually get on the terms we pick. We have the advantage of collecting data before our word selection, meaning we can have a rough estimate of how frequent occurrences of a given term are before selecting it for analysis. Somewhat similarly, we should avoid words that may be bleached significantly, but their bleaching likely happened far before the time frame we aim to examine. The word ‘addict’, for example, has certainly been trivialized in conversational speech, but this may not necessarily be a recent development or one spearheaded by online discourse [25].

Currently, ‘trauma’ and ‘trigger’ are two terms that seem to fit our loose criteria well, and terms that there is a workable amount of data on. In an initial sample of approximately 33,586 comments from Reddit, the word ‘trauma’ appeared 30 times and the word ‘trigger’ appeared 64 times. With our larger samples, we will have to keep examining new words until we feel we have a reasonable representation of the phenomenon we are attempting to capture.

3.2. Quantitative Analysis

3.2.1. Frequency Analysis

Our frequency analysis will be based on the variable of keyword per entry. For every day in the chosen date range, we will calculate the frequency by going through each entry in the output CSV and counting the number of times a given keyword shows up therein. We will also have to manually produce a list of forms of the keyword that have the same meaning as the keyword but are in a different form. For example “gaslight” and “gaslit” will be considered an

occurrence of the same word. The average number of times the keyword occurrence per entry will be totaled. This will provide us with the change in frequency of how often the keyword appears per entry. We can then use this information for later analysis.

3.2.2. Context of Usage: N-Grams

Context of usage is more of a challenge for analysis and we have two approaches. The first is to leverage n-grams to extract context. Our use of n-grams will be similar to co-word analysis, which is used to identify relationships between ideas [41]. For our purposes, an n-gram refers to a set of tokens of size n in order from a list. For example, given the text “Hello world of text,” the 2-grams would be (hello, word), (world, of) and (of, text). This will be the definition moving forward.

We will only look at n-grams surrounding the keyword, therefore for each entry containing the keyword we will perform our n-gram operations. Our n-gram approach will require tokenization and preprocessing of words in the input text. For each piece of text from an entry we will first remove all punctuation and then use NLTK to tokenize, ignoring case. After this is done for an entry we can find a n-gram around the keyword. For example, if using a 2-gram we will have (“left-word”, “keyword”) and (“keyword”, “right-word”). From this we can perform quasi-analysis to draw conclusions about the change in usage of the word over time [38].

We will also complete this analysis in date ranges to get the frequency of certain n-grams over time. This will allow us to see the change in what words are being used with the keyword over time.

3.2.3. Context of Usage: Word Embeddings

The second approach involves creating word embeddings for subsets of the text data. This will require choosing a list of words related to the keyword. This approach is based on distributional semantics, which is the idea that the distance between two words' embeddings can be used as a proxy for how close in meaning the two words are [44]. This approach also requires that we choose a list of words whose embeddings' distances to the keyword's embedding will be used as a proxy for change in meaning. We will rely on a pre-trained Word2Vec model for usage as a starting point for all training. By the end of this process we will have multiple Word2Vec models trained on data with the keyword from different time slices.

We will split entities containing the keyword into slices with an equal number of entries in each slice. We use the number of entries instead of date ranges to determine the slices because if we used dates we may run into a case where there are not enough slices with the keyword to affect the positions of the keyword in the embedding space in a meaningful way.

For each entry in each slice we will tokenize the input data ignoring case, punctuation, and tense using NLTK. We will then use this data as training input for the pre-trained Word2Vec model. After training, we will save the model and move onto the next slice.

Once a model has been trained for each slice, we will iterate through the list of meaning proxies words and check the change in cosine distance to the keyword. If there was a meaningful change in distance in the embedding space it will signify that the meaning of the word altered between slices.

3.3. Qualitative Analysis

3.3.1. Co-word Analysis

The bigrams we extract from our sample will primarily function as large-scale indicators of context. Broadly, the bigrams can indicate the most common contexts in which each word is used. We can graph modal categories of bigrams for each key term to develop a general idea of what words are used in relation to them, separated by year. This will indicate to us whether there has been any kind of diachronic evolution in the words used around our key terms. This is particularly useful if there is a unique co-word that seems to be paired very often with our key terms. For example, if there is an excessive use of the word ‘trigger’ in the phrase ‘trigger warning’ in our data, the bigrams will indicate that, and we can determine as a group how we will approach the kind of anomaly that may create in our data. For comments that we find ambiguous in isolation or that have unexpected co-words we will examine them in the context of their parent thread to form a better understanding of the context in which they are presented.

3.3.2. Embedding Model and Manual Analysis

The embedding-model will show us an estimate of the words that our key terms are semantically similar to over different time periods. We have to determine what these semantically similar words indicate about the contexts in which the key terms are being used. This will likely be a group process, where we determine how to characterize the different clusters produced by the embedding model into different codes. These codes will serve as a basis for the next part of our qualitative analysis.

The second qualitative use of the embedding models will be to provide sampling for a manual analysis. Word occurrences that the model determines as semantically similar will form clusters in vector space, which will indicate broadening contextual usage. To supplement the

quantitative analysis of the embedding-model results, we will need to randomly select a smaller sample of posts from each cluster in vector space to analyze manually [27], [43]. This will not only give us a more thorough understanding of the results of the embedding-model, but will also allow us to qualitatively evaluate if the change detected by the model is actually representative of the semantic broadening that we predicted.

Our manual analysis will involve assigning codes to our data points by reading the text of each post and deciding which of the previously determined codes they fall under. Since we will be using systematic coding for this rather than open coding, we will need to have multiple people code the same data points independently. When assigning codes to data for qualitative analysis, it is important to ensure that the researchers assigning these codes agree in order to ensure that the data can correctly be characterized by the assigned codes [45]. To do this, after data collection, at least two people will separately code some portion of the data. After this, we will measure their inter-coder reliability (ICR), which is a measure of agreement between coders regarding how to code data. There are several metrics for quantifying ICR, and there is conflicting advice regarding the coding process, but these details can be determined later.

There is also no clear path of action to take regarding codes with low ICR. Options include removing low-performing codes, creating new codes and coding the data with them again, and doing nothing but taking the ICR results into account down the line. Again, this can be determined later, depending on the data we collect and any findings we make.

3.4 Validity of Methodology

For our analysis of online semantic change, we chose to use word embeddings and base our analysis on the comparison of word meanings across both different contexts and time periods; the methodology treats word embeddings as imperfect but mostly informative

representations of word meaning. Out of the three different word embedding techniques considered (PMI, LSA, Word2Vec), we ultimately settled on Skip-Gram Word2Vec because of its ability to capture the meaning of slang and infrequently occurring words within datasets.

Reddit was chosen due to its accessibility and large user base; while many social media platforms have moved away from forum-style discussions Reddit is still very active, is interconnected with other social media platforms like Twitter and Instagram, and contributes to discussions of online pop culture. Understanding the possibilities of skewed data through uneven distribution, we have proposed collecting data by date range and random sampling. We see this decision as maintaining the external and ecological validity of our research, ensuring that the data collected is representative of how the selected terms are being used and is suitable for analysis. To address the challenge of comparing word embeddings from different time periods, the methodology adopts a workaround process involving incrementally training Word2Vec models for each time period, allowing for the direct comparison of word embeddings [36]. Random sampling will be done by generating and saving lists of comment IDs.

We consider our methodology to demonstrate a comprehensive consideration to detecting semantic change while upholding standard for internal, external, and ecological validity through intentional choices in determining data sources, collection methods, analysis technique, and methods of interpretation.

Chapter 4: Current Progress

This semester we made considerable progress on the base code, developing the initial approach and creating a solution to account for date ranges. We later found that comments are sequentially assigned IDs based on their timestamps and that we could query objects by their IDs in batches of 100, meaning going forward we can randomly generate IDs to sample comments from a given date range. We made a proof-of-concept script to implement this approach, collecting about 50,000 comments, and then tested this data to verify that there were no issues and that our initial assumptions were correct. There was an anomaly in the distribution of comment IDs, however, which we are continuing to work on resolving. We also built out bi-gram and n-grams to be used for analysis, improving the speed of the script, and continued to develop the scale data collection program, also improving speed. The overall program is not finished but is approaching being done.

In terms of word embeddings, we were able to establish a basis for a working word embedding script. Although we have future plans to explore other word embeddings, our working script uses Word2Vec to train the word embedding. The script works by taking a CSV file, preprocesses the text from the given CSV, and then trains a word embedding from the preprocessed text. This file is then stored locally in the format of a .bin file. The word embedding script is split into three Python file components: a preprocessing file to deal with preprocessing text, a training file that has the code to create and train a word embedding with the proper parameters, and a main file that merges the functionality of these two file together.

Currently, the preprocessing file has a simple methodology–it cleans the text from unnecessary punctuation and symbols, converts all the words to lowercase, and then stores the

words in a list that allows the word embedding to read and train from that list. In the future, we aim to modify this preprocessing file to match the kind of preprocessing we want on our data.

The training file is fairly short and uses the built in functions provided in the Word2Vec library to create and train the word embedding. A function is defined to create and train the word embedding, which is then used in the main file to actually execute the code. The ideology behind having a separate file for word embedding creation and training is to maintain organization. This will also make any modifications easier to track and implement in the future. Within the training file, the word embedding is initialized by passing parameters like minimum word count, vector size, and thread count, for example. Currently, our word embedding is initialized with the following parameters: min_count, vecotr_size, window, workers, sg, alpha, and min_alpha.

Min_count is a parameter of type int that ignores all the words with a word count less than that number. For example, if the min_count was set to 4, all the words in the dataset that appear less than 4 times are ignored.

Vector_size is a parameter of type int that specifies the dimensionality of the word vectors. In our case, this parameter is set to 300–typically, this value is around 100-300 for big datasets. In simple terms, this parameter specifies that we want vectors in 300D in the same way one may say vectors are in 2D or 3D.

Window is a parameter of type int that specifies the maximum distance between the current and predicted word within a sentence. In simple terms, if the window parameter is set to 4, the context in which the word embedding will be w-4, w-3, w-2, w-1, CENTER_WORD, w+1, w+2, w+3, w+4.

Workers is a parameter of type int that specifies how many threads will be used to train the model. Threads allow for concurrent processing which allow for faster training time for the model.

Sg is a parameter which is either 0 or 1 that specifies the type of training algorithm. Passing 0 will specify CBOW while passing 1 will specify skip-gram.

Alpha and min_alpha are parameters of type float that control the learning/training rate of the model.

There are many parameters we can choose to add in order to customize the word embedding to our own functionality and liking, which we will continue to tweak and explore in the future.

Lastly, the main file merges the two files together into a functioning script. In order to run the script, the command will expect three parameters: the path to the main file, the path to the CSV file, and the name of the column in the CSV you want to pull data from. In general, the command is: python3 [insert path of w2vec.py] -c [insert path of csv] -t [insert column of csv].

Although this is huge progress compared to the beginning of the semester when we did not have a working word embedding script, we still have a lot to build upon, as this is a very simple working script. As we continue to mold and shape our methodology, modifications to the script are expected to account for the changes.

This current script runs on the CSV file produced from our initial data collection. In terms of future changes, one modification we can look to make is allowing our script to accept different forms of data, as our script will not work if the format of the data is not a CSV, Additionally, this script can print out the cosine similarity of target words in the terminal, which we can look to modify. Instead of printing that information in the terminal, we could have the

script write that information into a .txt file, or have it be represented in a more practical and digestible format.

We're also currently working on the possibility of collecting data from Tumblr. The Tumblr API is currently free, however it does not provide access to what we would need to conduct our analysis. The guidelines they set for how long data can be collected would also make it difficult to collect a workable sample. They have the option of requesting a waiver over email for more access, or for an exception to their guidelines, so we've started to draft a request to them that presents our research plans and proposal. We are skeptical about how successful we will be, but if we can open up the possibility of collecting data from two different sites, it will give our data much more perspective.

References

[1] G. McCulloch, *Because internet: understanding the new rules of language*. New York: Riverhead Books, 2019.

[2] M. Polinsky, "Sign Languages in the Context of Heritage Language: A New Direction in Language Research." [Online]. Available: https://www.jstor.org/stable/pdf/26478225.pdf?refreqid=fastly-default%3Ac13289145bc02244b8b73e48faafdb5b&ab_segments=&origin=&initiator=&acceptTC=1

[3] A. Omaki and J. Lidz, "Linking Parser Development to Acquisition of Syntactic Knowledge," 2013.

[4] A. Fang and H. Zhu, "Measuring the Stigmatizing Effects of a Highly Publicized Event on Online Mental Health Discourse," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–18. doi: 10.1145/3544548.3581284.

[5] R. Al-Maroof *et al.*, "The acceptance of social media video for knowledge acquisition, sharing and application: A com-parative study among YouTube users and TikTok Users' for medical purposes," *10.5267/j.ijdns*, pp. 197–214, 2021, doi: 10.5267/j.ijdns.2021.6.013.

[6] Hauser, "Essentials of Linguistics, 2nd Edition - 2nd Edition".

[7] P. DiMaggio, E. Hargittai, W. R. Neuman, and J. P. Robinson, "Social Implications of the Internet," *Annu. Rev. Sociol.*, vol. 27, no. 1, pp. 307–336, Aug. 2001, doi: 10.1146/annurev.soc.27.1.307.

[8] A. Szymkowiak, B. Melović, M. Dabić, K. Jeganathan, and G. S. Kundi, "Information technology and Gen Z: The role of teachers, the internet, and technology in the education of young people," *Technology in Society*, vol. 65, p. 101565, May 2021, doi: 10.1016/j.techsoc.2021.101565.

[9] V. R. Bhargava and M. Velasquez, "Ethics of the Attention Economy: The Problem of Social Media Addiction," *Bus. Ethics Q.*, vol. 31, no. 3, pp. 321–359, Jul. 2021, doi: 10.1017/beq.2020.32.

[10] E. Neophytou, L. A. Manwell, and R. Eikelboom, "Effects of Excessive Screen Time on Neurodevelopment, Learning, Memory, Mental Health, and Neurodegeneration: a Scoping Review," *Int J Ment Health Addiction*, vol. 19, no. 3, pp. 724–744, Jun. 2021, doi: 10.1007/s11469-019-00182-2.

[11] G. J. Robson, "The threat of comprehensive overstimulation in modern societies," *Ethics Inf Technol*, vol. 19, no. 1, pp. 69–80, Mar. 2017, doi: 10.1007/s10676-016-9414-0.

[12] M. Bąk-Sosnowska and T. Holecki, "Overstimulation and its consequences as a new challenge for global healthcare in a socioeconomic context," *Pomeranian Journal of Life Sciences*, vol. 68, no. 1, pp. 52–55, Mar. 2022, doi: 10.21164/pomjlifesci.811.

[13] H. Le, "Blasé Attitude, Hyperreality, and Social Media," *cb*, vol. 2, no. 1, Nov. 2020, doi: 10.31542/cb.v2i1.1990.

[14] J. Y. Cuan-Baltazar, M. J. Muñoz-Perez, C. Robledo-Vega, M. F. Pérez-Zepeda, and E. Soto-Vega, "Misinformation of COVID-19 on the Internet: Infodemiology Study," *JMIR Public Health Surveill*, vol. 6, no. 2, p. e18444, Apr. 2020, doi: 10.2196/18444.

[15] Á. Horgan and J. Sweeney, "Young students' use of the Internet for mental health information and support," *Journal of Psychiatric and Mental Health Nursing*, vol. 17, no. 2, pp. 117–123, Mar. 2010, doi: 10.1111/j.1365-2850.2009.01497.x.

[16] A. Pavlova and P. Berkers, "Mental health discourse and social media: Which mechanisms

of cultural power drive discourse on Twitter," *Social Science & Medicine*, vol. 263, p. 113250, Oct. 2020, doi: 10.1016/j.socscimed.2020.113250.

[17] J. Blair and S. Abdullah, "Supporting Constructive Mental Health Discourse in Social Media," in *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, New York NY USA: ACM, May 2018, pp. 299–303. doi: 10.1145/3240925.3240930.

[18] Y. Luo, D. Jurafsky, and B. Levin, "From Insanely Jealous to Insanely Delicious: Computational Models for the Semantic Bleaching of English Intensifiers," in *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 1–13. doi: 10.18653/v1/W19-4701.

[19] Merriam-Webster, "Definition of INSANELY." Accessed: Feb. 22, 2024. [Online]. Available: https://www.merriam-webster.com/dictionary/insanely

[20] N. Baes, "The semantic inflation of 'trauma' in psychology", [Online]. Available: https://intapi.sciendo.com/pdf/10.58734/plc-2023-0002

[21] J. Haiman, "Ritualization and the development of language," in *Perspectives on Grammaticalization*, W. Pagliuca, Ed., in Current Issues in Linguistic Theory. , John Benjamins Publishing Company, 1994, pp. 3–28. doi: 10.1075/cilt.109.07hai.

[22] J. Bybee and S. Thompson, "Three Frequency Effects in Syntax," *BLS*, vol. 23, no. 1, p. 378, Sep. 1997, doi: 10.3765/bls.v23i1.1293.

[23] G. Chierchia, "Broaden Your Views: Implicatures of Domain Widening and the 'Logicality' of Language".

[24] M. Israel, "Ever: polysemy and polarity sensitivity".

[25] N. Haslam *et al.*, "Harm inflation: Making sense of concept creep," *European Review of Social Psychology*, vol. 31, no. 1, pp. 254–286, Jan. 2020, doi: 10.1080/10463283.2020.1796080.

[26] L. Squires, "From TV Personality to Fans and Beyond: Indexical Bleaching and the Diffusion of a Media Innovation," *J Linguistic Anthropol*, vol. 24, no. 1, pp. 42–62, May 2014, doi: 10.1111/jola.12036.

[27] A. S. Calude, A. Anderson, and D. Trye, "Intensifying expletive constructions and their use on social media: Innovative functions of the hashtag #wokeAF in English tweets," *Discourse, Context & Media*, vol. 56, p. 100741, Dec. 2023, doi: 10.1016/j.dcm.2023.100741.

[28] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic word embeddings and semantic shifts: a survey," in *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1384–1397. Accessed: Jan. 14, 2024. [Online]. Available: https://aclanthology.org/C18-1117

[29] M. Sahlgren, "The Distributional Hypothesis," *The Italian Journal of Linguistics*, 2008, Accessed: Nov. 08, 2023. [Online]. Available: https://www.semanticscholar.org/paper/The-Distributional-Hypothesis-Sahlgren/3cd2c9b3f05e2714e8ad1db7b5d0a7be4eb15da2

[30] M. Krishna Siva Prasad and P. Sharma, "Similarity of Sentences With Contradiction Using Semantic Similarity Measures," *The Computer Journal*, vol. 65, no. 3, pp. 701–717, Mar. 2022, doi: 10.1093/comjnl/bxaa100.

[31] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic Word Embeddings Reveal

Statistical Laws of Semantic Change," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1489–1501. doi: 10.18653/v1/P16-1141.

[32] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. 2021. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/6.pdf

[33] D. Chandrasekaran and V. Mago, "Evolution of Semantic Similarity—A Survey," *ACM Comput. Surv.*, vol. 54, no. 2, p. 41:1-41:37, Feb. 2021, doi: 10.1145/3440755.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space." arXiv, Sep. 06, 2013. doi: 10.48550/arXiv.1301.3781.

[35] E. Altszyler, M. Sigman, and D. Fernández Slezak, "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database," Oct. 2016.

[36] Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov, "Temporal Analysis of Language through Neural Language Models," in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, C. Danescu-Niculescu-Mizil, J. Eisenstein, K. McKeown, and N. A. Smith, Eds., Baltimore, MD, USA: Association for Computational Linguistics, Jun. 2014, pp. 61–65. doi: 10.3115/v1/W14-2517.

[37] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[38] S. Hagen and D. De Zeeuw, "Based and confused: Tracing the political connotations of a memetic phrase across the web," *Big Data & Society*, vol. 10, no. 1, p. 205395172311631, Jan. 2023, doi: 10.1177/20539517231163175.

[39] "Reddit User Agreement," Reddit. Accessed: Nov. 04, 2023. [Online]. Available: https://www.redditinc.com/policies/user-agreement

[40] "reddit.com: api documentation." Accessed: Nov. 06, 2023. [Online]. Available: https://www.reddit.com/dev/api/#listings

[41] "Knowledge discovery through co-word analysis - ProQuest." Accessed: Feb. 21, 2024. [Online]. Available: https://www.proquest.com/docview/220452924?parentSessionId=%2FCsFLyBybDwrjluHNuuS%2BtxbY19J4t3vMtFCQRk8Guk%3D&sourcetype=Scholarly%20Journals

[42] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study," *J Med Internet Res*, vol. 22, no. 10, p. e22635, Oct. 2020, doi: 10.2196/22635.

[43] A. Kutuzov, E. Velldal, and L. Øvrelid, "Contextualized embeddings for semantic change detection: Lessons learned," *Northern European Journal of Language Technology*, vol. 8, no. 1, Aug. 2022, doi: 10.3384/nejlt.2000-1533.2022.3478.

[44] X. Ferrer, T. van Nuenen, J. M. Such, and N. Criado, "Discovering and Categorising Language Biases in Reddit." arXiv, Aug. 13, 2020. doi: 10.48550/arXiv.2008.02754.

[45] C. O'Connor and H. Joffe, "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines," *International Journal of Qualitative Methods*, vol. 19, p. 160940691989922, Jan. 2020, doi: 10.1177/1609406919899220.