The Problem of Semantic Shift in Longitudinal Monitoring of Social Media
A Case Study on Mental Health During the COVID-19 Pandemic

Keith Harrigian
kharrigian@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Mark Dredze
mdredze@cs.jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

ABSTRACT
Social media allows researchers to track societal and cultural changes over time based on language analysis tools. Many of these tools rely on statistical algorithms which need to be tuned to specific types of language. Recent studies have shown the absence of appropriate tuning, specifically in the presence of semantic shift, can hinder robustness of the underlying methods. However, little is known about the practical effect this sensitivity may have on downstream longitudinal analyses. We explore this gap in the literature through a timely case study: understanding shifts in depression during the course of the COVID-19 pandemic. We find that inclusion of only a small number of semantically-unstable features can promote significant changes in longitudinal estimates of our target outcome. At the same time, we demonstrate that a recently-introduced method for measuring semantic shift may be used to proactively identify failure points of language-based models and, in turn, improve predictive generalization.

CCS CONCEPTS
• Information systems → Social networks; • Applied computing → Health care information systems; • Computing methodologies → Semi-supervised learning settings; Modeling methodologies.

KEYWORDS
semantic shift, longitudinal monitoring, mental health

ACM Reference Format:
Keith Harrigian and Mark Dredze. 2022. The Problem of Semantic Shift in Longitudinal Monitoring of Social Media: A Case Study on Mental Health During the COVID-19 Pandemic. In 14th ACM Web Science Conference 2022 (WebSci ’22), June 26–29, 2022, Barcelona, Spain. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3501247.3531566

1 INTRODUCTION
Across multiple disciplines, studies of social media text and metadata have yielded valuable insights into population-level dynamics (e.g., consumer habits [79], voting patterns [8]). In several cases, the outcomes have enabled policy makers to more effectively anticipate and respond to concerns amongst their constituents [17, 68]. Now, as the world is presented with new and evolving global crises – e.g., COVID-19, climate change, and racial inequity – researchers look to build upon the utility of these past analyses to inform decision-making that is almost certain to have enduring ramifications [87].

Methods based on machine learning (ML), natural language processing (NLP), and web mining build the foundation of these efforts, offering an opportunity to answer questions that cannot be easily addressed using traditional mechanisms alone [70, 73]. At the same time, researchers in these computational communities are aware of how brittle these methods can be. The challenges of transfer learning and domain adaptation are well known [13, 77], with various algorithmic techniques having since been developed to enhance model robustness and improve generalization within novel data distributions [50, 53]. Yet, how these problems and their proposed solutions affect conclusions within longitudinal studies remains largely absent from applied analyses.

Indeed, longitudinal studies almost ubiquitously follow the same formulaic approach. First, acquire ground truth for a target concept within a small sample of data (e.g., regular expressions to identify medical diagnosis disclosures [25], follower networks indicating political leaning [3]). Next, train a statistical classifier on this data with the objective of re-identifying language associated with the target concept. Finally, apply the trained classifier to a new population of individuals across multiple time steps (e.g., annually, weekly). The first two stages of this modeling procedure have been explored extensively [89], but studies validating the final step have been sparse, due largely to the difficulties of obtaining temporally granular ground truth for many high-level concepts [20, 97].

A lack of analyses of temporal robustness of these models belies the seriousness of the problem: language shifts over time – especially on social media [15, 61] – and statistical classifiers degrade in the presence of distributional changes [28, 50]. Three types of distributional change are of particular concern for classifiers applied over time: 1) new terminology is used to convey existing concepts; 2) existing terminology is used to convey new concepts; and 3) semantic relationships remain fixed, but the overall language distribution changes. The latter challenge frequently manifests when major social events cause large-scale shifts in the topic of online conversation (e.g., discussion of healthcare increases during a pandemic, discussion of a political leader increases near an election). Unfortunately, these are often the types of events we seek to study.

To better understand this gap in the literature, we conduct a case study on estimating changes in depression prevalence during the COVID-19 pandemic, a timely analysis with value to the medical...
and public health communities which has thus far has procured incongruous results across studies [14, 39]. We draw inspiration from research on detecting distributional shifts in language over time [34, 49], focusing our attention on a recently-introduced method that leverages word embedding neighborhoods to identify semantic shift between multiple domains [44]. We find that semantically-informed feature selection can improve classifier generalization when semantic noise and predictive power are interwoven. More importantly, we provide evidence that semantic shift can introduce undesirable variance in downstream longitudinal monitoring applications, despite having an indistinguishable effect on historical predictive performance. Altogether, our study serves as a cautionary tale to practitioners interested in using social media data and statistical algorithms to derive sensitive population insights.

2 MOTIVATION

When the COVID-19 pandemic began in March 2020, healthcare professionals warned of an impending mental health crisis, with economic uncertainty [42], loss of access to care [96], and physical distancing [39] expected to reduce mental wellness. Given the inherent difficulties of measuring mental health at scale using traditional monitoring mechanisms, the healthcare community called upon computational scientists to leverage web data to provide evidence for optimizing crisis mitigation strategies [84]. Computational researchers responded by analyzing search queries regarding anxiety and suicidal ideation [5, 6], developing novel topic models to gather an understanding of the population’s concerns [56], and applying language-based classifiers to streams of social media text [93].

Unfortunately, these inquiries failed to provide unanimous insights that could be used with any confidence to manage the ongoing situation [98]. For instance, application of a neural language classifier to the general Reddit population estimated over a 50% increase in depression after the start of the pandemic [93], despite an analysis of topic distributions within three mental health support subreddits finding evidence to suggest the opposite [12]. Similarly, multiple keyword-based analyses using Google Trends data suggested anxiety increased relative to expected levels [5, 81], while others suggested anxiety levels actually remained stable [55].

Two years later, our understanding of COVID-19’s effect on mental health is still evolving. For example, in a survey conducted early in the pandemic by the Centers for Disease Control and Prevention (CDC), 10.7% of respondents reported having thoughts of suicide in the previous 30 days [27] (a 2x increase over the expected rate). Later in 2020, data suggested suicide rates remained stable or even fell after the start of the COVID-19 pandemic [2]. Some argued this drop was the result of a “pulling together” effect [6], an outcome that had been observed previously during times of crisis [21, 45]. However, upon closer inspection, it became clear that this trend was subject to Simpson’s paradox [54]. Reductions in suicide rate were observed amongst white folk, while a significant increase was observed amongst ethnic and racial minorities [64], with whom stress-inducing factors such as financial instability and food insecurity are more common. These nuances illustrate one dimension of difficulty associated with monitoring mental health – heterogeneity. In this paper, we will highlight an understudied dimension that affects analyses of web data – semantic shift.

2.1 Understanding the Uncertainty

In general, it is common for web-based research utilizing different datasets and analysis techniques to arrive at varying measurements of a specific phenomenon [52, 75]. However, failure to understand why these discrepancies emerge critically constrains our ability to instill confidence in future use of computational monitoring methods. In the case of the aforementioned computational studies, we argue the primary distinction is the manner in which linguistic units are aggregated and transformed into downstream insights.

Indeed, a review of mental-health related n-gram usage\(^1\) over time (and inspection of the posts in which they are found) highlights how inclusion of semantically unstable terms could confound results (see Figure 1). For example, spikes in usage of the term “suicide” in August 2019 are actually a response to Jeffrey Epstein’s death. Meanwhile, increased usage of “panic,” “eviction,” “vulnerable,” and “isolated” in March 2020 primarily corresponds to discussion of pandemic-specific circumstances (e.g., toilet paper panic, eviction moratorium, medically vulnerable populations).

Keyword-based monitoring methods such as those from Stjilija and Mishara [81], which aggregate counts from a pool of several lexical units together, are vulnerable to measurement error if the underlying semantic distribution of any subset of terms changed during the course of the given analysis period. Statistical language models have a higher capacity to disambiguate contextual usage by leveraging thousands of lexical units simultaneously to arrive at a final inference [12, 93]. However, they provide no guarantee that this disambiguation is done correctly in the presence of dramatic distributional shifts [33]. These challenges raise two important questions:

1. To what extent has semantic shift affected results regarding mental health in the published literature?
2. Is it possible to obtain more reliable longitudinal estimates by explicitly invoking knowledge of semantic shift when training statistical algorithms?

\(^1\)Pointwise Mutual Information of each term within historical samples of depressed individuals was used to determine mental health relevance.
3 RELATED WORK
In the machine learning literature, semantic shift often induces what is known as concept shift [62, 92]. That is, we experience a mismatch between the training and deployment conditional distributions \( p(y | x) \). These shifts occur naturally in a variety of situations where time is involved [48, 76]. For example, consider the word “lonely” becoming a weaker indicator of depression during Valentine’s day when loneliness is expressed more frequently in a romantic context. Concept shift is often addressed in a two-stage “detect, then react” manner [95]. During the detection phase, the goal is to measure similarity between source and target distributions, raising a flag when the divergence surpasses a predetermined threshold [86, 91]. When representative labels are available for the target distribution, we can simply select a mixture of source and target data to train a new, more appropriate model [90]. In the much more common case when labels for the target distribution are not available, methods attempt to identify subsets of source data that contain patterns found within the target data [51] or build an ensemble of experts weighted by their appropriateness to the target distribution [57, 59]. Alternative methods operate on the target predictions instead, applying post-hoc corrections to re-calibrate the estimated probabilities [83]. Our study is more similar to the former; we use unlabeled deployment data to inform feature selection.

4 MEASURING SEMANTIC SHIFT
The type of manual lexical analysis discussed in §2.1 is not feasible to perform at scale for statistical language models that often have vocabularies with thousands of terms. Fortunately, a substantial pool of prior work has proposed methods for algorithmically quantifying semantic shift between language domains [34, 58]. We choose to leverage a method introduced recently by Gonen et al. [44], which has not only outperformed several state-of-the-art alternatives in preliminary studies [47], but also shown promise for use by applied practitioners. Core advantages of this methodology include its interpretability, robustness to stochasticity, ease of implementation, and low computational overhead.

Gonen et al.’s method [44] assumes that semantically stable language has similar sets of neighboring lexical units within word embedding spaces of different domains. More formally, for two domains \( \mathcal{P} \) and \( \mathcal{Q} \), the semantic stability \( S \) of a lexical unit (e.g., word, n-gram) \( w \) can be measured as:

\[
S(w; \mathcal{P}, \mathcal{Q}) = \frac{\text{nb}_{\mathcal{P}}(k)(w) \cap \text{nb}_{\mathcal{Q}}(k)(w)}{k},
\]

where \( \text{nb}_{\mathcal{X}}(k)(w) \) (i.e., the neighborhood of \( w \) in \( \mathcal{X} \)) denotes the top-\( k \) set of lexical units nearest to lexical unit \( w \) in word-embedding vector space \( \mathcal{X} \) based on a vector distance metric of the modeler’s choosing. Hyperparameters include the neighborhood size \( k \), the minimum frequency of \( n \)-grams used for building each neighborhood \( \text{cf}_{\mathcal{X}} \), the minimum frequency of \( n \)-grams input to the semantic shift calculation \( \text{cf}_{\text{shift}} \), the distance function used for measuring a word’s neighborhood, and the embedding model architecture. For the purpose of measuring semantic shift longitudinally, we can think of independent, discrete time periods as the domains \( \mathcal{P}, \mathcal{Q} \). We include our parameterization choices in Appendix B.

5 DATA
To comprehensively understand how semantic shift may influence downstream longitudinal analyses, we leverage datasets which come from multiple social media platforms, span a wide range of time periods, and leverage different annotation/sampling mechanisms. As mentioned before (§2), we focus on the task of estimating depression prevalence, an important undertaking within the public health community [40] due to the substantial burden on individuals, communities, and society [19, 35, 60]. While we consider this specific use case, our analyses are generally applicable to longitudinal monitoring of social media.

Institutional Oversight This research was deemed exempt from review by our Institutional Review Board (IRB) under 45 CFR §46.104. All datasets in this study are either publicly accessible (via authenticated application programming interfaces) or available through secure distribution mechanisms (i.e., non-commercial data usage agreements). Given the sensitive nature of mental health data, we abide by additional protocols enumerated in Benton et al. [10] to govern all data collection, storage, and analysis. We discuss the ethics of conducting this type of research (i.e., sensitive attribute tracking) in §8.

5.1 Labeled Data
To numerically quantify the effect semantic shift has on predictive generalization, we consider four widely adopted datasets containing ground truth annotations of individual-level depression status. To diversify our data sample and understand platform-specific differences, we consider two Twitter datasets – 2015 CLPsych Shared Task [26], Multi-Task Learning [11] – and two Reddit datasets – Topic-restricted Text [94], Self-reported Mental Health Diagnoses (SMHD) [22]. Each dataset relies on a form of distant supervision; the Topic-restricted Text dataset assumes original posts made in the r/depression subreddit serve as a proxy for a depression diagnosis, while the remaining three datasets use regular expressions to identify self-disclosures of a depression diagnosis. This annotation procedure remains widely used to train classifiers for monitoring population-level trends due to challenges inherent in acquiring sufficient samples of annotated data [18, 29], but remains prone to introducing label noise and other sampling artifacts [36].

| Dataset           | Platform      | Dates       | # Users |
|-------------------|---------------|-------------|---------|
| CLPsych [26]      | Twitter       | 2012 - 2014 | C: 477  |
|                   |               |             | D: 477  |
| Multi-Task Learning [11] | Twitter       | 2012 - 2016 | C: 1,400 |
|                   |               |             | D: 1,400 |
| SMHD [22]         | Reddit        | 2013 - 2018 | C: 127,251 |
|                   |               |             | D: 14,139 |
| Topic-Restricted Text [94] | Reddit        | 2016 - 2020 | C: 107,274 |
|                   |               |             | D: 9,210  |
| 1% Stream         | Twitter       | 1/2019 - 7/2020 | All: 25,379 |
| Pushshift.io [7]  | Reddit        | 1/2019 - 7/2020 | All: 40,671 |

Table 1: Summary statistics for labeled (top) and unlabeled (bottom) datasets. Labeled dataset statistics are further broken out as a function of (C)ontrol and (D)epression groups.
5.2 Unlabeled Data

Our primary interest is understanding the practical effects of semantic shift in longitudinal monitoring applications. Accordingly, we collect large samples of text data from both Twitter and Reddit to use for extrinsic model evaluation. Our sampling procedures are inspired by those used in prior COVID-19 related work [78, 93].

We acquire raw Twitter data from the platform’s streaming API, a 1% sample of all public tweets available for non-commercial research use. We isolate all original tweets (i.e., no retweets) that include an ‘en’ language metadata attribute and are further classified as being written in English based on automatic language identification [63]. To facilitate application of our statistical classifiers, which require multiple documents from each individual to make accurate inferences, we further isolate individuals with at least 400 posts across the entire study time period (January 1, 2019 through July 1, 2020).

We sample Reddit data within the same time period using the Pushshift.io archive [7], which, unlike the Twitter streaming API, provides access to nearly all historical Reddit data [38]. We begin data collection by identifying all users who posted a comment in one of the 50 most popular subreddits9 between May 25, 2020 and June 1, 2020. Of the 1.2 million unique users identified by this query, roughly 200k were identified to have posted at least once per week during January 2019 and to not exhibit clear indicators of bot activity (e.g., repeated comments, username indicators, abnormal activity volume). We collect the entire public comment history from January 1, 2019 through July 1, 2020 for a random sample of 50k users in this cohort and perform additional filtering to isolate English data and users who have at least 200 posts across the study time period. Summary statistics for all datasets are provided in Table 1.

6 PREDICTIVE GENERALIZATION

Our ultimate goal is to understand how the presence of semantic shift affects downstream outcomes obtained from longitudinal analyses of social media data. Critical to the success of this goal is a methodology for controlling a statistical classifier’s access to semantically unstable features when making inferences on unseen data. In this initial experiment, we demonstrate that Gonen et al.’s method for measuring semantic shift [44] can be adapted with minimal effort to curate vocabularies with constrained levels of semantic stability. Further, we demonstrate that these vocabularies often improve predictive generalization and outperform alternative feature selection methods despite lacking an explicit awareness for the target classification outcome.

6.1 Methods

We design our experiment with the intention of replicating a standard deployment paradigm seen within longitudinal analyses. Language classifiers are fit on historical accumulations of annotated data and evaluated iteratively within future one-year-long time windows (see Table 2). The influence of semantic shift on generalization is measured by comparing predictive performance (F1-score) of classifiers trained using a subset of semantically-stable terms to performance of classifiers trained using alternative feature selection methods which lack awareness of semantic shift altogether. We outline the full experimental design in Appendix C.

Feature Selection Semantic stability scores $S$ are computed for each source (training) and target (evaluation) time period combination using Gonen et al.’s method [44]. We vary vocabulary size in linear, 10-percentile intervals until all available tokens are used for training the language classifier. All vocabulary selection methods are enumerated below, chosen to encompass a variety of common strategies (naïve and statistical) for reducing dimensionality and enhancing model performance.

- **Cumulative**: Frequency $> 50$ in the source time period
- **Intersection**: Frequency $> 50$ in the source & target time periods
- **Frequency**: Top $p\%$ of $n$-grams with highest frequency
- **Random**: Randomly selected terms, $p\%$ of the total available vocabulary
- **Chi-Squared**: Top $p\%$ of $n$-grams with highest chi-squared test statistic [74]
- **Coefficient**: Top $p\%$ of $n$-grams with highest absolute logistic regression weight within the training data
- **Overlap**: Top $p\%$ of $n$-grams with the highest semantic stability score $S$
- **Weighted (Overlap)**: Top $p\%$ of $n$-grams with the highest $50/50$ weighted combination of Coefficient and Overlap scores

All feature selection methods below Frequency (inclusive) are a subset of the Intersection method. We introduce Weighted (Overlap) to balance the predictive value of a given feature and its semantic stability, theorizing that a vocabulary based solely on semantic stability may come at the cost of significant predictive power, while a vocabulary solely based on within-domain predictive power will be vulnerable to generalization issues.

6.2 Results

6.2.1 Stability Analysis. To validate our implementation of Gonen et al.’s method [42] and build context for our classification results, we first manually inspect a sample of learned semantic stability scores for each dataset. On a distribution level, we see that stability scores tend to decrease as the gap between training and evaluation time periods increases – evidence of increased semantic shift over time. Additionally, we note that semantic stability scores within the Twitter datasets are generally lower than scores within than the Reddit datasets. These platform-specific differences align with our prior understanding of each platform’s design, with Twitter tending to foster conversations motivated by current events (i.e., personal and global conflict) and Reddit offering individuals an opportunity to connect through shared interests that evolve over longer time periods [69].

For all datasets, common nouns and verbs make up the majority of terms with the highest semantic stability scores (e.g., eat, bring, give, city, room, pain). These types of tokens arise only infrequently within the lower tier of semantic stability scores, typically a result of isolated conflation with current events/pop culture – names of video games (e.g., blackout, warzone), television characters and celebrities (e.g., sandy, gore, rose), and athletic organizations (e.g.,

---

9Based on total number of subscribers as of 6/01/2020. Statistics sourced from https://subredditstats.com
| Dataset    | Train       | Test       | Naive Cumulative | Naive Intersection | Naive Frequency | Statistical Chi-Squared | Statistical Coefficient | Semantic Overlap | Semantic Weighted |
|------------|-------------|------------|------------------|--------------------|-----------------|-------------------------|-------------------------|-------------------|------------------|
| CLPsych    | 2012-2013   | 2013-2014  | 0.656            | 0.677              | 0.676           | 0.687                   | 0.677                  | 0.715*            | 0.696            |
| Multi-Task Learning | 2012-2013 | 2013-2014  | 0.746            | 0.759              | 0.759           | 0.761                   | 0.757                  | 0.779*            | 0.772            |
|            | 2014-2015   | 2015-2016  | 0.703            | 0.760              | 0.760           | 0.762                   | 0.758                  | 0.781*            | 0.765            |
|            | 2014-2015   | 2015-2016  | 0.699            | 0.775              | 0.773           | 0.777                   | 0.772                  | 0.781             | 0.786            |
|            | 2012-2014   | 2013-2014  | 0.778            | 0.779              | 0.778           | 0.781                   | 0.783                  | 0.781             | 0.786            |
|            | 2015-2016   | 2015-2016  | 0.788            | 0.800              | 0.800           | 0.789                   | 0.792                  | 0.789             | 0.791            |
| Twitter    | 2012-2014   | 2013-2014  | 0.799            | 0.800              | 0.800           | 0.800                   | 0.806                  | 0.802             | 0.806            |
| Topic      | 2016-2017   | 2017-2018  | 0.659            | 0.662              | 0.661           | 0.660                   | 0.660                  | 0.661             | 0.661            |
| Restricted Text | 2018-2019 | 2019-2020  | 0.670            | 0.669              | 0.668           | 0.668                   | 0.668                  | 0.666             | 0.666            |
|            | 2019-2020   | 2020-2020  | 0.670*           | 0.665              | 0.663           | 0.665                   | 0.667                  | 0.666             | 0.666            |
| Reddit     | 2013-2014   | 2014-2015  | 0.799            | 0.798              | 0.803           | 0.799                   | 0.799                  | 0.799             | 0.799            |
| SMHD       | 2013-2014   | 2015-2016  | 0.801            | 0.800              | 0.800           | 0.805                   | 0.801                  | 0.802             | 0.802            |
|            | 2016-2017   | 2017-2018  | 0.792            | 0.792              | 0.793           | 0.798                   | 0.797                  | 0.792             | 0.799            |
|            | 2017-2018   | 2018-2019  | 0.799            | 0.800              | 0.800           | 0.803                   | 0.804                  | 0.804             | 0.808            |
|            | 2013-2015   | 2015-2016  | 0.797            | 0.795              | 0.798           | 0.799                   | 0.798                  | 0.801             | 0.799            |
|            | 2016-2017   | 2017-2018  | 0.786            | 0.785              | 0.787           | 0.790                   | 0.790                  | 0.788             | 0.791            |
|            | 2017-2018   | 2018-2019  | 0.796            | 0.796              | 0.802           | 0.799                   | 0.804                  | 0.804             | 0.807            |
|            | 2013-2016   | 2016-2017  | 0.790            | 0.790              | 0.791           | 0.792                   | 0.793                  | 0.792             | 0.794            |
|            | 2017-2018   | 2018-2019  | 0.798            | 0.796              | 0.804           | 0.798                   | 0.804                  | 0.806             | 0.808            |
|            | 2013-2017   | 2017-2018  | 0.799            | 0.797              | 0.804           | 0.800                   | 0.803                  | 0.808             | 0.810            |

Table 2: Average F1 score for the best performing vocabulary size of each feature selection method. Bolded values indicate top performers within each test set, while asterisks (*) indicate significant improvement over alternative classes of feature selection (i.e., Naive vs. Statistical vs. Semantic). Semantically-informed vocabulary selection matches or outperforms alternatives in nearly all instances, despite lacking knowledge of target outcome.

Finally, we note that the semantically-informed vocabulary selection methods not only offer reasonably wide operating windows (usually 20 to 50% of the total vocabulary size), but also tend to correlate with performance within source time periods. This latter detail suggests that semantically-stable vocabulary selection can be adequately performed in the absence of validation samples from a target time period, a necessity for most longitudinal analyses. We leave hyperparameter optimization for this methodology as an area for future exploration.

7 PRACTICAL EFFECTS OF SEMANTIC SHIFT

Having demonstrated that semantically-aware vocabulary selection methods achieve comparable performance to alternative techniques using a fraction of features and can even improve predictive generalization outright, we turn our attention to understanding the practical effects semantic shift has in longitudinal modeling applications. Specifically, we leverage our ability to systematically constrain a language classifier’s access to semantically volatile terms to evaluate how estimates of depression prevalence vary in the presence of semantic shift. Ultimately, we find that small changes in the vocabulary of a language classifier can promote large deviations in downstream outcomes, despite offering little to no indication of concern within historical data samples.
7.1 Methods

We leverage a similar experimental design to that from §6, making small methodological changes to acutely focus on understanding the practical effect of semantic shift in a deployment scenario. For example, we now model the entire time span of each annotated dataset without additional temporal splitting. Furthermore, semantic vocabulary selection is performed using embeddings learned from pairs of labeled and unlabeled datasets (e.g., CLPsych and the 1% Twitter Stream) instead of discrete time window pairs within each of the labeled datasets. The full list of modifications is provided in Appendix D; we release our code to support future research and allow others to reproduce our analysis.

To control for seasonal effects, we focus on estimating the year-over-year change in the prevalence of depression-indicative language amongst individuals in each of the unlabeled social media samples. Each unlabeled data sample is split into two distinct time periods – March 1, 2019 to July 1, 2019 (Pre-Pandemic) and March 1, 2020 to July 1, 2020 (During-Pandemic). Classifiers are applied to each individual in the unlabeled temporal samples who meets a minimum post volume criteria – 200 for Twitter and 100 for Reddit. We compute the prevalence of depression as the proportion of users in the unlabeled sample who have a predicted probability of depression greater than 0.5. We then measure the difference in estimated prevalence between the two time periods as a function of the underlying model vocabulary.

7.2 Results

7.2.1 Language Analysis. Semantic stability scores for each of the raw data samples (pre/during COVID-19) align with intuition regarding sociopolitical events of the era. Many of the n-grams with the lowest semantic stability are related to the pandemic: “viral,” “masks,” “transmission,” “isolation,” “zoom,” “lockdown.” Of terms that over-index in historical usage amongst individuals with depression, the least semantically stable include: “panic,” “cuts,” “isolated,” “strain,” “vulnerable,” and “doctors.” Each of these terms becomes closely aligned with pandemic-related phenomena that is not explicitly linked to mental-health. We provide an overview of the changes in Table 3; additional examples are included in Table 4 (see appendix).

7.2.2 Prevalence Estimates. Turning our attention to the statistical classifiers, we observe that predictive performance as a function of the underlying vocabulary is nearly indistinguishable for vocabularies of size 40% and higher. However, as shown in Figure 2, we identify significant differences in the estimated change in population-level depression prevalence as a function of the model’s underlying vocabulary. In some cases, these differences are relatively minor and lead to the same general conclusions. In other cases, we arrive at entirely different statements regarding the directional change of depression prevalence (i.e., increase instead of decrease) and absolute change (i.e., nearly 10% in the case of the minimum and maximum Multi-Task Learning estimates).

| Term       | 2019 Context          | 2020 Context          |
|------------|-----------------------|-----------------------|
| Panic      | Emotion (i.e., Fear)  | Panic Buying, Misinformation |
| Cuts       | Physical              | Economic              |
| Isolated   | Feeling Detached      | Quarantine            |
| Strain     | Discomfort/Pressure   | Virus                 |
| Vulnerable | Emotion               | At-risk Populations   |
| Doctors    | Personal Experience   | Frontline Workers, PPE |

Table 3: Change in the most prevalent context from 2019 to 2020 for a handful of terms which historically over-indexed in usage amongst individuals living with depression.

4https://github.com/kharrigian/semantic-shift-websci-2022
8 ETHICAL CONSIDERATIONS

Population monitoring at scale warrants an ethical discussion. We discuss some trade-offs between risk and reward, specifically when studying sensitive characteristics (e.g., mental health status) from social media. We direct the reader to work from Conway and O’Connor [24] and Golder et al. [43] for an expanded review.

8.0.1 Risks. Two serious risks arise from measuring personal characteristics using social media data: 1) discrimination, and 2) measurement error. The former is a challenge associated with any approach used to acquire information about human characteristics or behavior, whether inferred by an algorithm or not. Knowledge of personal attributes could be used by educational institutions to make biased admissions decisions, by law enforcement to track individuals without cause, or by political/government entities to target vulnerable individuals. These concerns are particularly poignant with regards to stigmatized characteristics, such as mental illness. Discriminatory actions based on these characteristics could have long-lasting financial and social consequences – e.g., difficulty obtaining loans, increased insurance premiums, and exclusion from certain communities. While statistical models are not the only method for gathering this information, they can be used in some situations where other approaches are infeasible [73]. With respect to the second challenge, we draw the reader’s attention to substantial evidence that demonstrates language models trained on social media datasets perform disproportionately amongst different demographic groups [1] and maintain historical social biases [16]. These systematic errors in models of mental health may further exacerbate social stratification in an opaque and elusive manner [9].

8.0.2 Rewards. We must also take care not to ignore the tremendous need for these methods and the benefits they bring. The same technology used to ostracize vulnerable individuals could also be used to provide those individuals with social services. Likewise, access to reasonably accurate classifiers with well-defined bounds of uncertainty could help small organizations acquire sufficient data to optimize resource allocation without needing to invest in the cost-prohibitive infrastructure necessary to execute traditional monitoring at scale (e.g., random digit dialing, online surveys) [80, 88]. These opportunities come with a variety of additional advantages over traditional population monitoring mechanisms – social media monitoring preempts the need to use downstream outcomes that are not useful in situations that require immediate decisions (e.g., latent changes in suicide rate), addresses certain forms of sample bias (e.g., selection bias introduced when individuals opt into a survey, disclosure bias that emerges when individuals are hesitant to discuss stigmatized topics with an interviewer), and provides the opportunity to make comparisons against retrospective baselines. Moreover, a significant amount of work focuses on methods to mitigate risk of discrimination [99] and adequately correct for sampling biases specific to social media data [41]. The large body of literature on social media monitoring in public health, for example, evidences the tremendous need for these technologies [72]. It is our responsibility to develop and deploy them in an ethically responsible manner.

8.0.3 Discussion. Practitioners must weigh these trade-offs in the context of their particular use case. In our use case, we note the goal of this study is not to make claims about a particular longitudinal trend or even demonstrate the prowess of a statistical modeling approach. Rather, our intention is to understand whether existing models can be trusted for measuring longitudinal trends at all in the presence of semantic shift, and if not, identify potential opportunities for practitioners to improve reliability of their models. The utility of such an exploration would be questionable if these types of models had not already been deployed in academia and beyond. However, one need only to look at research published within the last year regarding COVID-19 to see that machine learning classifiers are actively being used to understand a variety of social dynamics, ranging from mental health outcomes [37, 82] to transportation usage [67]. These analyses will form a foundation for public policy in the coming post-pandemic years. It is critical that we answer: Are these results reliable?

9 DISCUSSION

In this study, we demonstrated that semantic shift can be problematic in longitudinal monitoring applications, both in terms of pure predictive performance and our ability to estimate population-level outcomes. The method for measuring semantic stability introduced by Gonen et al. [44] and adapted for use as a feature selection method here is promising for reducing domain divergence and improving generalization over time. However, more research must be done to understand which deployment scenarios may obtain the most significant benefit from its use.

9.0.1 Limitations. The outcomes of this study are designed to spur conversation amongst practitioners, not necessarily to provide a panacea for addressing semantic noise in deployment scenarios. Indeed, we recognize our quantitative experiments are limited by the annotated data itself. For example, it remains to be seen whether semantically stable vocabularies are most useful over the course of certain time frames (e.g., decades instead of years), within a subset of social media platforms, or in the context of specific modeling tasks. Moreover, the labeled datasets may not be entirely conducive to the longitudinal classification experiments we performed in §6 [31, 85], with depression known to present episodically within individuals [4, 23]. Given these dataset constraints, we urge researchers to consider replicating our analysis using new datasets which feature different underlying temporal dynamics.

9.0.2 Next Steps. Beyond expanding the analysis to a more diverse array of datasets and target measures, we foresee substantial value in continued exploration of semantic stability’s effect on predictive generalization under alternative technical perspectives. For example, we note that our current study focuses solely on discrete time windows, an abstraction that is useful for simple monitoring applications, but too constraining for others. It would be of significant value to the longitudinal monitoring community to evaluate whether continuous time and diachronic embeddings offer advantages over their discretized counterparts [47, 50]. We also recognize that our implementation operates in two distinct stages (i.e., feature selection, model training), a setup which may inhibit performance. A better approach may involve leveraging knowledge of semantic shift to explicitly regularize coefficients at training time.
A PREPROCESSING

To promote consistent analysis, we use automatic language identification [63] to isolate English text in each dataset (labeled and unlabeled). Additionally, per the recommendations of De Choudhury and De [30], we exclude posts in the labeled datasets that either contain a match to a mental health related n-gram or are drawn from a subreddit explicitly dedicated to providing mental health support. This filtering is designed to encourage statistical models to learn robust linguistic relationships with depression as opposed to those introduced by sampling-related artifacts.

B TRAINING PARAMETERS

Vocabularies of a maximum 500k n-grams (unigrams, bigrams, and trigrams only) are learned using Gensim’s Phraser module, parameterized using a PMI threshold of 10 and minimum frequency of 5 [66]. For classification experiments, all posts sampled for a user from a given time period are tokenized using a Python implementation of Twokenizer [71], rephrased using the time period’s specific Phraser model, concatenated together, and translated into a single
document-term vector. Representations are transformed using TF-IDF weights learned at training time before being $\ell_2$-normalized. As a classification architecture, we use $\ell_2$-regularized logistic regression, optimizing parameters using limited-memory BFGS as implemented in Scikit-learn [74]. Inverse regularization strength $C$ is selected independently for each training sample such that F1 score is maximized within the development splits of a 10-fold cross validation run.

To measure semantic shift, the same vocabularies and Phrase detection models introduced above serve as the foundation for training unique Word2Vec models [66] for each dataset and time period. We leverage Gensim’s implementation of the continuous bag of words (CBOW) formulation of Word2Vec to learn 100-dimensional embeddings, training each model for 20 iterations using the default window and negative sampling sizes [65]. We obtain semantic neighborhoods for each $n$-gram $w$ using $k = 500$, $cfnb = 50$, and $cfshift = 50$. Alternative neighborhood sizes ($k = 250, 1000$) and frequency thresholds ($cfnb = 10, 25, 100$; $cfshift = 25, 100$) did not have a significant effect on downstream outcomes. In line with Gonen et al. [44], we measure vector similarity using cosine distance.

C GENERALIZATION: EXPANDED

C.0.1 Experimental Design. Each instance of an experimental run consists of multiple stages. In the first stage, users are randomly allocated into a 80/20 train/test split, with data from those in the training group used to learn word embeddings for each time period within a dataset (i.e., historical accumulations and future one-year windows). During the second stage, users in both the training and test splits are independently resampled to form subsets for training and evaluating the language classifiers. This secondary sampling step is constrained such that all users meet a minimum post threshold within each time period, the number of users within each training time period is equivalent, and that classes are balanced in both training and test subsets. The first and second stages are repeated 100 and 10 times, respectively, providing us with 1,000 experimental samples for each dataset. Classifiers are evaluated using data from users in the 20% test split sampled during the first stage of the experiment.

D PRACTICAL EFFECTS: EXPANDED

D.0.1 Experimental Design. In the first stage of the experimental procedure, we fit 10 embedding models for each of the labeled datasets – using randomly sampled subsets (80% size) of the complete dataset – and three embedding models for each of the unlabeled data samples – one using data from the entire Jan. 1, 2019 to July 1, 2020 time period, one using data from March 1, 2019 to July 1, 2019, and one using data from March 1, 2020 to July 1, 2020. The latter two unlabeled data models are used to qualitatively identify language which has undergone semantic shift since the start of the COVID-19 pandemic, while the former model is used in conjunction with the labeled dataset models to identify semantically stable vocabularies for training classifiers. To reduce computational expense, we randomly sample 20% of posts to train each of the unlabeled data embedding models, but otherwise maintain the same hyperparameters and training settings enumerated in §6. The second stage of the experiment proceeds as before, resampling amongst the users allocated to the training split of the annotated data to derive semantically stable vocabularies and train language classifiers. We perform the second stage 10 times, providing us with 100 classifiers for each labeled dataset.

D.0.2 Performance/Prevalence. Figure 3 shows predictive performance of classifiers within their source data domains compared to estimated change in depression prevalence for all datasets and vocabulary selection methods. We note that estimates of change in depression prevalence vary more dramatically within the Twitter data than the Reddit data. In all cases outside of training on the CLPsych Shared Task dataset [26], semantically-stable vocabularies tend to reduce the estimated change in prevalence. Randomly sampled vocabularies tend to achieve the lowest within-domain performance; as expected, they also generate estimated outcomes that vary less from the Cumulative and Intersection baselines compared to alternative feature selection methods.

D.0.3 Language Analysis. We include a sample of terms from the unlabeled Twitter and Reddit data samples which underwent significant semantic shift in Table 4 (next page). It becomes clear through this analysis that the COVID-19 pandemic is a primary driver of the shift. Several terms which were previously used to describe emotional states (e.g., panic, isolated, relief) are now used to describe pandemic-specific phenomena.
| Token | 2019 Context | 2020 Context |
|-------|-------------|--------------|
| Panic | rage, meltdown, anxiety, anger, barrage, migraine, phobia, outrage, manic, rush, asthma | hysteria, chaos, fear, misinformation, confusion, frenzy, paranoia, mayhem, insanity, fearmongering |
| Eviction | deportation, emergency, cancellation, negligence, reservation, injunction, immediate, incarceration, inaction, cancellations, termination | evictions, repayment, foreclosure, indefinite, immediate, injunction, moratorium, visitation, bankruptcy, deportation |
| Crisis | humanitarian crisis, crises, catastrophe, epidemic, disaster, threat, emergency, issue, income inequality | pandemic, crises, crisi, catastrophe, #pandemic, crisis, pand, disaster, epidemic, outbreak, pandem, pandemi, downturn |
| Corona | bourbon, margarita, mesa, distillery, carmel, lager, grove, hut, riverside, fireball, mayo, scottsdale, tempe | corona, covid, coro, virus, #corona, rona, #covid_19, #coronavirus, ebola, #covid, vir, #covid19, #wuhanvirus |
| Doctors | psychiatrists, medically, clinic, cps, accountants, police, miscarriages, malpractice, abortions, prescribe, counseling, procedure | midwives, #nurses, #doctors, epidemiologists, emts, front-line, #coronawarriors, frontliners, virologists, masks, ppe, respirators, docs, heroes |
| Zoom | zooming, zooms, tap, log, zoomed, hop, hover, dial, jump, slide, camera, plugged, optical, infrared, roll, nikon | skype, webex, #zoom, hangouts, #microsoftteams, webcam, facetime, telephone, livestream, classroom |
| Panic | rage, anger, despair, desperation, reflex, terror, adrenaline, silence, anxiety, dread, paranoia, laughter | hysteria, fear, panicking, paranoia, civil unrest, toilet paper, adrenaline, diarrhea, virus, corona, anxiety, shutdowns, #panicbuying |
| Cuts | cut, jumps, runs, cutting, pulls, moves, bounces, falls, turns, burns, drags, dips, breaks, bursts, rips, goes, bumps | cut, cutting, subsidies, budgets, deductions, revenues, checks, payments, breaks, deals, figures, loans, deposits, gains |
| Isolated | unpleasant, unstable, detached, unsafe, populated, invasive, unknown, confined, endangered, absent, vulnerable, insulated | quarantined, isolating, separated, enclosed, insulated, infectious, confined, active, populated, autonomous, vulnerable, detached |
| Vulnerable | susceptible, dangerous, prone, unstable, aggressive, hostile, disruptive, detrimental, receptive, fragile, damaging, sensitive | susceptible, dangerous, immunocompromised, infectious, isolating, elderly, disadvantaged, contagious, tolerant, likely, isolated, symptomatic |
| Relief | satisfaction, sadness, fatigue, disappointment, warmth, despair, desperation, payoff, excitement, dread, nausea, accomplishment | assistance, stimulus, aid, bailout, funding, compensation, temporary, medicaid, fund, bailouts, cheque, unemployment, benefits, donations, payout |
| Strain | inflammation, deficiency, dose, stress, pressure, calcium, medication, concentration, tissue, nausea, receptors, doses, acne, effect, discomfort | disease, illness, infections, symptom, mutation, virus, outbreak, pneumonia, infection, strains, influenza, epidemic, dependency, diseases, allergy, flu, coronaviruses |
| Testing | test, tests, qa, certification, training, scans, screening, monitoring, research, filtering, tested, imaging, treatments, coding, experiments | tests, contact tracing, screening, test, containment, infections, transmission, tracing, infection, ppe, ventilators, tested, lockdowns |

Table 4: Examples of terms which experienced significant semantic shift from 2019 to 2020.