The Problem of Semantic Shift in Longitudinal Monitoring of Social Media

A Case Study on Mental Health During the COVID-19 Pandemic

Anonymous Author(s)

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:

Anonymous Author(s). 2022. The Problem of Semantic Shift in Longitudinal Monitoring of Social Media: A Case Study on Mental Health During the COVID-19 Pandemic. In *Proceedings of 14th International ACM Conference on Web Science in 2022 (WebSci '22).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Across multiple disciplines, studies of social media text and metadata have yielded valuable insights into population-level dynamics (e.g., consumer habits [67], voting patterns [7]). In several cases, the outcomes have enabled policy makers to more effectively anticipate and respond to concerns amongst their constituents [16, 57]. 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 [73].

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 [59, 62]. At the same, 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 [12, 66], with various algorithmic techniques having since been developed to enhance model robustness and improve generalization within novel data distributions [44, 46]. 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 [23], follower networks indicating political leaning [2]). Then, 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 [75], but studies validating the final step have been sparse, due in large part to the inherent difficulties of obtaining temporally granular ground truth for many high-level concepts [19, 79].

A lack of analyses of temporal robustness of these models belies the seriousness of the problem: language shifts over time – especially on social media [14, 52] – and statistical classifiers degrade in the presence of distributional changes [25, 44]. Three types of semantic 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 [13, 35]. We draw inspiration from research on detecting distributional shifts in language over time [31, 43], focusing our attention on a recently-introduced method that leverages word embedding neighborhoods to identify semantic

WebSci '22, June 26–29, 2022, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

shift between multiple domains [40]. 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 MONITORING MENTAL HEALTH DURING A PANDEMIC

When the COVID-19 pandemic began in March 2020, healthcare professionals warned of an impending mental health crisis to follow, with economic uncertainty [38], loss of access to care [78], and physical distancing [35] expected to reduce mental wellness amongst the population at large. Given the inherent difficulties of measuring mental health at scale using traditional monitoring mechanisms, the healthcare community called upon computational scientists to leverage passive data (i.e., social media) to provide evidence for optimizing crisis mitigation strategies [71]. Computational researchers were quick to respond, lending their expertise to analyze web search queries regarding anxiety and suicidal ideation [4, 5], develop novel topic models to gather an understanding of the population's concerns [48], and apply language-based mental health classifiers to emerging streams of social media text [76].

Unfortunately, these inquiries failed to provide unanimous insights that could be used with any confidence to manage the ongoing situation [80]. 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 [76], despite an analysis of topic distributions within three mental health support subreddits finding evidence to suggest the opposite [11]. Similarly, multiple keyword-based analyses using Google Trends data suggested anxiety increased relative to expected levels [4, 69], while others suggested anxiety levels actually remained stable [47].

In general, it is common for web-based research utilizing different datasets and analysis techniques to arrive at varying measurements of a specific phenomenon [45, 65]. 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 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 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).[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).

[1]Pointwise Mutual Information of each term within historical samples of depressed individuals was used to determine mental health relevance.

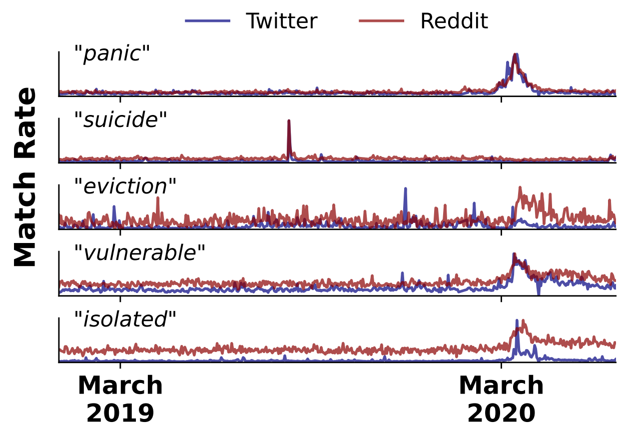

Figure 1: Proportion of posts containing a subset of depression-indicative *n*-grams across both Twitter and Reddit.

Keyword-based methods such as those from Stijelja and Mishara [69], which aggregate counts from a pool of several lexical units together, are vulnerable to noisy results if the underlying semantic/contextual 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 knowledge of thousands of lexical units simultaneously to arrive at a final inference [11, 76]. However, they provide no guarantee that this disambiguation is done correctly in the presence of dramatic distributional shifts, regardless of their algorithmic complexity [30]. These challenges raise two important questions: 1) to what extent has semantic shift affected results regarding mental health in the published literature, and 2) is it possible to obtain more reliable longitudinal estimates by explicitly invoking knowledge of semantic shift when training statistical algorithms?

3 MEASURING SEMANTIC SHIFT

The type of lexical-unit analysis applied above 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 [31, 49]. We choose to leverage a method introduced recently by Gonen et al. [40], which has not only outperformed several state-of-the-art alternatives in preliminary studies [42], but also shown promise for use by applied practitioners. Core advantages of this methodology include its interpretability, robustness to stochasticity, ease of implementation via open-source libraries, and low computational overhead.

Gonen et al.'s method [40] 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 Q, the semantic stability S of a lexical unit (e.g. word, n-gram) w can be measured as:

$$S(w;\mathcal{P},Q) = \frac{\mathrm{nb}_{\mathcal{P}}^{(k)}(w) \cap \mathrm{nb}_{Q}^{(k)}(w)}{k},$$

Dataset	Platform	Dates	# Users
CLPsych [24]	Twitter	2012 - 2014	C: 477 D: 477
Multi-Task Learning [10]	Twitter	2012 - 2016	C: 1,400 D: 1,400
SMHD [20]	Reddit	2013 - 2018	C: 127,251 D: 14,139
Topic-Restricted Text [77]	Reddit	2016 - 2020	C: 107,274 D: 9,210
1% Stream	Twitter	1/2019 - 7/2020	All: 25,379
Pushshift.io [6]	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.

where $\mathrm{nb}^{(k)}_{\mathcal{X}}(w)$ (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 X based on an arbitrarily chosen vector distance metric. Hyperparameters include the neighborhood size k, the minimum frequency of n-grams used for building each neighborhood $\mathrm{cf}_{\mathrm{nb}}$, the minimum frequency of n-grams input to the semantic shift calculation $\mathrm{cf}_{\mathrm{shift}}$, 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 $\mathcal{P}, \mathcal{Q}$.

4 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 the task of estimating depression prevalence, an important undertaking within the public health community [36] due to the substantial burden on individuals, communities, and society [18, 32, 50]. 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. [9] to govern all data collection, storage, and analysis.

4.1 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 [24], Multi-Task Learning [10] – and two Reddit datasets – Topic-restricted Text [77], Self-reported Mental Health Diagnoses (SMHD) [20]. 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 [17, 26], but remains prone to introducing label noise and other sampling artifacts [33].

To promote consistent analysis, we use automatic language identification [53] to isolate English text in each dataset. Additionally, per the recommendations of De Choudhury and De [27], we exclude posts 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.

4.2 Unlabeled Data

The annotated datasets are not representative of data used in real-world settings; individuals who disclose personal attributes vary non-trivially compared to individuals with no such disclosure [51, 64]. As our ultimate interest is understanding the practical effects of semantic shift in longitudinal monitoring applications, we collect large samples of text data from both Twitter and Reddit to use for extrinsic model evaluation.

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 [53]. 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 [6], which, unlike the Twitter streaming API, provides access to nearly all historical Reddit data. We begin data collection by identifying all users who posted a comment in one of the 50 most popular subreddits[2] between May 25, 2020 and June 1, 2020. Of 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.

5 PREDICTIVE GENERALIZATION

Our ultimate goal is to understand how the presence of semantic shift affects downstream outcomes obtained from longitudinal

[2]Based on total number of subscribers as of 6/01/2020. Statistics sourced from https://subredditstats.com

				Naïve			Statistical		Semantic	
	Dataset	Train	Test	Cumulative	Intersection	Frequency	Chi-Squared	Coefficient	Overlap	Weighted
Twitter	CLPysch	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	0.703	0.760	0.760	0.762	0.758	**0.778***	0.765
			2015-2016	0.699	0.775	0.773	0.777	0.772	**0.783***	0.772
		2012-2014	2014-2015	0.778	0.779	0.778	0.781	0.783	0.781	**0.786**
			2015-2016	0.788	0.787	0.787	0.789	**0.792**	0.789	0.791
		2012-2015	2015-2016	0.799	0.800	0.800	0.800	**0.806**	0.802	**0.806**
Reddit	Topic Restricted Text	2016-2017	2017-2018	0.659	**0.662**	0.661	0.660	0.660	0.661	0.661
			2018-2019	**0.670**	0.669	0.668	0.668	0.668	0.666	0.668
			2019-2020	**0.670***	0.665	0.663	0.665	0.667	0.666	0.667
		2016-2018	2018-2019	0.667	0.672	0.671	0.671	0.669	**0.674**	0.669
			2019-2020	0.672	0.674	0.674	0.674	0.674	**0.675**	0.674
		2016-2019	2019-2020	0.667	0.668	0.669	0.668	0.668	**0.674***	0.670
	SMHD	2013-2014	2014-2015	0.799	0.798	**0.803**	0.799	0.799	0.799	0.799
			2015-2016	0.801	0.800	0.800	**0.805**	0.801	0.802	0.802
			2016-2017	0.792	0.792	0.793	0.798	0.797	0.792	**0.799**
			2017-2018	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	0.786	0.785	0.787	0.790	0.790	0.788	**0.791**
			2017-2018	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	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.

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 [40] 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.

5.1 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). 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[3], 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. We include all relevant training parameters for the embedding models and language classifiers in the appendix. Classifiers are evaluated using data from users in the 20% test split sampled during the first stage of the experiment.

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. Semantic stability scores S are computed for each source (training) and target (evaluation) time period combination using Gonen et al.'s method [40] and the embeddings learned during the first stage of the experiment. 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

[3] 200 for Twitter, 100 for Reddit; Reddit comments contain 2x as many tokens on average.

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 [63]
- **Coefficient**: Top p% of n-grams with highest absolute logistic regression weight within the resampled 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 after *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.

5.2 Results

To validate our implementation of Gonen et al.'s method [38] 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 [58].

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., twins, braves, cal). Hashtags are frequently found in the lower semantic stability tier for the Twitter datasets, a reflection of the diversity of conversations in which they are used. Broadly, most of the observed semantic shift can be described as changes in the popularity of different word senses [41]. Although this suggests that contextual language models [29] would be well-suited to mitigating the effect of semantic shift in longitudinal analyses, emerging research suggests this is not necessarily true in the absence of additional tuning [30, 52].

We find that classifiers trained using vocabularies derived with a knowledge of semantic stability achieve equal or better predictive performance than alternative feature selection techniques in the majority of classification settings (Table 2)[4]. Semantic stability tends to be more useful for generalization within the Twitter datasets than the Reddit datasets, likely due to the aforementioned platform-specific distributions. In all cases, joint use of semantic stability and coefficient weights to derive feature selection scores (i.e., *Weighted (Overlap)*) matches or moderately improves performance over use of coefficient weights in isolation. 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.

6 PRACTICAL EFFECTS OF SEMANTIC SHIFT

Having demonstrated above 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.

6.1 Methods

We leverage a similar experimental design to that from §5, making small methodological changes to more acutely focus on understanding the practical effect of semantic shift in a deployment scenario. For example, we maintain the two-stage sampling procedure from above, but 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.

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

[4] We exclude the Random method to save space, but note it performed significantly worse across all settings as expected.

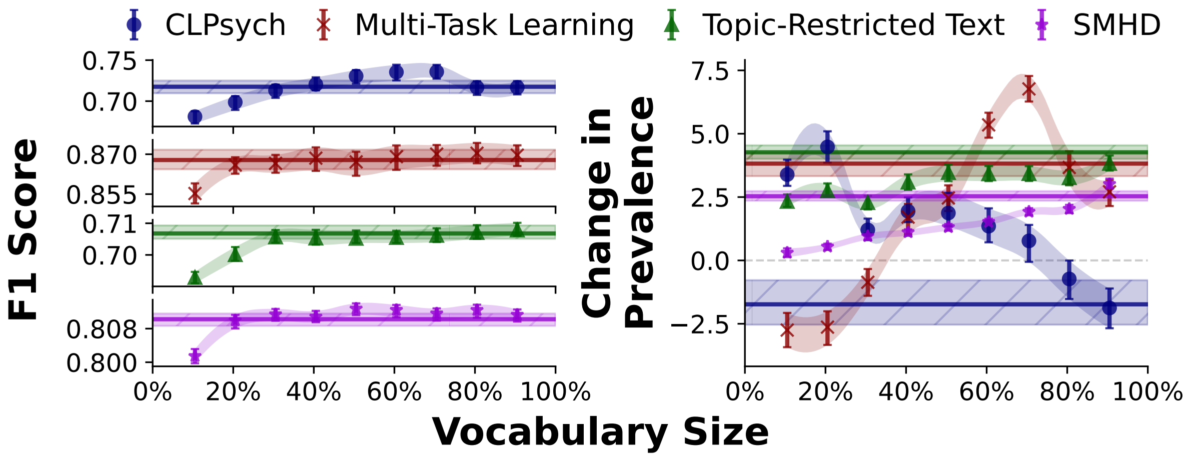

Figure 2: Horizontal bars denote each dataset's estimate under the naïve, *Intersection* baseline. Curves denote performance over varying sizes of vocabulary selected based on semantic stability S. (Left) Average F1 score within held-out samples drawn from each dataset's complete time period. Performance is largely indistinguishable for several of the vocabulary sizes. (Right) Estimated change in depression prevalence as a function of vocabulary. Estimates vary significantly despite having statistically equivalent predictive performance on source data.

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 §5. 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.

To control for seasonal effects, we focus on estimating the year-over-year change in 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 (Post-Pandemic). Classifiers are applied to each individual in the unlabeled temporal samples which meet 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.

6.2 Results

Semantic stability scores for each of the raw data samples (pre/post 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." In addition to a slew of named entities, Gonen et al.'s method [40] also identifies several terms associated with the Black Lives Matter movement and civil unrest: "peaceful", "floyd,", "crow", "looting," "riot." Of terms that over-index in historical usage amongst individuals with depression, the least semantically stable include: "panic," "cuts," "strain," "isolated," "death," "crisis," and "doctors." Each of these terms becomes closely aligned with pandemic-related phenomena that is not explicitly linked to mental-health. We provide examples of these shifts in Table 3 (see appendix).

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).

7 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 [22] and Golder et al. [39] for an expanded review.

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 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 [62]. 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 [15]. These systematic errors in models of mental health may further exacerbate social stratification in a opaque and elusive manner [8].

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 manipulate and 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) [68, 74]. 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, new work focuses on methods to mitigate risk of discrimination [81] and adequately correct for sampling biases specific to social media data [37]. The large body of literature on social media monitoring in public health, for example, evidences the tremendous need for these technologies [61]. It is our responsibility to develop and deploy them in an ethically responsible manner.

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 [34, 70] to transportation usage [56]. 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?

8 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. [40] 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.

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 §5 [28, 72], with depression known to present episodically within individuals [3, 21]. Given these dataset constraints, we urge researchers to consider replicating our analysis using new datasets which feature different underlying temporal dynamics.

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 [42, 44]. 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.

REFERENCES

[1] Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and Racial Fairness in Depression Research using Social Media. In *EACL*.

[2] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

[3] Jules Angst, Alex Gamma, Wulf Rössler, Vladeta Ajdacic, and Daniel N Klein. 2009. Long-term depression versus episodic major depression: results from the prospective Zurich study of a community sample. *Journal of affective disorders* (2009).

[4] John W Ayers, Eric C Leas, Derek C Johnson, Adam Poliak, Benjamin M Althouse, Mark Dredze, and Alicia L Nobles. 2020. Internet searches for acute anxiety during the early stages of the COVID-19 pandemic. *JAMA Internal Medicine* (2020).

[5] John W Ayers, Adam Poliak, Derek C Johnson, Eric C Leas, Mark Dredze, Theodore Caputi, and Alicia L Nobles. 2021. Suicide-related internet searches during the early stages of the COVID-19 pandemic in the US. *JAMA network open* (2021).

[6] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *ICWSM*.

[7] Nicholas Beauchamp. 2017. Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science* (2017).

[8] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *FAccT*.

[9] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Workshop on Ethics in NLP*.

[10] Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *EACL*.

[11] Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. 2020. Quantifying the Effects of COVID-19 on Mental Health Support Forums. In *NLP for COVID-19*.

[12] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*.

[13] Michael Johnathan Charles Bray, Nicholas Omid Daneshvari, Indu Radhakrishnan, Janel Cubbage, Michael Eagle, Pamela Southall, and Paul Sasha Nestadt. 2021. Racial differences in statewide suicide mortality trends in Maryland during the coronavirus disease 2019 (COVID-19) pandemic. *JAMA psychiatry* (2021).

[14] Igor Brigadir, Derek Greene, and Pádraig Cunningham. 2015. Analyzing discourse communities with distributional semantic models. In *ACM Web Science*.

[15] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *ICML*.

[16] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet* (2015).

[17] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* (2020).

[18] Sung Man Chang, Jin-Pyo Hong, and Maeng Je Cho. 2012. Economic burden of depression in South Korea. *Social psychiatry and psychiatric epidemiology* (2012).

[19] Daejin Choi, Steven A Sumner, Kristin M Holland, John Draper, Sean Murphy, Daniel A Bowen, Marissa Zwald, Jing Wang, Royal Law, Jordan Taylor, et al. 2020. Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. *JAMA network open* 3, 12 (2020), e2030932–e2030932.

[20] Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *COLING*.

[21] Stephan Collishaw, Barbara Maughan, Robert Goodman, and Andrew Pickles. 2004. Time trends in adolescent mental health. *Journal of Child Psychology and psychiatry* (2004).

[22] Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology* (2016).

[23] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *CLPsych*.

[24] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *CLPsych*.

[25] Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *JAIR* (2006).

[26] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *ACM Web Science*.

[27] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

[28] Orianna DeMasi, Konrad Kording, and Benjamin Recht. 2017. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one* (2017).

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[30] Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2021. Time-Aware Language Models as Temporal Knowledge Bases. *arXiv* (2021).

[31] Mark Dredze, Tim Oates, and Christine Piatko. 2010. We're not in kansas anymore: detecting domain changes in streams. In *EMNLP*.

[32] William W Dressler. 1991. *Stress and adaptation in the context of culture: Depression in a southern black community.*

[33] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *CHI*.

[34] Alex Fine, Patrick Crutchley, Jenny Blase, Joshua Carroll, and Glen Coppersmith. 2020. Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using NLP applied to social media data. In *NLP + CSS*.

[35] Sandro Galea, Raina M Merchant, and Nicole Lurie. 2020. The mental health consequences of COVID-19 and physical distancing: the need for prevention and early intervention. *JAMA internal medicine* (2020).

[36] Alan J Gelenberg. 2010. The prevalence and impact of depression. *The Journal of clinical psychiatry* (2010).

[37] Salvatore Giorgi, Veronica Lynn, Keshav Gupta, Farhan Ahmed, Sandra Matz, Lyle Ungar, and H Andrew Schwartz. 2021. Correcting Sociodemographic Selection Biases for Population Prediction from Social Media. (2021).

[38] Danijela Godinic, Bojan Obrenovic, Akmal Khudaykulov, et al. 2020. Effects of economic uncertainty on mental health in the COVID-19 pandemic context: social identity disturbance, job uncertainty and psychological well-being model. *Int. J. Innov. Econ. Dev* (2020).

[39] Su Golder, Shahd Ahmed, Gill Norman, and Andrew Booth. 2017. Attitudes toward the ethics of research using social media: a systematic review. *JMIR* (2017).

[40] Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *ACL*.

[41] Christian Haase, Saba Anwar, Seid Muhie Yimam, Alexander Friedrich, and Chris Biemann. 2021. SCoT: Sense Clustering over Time: a tool for the analysis of lexical change. In *EACL*.

[42] William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *ACL*.

[43] Xiaolei Huang and Michael Paul. 2018. Examining temporality in document classification. In *ACL*.

[44] Xiaolei Huang and Michael Paul. 2019. Neural Temporality Adaptation for Document Classification: Diachronic Word Embeddings and Domain Adaptation Models. In *ACL*.

[45] Luke Hutton and Tristan Henderson. 2015. Toward reproducibility in online social network research. *IEEE Transactions on Emerging Topics in Computing* 6, 1 (2015), 156–167.

[46] Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. ACL.

[47] Duleeka Knipe, Hannah Evans, Amanda Marchant, David Gunnell, and Ann John. 2020. Mapping population mental health concerns related to COVID-19 and the consequences of physical distancing: a Google trends analysis. *Wellcome Open Research* (2020).

[48] Jing Xuan Koh and Tau Ming Liew. 2020. How loneliness is talked about in social media during COVID-19 pandemic: text mining of 4,492 Twitter feeds. *Journal of psychiatric research* (2020).

[49] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *COLING*.

[50] Jean-Pierre Lépine and Mike Briley. 2011. The increasing burden of depression. *Neuropsychiatric disease and treatment* (2011).

[51] Tom Lippincott and Annabelle Carrell. 2018. Observational Comparison of Geo-tagged and Randomly-drawn Tweets. In *PEOPLES*.

[52] Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. TimeLMs: Diachronic Language Models from Twitter. *arXiv preprint arXiv:2202.03829* (2022).

[53] Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *ACL: System Demonstrations*.

[54] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv* (2013).

[55] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.

[56] Syed Ahnaf Morshed, Sifat Shahriar Khan, Raihanul Bari Tanvir, and Shafkath Nur. 2021. Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis. *Journal of Urban Management* (2021).

[57] Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. 2013. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *JMIR* (2013).

[58] Bill Noble, Asad Sayeed, Raquel Fernández, and Staffan Larsson. 2021. Semantic shift in social networks. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. 26–37.

[59] Alicia L Nobles, Caitlin N Dreisbach, Jessica Keim-Malpass, and Laura E Barnes. 2018. " Is This an STD? Please Help!": Online Information Seeking for Sexually Transmitted Diseases on Reddit. In *ICWSM*.

[60] Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

[61] Michael J Paul and Mark Dredze. 2017. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services* (2017).

[62] Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing*.

[63] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *JMLR* (2011).

[64] Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role

of personality, age, and gender in tweeting about mental illness. In *CLPsych*.

[65] Dwaipayan Roy, Mandar Mitra, and Debasis Ganguly. 2018. To clean or not to clean: Document preprocessing and reproducibility. *Journal of Data and Information Quality (JDIQ)* 10, 4 (2018), 1–25.

[66] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *NAACL: Tutorials*.

[67] Jose Ramon Saura, Ana Reyes-Menendez, and Pedro Palos-Sanchez. 2019. Are black Friday deals worth it? Mining Twitter users' sentiment and behavior response. *JOItmC* (2019).

[68] Lance Garrett Shaver, Ahmed Khawer, Yanqing Yi, Kris Aubrey-Bassler, Holly Etchegary, Barbara Roebothan, Shabnam Asghari, and Peizhong Peter Wang. 2019. Using Facebook advertising to recruit representative samples: feasibility assessment of a cross-sectional survey. *JMIR* (2019).

[69] Stefan Stijelja and Brian L Mishara. 2020. COVID-19 and Psychological Distress—Changes in Internet Searches for Mental Health Issues in New York During the Pandemic. *JAMA internal medicine* (2020).

[70] Tom Tabak and Matthew Purver. 2020. Temporal Mental Health Dynamics on Social Media. In *NLP for COVID-19*.

[71] John Torous, Keris Jän Myrick, Natali Rauseo-Ricupero, and Joseph Firth. 2020. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. *JMIR mental health* (2020).

[72] Adam Tsakalidis, Maria Liakata, Theo Damoulas, and Alexandra I Cristea. 2018. Can we assess mental health through social media and smart devices? Addressing bias in methodology and evaluation. In *ECML PKDD*.

[73] Mihaela van der Schaar, Ahmed M Alaa, Andres Floto, Alexander Gimson, Stefan Scholtes, Angela Wood, Eoin McKinney, Daniel Jarrett, Pietro Lio, and Ari Ercole. 2021. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning* 110, 1 (2021), 1–14.

[74] Jerry J Vaske. 2011. Advantages and disadvantages of internet surveys: Introduction to the special issue. *Human Dimensions of Wildlife* (2011).

[75] Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *AAAI*.

[76] JT Wolohan. 2020. Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic. In *NLP for COVID-19*.

[77] JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: attending to self-stigmatized depression with NLP. In *LCCM*.

[78] Hao Yao, Jian-Hua Chen, and Yi-Feng Xu. 2020. Patients with mental health disorders in the COVID-19 epidemic. (2020).

[79] Reza Zafarani and Huan Liu. 2015. Evaluation without ground truth in social media research. *CACM* (2015).

[80] Jon Zelner, Julien Riou, Ruth Etzioni, and Andrew Gelman. 2021. Accounting for uncertainty during a pandemic. *Patterns* (2021).

[81] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *NAACL HLT*.

A 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 [55]. For classification experiments, all posts sampled for a user from a given time period are tokenized using a Python implementation of Twokenizer [60], 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 ℓ_2-normalized. As a classification architecture, we use ℓ_2-regularized logistic regression, optimizing parameters using limited-memory BFGS as implemented in Scikit-learn [63]. 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 [55] 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 [54]. We obtain semantic neighborhoods for each n-gram w using $k = 500$, $\text{cf}_{\text{nb}} = 50$, and $\text{cf}_{\text{shift}} = 50$. Alternative neighborhood sizes ($k = 250, 1000$) and frequency thresholds ($\text{cf}_{\text{nb}} = 10, 25, 100$; $\text{cf}_{\text{shift}} = 25, 100$) did not have a significant effect on downstream outcomes. In line with Gonen et al. [40], we measure vector similarity using cosine distance.

B PRACTICAL EFFECTS: EXPANDED

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 [24], 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.

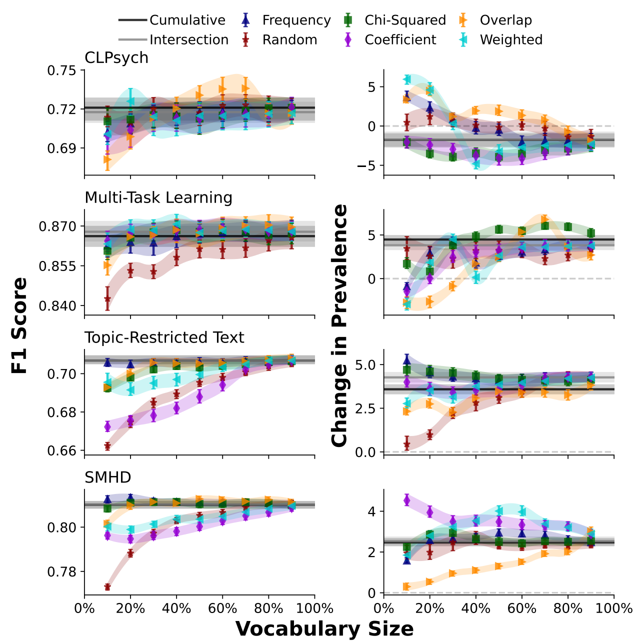

Figure 3: Comparison of within-domain predictive performance and estimated change in depression prevalence as a function of the underlying vocabulary.

We include a sample of terms from the unlabeled Twitter and Reddit data samples which underwent significant semantic shift in Table 3 (next page). It becomes clear through this analysis that the COVID-19 pandemic, combined with the Black Lives Matter social movement later in 2020, are primary drivers 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
Twitter	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, cris, pand, disaster, epidemic, outbreak, pandem, pandemi, downturn
	Corona	bourbon, margarita, mesa, distillery, carmel, lager, grove, hut, riverside, fireball, mayo, scottsdale, tempe	carona, covid, coro, virus, #corona, rona, #covid_19, #coronavirus, ebola, #covid, vir, #covid2019, #wuhanvirus
	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
	Floyd	dwight, wesley, sheldon, eddie, andy ruiz, mayfield, wade, ronnie, willie, holloway, george, henry, albert	floyd's, floyd's, #georgefloyd, floyds, ahmaud arbery, #georgefloyd's, #ahmaudarbery, arbery, zimmerman, carlin, geor, pell
Reddit	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
	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
	Looting	spawning, loot, raiding, farming, grinding, camping, sniping, reloading, spawn, mobs, afk, fighting, hunting	rioting, looters, rioters, riots, arson, vandalism, protesting, blm, protest, peaceful, protestors, protesters, riot, protests, violence
	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
	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 3: Examples of terms which experienced significant semantic shift from 2019 to 2020. The COVID-19 pandemic and Black Lives Matter movement are primary drivers of the change in usage.