<Note: Author's copy>

An interpretable model for real-time tracking of economic indicators using social media data

NEETI POKHRIYAL, Dartmouth College
BENJAMIN VALENTINO, Dartmouth College
SOROUSH VOSOUGHI, Dartmouth College

Measures of public opinion on economic matters are vital in creating accurate official statistics needed to design appropriate policy interventions and shape private investment decisions. This is traditionally done using public opinion polls and representative surveys. However, this process is time and money intensive and currently suffers from reduced public participation. As a result, official statistics are usually delayed and are not frequently available at the resolution needed for better policy interventions. Hence, researchers have looked into anonymized digital data to continuously sense information about public behaviors, especially those related to consumer sentiment index and unemployment insurance claims. However, past studies relied on linear models with simplistic assumptions and thus provided limited extrapolatory power and no insights as to why these predictive methods work. Worryingly, the strong correlations reported in these studies disappeared when the original models were tested with newer social media data.

We propose a novel interpretable machine learning model, called Group Additive Gaussian Processes, to provide accurate and near real-time estimates of economic indicators about public behaviors using social media data, along with insights into the model behavior. Our model exploits the underlying structure in data and encodes interpretability in the modeling framework. It is based on Gaussian Process (GP) regression, which provides a robust non-parametric Bayesian learning framework that produces calibrated uncertainty measures along with its predictions. A key challenge in the learning task is learning from limited training data. We demonstrate how our model not only learns but also generalizes well in these scenarios. Through extensive evaluation we show how our model performs on two important indicators of economic health - consumer confidence index and unemployment insurance claims data. Further, we demonstrate how our model can reduce the need to conduct surveys by producing highly accurate and frequent estimates in between the surveying periods.

CCS Concepts: • **Computing methodologies** → **Gaussian processes**; • **Applied computing** → **Sociology**; • **Information systems** → *Data mining*.

ACM Reference Format:
Neeti Pokhriyal, Benjamin Valentino, and Soroush Vosoughi. 2021. An interpretable model for real-time tracking of economic indicators using social media data. *ACM/IMS Trans. Data Sci.* 1, 1, Article 1 (January 2021), 32 pages. https://doi.org/10.1145/3498332

1 INTRODUCTION

Sample surveys have been the *de facto* instrument for understanding the social, political and economic realities of the population. Social scientists have long used these surveys to gather

Authors' addresses: Neeti Pokhriyal, Dartmouth College, neeti.pokhriyal@dartmouth.edu; Benjamin Valentino, Dartmouth College, benjamin.a.valentino@dartmouth.edu; Soroush Vosoughi, Dartmouth College, Soroush.Vosoughi@dartmouth.edu.

2577-3224/2021/1-ART1
https://doi.org/10.1145/3498332

information from a subset of population and generalize their results to larger populations of interest. However, traditional surveys are time consuming and money intensive exercises and suffers from dwindling public participation [5], especially in phone surveys. As a result, these statistics are available at low frequency and after significant time lags, and not available at the granularity that would aid informed decision making.

Researchers have explored the use of social media (SM) data, mainly from Twitter and Google trend data, to either supplant or supplement traditional survey data to measure various social and economic indicators frequently and cheaply. Models have been proposed to gauge consumer confidence and labor flows using SM data [1, 30, 33] mostly as a way to reduce surveying costs. However, the strong correlations disappeared when those models were tested with newer and longer data [10, 32]. Studies demonstrated high correlation for few initial years followed by deteriorated results, raising serious concerns about the viability of SM data for estimating macro-economic indicators. Owing to the cheap and frequent availability of SM data and their potential to produce significant cost savings at current times of financial chaos warrants another look into the feasibility of these methods.

Our goal is to predict macroeconomic indicators from surveys using SM data by proposing novel methods that overcome the deficiencies of existing works. We attribute the inconsistencies of past methods to two factors: one, simplistic linear assumption used to model the relationship between the sentiment found in SM data and economic indicators and second, the lack of interpretability or insights into the modeling framework.

A key challenge in learning tasks where labeled data comes from surveys is that they are time and money intensive and are sparse (available at monthly, quarterly or annual periodicity). For example, index of consumer sentiment (ICS), a leading indicator of consumer confidence, is available at monthly granularity, so even taking decadal data amounts to just 120 data points. Additionally, this task is complicated by the fact that learning is usually accomplished by recent past data (usually couple of years) due to its temporal contextual relevancy for prediction. Thus, even though we may have 10 years of data available, we learn only using past 2 years of data (just 24 data points). To efficiently learn from limited labeled data, we need models that can **extrapolate** well even from small training data.

We propose an interpretable machine learning model called **Group Additive Gaussian Processes (Group AGPs)** with two goals: to learn efficiently from **small training data** and to provide **interpretability**. Our model is inspired by Gaussian Processes (GPs) [41]. GPs belong to the class of Bayesian non-parametric models, where no assumptions are made on the functional form of the relationships between covariates and targets, and thus, these methods are known to learn highly non-linear boundaries. GPs provide a flexible and tractable prior over functions and can learn the underlying structure of data as determined by its kernel function.

Group AGP is a hierarchically additive GP model that attempts to learn the underlying structure of the data and provides both interpretability and extrapolation power, at time points far from the training data. We assume that our underlying generative process can be hierarchically decomposed into lower-order additive kernels, where order is number of dimensions in the kernel and each kernel accounts for specific structure in data. Our decomposition abstracts the granularity for interpretation at feature level, which is shown to be relevant to the domain audience [29]. Since our model is based on GPs, it provides measures of **calibrated uncertainty** along with its estimates. Quantifying these uncertainties is important as the linguistic uncertainties and noises in large-scale SM data might translate to uncertainty in the estimating the signal for consumer sentiment [51]. This uncertainty provides a measure of our model's trust in giving that prediction.

Our motivation for hierarchical decomposition comes from the fact that many real life scenarios can be understood as having a **hierarchical additive structure**. As an example, the price of a

house can be decomposed as an addition of two components: price of the cost of building the house and price related to neighborhood factors. The first component can, in turn, be thought of as an interaction of two factors: the size (sq. ft.) of the house and the cost of building material per sq. ft. The second component can, in turn, be studied as an interaction of a number of factors, such as prices of the nearby houses; proximity to good schools, shopping, entertainment etc. Thus the price of a house can be understood as having two additive components where each component models the interaction among its constituent factors.

The hierarchical **structure discovery** in our model is guided by the interplay of these criteria - maximizing the log-likelihood of training data and minimizing the log-loss on test data; minimizing the complexity of the model and maximizing the interpretability of the model. We provide a novel metric to quantify the interpretability of our model. Through extensive evaluation on Reddit data and Google purchase history data, we show that one doesn't have to sacrifice predictive accuracy for interpretability.

Our contributions are summarized below:

(1) We propose a novel hierarchically additive GP model called **Group Additive GPs (Group AGPs)** that provides descriptive interpretability and better accuracy than state-of-the-art models to estimate economic indicators. Our model exploits the underlying structure in data and provides better generalization ability, which is an important concern when learning with small training data as is the case here.
(2) We present a novel metric to quantify the notion of interpretability in a hierarchical Bayesian model. We show that our interpretations are based on features and are, thus, meaningful to the practitioners as they take prior knowledge into account and can be analyzed and put into practice. Through detailed experimentation, we show how our model uncovers three broad clusters, related to jobs search; economy/ recession and lending/finance as important topics affecting its predictive accuracy for consumer confidence index, one of the economic indicators used in this work.
(3) Our model provides measures of calibrated uncertainty along with its estimates and also assists in interpretability analysis. Uncertainty provides a measure of our model's trust in giving that estimate and might assist in algorithmic decision making by telling practitioners where and how much to trust the individual predictions.

1.1 A note about interpretability

We use the following definitions of **interpretability** - An interpretable ML model is defined as "the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model" [29]. Relevancy is defined if the model provides insights for a specific problem and for a particular audience. Recent works point out that interpretations of model outcomes must be based on features that are meaningful to the practitioners, so they can be analyzed and put into practice [28]. We show how our model embeds the notion of interpretability in the feature space while conceptualizing the model and thus provide inferences that are meaningful to the practitioners. This contrasts with the explain-ability that is sought as a post hoc step especially in deep learning systems [6, 20]. At times, post hoc analysis either has uncovered model interpretations that a practitioner knows to be incorrect [43]; or are not tailored to the domain experts as they do not take prior domain knowledge into account [28].

1.2 Economic indicators used in this work

We focus on two indicators - **consumer confidence index** or **index of consumer sentiment (ICS)**, one of the most widely cited economic surveys in the United States; and **unemployment**

insurance claims data (UI), another important indicator describing the health of the economy. Accurately measuring consumer confidence is pivotal to governments, businesses, media and individuals as it describes both current economic conditions and captures consumer's hope for near future [12], and influences whether the economy expands or contracts, as consumer spending drives 70% of the GDP [11]. The initial claims data, along with other employment data, is closely watched by market analysts, as it determines the strength of the labor market. Both economic indicators have been studied in conjunction with correlating them to signal in SM data.

1.3 Social media data for understanding economic indicators

Different types of SM data (from social networking applications like Twitter, Reddit, Facebook etc.) and Google search history data, have been used to construct official indicators of consumer confidence. Researchers have mostly used the sentiment of various phrases related to job(s) or economy or mentions of purchase intentions in search engine or SM posts. There are a number of advantages in studying economic indicators using SM data. First, these datasets are collected continuously and provide frequent sensing of public perceptions than administrative and polling data. Such timely data might offer market analysts with almost real-time information, and assist in when economic decisions must be made needed prior to the availability of official indicators. Second, SM are available at a lower (or no) cost compared to traditional polling, which often cost tens or hundreds of thousands of dollars. Third, analyzing SM offers a unique opportunity to glean signals from personal conversations and web searches, which might assist in capturing additional signs/parameters about the population [1].

The general approach to understanding SM data involves extracting quantitative metrics, mostly sentiments, from its content. If the goal is to demonstrate that SM can supplement traditional surveys, then metrics from SM is added to the statistical models used to construct survey indices and their net gain is studied. For example, models that include information from SM have been shown to produce more accurate estimates of election outcomes than than models that include only traditional poll data [48]. If the goal is to show that social media can potentially be used to substitute or supplement surveys/polls, then longitudinal correlations between two data sources are studied [30]. Our work explores the latter idea.

2 RELATED WORKS

We describe the related works along two areas: one which describes the existing works related to estimating consumer confidence and other economic indicators using SM; and second, related to the methods in additive GPs literature, which has inspired our work.

2.1 Social media to predict economic indicators

Researchers have extracted sentiments and content features from Twitter, Facebook, news and Google search queries regarding purchase history etc. and used them to model **ICS** with mixed results [15, 30, 33, 35, 46]. Though some studies highlight a good correlation prior to 2012, but a comprehensive study states that the relation disappeared when more recent data was included in the analysis [10].

Most existing studies use a single measure (mostly a sentiment feature) and study correlations to forecast CCI values. Most of these studies are based on the sentiment of few keywords, such as keywords like jobs, job on Twitter. These analyses utilize methods that look for correlations over a short period of time or with coarser granularity of data, e.g. monthly. Some works [1] have employed n-grams models to analyze the textual content in social media and used these to learn a pattern of relationships between SM text and survey responses for prediction purposes. Some works [35] are autoregressive in nature, with some components that factor SM data. However these

models have a linearity assumption which fails to learn the complex relationship between high dimensional SM data and economic indicators. Additionally, existing works have failed to assess the temporal stability of their estimates over longer periods of time.

Researchers have used Reddit data for a variety of applications such as understanding online user behavior [2], (dis)information propagation, health informatics [4] etc. In the context of studying public opinion on economic issues, researchers have recently started using it [38] using GP regression to estimate CCI.

2.2 Additive GPs

Gaussian Processes have been studied extensively in geostatistics (as kriging), meterology etc, as they provide high flexibility in modeling as well as a principled way of learning the hyperparameters [41].

An **additive GP** with kernels that fully decompose additively is given as:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{\mathbf{i=1}}^{\mathbf{D}} \mathbf{k_i}(\mathbf{x_i}, \mathbf{x_i'}) \tag{1}$$

where $\mathbf{x}$ is the vector in D-dimensional space and $k_{i\,i=1}^{D}$ is a sub-kernel. When each of these subkernels operate on single dimensions, these are considered as Generalized Additive Models (GAMs). Works in the area of additive GPs assert that decomposing a function into additive components provides better extrapolatory power [17] away from the training instances. The problem with GAM GPs is that they only model first order interactions, which is an unrealistic assumption in real problem settings. An extension of this is to consider higher order interactions [17], however the number of combinations grows exponentially. Thus, learning the structure of kernel is extremely challenging. Researchers have resorted to learning the structure of the kernels via enumeration methods (where sub-kernels consider every combination of feature interactions up to a degree) [17], search methods (where possible decompositions are traversed by a search algorithm) [40], and projection-pursuit (where a projected-additive GP is learned by iteratively optimizing projection directions from regression residuals) [13]. While search methods are burdened by combinatorial search, random projection based methods impart no interpretability into the model.

Recent works [18, 23, 44] have explored additive GP where kernels operate on a subset of dimensions, which are obtained either randomly or via an informed search in model space. Compared to our proposed model, these works solely focus on Bayesian optimization and use a heuristic search strategy while still working with smaller dimensional spaces (few tens).

3 MODEL

Notation: We denote the target variable at time t, as $y_t \in \mathbb{R}$, and the independent covariates as $\mathbf{x}_t \in \mathbb{R}^d$. We use a bold lower-case letter to denote a vector, a lowercase letter to denote a scalar value and an uppercase letter to denote a matrix.

We assume that the covariate vector $\mathbf{x}_t$ is a concatenation of c feature vectors, i.e., $\mathbf{x}_t \equiv (\mathbf{x}_t^{(g_1)}; \mathbf{x}_t^{(g_2)}; \ldots; \mathbf{x}_t^{(g_c)})$, where each vector, $\mathbf{x}_t^{(g_i)}$ corresponds to a d_{g_i} length feature vector, indexed by the set $g_i \subseteq \{1, \ldots, d\}$, such that $\sum_{i=1}^{c} d_{g_i} = d$ and, further, $g_1 \cap g_2 \cap \ldots \cap g_j = \phi$. This intuitively corresponds to decomposing the multivariate feature space into c distinct groups, which share none of the features among them. In this paper, as will be discussed later, these c groups correspond to clusters in the data and is empirically determined.

We further assume that each group is the result of an interaction of multiple group-specific effects. We model each $\mathbf{x}_t^{(gi)}$ as a concatenation of l different effects. We assume $\mathbf{x}_t^{(g_i)} \equiv (\mathbf{x}_t^{(sg_{i1})}; \mathbf{x}_t^{(sg_{i2})}; \ldots; \mathbf{x}_t^{(sg_{il})})$,

such that $\sum_{j=1}^{l} d_{g_{ij}} = d_{g_i}$ and $sg_{i1} \cap sg_{i2} \cap \ldots \cap sg_{il} = \phi$. Again, these conditions enforce that the each group is modeled as an interaction of l effects. Each effect operates on distinct feature spaces. We show this hierarchical composition enables us to model the groups and the interactions within each group.

We pose the problem of predicting the target at time point T, y_t as a regression problem using the independent covariates, $\mathbf{x}_t$.

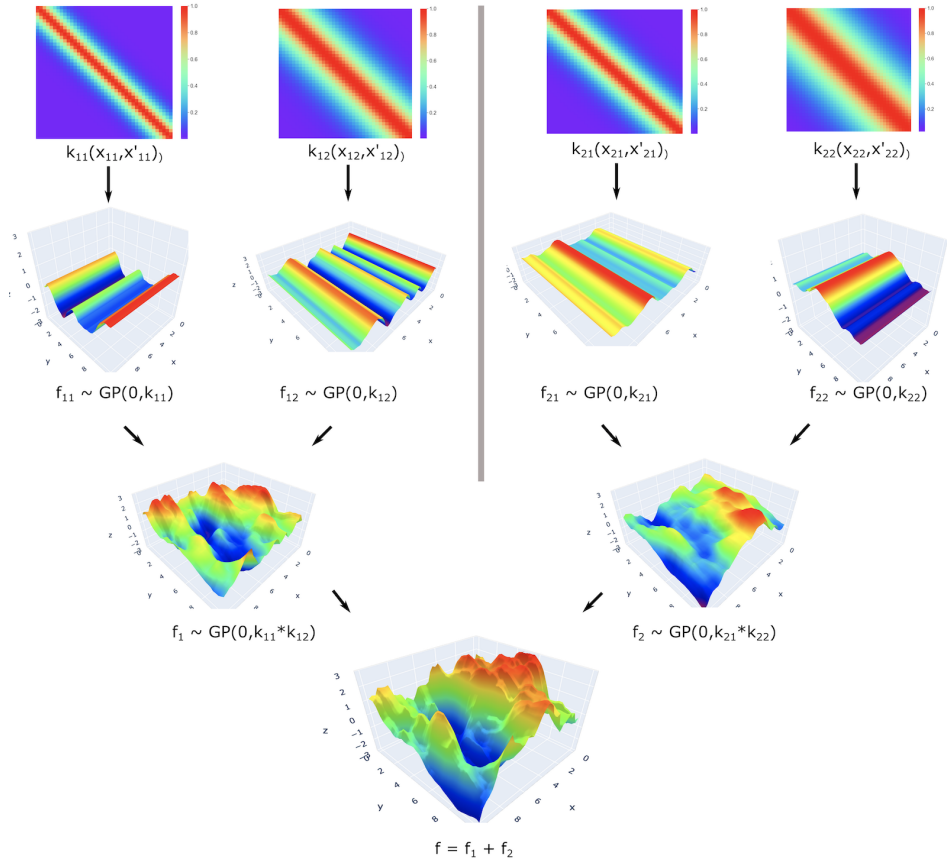


Fig. 1. A pictorial description of the Group AGP model. From top to bottom: two sets of kernels, k_{11} and k_{12}; and k_{21} and k_{22} are chosen. 1D functions can be drawn from each of the kernels, as shown by f_{11}, f_{12}, f_{21} and f_{22}. The functions f_1 and f_2 are drawn from GP priors, whose kernel functions are defined as $k_{11} * k_{12}$ and $k_{21} * k_{22}$, respectively. Finally, f is computed as an addition of f_1 and f_2.

3.1 Background: Gaussian Process Regression with Matern Kernel

A GP is a stochastic process, indexed by $\mathbf{x} \in \mathbb{R}^d$ (ignoring the time index t for simplicity), and is completely specified by its mean $m(\mathbf{x})$ and its covariance/kernel function $k(\mathbf{x}, \mathbf{x}')$, as shown in: $f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. The mean is usually assumed to be 0 ($m(\mathbf{x}) = 0$), and covariance between any two evaluations of $f(\mathbf{x})$ is $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$, where $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$. The covariance function defines the notion of nearness or similarity in GPs, and thus encodes different types of nonlinear relationships between the covariates and targets. GPs belong to the

class of Bayesian non-parametric models, where no assumptions are made on the functional form of the relationships between covariates and targets, and thus, these methods are known to learn highly non-linear boundaries. For details of GP for machine learning, the reader is directed to [41].

At the core of our proposed model lies the GP regression (GPR), which is stated as $y \sim \mathcal{N}(f(\mathbf{x}), \sigma_n^2)$. Given training data, $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ and a GP prior on $f()$, the posterior distribution of y_* (for an unseen input vector, $\mathbf{x}_*$), is a Gaussian distribution, with the following mean and variance which follows from the conditional and marginal properties of the multivariate Gaussian distribution:

$$\bar{y}_* := \mathbb{E}[y_*] = \mathbf{k}^\top (K + \sigma_n^2 I)^{-1} \mathbf{y} \tag{2}$$

$$\sigma_*^2 := \text{var}[y_*] = k_* - \mathbf{k}^\top (K + \sigma_n^2 I)^{-1} \mathbf{k} + \sigma_n^2 \tag{3}$$

Here, $\mathbf{y} = [y_1, y_2, \ldots]^\top$, and K is a matrix which contains the covariance function evaluation on each pair of training data, i.e., $K[i, j] = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}$ is a vector of the kernel computation between each training data and the test point, i.e., $\mathbf{k}[i] = k(\mathbf{x}_*, \mathbf{x}_i)$, $k_* = k(\mathbf{x}_*, \mathbf{x}_*)$, and I is an identity matrix. The marginal likelihood of the training data set for GPR is given by:

$$p(\mathbf{y}|X) = \mathcal{N}(\mathbf{0}, K + \sigma_n^2 I) \tag{4}$$

where X is the $(N \times d)$ data matrix.

We use the **Matern 3/2** class of covariance functions given by:

$$k_{\upsilon=3/2}(r) = \sigma_f^2 (1 + \frac{\sqrt{3}r}{\ell}) exp(-\frac{\sqrt{3}r}{\ell}) \tag{5}$$

where $r = \|\mathbf{x} - \mathbf{x}'\|_2$, ℓ denotes the characteristic length scale and σ_f^2 denotes the signal variance associated with the covariance function. This covariance function is a product of an exponential and a polynomial of order 1, and is suited for learning non-smooth behavior, such as those exhibited by financial time-series [19, 26]. We also empirically determined that Matern 3/2 provided better results than other kernel choices, like squared-exponential.

The above formulation assumes that a single, global characteristic length scale l is sufficient to capture the associations in all feature dimensions, which is a restrictive condition. To overcome this, a kernel that employs feature-specific characteristic length scales for each of the input dimension is used. Such a kernel is called an **automatic relevance determination (ARD)** kernel, and has been used as feature selection in econometric data [19].

Uncertainty in GP: GP regression (GPR) provides a measure of uncertainty along with the predictions via the variance in (3). The uncertainty is a measure of trust provided by our model, and might assist in algorithmic decision making by telling where (time points) and how much to trust the individual predictions.

3.2 Proposed model: Group-Additive GPs' (Group AGPs) description

We propose a novel Group-Additive GP (Group AGP) model, which hierarchically decomposes the underlying function as a sum of GPs, each operating on a group, and, further, models each group as an interaction of effects. We show how such a hierarchical decomposition exploits clustering to learn the structure of data. This additive structure imparts the Group AGPs framework increased flexibility and interpretability.

The **Group AGP model** decomposes the latent function value to be learnt, $f(\mathbf{x})$ for a given input D-dimensional $\mathbf{x}$ using the following additive form:

$$f(\mathbf{x}) = f_{g_1}(\mathbf{x}^{(g_1)}) + f_{g_2}(\mathbf{x}^{(g_2)}) + \ldots + f_{g_c}(\mathbf{x}^{(g_c)}) \tag{6}$$

where $f_{g_i}(\mathbf{x}^{(g_i)})$ is the group-specific latent function, which are taken to be independent stochastic processes, then the covariance function of $f(x)$ will be of the form:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{c} k_{g_i}(\mathbf{x}^{(g_i)}, \mathbf{x}^{(g_i)'}) \tag{7}$$

where $k_{g_i}(\mathbf{x}^{(g_i)}, \mathbf{x}^{(g_i)'})$ is a kernel function operating on each $\mathbf{x}^{(g_i)}$ (subset of features). This kernel function is hierarchically modelled as the following product form:

$$k_{g_i}(\mathbf{x}^{(g_i)}, \mathbf{x}^{(g_i)'}) = \prod_{j=1}^{l} k_{g_{ij}}(\mathbf{x}^{(g_{ij})}, \mathbf{x}^{(g_{ij})'}) \tag{8}$$

where $k_{g_ij}(\mathbf{x}^{(g_{ij})}, \mathbf{x}^{(g_{ij})'})$ is the kernel function for the j^{th} group.

Thus $f(\mathbf{x})$ is composed of c additive groups; and each additive group hierarchically models the interactions within itself. We refer to each $f_{gi}(\mathbf{x}^{(\mathbf{gi})})$'s as the **additive** groups and grouping of different dimensions into these groups as the **decomposition** of the function.

Example: To predict the consumer confidence index from SM, the additive groups could capture the sentiment and the content extracted from clusters that contain important topics of discussion affecting consumer confidence like growth of economy or jobs. In Group AGP formalism, each cluster is modeled as an interaction of content and sentiment effects; and the number of clusters correspond to the number of additive components (denoted by c in (7)).

A pictorial depiction of our model for two additive groupsis shown in Figure 1. Here, the kernels k_{11} and k_{12} can correspond to the content and sentiment values for a cluster about economy; and k_{21} and k_{22} can correspond to the content and sentiment values for a cluster about jobs. Then, f_1 shows a 1D function drawn from a GP, whose kernel is defined as a product of k_{11} and k_{12}. f_2 shows a 1D function drawn from a GP, whose kernel is defined as a product of k_{21} and k_{22}. Next, f is shown as an addition of two functions f_1 and f_2. The number of additive components has an important bearing in our model as it embeds interpretability along the feature space in modeling phase, which is preferred than *post hoc* attempts at explaining the model behavior to practitioners [28]. Our proposed decomposition into additive structures is beneficial in domains, where the dimensionality of the feature space is large, as in the case of natural language processing.

3.3 GroupAGP training and inference

The posterior distribution for a test input, $\mathbf{x}_*$ is obtained using (2) and (3), with the key difference being the use of the group additive covariance function given in (7) and (8). In Group AGP, the kernel matrix is the sum of kernel functions of individual additive components. **Additive decomposition** has been shown to better learn the structure of the underlying data compared to local kernels (such as squared exponential kernel) and allows better generalization capability away from the training samples [17]. Thus, our methodology is tailored to work with small training data regime, which are frequently encountered in many real world problems, where one has to learn from limited labeled data, such as expensive surveys, and learning is guided by temporal and contextual constraints.

The **hyper-parameters** of the Group AGP model consist of the length-scale (ℓ) parameter for the individual Matern kernel function for each effect, i.e., $k_{g_{ij}}(\mathbf{x}^{(g_{ij})}, \mathbf{x}^{(g_{ij})'})$, and one signal-variance (σ_f^2) parameter for each of the c groups. We fix the signal-variance parameter for all but one Matern kernel function within a group to 1. Thus there will be $(cl + c)$ kernel hyper-parameters and one additional likelihood noise, σ_n^2. The hyper-parameters are estimated by maximizing the marginalized log-likelihood of the training data (See (4)) using conjugate gradient descent.

3.4 GroupAGP model selection

In the previous section, we have described how to use Group AGPs to estimate the value of a given target for a given kernel structure based on c (number of additive groups). The space of models is large and choosing the right number of c requires a method to compare models. We specify a Group AGP model with c clusters as M_c, and perform model selection on c. [1]

Metrics for model selection: Rather than choosing basing our model evaluation on a single criteria, for example model complexity. Our discovery of the best model is guided by an interplay of three criteria - it should explain the observed data, reduce the error on the test set and provide the best interpretability score. To provide a great fit on the observed data, we want a model that maximizes the log-likelihood of the training data and penalizes the model's complexity. For the second criteria, we measure the error on unseen test point via mean log loss, that takes the absolute difference between the prediction and the target, while also incorporating the variance associated with the prediction. For the third criteria, we provide a novel scoring method called interpretabilty score that quantifies the ease of interpretation of what different additive components encode.

Now, we explain how the three criteria for model selection are defined:

(1) **Maximize the fit on observed data**: We use bayesian model selection to select the "right" model among a set of different hypotheses. Suppose for a set of observed data $\mathcal{D} = (X, y)$ and a set of models, $\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_c$.$\mathcal{M}_2$ means an additive model with 2 groups. First, we compute an approximation of the model evidence ($p(y|\mathbf{X}, \mathcal{M}_i)$), using maximum likelihood estimates as follows:

$$p(y|\mathbf{X}, \mathcal{M}_i) \approx p(y|\mathbf{X}, \theta_i \mathcal{M}_i) \tag{9}$$

$$p(y|\mathbf{X}, \theta_i \mathcal{M}_i) = -0.5 * y^T \mathbf{K}_{\mathcal{M}_i}^{-1} - \mathbf{0.5} * \mathbf{log}(2\pi)^{\mathbf{n}} |\mathbf{K}_{\mathcal{M}_i}| \tag{10}$$

But, the log marginal likelihood does not penalize complexity of the model and thus usually favor complex models. Here, model complexity is defined as the total number of hyper-parameters in our model.

Thus, to find a good balance between the fit and the model complexity, we use **Bayesian Information Criteria (BIC)** as measure of the model's fit on observed data; and is given as follows:

$$BIC(\mathcal{M}_i) = log p(y|\mathbf{X}, \theta_i \mathcal{M}_i) - 0.5 * |\theta_i| \log(|\mathcal{D}|) \tag{11}$$

where the first term corresponds to the marginal log-likelihood of the data and second term encodes model complexity. BIC is often used for model selection in GPs because of its simplicity and performance [16] [2]

(2) **Loss on test data**: To measure the quality of predictions of our model, we use the **mean log loss** (MLL) given as follows:

$$MLL = -\log p(y_*/\mathcal{D}, \mathbf{x}_*) = \mathbf{0.5} * \log(2\pi\sigma_*^2) + \frac{(\mathbf{y}_* - \bar{\mathbf{f}}(\mathbf{x}_*))^2}{2\sigma_*^2} \tag{12}$$

where $\sigma_*^2 = var[f_*] + \sigma_n^2$ is the predictive variance for GPR as given in (3). Compared to metrics like root mean squared error (RMSE), MLL normalizes the squared error loss by the variance of the target value. Lower values of MLL are better.

[1] Our model selection method can also be explored for kernel selection. However, since we have empirically determined that Matern kernel performs better than squared exponential kernel, the kernel type does not feature in model selection [38], and our kernel choice is always Matern, as described in 5.

[2] Although, we note that BIC assumes that the data are i.i.d. given the model parameters. This assumption is not true of GPs. Thus, BIC is not the optimal measurement of model fit in GPs.

(3) **Interpretability**: We describe a novel metric of quantifying interpretability called interpretability score (IS). As a notion of interpretability of a model $\mathcal{M}_c$, where c is the number of groups, we want to measure the ease of explanation of what each group encodes.
IS quantifies how each cluster is distinct from others. We assume that if the clusters encode distinct information, then the model output is easier to be interpreted by domain experts. So we quantify how many words are pair-wise common across the different cluster centroids. To get IS for $\mathcal{M}_c$, we take top-n posts that lie closest to each cluster centroid; and get the top-k words that describe the centroid (by removing stop-words, stemming etc).
Next, we define a matrix, M_c^t, where c is the number of clusters chosen to make prediction at time point t. Let c^i be the i^{th} cluster centroid; and k denote the top-k words representing the cluster centroid. Thus, each cluster, $\mathbf{c^i}$ is represented by a vector of top-k words, denoted by w_k^i

$$\mathbf{c^i} = \mathbf{w_1^i}, \mathbf{w_2^i}, ..., \mathbf{w_k^i} \tag{13}$$

We define $S^{i,j}$ as set intersection between any two cluster centroids, and is defined as $|c^i \cap c^j|$, between any i, j cluster.
Next, we define a matrix M_c as follows:

$$M_c = \begin{bmatrix} S^{0,0} & S^{0,1} & .... & S^{0,c} \\ S^{1,0} & S^{1,1} & .... & S^{1,c} \\ .. & .. & .... & .. \\ .. & .. & .... & .. \\ S^{c,0} & S^{c,1} & .... & S^{c,c} \end{bmatrix} \tag{14}$$

where $0 <= S^{i,j} <= c$. The value of $S^{i,j} = 0$ means that nothing is common and $S^{ij} = c$ means that all top-k words are common between c^i and c^j.
We define interpretability score as the sum of the upper diagonal entries of the matrix M_c (as it is symmetric), normalized by the maximum value of $S^{i,j}$.

$$score_{M_c} = 1 - \frac{\sum_{i,j=0}^{c} S^{i,j}}{0.5 * k * c * (c+1)} \tag{15}$$

Higher values of $score_{M_c}$ means more interpretable model.

Thus the interplay of these three factors help us choose the best Group AGP model. All our model's evaluation and comparative analysis is done with the selection model.

3.5 GroupAGP model interpretability analysis

Here, we go into more detail about our model's interpretability. Through interpretability we want to answer the following two questions and also describe how our model answers them:

(1) What is the contribution of each additive group towards the final prediction?
To find the contribution of each of the additive groups towards the final prediction, we use the posterior means and variance of each group. For instance, to find how much the first additive group ($f_{g_1}(\mathbf{x}_*^{(\mathbf{g_1})})$ contributes to the overall mean and variance, the **conditional posterior distribution**, $f_{g_1}(\mathbf{x}_*^{g_1})|f(\mathbf{x}_*) \sim \mathcal{N}(\mu_*^{g_1}, \Sigma_*^{g_1})$, can be obtained as [18]:

$$\mu_*^{g_1} = (\mathbf{k}_*^{(g_1)})^\top (K + \sigma_n^2)^{-1}\mathbf{y} \tag{16}$$

$$\Sigma_*^{g_1} = k_*^{(g_1)} - (\mathbf{k}_*^{(g_1)})^\top (K + \sigma_n^2)^{-1}\mathbf{k}_*^{(g_1)} \tag{17}$$

Similarly, one can express the contribution of other additive groups to Group AGP model's final posterior mean and uncertainty. The idea is that Group AGPs posterior mean will be the

Subreddit	Posts	Comments	Users
Economy	114040	361725	16313
Finance	198176	255006	53680
Jobs	267767	1384548	125422

Table 1. Reddit data description. Data was available from March 2008 until August 2019 for all the subreddits.

sum of individual posterior means of each of the additive groups and, similarly its posterior variance will be the sum of posterior variances of each of the additive groups.
The above formulation can also be used to determine the relevance of a group by checking if the corresponding posterior variance is 0. This means that the marginal likelihood selected a model that did not depend on the features of that group. This provides **automatic grouped feature selection** for our model, and induces **sparsity**.

(2) What is the contribution of each individual effect within the additive group?
We compare the length scales of the kernels for each interacting effects. We take the inverse of the length scale as **relevance** for that effect, conforming to literature [19]. Higher values of relevance translates to that effect being more important for our model.

4 METHODOLOGY

This section details how the Reddit and Google Trends (GT) data are transformed into multivariate features, which can be input to Group AGP. Reddit data entails extracting quantitative metrics or features that capture the content, psychological as well as sentiment features from SM data, clustering and aggregating them for all the posts at the necessary level of granularity. The purchase history data that we get via GT is already compiled and in a form that can be easily input to our model. We describe both data sets separately. A process overview of how the training and testing data are created, is described in Figure 2.

4.1 Reddit Data transformation

4.1.1 ***Description of Reddit data.*** We focus on Reddit as the SM platform. It is based on a user-to-topic subscription model and is divided into subreddits. These are content feeds focused on particular topics such as finance or jobs, etc. and are moderated by users for relevancy to that subreddit. User interaction occurs when users post something of interest (personal narrative, news etc.) and others comment on the post.

We focus our analysis on three subreddits namely **economy**, **jobs** and **finance**, as these contain "chatter" of our interest[3]. Researchers who have used Reddit for understanding social behaviors have, primarily, focused their analysis on specific subreddits pertaining to their tasks [4, 9, 52].

We study the texts written on the relevant subreddits in the form of **posts**. Each post is specified by a timestamp and its author. Other users can either upvote, downvote or comment on previous posts. For each post, we extract two types of features based on **content and sentiment** aspects.

Data extraction details: Reddit data was extracted using Google BigQuery[4]. Table 1 describes the time extents and characteristics of the relevant subreddits. The data spanned from March 2008 until August 2019. An important part in the data preprocessing timeline is removing the noise that is inherent to SM, which we shortly discuss.

[3]It is important to note that our methodology is scalable and data from more subreddits can easily be added into the model
[4]https://console.cloud.google.com/bigquery

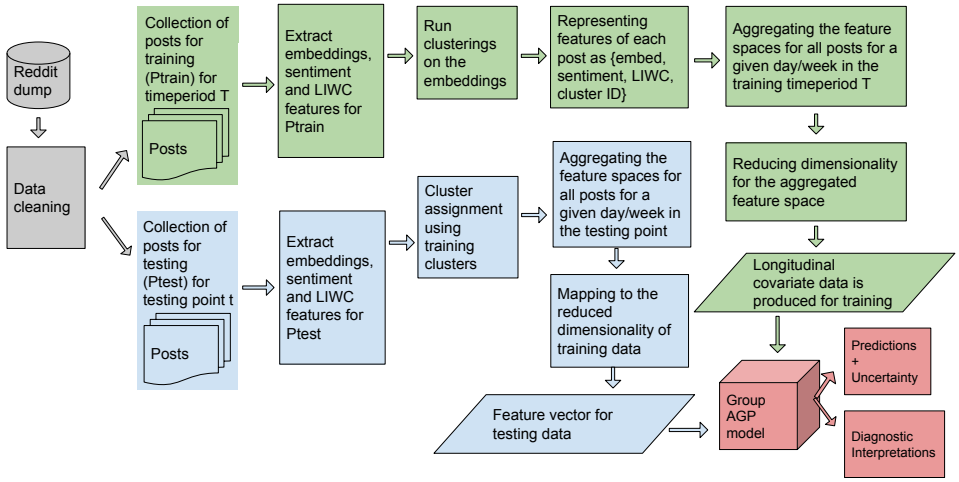


Fig. 2. The data processing pipeline showing a run of our algorithm, which describes how training and testing data are created from Reddit data. The green boxes mark the steps followed for creating training data, while the blue boxes describe steps for getting the testing data, which are then in fed to our Group AGP model (shown in red). The output of the model are predictions with associated uncertainties' as well as diagnostic interpretations. The gray boxes describe general pre-processing steps.

Summarized Dimensions	LIWC categories
linguistic	`funct, pronoun, ppron, i, we, you, shehe, they, ipron, article, verb, auxverb, adverb, preps, conj, negate, quant, number`
psych	`affect, posemo, negemo, anx, anger, sad`
cogperp	`cogmech, insight, cause, discrep, tentat, certain, inhib, see, hear, feel`
personal	`work, leisure, home, money, relig, death, social, family, friend`
informal	`swear, assent, nonfl, filler`

Table 2. Description of how the individual LIWC categories are summarized into 5 dimensions of our analysis

4.1.2 **Sentiment Extraction.** To get sentiment features we use both the **absolute sentiment** as well as the **psychological features** extracted from posts. To get the sentiment of posts, we use VADER, which has been shown to perform well in SM texts [21], as it takes in account acronyms, emoticons, slang etc, which are ever present in SM. For each post, we calculate the positive, negative and absolute sentiment ($|pos - neg|$).

Linguistic Inquiry and Word Count (LIWC): We use LIWC to extract features that correspond to emotional, structural and psychometric components present in the text [49]. LIWC classifier categorizes the words in each post into 64 categories related to social, cognitive, personal, informal language etc. For our analysis, we group the existing LIWC categories into 5 broad categories or summarized dimensions, named as **linguistic; psych; cogperp; personal and informal** to aid in downstream analysis of posts. Table 2 describes how individual LIWC categories are summarized into 5 dimensions. The descriptions of LIWC categories are taken from [36].

4.1.3 **Content Extraction.** Though most of the previous research focus solely on sentiment features, we extract both the sentiments and text content of the posts. We argue that content features are important and empirically show that these features alone capture the overall trend of the consumer sentiment. The rationale behind using content features is that SM users generate posts on temporally relevant and important issues, since the main motivation for posting is to elicit responses from the community. This is especially true for economic variables, as SM users can be expected to write and reflect more about recent economic happenings when economic projections are changing. This is similar to survey respondents' indicating that the economy is their most important issue during a recession or stock market crash [33].

For content extraction we use the concept of **embeddings**, where each text is mapped into a vector such that semantically similar texts have more similar vector representations than dissimilar ones [25]. We employ **sentence-BERT (SBERT)** to get the embeddings of each post [42]. SBERT is a modification of the pretrained BERT network that efficiently derives semantically meaningful sentence embeddings compared to BERT [14].

4.1.4 **Handling noisy Reddit data.** Reddit data are corrupted with noise in the form of advertisements, chatter in non-english languages and postings by bots, etc. To remove noisy posts, we perform **clustering** on the SBERT embeddings of the posts using the Kmeans clustering algorithm. When the cluster centroids are **visualized**, we notice that some clusters distinctly correspond to noise i.e either they are non-English language posts; or advertisement related postings etc; so we remove these clusters for downstream tasks. We also notice that repeated clustering on the remaining clusters followed by visual inspection of the centroids helped in further removing noisy clusters. While this method may not guarantee that all noise would be removed, it is a computationally cheap, which is an important consideration when working with millions of posts, and provided good performance in our case.

4.2 Creating daily covariates from Reddit data

The above subsections describe how content embeddings, linguistic and sentiment features are extracted corresponding to each post timestamped at t, say p_t. Lets denote the three feature spaces for p_t are stacked together as a single vector, $\mathbf{x}_{p_t} = embed_{p_t}, ling_{p_t}, sent_{p_t}$.

Next, we create a collection of posts spanning a time period of T (order of years), denoted by $\mathcal{P}_T = p_1, p_2, ..., p_t$, where $t \in T$, and perform clustering on their embeddings. Though any clustering algorithm could be employed, we used Kmeans clustering with c clusters. This results in each post in $\mathcal{P}_T$ to be associated with a cluster ID along with its feature space. Now the feature space representation for each post, p_t, in $\mathcal{P}_T$ is given as $\mathbf{x}_{p_t} = embed_{p_t}, ling_{p_t}, sent_{p_t}, c_i$, where c_i denotes the i^{th} cluster.

We will shortly describe how T and c are set. Our rationale behind clustering is two-fold: 1) It helps to unravel the different topics existing in the SM data. 2) It delineates the sentiment and psychological features associated with the posts by the topics. We, later, show how these clusters correspond to building interpretability into our model.

4.2.1 ***Aggregating features and dimensionality reduction.*** We need to aggregate covariate vector for all the features of the posts to the granularity (daily for predicting consumer confidence, and weekly for predicting unemployment claims) at which the targets for regression are available. Since there are numerous posts in a single day, the next step is to aggregate the feature representations of those posts to daily granularity. Since we have cluster IDs associated with each post, we first group all the posts by their cluster IDs, and then aggregate the features within each group.

Let us denote d as the index for daily time granularity, and the three feature vectors are given as $embed_d, ling_d, sent_d$, corresponding to the aggregated embedding, linguistic, and the sentiment features, respectively. We need to compute the cluster-wise feature vectors, denoted by $embed_{cd}, ling_{cd}$ and $sent_{cd}$, where c denotes the cluster index.

To get $embed_{cd}$ and $ling_{cd}$, we, first, assign all the posts during that day to their clusters and, then, average the embedding and linguistic vectors within each cluster. Mathematically, this can be written as follows:

$$embed_{cd} = \frac{1}{N_{cd}} \sum_{i=1, c_{p_{id}}=c}^{N_{cd}} embed_{p_{id}} \tag{18}$$

$$ling_{cd} = \frac{1}{N_{cd}} \sum_{i=1, c_{p_{id}}=c}^{N_{cd}} ling_{p_{id}} \tag{19}$$

where p_{id} denotes a post during a day d, N_{cd} denotes the number of posts assigned to the c^{th} cluster and $c_{p_{id}}$ is the index of the cluster that p_{id} belongs to.

The sentiment vector for a day, $sent_{cd}$ is the relative sentiment of all posts during that day. After all the posts are assigned to their respective clusters, we calculate $sent_{cd}$, as follows: An absolute sentiment for a post is given as $|pos - neg|$, where pos and neg are the positive and negative sentiment values assigned to that post by VADER.A post is marked positive if the absolute value is > 0; else it is negative. Next, $sent_{cd}$ is given as the number of positive posts minus the number of negative posts; and is normalized by the sum of the total positive and negative posts. Intuitively the relative sentiment for a day distills the prominent sentiment (either positive or negative) among all the posts during that day.

We concatenate the relative sentiment and aggregated linguistic features for conciseness and are, together, referred as sentiment features henceforth. Lastly, we also reduce the dimensionality of aggregated embeddings vector using PCA, and take top 10 dimensions as they capture most of the data's variance. Thus, the aggregated covariate data for a day is composed of reduced averaged embeddings and aggregated sentiment vectors.

4.3 Creating daily covariates from Google Trend (GT) data

Using GT, we are interested in the normalized search volume of goods and services that are traditionally studied in relation to consumer spending in the US. Following works by [7, 34, 35, 45, 50], we queried the GT website using search queries that match the categories defined in the US Bureau of Economic Analysis (BEA). Some examples of BEA categories to search queries include: *durable goods – computer and electronics*, *non durable goods – food and drink retailers*, and *services – health insurance*.

Corresponding to each category, the GT website was queried to extract longitudinal data detailing the normalized volume of queries on the Google search. As noted in the results section, data was extracted at weekly or monthly levels to match the temporal granularity of the corresponding target to be predicted.

4.4 GroupAGP framework description

In this section, we detail how Group AGP model is evaluated. Both the covariates and targets are timeseries data. We denote the time series of covariates as $\mathbf{X}_t \equiv \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$, where each $\mathbf{x}_t \in \mathbb{R}^d$; d refers to the number of features extracted from SM data.

The target time series is denoted as $\mathbf{y}_t \equiv y_1, y_2, \ldots, y_t$, where each $y_i \in \mathbb{R}$. We denote a time series starting at index t_{1+1} and ending at index t_2 as $\mathbf{X}_{t_1:t_2}$.

Algorithm details: Ours is a regression task, where the longitudinal covariates are $\mathbf{X}_t$ and the targets are $\mathbf{y}_t$. To make prediction at a given time step t, $\hat{y}_t$, we train our Group AGPs model on $\mathbf{X}^{train}$ and $\mathbf{y}^{train}$. The training data are the historical data derived from fixed length window of size w. The learning in our method is accomplished using data residing in window w, which is advantageous as our estimates are governed by only recent relationship between covariates and targets that is encoded within the window.

We introduce the notion of a **prediction step** (indicated by Δ), which encodes how far ahead can our model reliably predict. Thus to predict at time point t, rather than training till time point $t-1$, our training data are lagged further by Δ time steps, ranging from $t-\Delta-w; t-\Delta$.

The training data are used to estimate the kernel hyper-parameters of our Group AGP model, denoted as θ_t. At inference, the training data along with the input at t are used to estimate the target, denoted as $\hat{y}_t$, which is a Gaussian distribution with mean $\bar{y}_t$ and variance σ_t^2. We use the variance, σ_t^2, as the model uncertainty at t. To make predictions at $t+1$, the training window is shifted to the right by 1, while maintaining its size as w. The method to produce estimates for the entire time series is described as the routine `Monitor` in Algorithm 1.

The **`MonitoringWithMissingTargets`** routine in Algorithm 1 describes how our model is used to evaluate reducing the frequency of survey data i.e. some of the y_t values are not available or missing. To fill in the missing value at t', we use the historical data till $t'-1$ for training Group AGP model and predictions are derived at t'. The mean of the predictions $\bar{y}_{t'}$ is now our estimates at t'. This procedure is repeated for all the unavailable target values. More details about this experimentation is given in Experiment section.

Algorithm 1: Algorithms for predicting economic indicators using longitudinal covariates

Procedure `GroupAGP`$(\mathbf{x}, \mathbf{X}^{Train}, \mathbf{y}^{Train})$
  $\theta \leftarrow$ `train`$(\mathbf{X}^{train}, \mathbf{y}^{train})$
  $\bar{y}, \sigma^2 \leftarrow$ `predict`$(\mathbf{X}^{train}, \mathbf{y}^{train}, \mathbf{x}, \theta)$
  return $\hat{y}_t = (\bar{y}, \sigma^2)$

Algorithm `MonitorWithMissingTargets`$(\mathbf{X}_{0:T}, \mathbf{y}_{0:T}, w)$
  for $t = (w+1) : T$ **do**
    if `isMissing`(y_t) **then**
      $\bar{y}_t, \sigma_t^2 \leftarrow$ `GroupAGP`$(\mathbf{x}_t, \mathbf{X}_{t-w:t-1}, \mathbf{y}_{t-w:t-1})$
      $y_t \leftarrow \bar{y}_t$
    end
  end
  return $\mathbf{y}_{0:T}$

5 EXPERIMENT DESIGN

We describe the targets (ICS and UI) and different types of experiment setup. All experiments were done with two types of covariate data: Reddit and GT, and two targets as described below:

5.1 Survey and claims data description

In this section, we describe the two types of targets used in this work - consumer confidence index and the unemployment insurance claims data as follows.

Consumer Confidence Data: The *Index of Consumer Sentiment* (ICS) collected by the Institute for Social Research at the from University of Michigan is an important survey that measures U.S. consumer confidence. Researchers have extensively studied it in social sciences, as well as in prior works on SM forecasting. The survey consists of responses aggregated to a single index of five questions based on an individual's personal finances, their outlook for the economy in near future and their recent buying decisions. A nationally representative sample of approximately 500 respondents is drawn, and the poll is administered via telephone interviews. Although this indicator is released at monthly, the daily time-series is also available. We perform our analysis at both daily and monthly granularity. While the monthly ICS data is available from 1978, the daily ICS data is only available from January 2008 to May 2017. It is important to note that there are error bounds corresponding to target data too. With monthly ICS, the 95% confidence interval is +/- 3.29%, while the confidence interval with daily data is not available to us.

Unemployment claims revised: In [1], features are extracted from Twitter data between 2011 - 2013 related to job losses, and it is correlated with the revised claims of unemployment insurance (UI). The researchers argue that it is important to focus on understanding these UI claims, as they state the health of job flows, which is of importance to economists, market participants, and policymakers. Additionally since the UI claims data is available at high frequency, the researchers claim that it is a "good" indicator to test the performance of social media derived signal. They describe a linear model to predict the unemployment signal, with independent variable as the employment situation.

5.2 Performance metrics

As a measure of our model performance and its comparison with existing works, we employ the following metrics:

(1) Root mean square error (RMSE) and Mean log loss (MLL): RMSE is the square root of the mean absolute errors for a set of predictions. This is summarized over the test set. However, this quantity does not take care of the variance of the predictions in the test set, so we use the mean log loss (MLL). MLL is not used as a metric when our model's is compared with existing works, since it is a measure for Bayesian methods and existing works are not Bayesian, hence RMSE is only reported for those experiments. Additionally both metrics vary similarly in our experiments.
(2) Detrended cross-correlation analysis (DCCA): DCCA is known to be a robust measure of cross-correlations between two different but equal length time series, in the presence of non-stationarity and outliers [37].

5.3 Different types of experimental setup

We conducted experiments to assess three important characteristics of our model, described as follows:

5.3.1 **Model selection.** : We perform model selection by changing number of additive components, and studying the performance of each model using three criteria: MLL, Bayesian Information Criteria (BIC) and interpretability score (IS).

5.3.2 **Model's predictive accuracy and descriptive nature of interpretations.** Reddit data is smoothed at 28 days, Δ is kept at 28. While GT data is studied at monthly granularity for ICS. We test our model performance as well as use the interpretability formulations in Equation 17 to retrieve the descriptive statistics corresponding to each target.

5.3.3 **Comparison and baselines.** We compare our Group AGP model with three existing methods [1, 30, 34],which are representative of the existing works in this area [7, 8, 10, 33, 35]. [30] uses the ratio of sentiment in the past few days as a feature to predict ICS in a linear model. Later works [10, 33] have studied their work with longitudinal data. [34] which is similar to [7, 8, 35] uses an autoregressive linear model with additional features extracted from SM or GT data for both ICS and UI. [1] also has a linear model with content signals extracted from Tweets to predict UI.

Additionally, we also perform the following baseline studies: Group AGP (Rand): The additive components of our model correspond to random subsets of feature spaces, to see the net gain achieved by a hierarchical additive decomposition that is guided by natural clusters in data. Concat model: A GPR model with a Matern kernel is used with the feature space (both content and sentiment) of all the clusters concatenated together. Concat+ARD model: An automatic relevance detection (ARD) kernel is placed on concatenated feature space as described in Concat model. Sum model: Rather, than hierarchically modeling each additive group as a product of content and sentiment within it, we add the content and sentiment kernels in Sum model.

5.3.4 **Fine-granularity evaluation: A special scenario evaluating our model for reducing frequency of ICS.** We perform this important experiment as a special scenario to investigate the extent to which SM data can reduce the need for frequent surveys, while preserving a high degree of correlation with traditional survey measures. Our approach is to create a scenario when survey data is available every other month or after two-months; and then we use our model to estimate the targets for missing months.

Our experimental setup is as follows: In first scenario, we "remove" every alternate 28-day period of ICS data, after an initial training window of 2 years. The algorithm **`MonitorWithMissingTargets`** is used to estimate the ICS values for the missing days, one day at a time. After each day, the estimated value for that day is "fed back" into the training data for the subsequent prediction. Thus, in the following time steps, the model uses a mix of observed and estimated target values to produce the estimates. This process is repeated until the next survey data is available. The `MonitorWithMissingTargets` algorithm is applied again whenever there are subsequent unavailable survey targets. The estimated targets are compared with the ground truth values to calculate error. The same experimental setup is repeated for scenarios in which the survey data is available at a further reduced frequency, i.e., it is available for a 28 day period after every 56 days, then 84 days, and 112 days. Hence, for each window, the model must generate predictions using fewer days of survey data.

6 RESULTS AND DISCUSSION ON INDEX OF CONSUMER SENTIMENT

We describe our results for predicting Index of Consumer Sentiment (ICS) using Reddit and GT data. ICS is available at daily granularity. Analysis of Reddit data for ICS prediction is done at daily granularity. While GT data are available at monthly granularity, hence for its analysis ICS is upsampled to monthly granularity.

6.1 Reddit data for ICS

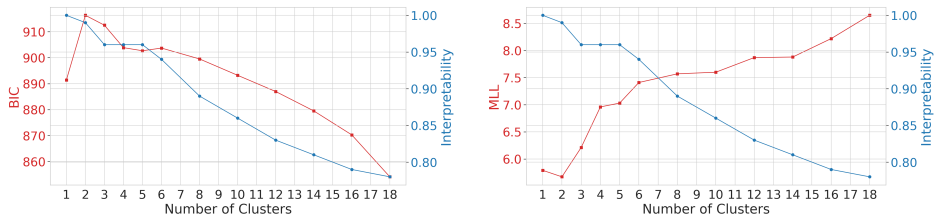
6.1.1 Model selection. As a criteria of model selection, we compare how the Bayesian Information Criteria (BIC), mean log loss (MLL) and interpretability vary for each model configuration and is demonstrated by Table 3. Studying the individual components of BIC, namely log likelihood and complexity, we observe that as the number of additive components increase the log likelihood of the model keeps on increasing while at the same time the mode complexity also increases. This highlights the fact that as the model becomes more and more complex, it is fitting the training data

Additive	RMSE	DCCA	MLL	Variance	Loglik	Complexity	BIC	Interpret.score
1	5.91	**0.82**	5.79	6.29	904.47	**26.22**	891.36	NA
2	**5.49**	0.78	**5.67**	5.02	939.26	45.88	**916.32**	**0.99**
3	5.53	**0.82**	6.21	3.92	945.28	65.54	912.51	0.96
4	5.82	0.79	6.96	3.80	946.46	85.20	903.86	0.96
5	5.65	0.77	7.03	3.52	955.10	104.86	902.67	0.96
6	5.67	0.80	7.41	3.39	965.96	124.52	903.70	0.94
8	5.80	0.81	7.57	3.19	981.43	163.85	899.51	0.89
10	5.76	0.81	7.60	3.11	994.80	203.17	893.22	0.86
12	5.77	0.81	7.87	3.07	1008.23	242.50	886.98	0.83
14	5.66	0.81	7.88	2.86	1020.44	281.82	879.52	0.81
16	5.74	0.77	8.22	2.86	1030.85	321.14	870.28	0.79
18	5.99	0.79	8.65	2.90	**1034.35**	360.47	854.12	0.78

Table 3. Results for model selection in Group AGP model to predict ICS using Reddit data: Comparing the different metrics of performance with different additive components in Group AGP. Log-likelihood and model complexity together constitute the BIC score. Variance is the average predictive variance for test data.

better, likely overfitting the training data. This is also empirically shown by increasing errors on test data.

It is important to note that since our model gives uncertainty along with each prediction, as the model becomes more complex its variance decreases while the RMSE increases, i.e. the model is more certain in making a erroneous prediction. MLL which captures both the RMSE and predictive uncertainty is, thus, a better metric for measuring model's generalization ability on unseen data. The third criteria is the interpretability score, which captures the ease of explainability along the feature space. We notice that it increases as the number of additive components increases. Thus, the optimal number of components at which these three criteria intersect is at 2. Thus our Group AGP model to predict ICS is given by 2 additive components corresponding to the 2 clusters in the SM data and each component is modeled as a product of its content and sentiment features. The interplay of these three criteria is also depicted in Figure 3. For optimal model, higher BIC, lower loss and higher interpretability scores are preferred.



(a) BIC vs. Interpretability (b) MLL vs. Interpretability

Fig. 3. Results for model selection in Group AGP model to predict ICS using Reddit data: Comparison of BIC (Bayesian Information Criteria), MLL (Mean Log Loss) and interpretability scores as a function of the number of clusters. For optimal model, higher BIC, lower loss and higher interpretability scores are preferred. Please note that the range of y-axes are different in all three subplots and thus are not intersecting, contrary to appearances.

6.1.2 Model's predictive accuracy and descriptive nature of interpretations. The predicted time-series given by our optimal model is shown in Figure 4. We notice that our model does capture the overall trend in ICS values, and also some of the finer longitudinal movements, like those between 2012 and 2015. Our model under-predicts for some of the noticeable peaks in Jan. 2015 and Jan 2017. We also observe that our model does not perfectly capture the significant drop in ICS during later half of 2011, but it does predict a noticeable dip around that time point, albeit with a delay. However, each of our predictions are marked with values of uncertainty given by our model.

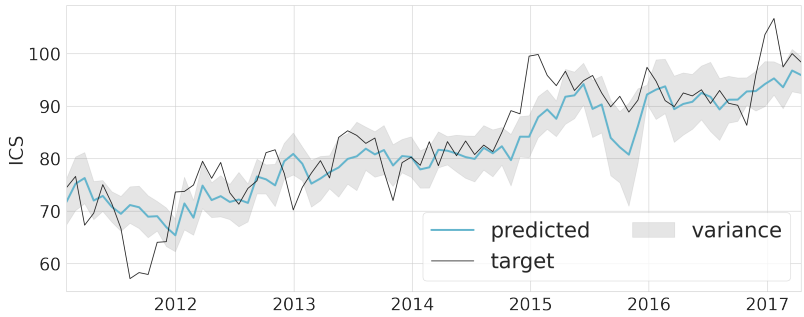


Fig. 4. Performance of Group AGP to predict ICS using Reddit data at daily granularity (smoothing window size - 28, Δ - 28)

As a quantitative measure of interpretation, we compare the interpretation scores of our Group AGP model with different numbers of clusters (c = 1,2,...,18). We find that as the number of clusters increases the interpretability scores declines consistently, suggesting that the clusters increasing become less distinct in their content. This might make the task of explaining what each of the additive components mean in our model cumbersome to the domain expert. However, these scores are contextualized by the target (i.e. macroeconomic indicator) of our model; and are thus data dependent.

As a qualitative measure of the interpretable clusters, we see the evolution of prominent topics throughout the years, which are obtained by performing kmeans clustering on the data. We noticed that three main clusters that emerged over the years. Each cluster is visualized by top-k words in the posts that lie closest to in the cluster centroid, and are given as follows:

(1) Cluster 1: `ago applied applying asked boss career contact cover degree department email employer entry field fit graduated hire hired hiring hr internship interview interviewed interviewing interviews letter linkedin manager phone quit received recruiter references resume school send sent skills told worked`
(2) Cluster 2: `acquire aid arrange banks borrowers collateral crisis economic faxing fee fiscal fulfill hassle installment instant lenders monetary payday repayment solution sufficient troubles unexpected unforeseen unplanned unsecured unwanted`
(3) Cluster 3: `america banks bonds capital cash crisis currency debt dollar economic fed funds gold inflation installment instant investors loan loans markets monetary mortgage payday price rates reserve stocks street tax trading unsecured wall wealth`

We see that over the years mainly three broad categories of topics emerge. We assign cluster 1 to jobs (availability/loss); cluster 2 to financial (losses/gains) and cluster 3 to economy (recession/growth) topics.

Contribution of each additive component Our modeling framework helps answer further questions such as which topics were deemed more important for predicting consumer confidence, and which topics contributed to the predictive uncertainty at different time points. We find the contribution of each of the additive components towards the final posterior mean and variance, as shown in Figures 5 and 6. Since our model selection has chosen 2 additive components, at each time point we highlight the clusters to which each of the additive components lie.

As clearly evident in Figure 5 for the years around 2011 - 2014, our model deemed jobs and economy as important topics affecting its predictive accuracy; while during the later years (2015 onwards) lending and finance topics gained prominence. It is also interesting to note that some topics have negative contribution to the overall predictive mean. Also interesting to notice is how we recover the daily ICS index, as a sum of the two additive components. Similarly, the anomalously high uncertainty observed just prior to 2016 can be attributed to the lending and finance cluster, as seen in Figure 6.

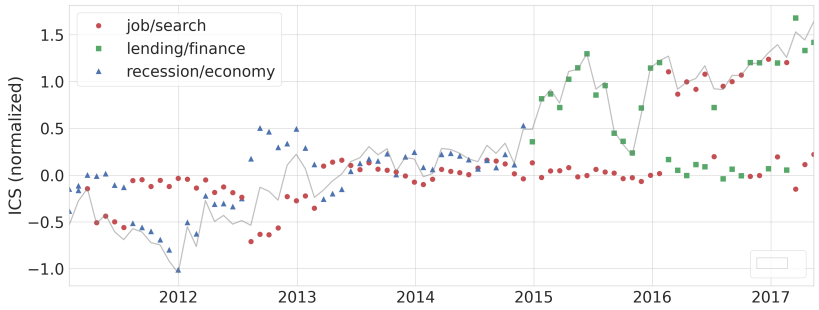


Fig. 5. Interpretation for ICS using Reddit: Depiction of the posterior means of each of the two additive components. The line is the predictive mean of ICS and is recovered as a sum of the two components.

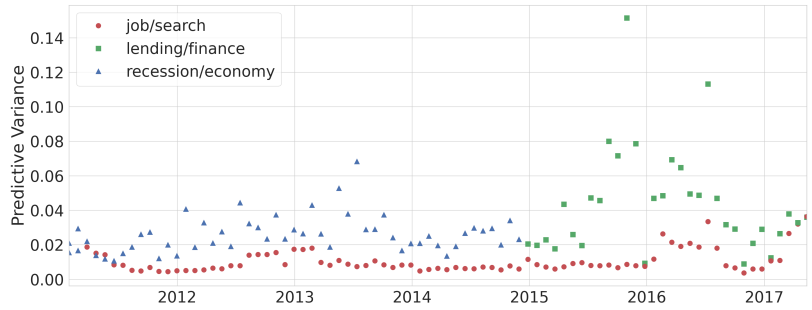


Fig. 6. Interpretation for ICS using Reddit: Depiction of the posterior variances of each of the two additive components compared against each other.

Relevance of individual effects: Content vs Sentiment Since each additive component in Group AGPs is modeled as an interaction of effects of content and sentiment. We can get further

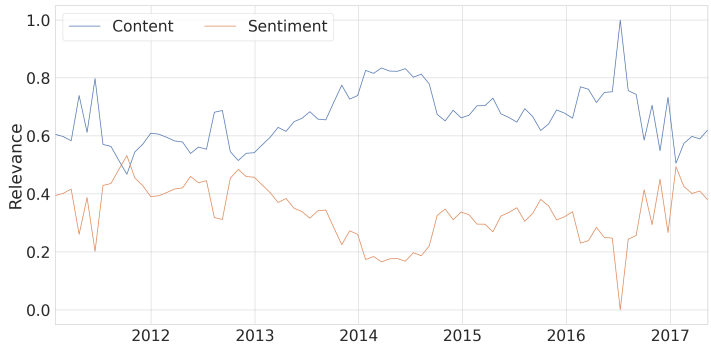


Fig. 7. Interpretation for ICS using Reddit: Depiction of how the average relevance of content and sentiment compare against each other.

insights into the model and answer questions regarding which of the effect (i.e. either content or sentiment) is deemed more important for prediction. Figure 7 shows the averaged relevance of content and sentiment effects. As evident our model mostly chooses content over sentiment in predicting consumer sentiment. This points out that one of the reasons previous studies failed to replicate the initial success might be because they used sentiment existing in the SM data, which has a weaker signal in predicting consumer confidence as compared to the content.

6.1.3 Baselines and comparisons. : As shown in Table 7, we note that the proposed model is consistently better than the existing methods, both in terms of RMSE and DCCA.

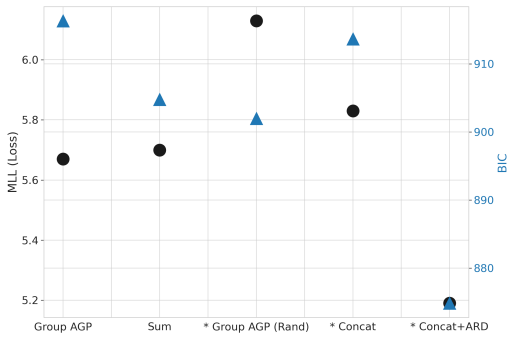


Fig. 8. Baselines to predict ICS using Reddit data: Comparison of different kernel configurations along the two dimensions of BIC (model fit) and MLL (loss). The model names on x-axis marked with asterisk denote models where no interpretable inferences can be drawn. For model selection, higher BIC and lower MLL is preferred.

The time-only baseline produces an RMSE of 7.31 (vs 5.49) and DCCA of 0.51 (vs 0.78); which is significantly higher than the results obtained with the Reddit data. Similar performance degradations were observed for other metrics. This bolsters the fact that using Reddit data does help to make better predictions for consumer confidence index, than just using past target values. Figure 8 shows how the different kernels compare against each other. Our Group AGP model has the highest BIC

with lower MLL loss, both of which are preferred. We show how "Group AGP (Rand)" not only provides poorer BIC and high loss; but also imparts no notion of interpretability. The "Concat" model has lower complexity as compared to Group AGP, however it has a higher loss on the test data, and no insight into interpretability. The "Concat (ARD)" model is impaired by huge complexity.

6.2 Google Trends data for ICS

6.2.1 Model selection. We studied model selection in terms of the number of additive components. As described earlier, for calculating ICS the US Bureau of Economic Analysis (BEA) studied consumer sentiment across three expenditure types: durable, non-durable and services. Each expenditure type is composed of distinct categories, which exemplifies a class of personal expenditure products. Each category is composed of a set of products, called sub-categories. For example, within durable goods, there is a category called "motor vehicle and parts", which has following sub-categories - auto parts, vehicle brands, vehicle shopping, automotive and auto financing. The GT tool is queried with the sub-categories. There are 4 categories each in durable and non-durable goods; while 7 categories in services, as used in literature [45].

For model selection in Group AGP, we worked with two models. One is a Group AGP model with three additive components, where each component is a kernel corresponding to either durable goods, non-durable goods or services. Each component is further modeled as an additive interaction of its constituent categories. We call this the "3-component" model. Our intuition behind this model is that since each category and sub-category are distinct, to get a comprehensive understanding of each personal expenditure product, all of its constituent categories and sub-categories need to be summed up. This can be contrasted in

The second model is the one where each of the categories are modeled as additive components, and each category is modeled as additive interaction of its constituent sub-categories. Since there are 15 categories, it is called a "15-component" model. The "3-component" model had an RMSE of 6.29 and DCCA of 0.93, with a BIC of -44.95. The "15-component" model has an RMSE of 6.81 and DCCA of 0.92, and BIC score was -84.30. Thus, the "3-component" model was selected for performance analysis.

6.2.2 Model's predictive accuracy and Descriptive nature of interpretations. Figure 9 shows the predictive accuracy of model. We see that our predictions do track the longitudinal ICS time-series, while providing uncertainties with them.

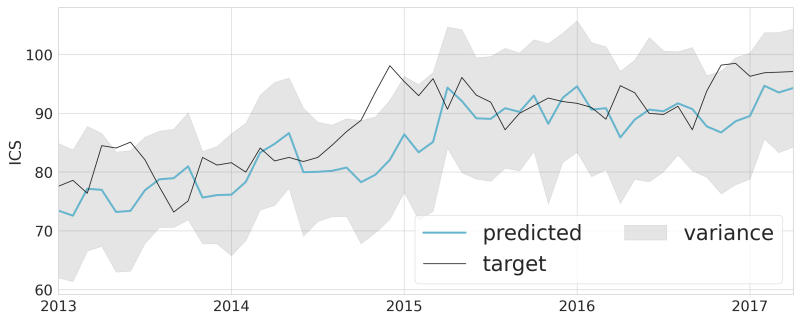


Fig. 9. Performance of Group AGP to predict ICS using Google Trends data at monthly granularity

Figures 10 and 11 shows the contribution of each of the additive components towards the final posterior mean and variance. Our model deems durable and non-durable goods to be better predictor of consumer confidence than services prior to mid-2015. After that services component start to play a significant role. Similarly our model attributes the high predictive variance time points to non-durable goods. These are interesting insights which needs further work by domain experts.

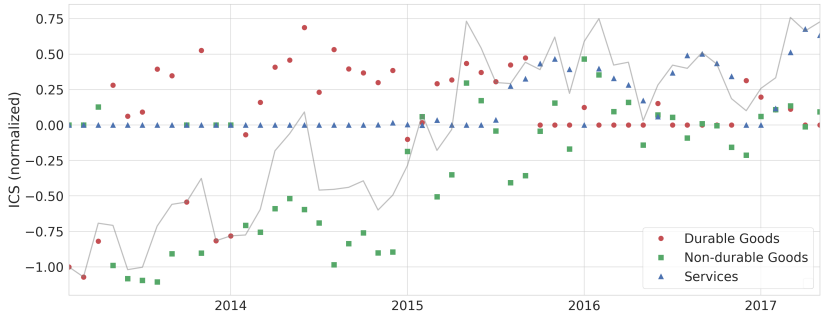


Fig. 10. Interpretation for ICS using Google Trends: Depiction of the posterior means of each of the three additive components. The line is the predictive mean of ICS and is recovered as a sum of the three components.

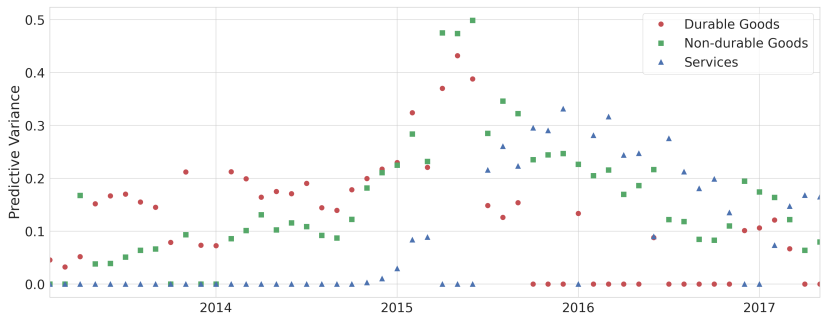


Fig. 11. Interpretation for ICS using Google Trends: Depiction of the posterior variances of each of the three additive components compared against each other.

6.2.3 Baselines and comparisons. As shown in Table 7, the proposed "3-component model" is significantly better than existing methods [1, 34], both in terms of RMSE and DCCA. While, the **concat (ARD)** model was marginally better than the "3-component model", both for RMSE (6.20) and DCCA (0.96), it had a lower BIC score. This shows that GT data has the potential to predict ICS with high accuracy. Since [30] works in sentiment, and there is no concept of sentiment in GT data, there are no results for this comparative work.

7 RESULTS AND DISCUSSION ON UNEMPLOYMENT INSURANCE (UI) CLAIMS DATA

We experimented with UI claims data, which has been studied in the past using SM data (Twitter and Google Trends) [1, 7, 8]. Data and method of extraction of covariates remain the same as with the ICS. Since the UI claims targets are available weekly, the covariates from Reddit and GT are aggregated to weekly granularity. Similar training and testing procedures are employed as in the prediction of ICS.

7.1 Reddit for UI

7.1.1 Model Selection. The detailed performance metrics for model selection is given in Table 4. The optimal model selection, which has the highest BIC, lowest loss and best interpretability score is given with 1 cluster or the training data. Thus, our modeling framework is reduced to a product kernel of content and sentiment effects.

We attribute it to the fact that UI data are a direct measurement of the unemployment claims filed by people and does not have a first order dependence on other factors (like finance etc); and thus one cluster is enough to capture the content that is predictive of UI indicator. As mentioned earlier, the evaluation methodology is based on a sliding window protocol, i.e., to make prediction on a test point, t, we train on a historical window of data (two years); and, further, to make predictions at $(t + 1)$, we slide our training window by 1. Since clustering is performed for each training window, the cluster contents are updated every time. So, even though, at each test point our Group AGP model selects 1 canonical cluster, its content varies. However, we do not expect a drastic change of content with subsequent time points, unless a major economic shock, like recession affects the market. This study for UI data spans years from 2011-2017 during which no recessions were reported[5].

Empirically, we also notice that the top words of the cluster centroid at each time step of our model remains mostly same, indicating the existence of a singular semantic theme within the subreddits that is predictive of UI claims data. The top-k words averaged over all time points were: `job, company, would, interview, work, position, get, like, time, know, really, experience, offer, got, resume, jobs, good, looking, back, people`, which is indicative of people looking for new employment.

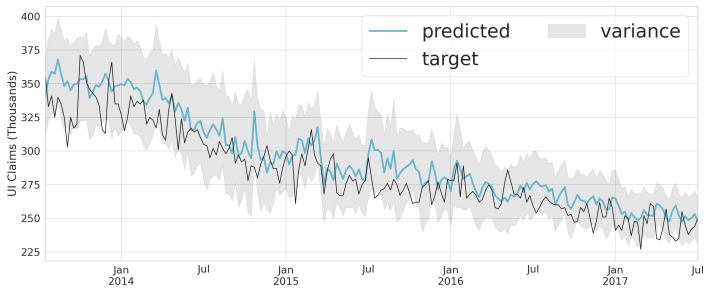


Fig. 12. UI claims using Reddit: Model performance - Comparing the model predictions with observed targets at weekly granularity

7.1.2 Model's predictive accuracy and descriptive nature of interpretations. The Group AGP model predictions for UI data are shown in Figure 12. We note that the model predictions tracks the UI consistently, and the predictive uncertainty shrinks over time. Figure 13 shows the relative relevance of the content and sentiment for prediction at each time point. We notice that sometimes our model deems sentiment to be more important and other times content to be important; contrary to the case of consumer sentiment prediction where our model, consistently, deemed content to be of more relevance than sentiment.

[5]https://fred.stlouisfed.org/series/ICSA

Components	RMSE	DCCA	MLL	Variance	Loglik	Complexity	BIC
1	16.61	0.97	**4.28**	196.83	-18.59	18.54	**-27.86**
2	17.52	0.93	4.34	194.59	-18.04	32.44	-68.52
4	18.90	0.94	4.60	144.39	-12.36	60.25	-84.97
6	18.07	0.95	4.58	131.41	-11.50	88.06	-111.05
8	19.56	0.93	4.78	123.55	-9.65	115.87	-135.18

Table 4. UI claims using Reddit: Model Selection - Comparing the different metrics of performance with different additive components in Group AGP. Log-likelihood and model complexity together constitute the BIC score. Variance is the average predictive variance for test data.

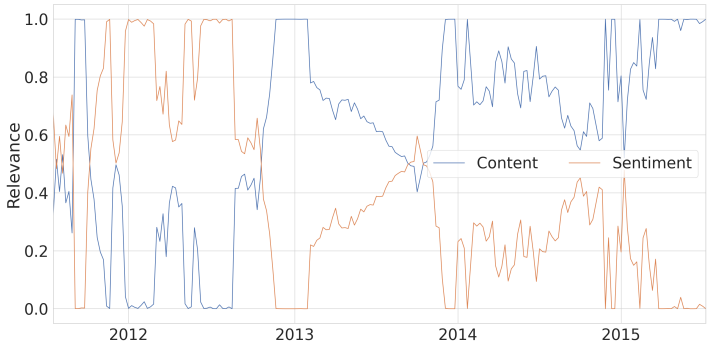


Fig. 13. UI claims using Reddit: Interpretation Analysis - Depiction of how the average relevance of content and sentiment compare against each other.

7.1.3 Baselines and comparisons. Table 7, we note that the proposed model is better than the existing methods in terms of DCCA. However it is poorer than [34] in terms of RMSE. We attribute this to the fact that [34] uses an autoregressive model along with the features. Since Group AGP does not explicitly model autoregressive nature of the targets, it suffers in its performance here. However, adding a temporal kernel with Group AGP is straightforward, and will be explored as future work.

Table 5 shows the results of different kernel compositions, like sum (of content and sentiment kernels) and concatenation of the feature spaces (given as concat) and concatenation with ARD kernel (given as Concat_ARD). Again, we notice that Group AGP gives the best performance in terms of higher BIC and lower loss (MLL).

Kernel	RMSE	DCCA	MLL	Var	Loglik	Complexity	BIC
Group AGP	16.61	0.97	**4.28**	196.83	-18.59	18.54	**-27.86**
Sum	16.64	0.97	**4.28**	195.62	-18.48	23.17	-30.06
Concat	17.68	0.98	4.34	209.86	-21.30	13.90	-28.26
Concat_ARD	17.24	0.98	4.48	165.75	-6.65	88.06	-50.68

Table 5. UI claims using Reddit: Baselines comparisons of different kernel configurations along the dimensions of BIC (model fit) and MLL (loss).

7.2 Google Trends data for UI

GT has been studied in the past to predict UI [8, 50]. The idea behind the use of GT is that it is natural for an individual to look for jobs or search for unemployment claims on search engines once they are unemployed. We have used the same GT categories as past research [8, 50], i.e., "jobs" and "welfare and unemployment". Thus, GT tool is queried with these categories from 07/2011 to 07/2016, aligning with the UI claims target. Both data are at weekly granularity. The GT tool returns the normalized values of search queries belonging to the two categories, which translates to two features. Since there is not a rich or high dimensional feature space to be explored, our Group AGP model reduces to a simple additive model with two components, each modeling a single feature with Matern kernel. Thus, no model selection is explored in this case.

7.2.1 Model's predictive accuracy and descriptive nature of interpretations. Figure 14 shows the predictive performance of our Group AGP model. We notice that our predictions do track the targets, however it consistently overpredicts for the period between July 2014 to July 2015. Our model also provides high uncertainty with those predictions. We attribute this to the fact that the covariates solely come from Google searches, which are not representative of the unemployed population. Also individuals might be searching for welfare or social security claims even when employed or for other reasons. Figure 15 show that our model deems jobs searches as better predictor of UI claims compared to welfare and unemployment searches. We notice that jobs component closely track the UI claims most of the time except between the period of July 2015 - Jan 2016, and further analysis from domain experts is needed to explain the observations

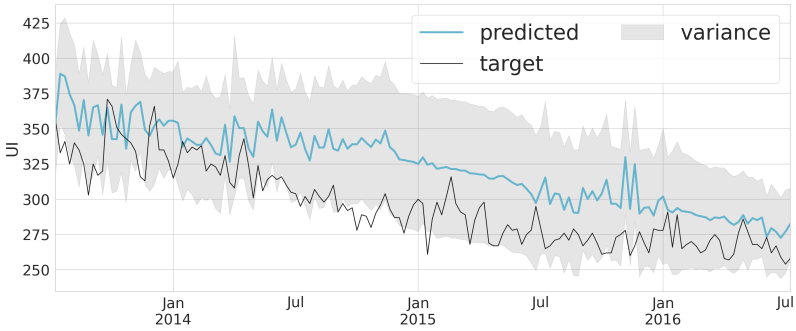


Fig. 14. UI claims using GT: Model performance of Group AGP to predict UI using Google Trends data at monthly granularity

.

7.2.2 Baselines and comparisons. We compared Group AGP with concat kernel as well as with related works [1, 30, 34]. The concat kernel has a RMSE of 32.50 and a DCCA of 0.88. As shown in Table 7, while Group AGP performs better than [1], it is outperformed by the model proposed by [34]. As noted earlier, in comparison with Reddit data, the autoregressive nature of [34] provides it with an advantage over models that ignore the temporal aspect.

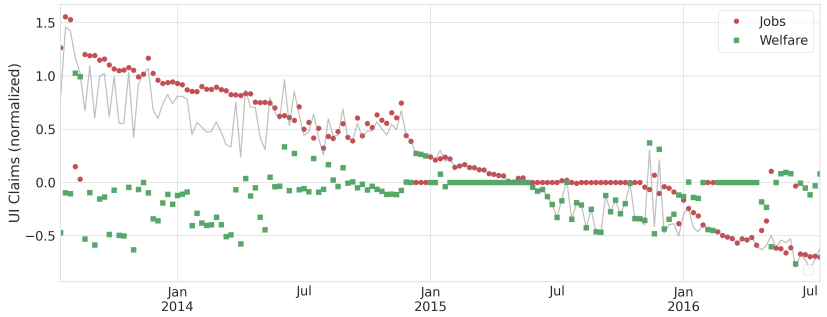


Fig. 15. UI claims using GT: Interpretation - Depiction of the posterior means of each of the two additive components. The line is the predictive mean of UI and is recovered as a sum of the three components.

Gap months	RMSE	DCCA
1	4.65	0.99
2	5.56	0.90
3	7.69	0.91
4	7.43	0.87
5	8.70	0.77

Table 6. Fine-granularity evaluation: Reducing frequency of ICS using Reddit data. Gap months denote the months when no survey is done.

7.3 Fine-granularity evaluation: A special scenario evaluating our model for reducing the frequency of ICS

Here, we show results on an important evaluation using Reddit data for ICS for reducing the frequency of ICS survey. Reddit is chosen for this analysis as it is available at fine daily granularity. Table 6 shows that Reddit data can be used to reduce the cost of surveys, by estimating accurate responses in periods when no survey was undertaken. The step in the Table 6 refers to the frequency of conducting the survey - Step of 2 means survey can be done every 2^{nd} month; step of 3 means every 3^{rd} month, so on. As we decrease the frequency of the surveys, the estimates during the off-month period decrease at a steady rate.

Figure 16 shows the performance of our model when survey data for alternate month was made unavailable. We see how our predictions closely track the targets, and is an encouraging result showing that Reddit data can to used to supplement surveys. In Figure 17, we take a snapshot the timeseries (July-Jun 2014), so that we focus on how our predictions and their associated uncertainties vary when surveys are run every second, third and fourth month respectively for the duration of our experimental data. While the estimates are close to the target for some of the months, they are off for others.

Comparing the left to the rightmost diagram in Figure 17), we notice that as more time elapses between surveys, the quality of the estimates suffer. This is because to make prediction at time t, the model relies on a window of past 2 years of data, where the missing survey points have been replaced by the model estimates. When the time elapsed between the surveys is large, the predictions rely mostly on model estimates. However, once the next survey is available, the model can again reliably predict the subsequent target values.

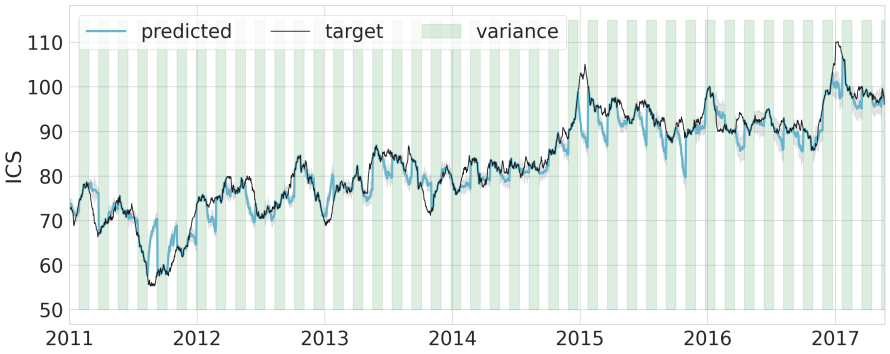
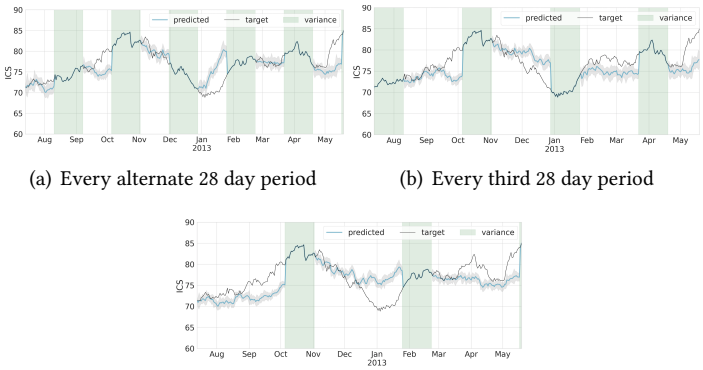


Fig. 16. Comparison of our predictions with survey data for experiments conducted to reduce the frequency of surveys by half. The shaded green regions marks the 28 day periods when surveys are available, which is every alternate time-period. Notice how our estimates closely follow the targets. A close-up of a random sample of this time-series is shown in Figure 10 (leftmost).



(a) Every alternate 28 day period (b) Every third 28 day period

(c) Every fourth 28 day period

Fig. 17. A snapshot of the time-series (for the year 2013) displaying targets, predictions and corresponding uncertainties for the set of experiments to reduce the survey frequency. The shaded green color marks the 28 day periods when surveys are available. The white regions mark the time period when our model was used to estimate survey values. Our predictions are shown as blue. It is important to see how the model estimates and uncertainties compare with the targets at these time-periods. From L to R - surveys available for (leftmost) every alternate 28 day period, (middle) for every third 28 day period and (rightmost) surveys available for every fourth 28 day period. As we go from left to right, the frequency of surveys decreases. Notice how uncertainty varies with the predictions in the off-months when survey is not done.

7.4 Discussion on the effect of different types of data for understanding ICS and UI

While previous works have use Twitter for estimate ICS, we use Reddit data. There is an important difference in communicative dynamics between the two platforms - Twitter is more appropriate to study social connections and diffusion or viral-ity of information [39], whereas Reddit provides a discussion platform through which individuals can discuss their plans of buying commodities or houses, job losses or gains and general chatter about economy. Reddit provides moderated and dedicated communities in the form of subreddits for discussion, which are not present on Twitter.

	Consumer Confidence Index (ICS)				Unemployment Claims (UI)				
	Reddit		Google Trends		Reddit		Google Trends		
Method	RMSE	DCCA	RMSE	DCCA	RMSE	DCCA	RMSE	DCCA	Uncertainty
[30]	6.79	0.48	-	-	30.40	0.92	-	-	×
[34]	6.20	0.77	24.16	0.53	12.83	0.95	12.73	0.98	×
[1]	5.86	**0.79**	23.14	0.59	24.74	0.92	33.95	0.95	×
Proposed	**5.49**	0.78	**6.29**	**0.93**	16.61	**0.97**	32.71	0.89	√

Table 7. Comparison of state-of-the-models with our proposed model, across the two indicators - ICS and UI and two data sources - Reddit and Google Trends. Compared to existing methods only our proposed model provides calibrated uncertainty values corresponding to its estimates.

Specific subreddits have been successfully studied for understanding opinions related to important niche topics, like mental illness, sexual violence etc [22, 24, 27, 31, 47]. Since ICS is ultimately intended to understand human behavior, we explored Reddit as a publicly available data source that is captures public optimism or pessimism in jobs and economic affairs.

Both Reddit and GT data capture similar but also disparate information about individual's perception towards consumer sentiment. Compared to Reddit data, GT data are consolidated from the web search queries. For example, an individual is likely to announce a new job or economy related sentiment on Reddit, but would do web search for specific purchase intentions. While Twitter/Reddit data are public, raw individual Google searches are not publicly available and only aggregated normalized frequencies are available through the GT tool, thereby making certain case-studies or uses of the data difficult or impossible.

7.5 Complexity analysis of Group AGP model

Being non-parametric in nature, the Group AGP model incurs a $O(T^3)$ computational cost to obtain a single prediction, where T is the number of the training instances. The primary driver for this cost is the inversion of the $(T \times T)$ kernel matrix in (2) and (3). This is clearly more expensive than other linear models explored in this paper (typically $O(1)$), however this is the trade-off for better performance.

Since the hierarchical decomposition in Group AGP does not change the number of training instances T used to obtain a single prediction, so a major cost of complexity is still dominated by $O(T^3)$. The model does introduce a number of hyper-parameters (length-scales and variances), but these scale linearly with number of components, c. Since $c << T$, the hierarchical decomposition of Group AGP does not significantly increase its complexity, which is still in the order of $O(T^3)$. Additionally, our model's complexity does not increase with time, as we use a sliding window with fixed-length T. The estimation of the model hyper-parameters is an iterative optimization procedure, in which each step has complexity ($O(T^3)$). However, the number of steps needed for convergence is data-dependent.

7.6 Limitations

Despite their advantages, sensing with SM data has some pitfalls. Some of the most important limitations include the lack of representativeness of SM users compared to the general population and the lack of demographic information present in social media that could help rectify the bias. When researchers use social media data they either use a data dump (e.g. Reddit), or may be provided a sample of the data (e.g. Twitter). In both cases, studies based on these datasets suffer

from sampling bias. GT data does not reveal demographic information about individuals who generated the data, thus hindering the use of this data for society wide studies. While Reddit and GT data used, in this study, are anonymized, but all care should be exercised to preserve the privacy of users if disparate user-generated SM data are employed [3]. Keeping the limitations in mind, we suggest that Reddit and GT data can be used to supplement traditional survey responses to consumer confidence and labor flows, especially for near future; but may not completely replace these surveys.

8 CONCLUSIONS AND FUTURE WORK

We have proposed a novel computational framework called Group AGPs which hierarchically decomposes the underlying function to be learnt into additive groups corresponding to different clusters existing in SM data. Further, we model each cluster as interaction of content and sentiment effects. Our modeling framework is based on GP regression, which utilizes prior information and produce calibrated uncertainty bounds, which is a measure of trust in the estimates. We empirically show how such a decomposition helps to embed interpretability into the modeling phase as well as to produce accurate estimates of targets along with uncertainty bounds. We show how our model can be used to generate highly accurate estimates of two important economic indicators, the consumer confidence index and the unemployment claims using Reddit and Google Trends data. We also produce sound insights for practitioners into which cluster content or sentiment is deemed important for prediction.

Our model uncovers three broad clusters, related to jobs search; economy/recession and lending/finance. Our model deems jobs and economy as important topics from 2011 - 2014,; while from 2015 onwards lending and finance topics gained prominence in gauging consumer confidence index. We also show how our model highlights that content of the posts in Reddit data is more predictive of consumer confidence than sentiment alone. Though the results are shown for two macroeconomic indicators, but our methodology is generic and can be applied to broader class of survey indicators. We are also working to incorporate data from other SM platforms like Twitter into our model and to expand this work to broader categories of indicators in order to better understand the conditions under which information extracted from SM can substitute or compliment existing methods that generate these statistics.

ACKNOWLEDGMENTS

We gratefully acknowledge the financial support of the cybersecurity cluster programmatic fund from the Institute for Security, Technology and Society (ISTS), Dartmouth College, NH.

REFERENCES

[1] Dolan Antenucci, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2014. *Using Social Media to Measure Labor Market Flows.* NBER Working Papers. National Bureau of Economic Research, Inc.

[2] Duilio Balsamo, Paolo Bajardi, and Andre Panisson. 2019. Firsthand Opiates Abuse on Social Media: Monitoring Geospatial Patterns of Interest Through a Digital Cohort. In *The World Wide Web Conference*. 2572–2579.

[3] Ghazaleh Beigi and Huan Liu. 2020. A survey on privacy in social media: identification, mitigation, and applications. *ACM Transactions on Data Science* 1, 1 (2020), 1–38.

[4] Emma I. Brett, Elise M. Stevens, Theodore L. Wagener, Eleanor L.S. Leavens, Taylor L. Morgan, Whitney D. Cotton, and Emily T. Hébert. 2019. A content analysis of JUUL discussions on social media: Using Reddit to understand patterns and perceptions of JUUL use. *Drug and Alcohol Dependence* 194 (2019), 358 – 362.

[5] J Michael Brick and Douglas Williams. 2013. Explaining rising nonresponse rates in cross-sectional surveys. *The Annals of the American academy of political and social science* 645, 1 (2013), 36–59.

[6] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, and Raghuveer M Rao. 2017. Interpretability of deep learning models:

A survey of results. In *IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation*. 1–6.

[7] Hyunyoung Choi and Hal Varian. 2009. Predicting Initial Claims for Unemployment Benefits. Available at SSRN: https://ssrn.com/abstract=1659307.

[8] Hyunyoung Choi and Hal Varian. 2012. Predicting the Present with Google Trends. *The Economic Record* 88, s1 (2012), 2–9.

[9] 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 *Proceedings of the 27th International Conference on Computational Linguistics*.

[10] Fred Conrad, Johann Gagnon-Bartsch, Robyn Ferg, Michael Schober, Josh Pasek, and Elizabeth Hou. 2019. Social Media as an Alternative to Surveys of Opinions About the Economy. *Social Science Computer Review* 39, 4 (2019), 489–508.

[11] Richard Curtin. 2007. The University of Michigan's Consumer Sentiment Index. *Encyclopedia of Survey Research Methods* (2007).

[12] Irene Daskalopoulou. 2014. Consumer Confidence Index. *Encyclopedia of Quality of Life and Well-Being Research* (2014).

[13] Ian A. Delbridge, David S. Bindel, and Andrew Gordon Wilson. 2020. Randomly Projected Additive Gaussian Processes for Regression. In *International Conference on Machine Learning*.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

[15] Xianlen Dong and Johan Bollen. 2014. Computational models of consumer confidence from large-scale online attention data: crowd-sourcing econometrics. *CoRR* (2014).

[16] David Duvenaud. 2014. *Automatic model construction with Gaussian processes*. Ph.D. Dissertation. University of Cambridge.

[17] David K Duvenaud, Hannes Nickisch, and Carl E. Rasmussen. 2011. Additive Gaussian Processes. In *Advances in Neural Information Processing Systems*. 226–234.

[18] Jacob Gardner, Chuan Guo, Kilian Weinberger, Roman Garnett, and Roger Grosse. 2017. Discovering and Exploiting Additive Structure for Bayesian Optimization. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Vol. 54. 1311–1319.

[19] Sid Ghoshal and Stephen Roberts. 2016. *Extracting Predictive Information from Heterogeneous Data Streams using Gaussian Processes*. Technical Report. arXiv.org.

[20] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2018).

[21] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[22] Nur Shazwani Kamarudin, Vineeth Rakesh, Ghazaleh Beigi, Lydia Manikouda, and Huan Liu. 2018. A study of reddit-user's response to rape. (2018), 591–592.

[23] Kirthevasan Kandasamy, Jeff Schneider, and Barnabas Poczos. 2015. High Dimensional Bayesian Optimisation and Bandits via Additive Models. In *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37. 295–304.

[24] Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports* 10, 1 (2020), 1–6.

[25] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*.

[26] Bingqing Liu, Ivan Kiskin, and Stephen Roberts. 2020. An Overview of Gaussian process Regression for Volatility Forecasting. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 681–686.

[27] Lydia Manikonda, Ghazaleh Beigi, Huan Liu, and Subbarao Kambhampati. 2018. Twitter for sparking a movement, reddit for sharing the moment:# metoo through the lens of social media. *arXiv preprint arXiv:1803.08022* (2018).

[28] Wagner Meira, Antonio L. P. Ribeiro, Derick M. Oliveira, and Antonio H. Ribeiro. 2020. Contextualized Interpretable Machine Learning for Medical Diagnosis. 63, 11 (2020), 56–58.

[29] James W. Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.

[30] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *International Conference on Web and Social Media*.

[31] Albert Park and Mike Conway. 2018. Harnessing Reddit to Understand the Written-Communication Challenges Experienced by Individuals With Mental Health Disorders: Analysis of Texts From Mental Health Communities.

Journal of Medical Internet Research 20 (2018).
[32] Josh Pasek and Jake Dailey. 2018. Why Don't Tweets Consistently Track Elections?: How Big Data Informs Political Communication. In *Digital Discussions*, Natalie Jomini Stroud and Shannon McGregor (Eds.). 68–95.
[33] Josh Pasek, H Yanna Yan, Frederick G Conrad, Frank Newport, and Stephanie Marken. 2018. The Stability of Economic Correlations over Time: Identifying Conditions under Which Survey Tracking Polls and Twitter Sentiment Yield Similar Conclusions. *Public Opinion Quarterly* (09 2018).
[34] Viktor Pekar. 2020. Purchase Intentions on Social Media as Predictors of Consumer Spending. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (2020), 545–556.
[35] Viktor Pekar and Jane Binner. 2017. Forecasting Consumer Spending from Purchase Intentions Expressed on Social Media. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.
[36] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. Technical Report. University of Texas at Austin.
[37] Boris Podobnik and H. Eugene Stanley. 2007. *Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Non-stationary Time Series*. Technical Report. arXiv.org.
[38] Neeti Pokhriyal, Abenezer Dara, Benjamin Valentino, and Soroush Vosoughi. 2020. Social media data reveals signal for public consumer perceptions. In *Proceedings of the ACM International Conference on AI in Finance*.
[39] Shalini Priya. 2018. Where should one get news updates: Twitter or Reddit. *Online Social Networks and Media* (2018).
[40] Shaan Qamar and Surya T. Tokdar. 2014. Additive Gaussian Process Regression. arXiv:1411.7009
[41] Carl Edward Rasmussen. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*. Springer, 63–71.
[42] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.
[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.
[44] Paul Rolland, Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. 2018. High-Dimensional Bayesian Optimization via Additive Models with Overlapping Groups. In *Proceedings of 21st International Conference on Artificial Intelligence and Statistics*, Vol. 84.
[45] Torsten Schmidt and Simeon Vosen. 2009. *Forecasting Private Consumption: Survey-based Indicators vs. Google Trends*. Ruhr Economic Papers. RWI - Leibniz-Institut für Wirtschaftsforschung, Ruhr-University Bochum, TU Dortmund University, University of Duisburg-Essen.
[46] Michael F. Schober, Lauren Pasek, Josh Guggenheim, Cliff Lampe, and Frederick G. Conrad. 2016. Social Media Analyses for Social Measurement. *Public Opinion Quarterly* (2016).
[47] Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An Analysis of Domestic Abuse Discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2577–2583.
[48] Benjamin K. Smith and Abel Gustafson. 2017. Using Wikipedia to Predict Election Outcomes: Online Behavior as a Predictor of Voting. *Public Opinion Quarterly* (2017).
[49] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.
[50] Gijs te Brake. 2017. *Unemployment? Google it! Analyzing the usability of Google queries in order to predict unemployment*. Master's thesis. Universitat de Barcelona.
[51] Christoph Kilian Theil, Sanja Štajner, and Heiner Stuckenschmidt. 2020. Explaining financial uncertainty through specialized word embeddings. *ACM Transactions on Data Science* 1, 1 (2020), 1–19.
[52] Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In *Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis*.