INDIANA UNIVERSITY

Time-series analysis of sentiment in social media can predict individual and collective behavior of public health significance

by

Ian Wood

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
School of Informatics, Computing, and Engineering

April 2023

INDIANA UNIVERSITY

Abstract

School of Informatics, Computing, and Engineering

Doctor of Philosophy

by Ian Wood

Lexical sentiment analysis has been used to understand what is expressed in natural language, sometimes to better understand the psychological characteristics of an author, or to understand the author's stance towards an object, such as a product, person, or idea. This methodology has also been used to study temporal patterns in the mood of populations, or to understand the general use of language. In general, these applications study changes in the central tendency of the mood of entire populations. However, it is likely that collective moods may be composed of discordant parts. Can we use these sentiment tools to predict and understand health outcomes for both small cohorts and at the level of populations, and does the analysis of the distribution of sentiment reveal distinct components useful for those goals?

In this dissertation, I demonstrate how the use of natural language processing and lexical sentiment of social media timelines can be useful in predicting health outcomes for a small cohort of epilepsy patients. I develop a method based on the singular vector decomposition to discern characteristic components of the distributions of collective sentiment, associated with sub-populations and cohorts of interest. I demonstrate that the first singular component represents the base distribution of sentiment due to the frequency of sentimental words in natural language and show how further components can reveal meaningful patterns in sentiment over time. To show the predictive and analytical power of these components, I demonstrate their use in modeling sex searches as a proxy for human reproductive cycles and mortality during the Covid-19 pandemic.

Acknowledgements

[ACKNOWLEDGEMENTS TO BE WRITTEN]...

Contents

List of Figures

List of Tables

For/Dedicated to/To my...

Chapter 1

Introduction and Related Work

1.1 Motivation

Lexical sentiment analysis has been used to understand what is expressed in natural language – sometimes to better describe the psychological characteristics of an author [1], or to measure the author's stance towards an object, such as a product, person, or idea [2]. This methodology has also been used to study temporal patterns in the mood of populations, or to understand the general use of language [3, 4]. In general, the application of this methodology involves measuring changes in the central tendency of the mood of entire populations. However, it is likely that collective moods may be composed of discordant parts. For instance, an event such as an election or a competition may lead to the simultaneous appearance of happy and sad sub-populations, which central tendency measures of the whole population may miss. Furthermore, changes in the components of mood in time may allow us to identify cohorts and individuals with specific diseases or conditions. To study and address those situations, I propose to develop methods that can discern characteristic components of collective mood, potentially associated with

sub-populations and cohorts of interest and test them on problems of public health and social interest. I will treat measured sentiment in social media posts as a composition of sentiments over time, and define characteristic mood components and temporal patterns as the right and left singular vectors of this data matrix respectively. To test the hypothesis that central tendency in lexical sentiment analysis may miss important aspects of collective human behavior, and that methods that discriminate different contributions of collective mood can be more effective in predicting that behavior, I propose to answer the following questions (numbers) and hypotheses (letters):

1. Can meaningful *components* of collective mood states be extracted from time-series analysis of the sentiment of entire populations measured on social media?

   1.Ha A singular value decomposition of lexical sentiment measurements over time provides useful components representing collective mood states

   1.Hb These components are characteristic of sub-populations and phenomena of interest.

2. Are components of collective mood predictive of the future collective behavior of populations?

   2.Ha Components of collective moods can predict and explain the influence of culture on human reproductive cycles.

   2.Hb Components of collective mood can track mortality due to Covid-19.

3. Can characteristic temporal patterns associated with individuals or small cohorts be used to predict specific mental or medical conditions?

   3.Ha Patterns in sentiment measures are predictive of increased risk of sudden unexpected death in epilepsy.

1.2 Related Work

1.2.1 Sentiment Analysis

Sentiment Analysis, also called *Opinion Mining*, has been an area of increasing interest in the past two decades. People have long been interested in what others think. Among others, consumers want to know the opinions of trusted reviewers before making a purchase, politicians want to know the views of their constituents, and companies want to know how their products are seen by the market. With data increasingly available in digital form from public blogs, reviews, and social media; traditional opinion polling can be complemented by automatic information extraction and retrieval tools. Due to these interests, much of the computer science literature casts problems in this domain as classification or regression problems, determining whether a text holds a positive or negative opinion. As Pang and Lee note in their review of the subject, extracting subjective opinion offers a host of new challenges that objective information extraction may not face, including the use of sarcasm (particularly in political discussions), the determination of subjective opinion, and the polarity of a sentiment-laden word or phrase in context[5]. Liu in his book [2] casts the problem generally as one of extracting quintuples representing a target (e.g. a camera), an aspect of the target (e.g. the battery), the sentiment directed towards the target (e.g. positive or negative), the holder of the sentiment, and the time when the sentiment was expressed. Approaches to solving these classification problems involve selecting features based on likely expressions of sentiment; transforming those features based on syntactic rules; multi-stage classifications of topic, subjectivity, and sentiment; and natural language processing (NLP) techniques to identity product features and the authors of the opinion[2, 5].

Bing Liu defines sentiment analysis as "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes"[2], and Pang an Lee suggest that many use the term "broadly to mean the computational treatment of opinion, sentiment, and subjectivity in text." Work focusing on more linguistic or affective concerns thus also fall under this umbrella. Dictionaries of sentiment-laden words are commonly built through a variety of methods – such as expert curation and consensus, surveys, and supervised machine learning classifiers – for use in feature construction and selection in classification tasks [2, 5]. However, these tools and how they describe the affect of words can themselves be the subject of interest or interpreted as revealing underlying moods[3, 4, 6–9]. Implicit in this approach is an understanding of individual or population-level emotion as a latent state, observable through written text, similar to psychological mood states[10], or hidden states in a generative model for text [11, 12].

1.2.2 Available Sentiment Instruments

Most out-of-the-box sentiment tools are dictionary based: words and a numeric score describing their sentiment, and most focus on a single dimension of measured affect, ranging from negative to positive valence (happiness). A non-comprehensive list includes the General Inquirer[13], the Affective Norms for English Words (ANEW)[4, 6] along with many extensions and translations of the ANEW dictionary[14–16], Google Profile of Mood States (GPOMS)[9, 17], Peter Dodds' LabMT dictionary[8], SentiWordNet[18, 19], the Lingustic Inquiry and Word Count (LIWC)[1], VADER[20], and OpinionFinder[21]. There are many more, twenty-four of which are listed in Reagan et al.'s paper comparing six in depth[22], while a different, but overlapping set of twenty-four were compared on a

number of classification tasks in Ribeiro et al.'s Sentibench[23] . Due to this proliferation of tools I will only discuss those mentioned above, which cover historically important tools, like the General Inquirer and ANEW, tools developed for large-scale social media analysis like LabMT, as well as widely used tools like LIWC and VADER that are among the best across a number of the Sentibench three-class (negative, neutral, and positive) classification tasks[23].

The *General Inquirer*[13] was developed as a tool to organize nonnumerical data, and tag words in a text across various categories, and allow the text to be organized according to such tags. The system started as a general-purpose tool with a dictionary of categories over the 3000 most common English words and a few hundred words of interest to a behavioral scientist. The categories included "Persons", "Behavorial Processes", "Psychological States", and more for the purpose of content analysis to trace psychological themes over a series of group discussions It has since grown to include the "Harvard IV-4" and "Lasswell" content analysis dictionaries as well, for a total of 198 categories[24].

The *Affective Norms for English Words* (ANEW) includes ratings from 1 to 9 for 1034 words along three mood dimensions: *valence* from unhappy to happy, *arousal* from calm to excited, and *dominance* from controlled to in-control. These ratings were collected from surveys given to undergraduates in a psychology class using a 9-point Likert-like scale[6]. It has been used as a basis for a number of new dictionaries, including an extension to nearly 14,000 words[16], a translation to Spanish[14], European Portuguese[15], among others.

The *Google Profile of Mood States* (GPOMS) is an extension of the Profile of Mood

States (POMS), a test of self-report Likert-scale questions measuring 6 underlying dimensions of mood: Tension or Anxiety, Depression or Dejection, Anger or Hostility, Vigor or Activity, Fatigue or Inertia, and Confusion or Bewilderment[10]. GPOMS tries to translate the questionaire to a dictionary suitable for sentiment analysis of large-scale social media data. This tool extended the original 72 terms in the POMS questionnaire to a dictionary of 964 words by looking at co-occurrences in Google's 4, and 5-gram corpora. These terms correspond to moods across 6 categories: *calm, alert, sure, vital, kind*, and *happy* [17].

LabMT Hedonometer used Amazon's Mechanical Turk to send out ANEW-like surveys ranking 1,000s of words on a 9-point scale from *sad to happy*, collecting at least 50 ratings for each word. Initially, LabMT Hedonometer was comprised of 10,222 English words found by merging the 5,000 most used words in each of four corpora: Google Books, Twitter, music lyrics, and the New York Times[8]. This has since been extended to include 10 languages with about 10,000 words each collected across 24 corpora [7].

Linguistic Inquiry and Word Count, LIWC (pronounced "Luke"), is a software tool for text analysis whose first version was released publicly in 2001 and has been actively supported and widely used since[1, 25]. LIWC was developed by a number of judges who independently created lists of words, tested for consistent categorization between a majority of judges, uncommon words not present in a variety of corpora (blogs, novels, spoken language studies, etc.) were removed, internal consistency was evaluated with a corrected Cronbach's alpha calculation, and external validity was tested through psychological studies, including writing prompts for students. The version of the software used in this work, LIWC2015, has dictionaries containing nearly 6,400 words and produces outputs across about 90 categories, including positive and negative emotion, but

also pronouns, articles, congitive processes, time focus, personal concerns, and informal language among others [1].

SentiWordNet[18, 19] is a dictionary assigning words values from 0 to 1 along three dimensions: Objective, Positive, and Negative, such that all values sum to one for each word. SentiWordNet was built on *synsets*, groups of synonymous words, from *WordNet*[26] and the lexical relationships between them. A committee of ternary classifiers were trained in a semi-supervised fashion. Starting from a small set of positive or negative labeled seeds, labels were propagated to related synsets within various radii, and various supervised classifiers were trained on these sets. The final values for each word/synset are determined by the proportion of classifiers labeling the synset as objective, positive, or negative, with random walk dynamics further refining values[19].

VADER[20] is a tool for measuring the extent of positive or negative sentiment with more than a dictionary, and is readily available as part of the natural language toolkit for python. In addition to dictionary-based sentiment scores, VADER looks at other words in a sentence modifies sentiment scores based on 5 simple rules, namely the presence of exclamations (e.g. "!!!"), capitalization, adverbs (e.g. "very"), negations within the last three words before a sentiment-laden lexical feature (e.g. "not"), and contrastive conjunctions (e.g. "but").

OpinionFinder[27] is a full processing pipeline, first tokenizing a document, and then using a series of classifiers trained on various corpora, to find subjective statements, find speech events, identify opinion source, identify expressions of sentiment, and finally to identify the expression as positive or negative.

Other work suggests modifications of sentiment scores through context, such as

Karo Moilanen and Stephen Pulman's proposed compositional rules to modify sentiment scores from sentence parse trees[28]. More extensions to dictionary-based sentiment analysis involve techniques to build features for traditional machine learning classifiers (Naive Bayes, SVM, etc.) on top of or in lieu of lexical scores[20]. Such features include word modification, grammatical position, sentence-level, and document-level features[27]; adding semantic features identifying the type of entity discussed (person, place, etc.) [29]; using features from a hidden markov model latent dirichlet allocation analysis[30]; or comparing the parse trees of text from different classes through boosting methods[31]. These methods will not be explored in detail here, for most are not available out-of-the-box, and must be trained for specific tasks.

1.2.3 The Use of Sentiment Tools

A common application of sentiment analysis tools is towards automatic classification of reviews, in part because the data is often well structured with clear labels, e.g. number of stars, thumbs up, ratings, etc. One early application of this sort of analysis was Pang, Lee, and Vaithyanathan's classifications of movie reviews, finding that straightforward machine learning models like Naive Bayes, Support Vector Machines, and Maximum Entropy models could outperform simple dictionaries invented by human participants at determining positive or negative reviews.[32]. It is commonly noted, however, that sentiment classifiers trained in one domain will not necessarily do well when transferred to another. A more thorough review of papers concerned with sentiment analysis towards reviews can be found in Pang and Lee's review paper and Liu's book [2, 5].

The authors of LIWC cite psychological studies to demonstrate how various LIWC categories reveal underlying psychological states, including the use of more first-person singular pronouns when describing pain or trauma, the use of verb tense describing the

immediacy of an experience, the use of first-person plural pronouns to denote higher social status, and the use of prepositions and conjunctions as a proxy to the complexity of thought, among other examples[25]. A number of studies have used this tool as part of automated text analysis for different applications. Robinson, Navea, and Ickes found that some variables from LIWC calculated from students' self-introductions, followed by a Principal Component Analysis, are good predictors of a student's overall performance in a class — for example, the use of commas, quotes, and negative affect were positively correlated with final performance, while use of the present tense, first-person singular, home, and eating and drinking categories were negatively correlated[33]. LIWC has also been found to be accurate for the automatic classification of the positive vs negative affect of dream reports[34], used as descriptive statistics and as features for the classification of suicide notes in completed vs non-completed attempts[35], and used to track feelings of sadness, anxiety, and anger on social media during the September 11th World Trade Center attacks[36, 37].

When applied to large-scale social media data, sentiment dictionaries reveal patterns in population-level moods. Golder and Macy, applying LIWC to Twitter data, find diurnal and seasonal rhythms to positive and negative sentiment, with increased positive to negative sentiment in the morning decreasing through the day, and increasing with longer day-lengths [3]. Using GPOMS and Twitter data, Bollen, Mao, and Zeng were able to predict the overall direction of the Dow Jones Industrial Average [17]. Dodds and Danforth, using song lyrics and blogs, found that there is a strong correspondence between average ANEW valence and genre, blogger age, blogger location, and the day of the week (peaking on the weekend and bottoming on Wednesday)[4]. In Dodds et al.'s later work using the larger LabMT Hedonometer dictionary, many of these results are reproduced with finer detailand shown to be robust to tuning parameters to remove

more neutral terms[8]. Once extended to ten languages, LabMT Hedonometer shows a universal positive skew in the frequency distribution of human language [7]. Reagan et al. compare a number of tools to each other and their performance on a number of corpora, including LabMT Hedonometer, ANEW, and LIWC, finding that in general sentiment dictionaries tend to agree on positive and negative terms, but that ANEW, being a small dictionary, lacks coverage of much of the text, while LIWC and similar dictionaries offer less nuanced interpretation since the strength of terms in these dictionaries is limited to a binary choice between positive and negative (or neutral/uncategorized) rather than a metric or ordinal scale[22].

1.2.4 The Use of Social Media to Study Public Health

More broadly, social media and the use of online tools has been used to study a large variety of topics, from social protests [38], to the spread of information [39], to a variety of public health applications [40–42]. Studies have investigated how social media has been used to spread information regarding ebola [43], track flu epidemics [44, 45], or track dengue epidemics through search volume [46]. It should be noted, however, that due to changes in underlying systems, data over-fitting, or spurious correlations with common factors, big-data models that initially predict some effect well, such as Google Flu Trends's prediction of CDC flu reports [47], may not predict such phenomena accurately beyond training [48].

Sentiment analysis of social media in various forms has also been used to good effect for public health applications. Qualitative content analysis finds that most doctor reviews are positive [49] and that more positive reviews are associated with surgeons with high procedure volume [50]. Similar qualitative content analysis found mostly positive views of marijuana in related tweets [51] with increasing volume when marijuana was

legalized in two states [52]. Machine learning methods for sentiment classification were used to find that negative sentiment towards vaccines spreads more easily than positive sentiment in social networks [53] and users with like-sentiment tend to cluster in social networks [54]. In addition, machine learning methods have been able to accurately classify social media posts according to the mental conditions to which they relate[55], while ANEW and LIWC features have been useful in particular for classifying tweets related to depression [56]. However, many of these studies focus on qualitative content analysis, or building custom machine learning classifiers, while few have investigated the ability of sentiment measures to capture general trends for small cohorts. There is evidence that depression is related to phase shifts in mood, according to a study using self-reports [57]. If these moods can be accurately measured by sentiment tools on social media posts, we can better understand how mental illness relates to underlying dynamics in mood, and potentially direct those most at risk to appropriate help and resources.

1.2.5 Topic Modeling and Vector Space methods

Related research in Natural Language Processing investigates how to capture semantic topics from natural text, such as well known models like the Latent Dirichlet Allocation [58–60]. Some of the above methods, like LIWC, cover not only the affective categories of positive and negative emotion, but also semantic and cognitive categories like family, or work. A topic can be discussed in different affective terms, and choice of affective expression can be driven by a topic. To get to the try to access the underlying semantics of word use, various methods can be used to give words a vector representation based on the context in which the word appears.

Latent Semantic Indexing and Probabilistic Latent Semantic Indexing [61, 62] are methods to find underlying factors in natural text. More generally, Principal Component

Analysis (PCA), and Singular Value Decomposition (SVD) are useful methods both to explore data, parsimoniously decomposing the variance into a few components, and to give data the required properties (e.g. linear independence, whitening) for further analysis [63–66]. Related methods, like common factor models, similarly find parsimonious descriptions of covariance structures with explicit reference to hypothesized latent variables [64, 65]. More matrix factorization methods have recently been proposed to apply different constraints to data, like non-negativity [67–69], while Matrix Factorization Machines have provided a unifying framework to understanding these factorizations applied to a variety of tasks [70, 71]. Techniques like word2vec learn vector representations of words from context, and successfully capture semantic information, such that vector operations can operate like analogies [72–74]. Vector representations allow for geometric descriptions of topics as nearby clusters of words in a vector space [75]. Additionally tensor or multi-table factorization methods like multilinear singular value decomposition [76], the canonical polyadic decomposition [77], and multiple factor analysis [78, 79] have been used to extend matrix factorization to data with three or more sets of related variables.

PCA has also been applied to "compositional data", data whose variates are proportions of a whole, including the composition of blood, chemical, or geological compounds [66]. This sort of data has an imposed constraint – the sum of all variates for each sample must equal a constant, namely one. Aitchison argues that this constraint leads to two difficulties when applying PCA: first that actual data in the simplex is often curved in ways that linear principal directions do not capture, and second that he constraint biases the covariance structure towards negative values, as increases in one proportion must reduce at least one other. Aitchison proposes a transformation of the data to the log ratio of proportions to the harmonic mean of all proportions, which removes the

constraint on the covariance structure, and allows for principal directions curved in the original space [80]. Filmoser, Hron, and Reimann extend this method to handle robust estimation of the covariance matrix with outliers [81]. Others, like Kaciak and Sheahan, however, have had success without transformations, or even centering [82].

1.2.6 Key Methodology in Support of Research Questions

While the opinions and sentiment of individuals has received a great deal of attention, less has been paid to the mood of groups. It is unclear if the mood states of the individuals comprising a group are a good description of the overall mood of a group. In particular, we may conjecture that groups have a greater propensity to hold divergent or conflicting opinions towards the same object than individuals would, since an individual may find it inconsistent to hold such conflicting opinions, while two separate individuals holding opposing opinions encounter no such inconsistency. Additionally, there may be patterns in the moods of a group of people that are not apparent from any individual. Previous work looking at the large-scale moods of populations typically focus on mean sentiment, assuming that the mood of individuals in the population follows the mood of the group [3, 8]. In this work, I will try to characterize these collective moods by measuring the sentiment of groups of people posting on social media, treating these distributions as compositions of sentiment, and factorizing the composition of sentiment over time. The goal will be to find meaningful collective mood states and temporal patterns that correspond to other large-scale patterns of collective behavior. The consequences for individuals will also be examined by looking at the correspondence of individuals in smaller cohorts to these characteristic patterns.

With the variety of both sentiment tools, and matrix factorization methods, it is impractical to systematically study all of them. In the proposed dissertation, I will use the

following sentiment tools: ANEW, VADER, LIWC and LabMT Hedonometer. These tools are widely used in sentiment analysis literature, and have well-established foundations through psychological surveys, qualitative analysis followed by machine learning evaluation, expert curation, and distributed surveys respectively. Additionally, VADER and LIWC, were consistently among the best tools for 3-class polarity classification (negative, neutral, or positive emotion) across a number of corpora in a benchmark comparison study, which is useful since an increase in the rate of neutral valence mood in a group could still be an important state [23]. To find characteristic patterns in multimodal time series, I will focus on the singular value decomposition which is one of the most established matrix factorizations, and requires tuning few parameters.

1.3 Thesis Overview

The following subsections describe how I will approach the research questions and hypothesis outlined in 1.1.

1.3.1 Eigenmood Analysis

This chapter will present the *Eigenmood* methodology developed to uncover characteristic components of collective moods, particularly in multi-modal distributions of sentiment. Many patterns in collective mood can be overlooked by a focus on the central tendency of sentiment measured for some population. Populations may experience different moods simultaneously. Especially when populations are comprised of different groups, its collective mood may become complex, composed of a k-modal distribution of sub-population moods, and no longer easily described by the central tendency of sentiment. The *Eigenmood* methodology aims to provide a decomposition of the distribution

of mood values of large populations on social media—in this thesis, tested on Twitter. The goal is to understand the various, time-changing components that make up a collective mood, which we think of as *eigenmoods*.

This chapter will explore the use of a singular value decomposition to find component moods. Its use will be illustrated through artificial examples, including Gaussian mixture models with known mixture dynamics, as well as mixtures with a dominating distribution representing usual language use. This chapter will also apply *Eigenmood* to real social media data, and demonstrate that the first component captures the overall frequency of sentiment in language through comparison with external corpora. The dataset used includes a 10% sample of tweets from the US between September 2010 to February 2014, and between August 2016 and February 2018 from the Truthy and OSoMe projects [83, 84].

This chapter will address hypotheses 1.Ha, 1.Hb, and 2.Hb. Question 1.Ha will be explored through the use of artificial examples, showing what the method can recover. Our hypotheses are that SVD components will be able to recover underlying dynamics closely, and remove the influence of a dominating baseline distribution better than the subtraction of a mean distribution. Hypothesis 2.Hb will be explored through the predictive ability of sentiment measures and components against weekly mortality reports during the Covid-19 pandemic [85]. We find that VADER sentiment does improve the predictive power of ARIMA models to track changes in mortality patterns due to the pandemic in most of the 20 cities with the most per-capita cases in the first months of the pandemic. We show that k-model eigenmood We find that mean VADER sentiment does improves the predictive power of ARIMA models to track changes to mortality patterns due to the pandemic, while eigenmood components provide the best fit to in-sample data.

1.3.2 Human Sexual Cycles are Driven by Culture and Match Collective Eigenmoods

This chapter will apply *Eigenmood* methodology to study the collective sentiment behind yearly cycles of human reproductive interest. Human reproduction does not happen uniformly throughout the year and what drives human sexual cycles is a long-standing question. The literature is mixed with respect to whether biological or cultural factors best explain these cycles [86]. The biological hypothesis proposes that human reproductive cycles are an adaptation to the seasonal (hemisphere-dependent) cycles, while the cultural hypothesis proposes that conception dates vary mostly due to cultural factors, such as holidays. However, for many countries, common records used to investigate these hypotheses are incomplete or unavailable, biasing existing analysis towards Northern Hemisphere Christian countries. Here we show that interest in sex peaks sharply online during major cultural and religious celebrations, regardless of hemisphere location. This online interest, when shifted by nine months, corresponds to documented human births, even after adjusting for numerous factors such as language and amount of free time due to holidays. We further show that mood, measured independently on Twitter, contains distinct collective emotions associated with those cultural celebrations. Our results provide converging evidence that the cyclic sexual and reproductive behavior of human populations is mostly driven by culture and that this interest in sex is associated with specific emotions, characteristic of major cultural and religious celebrations, contradicting the biological hypothesis. This work is especially noteworthy because it demonstrates that data science—and *Eigenmood* in particular—can be used to test existing hypotheses in sociobiology and provide new theories to better anticipate human behavior of great public health interest.

In particular, to measure holiday mood, we applied ANEW to a 10% sample of Tweets from 2010 to early 2014. We found holiday moods by applying a singular value decomposition of the collective mood distributions over time, and selecting *eigenmoods* composed of two components corresponding to particular holidays. We found major holiday eigenmoods correspond to reproductive interest throughout the year, in multiple countries around the world, lending additional evidence towards a cultural explanation of human reproductive cycles. Additionally, we perform a Granger causality analysis and find that some mood components significantly Granger-cause sex searches, while others are significantly Granger-caused by sex searches.

This chapter will directly address hypothesis 2.Ha, that components of the mood that correspond with holidays associated with sex searches will be associated with sex searches throughout the rest of the year. Hypothesis 1.Hb will be explored through the emphasis of particular components during holidays and cultural events, particularly the cultural holidays associated with sex searches. Our hypotheses are that various components will be more closely associated with particular annual events than others, allowing a characterization of the annual events.

1.3.3 Using Sentiment in Small Cohorts

This chapter will address hypotheses 3.Ha through temporal analysis of sentiment on social media from small patient cohorts to investigate the feasibility of uncovering early warnings for the onset of sudden unexpected death in epilepsy (SUDEP), a frightening outcome for those with epilepsy that is poorly understood [87]. I explore whether there are characteristic patterns detected by sentiment measures in the time period before SUDEP, using a SUDEP patient cohort on Facebook—data obtained with consent of the families of the deceased patients via the Epilepsy Foundation of America. Specifically,

we have collected the facebook posts of a small set of 12 subjects who experienced SUDEP. First, we will use sentiment tools, particularly ANEW, VADER, and LIWC, to characterize changes in sentiment immediately before SUDEP. If SUDEP is driven by periods of life-change and stress, as anecdotal accounts suggest, and sentiment measures like LIWC can detect such experiences in social media posts, we should find sentiment measures with time series that correlate between subjects in time windows shortly before SUDEP. We find for subjects who post frequently on social media in the months before SUDEP their posts demonstrate an increase in verbosity, functional words, and certain sentiment measures like neutral Vader terms. We then examine the posts corresponding to patterns seen before SUDEP qualitatively, to understand the sorts of life events and topics discussed.

As a caveat, from such little data, we don't draw definitive results in terms of predictions. Additionally, the relationship between SUDEP and other seizures is unknown. It is possible that periods of our data that appear similar in sentiment pattern to SUDEP correspond to other seizures we don't know about. However, our results will aid our research group in recruiting more participants to explore behavior on social media around both SUDEP and seizures.

1.4 Conclusion

By finding the components of population sentiment, we can better understand both individual and collective moods, and make predictions that could inform public health professionals, as well as patients with particular conditions. Chapters 2 proposes and demonstrates the k-modal methodology to extract eigenmood components of collective sentiment. Chapter 3 demonstrates the utility of the approach by contradicting the

prevalent hypothesis of a long-standing sociobiology phenomenon in human reproduction, while providing compelling evidence for an alternative, cultural hypothesis. Chapter 4 uncovers characteristic patterns in the sentiment of small patient cohorts with a specific disease to make predictions about outcomes. This thesis demonstrates that there are meaningful patterns to collective mood that are distinct from individual mood states. These patterns will be shown to be useful in public health contexts, but will likely be of interest for a variety of problems, possibly including tracking political sentiment, economic activity, and other social concerns.

Chapter 2

Eigenmood Twitter analysis: Measuring collective mood variation

Much of this Chapter comes from a paper In Preparation [88] and includes writing from co-authors Marijn ten Thij, Luis M. Rocha, and Johan Bollen

2.1 Abstract

Sentiment analysis of large scale social media data provides a window into the collective emotions of millions of individuals, allowing cognitive and behavioral modeling at previously unseen scales. However, sentiment analysis of online content may confound prevailing term frequencies in a language with collective emotions. We demonstrate how to separate collective emotion signals from online sentiment with a singular value decomposition of diachronic sentiment matrices. We refer to such decompositions as

'eigenmoods' since their multi-dimensional features represent various components of collective emotions beyond the baseline sentiment of word use in a language. This approach can identify diverging 'eigenmoods' as separable components of collective affect in large groups of individuals and the variegated dynamics of individual mood states. Our results point towards the possibility of extracting collective and individual mood states from online text, disentangling prevailing term frequencies and ephemeral topical language changes from the underlying collective and individual emotions.

2.2 Introduction

Mood is an important driver of behavior, cognition, and language. The relationship between language and mood creates opportunities to measure mood from language. The analysis of natural language, in particular written text, for indications of mood or sentiment has therefore become a common, but non-trivial natural language processing (NLP) task called "sentiment analysis" (SA). A plethora of SA techniques to perform sentiment rating or classification of written text has emerged in the past decades, ranging from unsupervised lexicon matching [1, 3, 8, 16] techniques, often coupled with grammatical analytics to handle negation and hedging[20], to sophisticated supervized machine learning techniques [89].

Lexicon matching in particular remains a very common approach to gauge text sentiment: it is highly scalable, robust and well-validated, and straightforward to adapt to specific application areas through the design of dedicated lexicons. Generally, a group of test subjects assigns each word in a lexicon a score or rating for a number of emotional dimensions, for example, Valence, Arousal and Dominance [16], Happiness [8], or a variety of other psycho-social indicators [1]. By combining the lexicon values of the

words in a document we can determine its overall sentiment rating, e.g. the average of the sentiment loadings of its constituent words.

This technique can be applied to individual texts, the collection of texts generated by an individual as an assessment of their long-running emotional state[90, 91], for the texts generated by a cohort of individuals [92], and even entire societies from online diachronic collections of text, e.g. daily tweets posted in the US[8, 93]. SA approaches have in particular found applications in the analysis of social media data which offer an increasingly detailed large-scale and fine-grained record of the behavior of a considerable fraction (about one-seventh) of the world's population [94].

However, an interesting problem emerges when we consider the documented tendency of languages to exhibit very skewed word frequency distributions, i.e. the frequency of words is generally inversely proportional to its rank in the frequency table [95]. In fact, the first 100 most frequent words in the English language comprise more than 50% of all written English. Given the stability of these distributions over corpora and time, the frequency distribution of words in a language will result in sentiment ratings that are predominantly shaped by prevailing term frequencies in the language, instead of actual changes in individual or collective mood states. This problem also present itself when we consider specific socio-cultural events, such as New Years Eve. It is common to express specific greetings at such times, e.g. "Happy New Year", hence the positive sentiment of constituent words ('happy", "new", etc) of such greetings will confound language sentiment ratings with temporo-spatial lexical artifacts.

Here, we propose to use singular value decomposition (SVD) to separate the effects of prevailing term frequencies in a language from actual mood states. This approach enables us to separate mood changes from the dominant and uninformative term frequency

distributions in a language. We further show that such an analysis can be performed to texts that are in the regions for which we see the largest differences in the observed mood dimension. This way, we can quantitatively find texts that are the best examples of the change in the mood of the users in the data set. This approach can even be performed on the collection of an individuals' tweets over time to investigate whether or not there longitudinal changes have occurred in the individuals' mood profile[91].

We will demonstrate that sentiment analysis of social media is obscured by the perceived mood of regular language. Furthermore, we will demonstrate how this influence of regular language can be removed to reveal the precise components of mood that are associated with the phenomena of interest.

These eigenmoods are sets of components (singular vectors) that explain a significant proportion of the time-series variation of mood associated with a phenomenon of interest. For example, during the COVID-19 pandemic many tweets may contain the word "virus, which has a low valance rating, but they could be optimistic messages of hope or support. The goal of eigenmoods is to separate the former (decontextualized language) from the latter (signals of affect). As such, these eigenmoods reveal the otherwise hidden components of "underlying" mood signals, allowing for more fine-grained assessments of individual- and population-level emotions associated with health behaviors of interest.

Based on this evidence, the underlying hypothesis of our methodology is that removing the first singular vector is equivalent to removing the base mood contribution attributable to regular language use. Conversely, not removing this first singular vector obscures the study of phenomena regarding mood on social media, as mood variations central tendency merely reflects prevailing language use. Removing the first singular vector better reveals variation in the phenomenon of interest.

This opens op many opportunities to further investigate the effects of specific events on individual mood state, an approach typified in [91]. Furthermore, the development of analytical methods that can extract informative components of mood and high-quality sentiment signals from collections of tweets pertaining to the biomedical, and cognitive discourse of large populations may support the evaluation of global well-being and health.

2.3 Language overdetermines sentiment observations

First, we examine how prevailing term frequencies in a language affect sentiment analysis by applying the same sentiment analysis technique, i.e. Hedonometer[8], to three different corpora of Twitter content. Each corpus consist of large-scale collections of individual Twitter timelines, i.e. all tweets posted by a cohort of individuals over time, for groups of individuals that (1) were located in New York City during the COVID-19 pandemic $N = 14,130,720$ [96], (2) who reported a clinical diagnosis of depression $N = 593,993$, and (3) a random cohort of individuals $N = 1,982,311$ [92]. This allows the sentiment analysis between cohorts and within-subjects, illustrating the issues with applying sentiment analysis to large corpora of social media content to measure diachronic changes in public or individual sentiment.

Each individual tweet is subjected to a sentiment rating on the basis of the Hedonometer lexicon of $N =$ English words that were rated by N human subjects (Amazon Mechanical Turk). Each tweet's sentiment rating is calculated as the average Hedonometer ratings of its constituent words. Chapter 2.1 shows the resulting distribution of Tweet sentiment ratings. We find that the distribution of Hedonometer scores per tweet differ little across the three corpora, even though the equality of the distributions can be rejected on the basis of a pair-wise KS two sample test with Bonferroni correction [97],

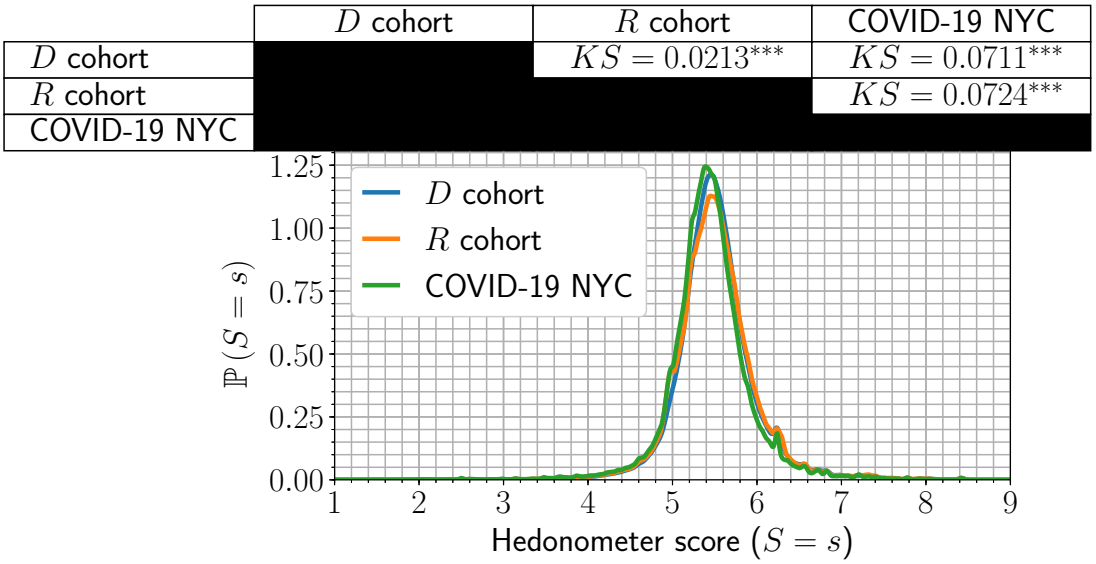


FIGURE 2.1: **Kernel Density Estimate (KDE) of the per document Hedonometer score for all considered data sets.** The table above the panel shows the outcomes of two sample KS-tests between data sets. To account for multiple comparisons, we use a Bonferroni correction.

see Chapter 2.1. This is not surprising, given the large amount of observations (tweets) in each corpus, and the fact that these data sets are obtained at different points in time with very different topic and sample composition criteria.

The distributions shown in Chapter 2.1 visually resemble a normal-distribution, possibly as a result of how the score per tweet is calculated as the sum of individual word lexicon ratings, resulting in an overall distribution that resembles a normal-distribution, centered around the mean of the Hedonometer word-score distribution [7, 8, 98]. The Anderson-Darling test for normality is rejected for all considered data sets (COVID-19 NYC: $A^2 = 128,939.377^{**}$, D cohort: $A^2 = 10,358.988^{**}$, R cohort: $A^2 = 31,346.166^{**}$). Subsequently performing a one-sided Mann-Whitney U test, we find that there is a positivity bias in the observed mood (COVID-19 NYC: $U = 27,417,406,955,105^{***}$, D cohort: $U = 210,506,706,133^{***}$, and R cohort: $U = 2,707,934,370,946^{***}$), a phenomenon that has been observed previously across several sources [7, 99].

2.4 Mood distribution over time

We observe nearly identical distributions of twitter sentiment, leading to comparable estimates of expected means for all three corpora, which concist of entirely different collections of social media posts by different cohorts of individuals. Most common sentiment analysis tools, like the Hedonometer, would likely fail to uncover potentially important sentiment or mood differences between and within individuals and cohorts, even between disparate corpora that are focused on different topics, localities, and samples.

Applications of sentiment analysis to measure public mood states from online text data, such as a stream of Tweets, are generally diachronic, i.e. they are aimed at detecting changes in public sentiment or mood over time. For that reason here we also examine the variability of the mood distribution over time in the mentioned corpora using the same sentiment analysis tool.

Rather than drawing conclusion from changes in the estimated average of daily sentiment, here we determine the 83% confidence intervals (CIs) of the observed sentiment scores on a daily basis by finding the 8.5th percentile to the 91.5th percentile of sentiment scores

Statistically significant differences can be inferred at the level $\alpha < 0.05$ when two 83% CIs do not overlap [100]. For all data sets, we find that the overlap of the daily CIs is at least 60% of the complete CI (COVID-19 NYC: $\bar{o_d} = 92.65\%$, $\min_{o_d} = 83.07\%$; D cohort: $\bar{o_d} = 75.25\%$, $\min_{o_d} = 59.55\%$; R cohort: $\bar{o_d} = 86.54\%$, $\min_{o_d} = 78.26\%$). Given this large overlap of day-to-day CIs of mood distributions, we can not reject the null-hypothesis of no significant differences between daily mood values, and can infer no actual changes in observed daily sentiment values. In other words, when we consider the variance of mood estimations on a daily basis, very few days are found to have a mood

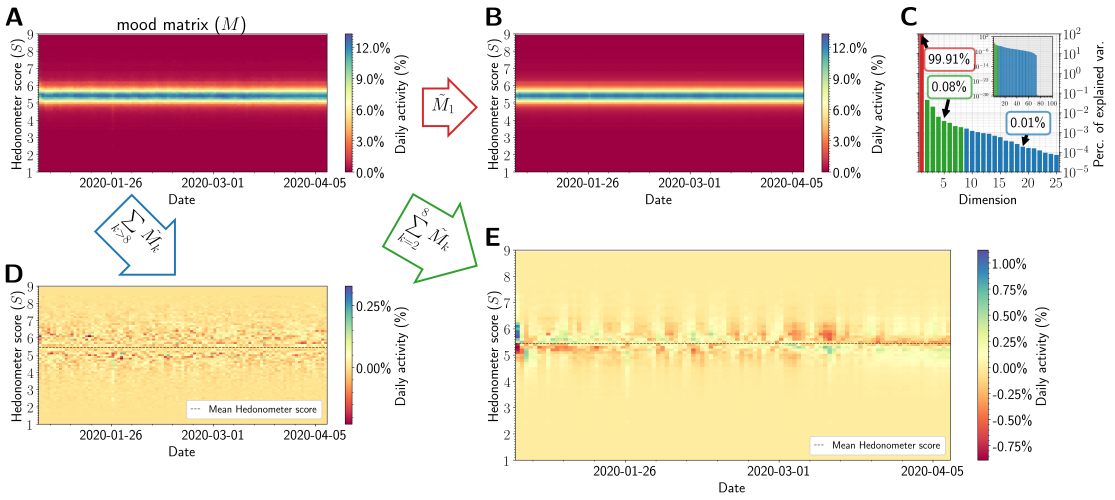


FIGURE 2.2: Caption

value that differs significantly from that of other days, indicating no changes in mood using this common sentiment analysis application.

2.4.1 Unraveling the diachronic eigenmood components of mood distributions

We can collate the mentioned distributions of daily mood values into a timemood matrix whose entries are the number of tweets observed at a given discrete time interval and given sentiment value. For our current analysis, we bin tweets at daily intervals and mood bins at intervals of $0.1 \in [0.95, 9.05]$ (Hedonometer scores are the average word ratings per tweet ranging from 1 to 9). Panel A of Chapter 2.2 visualizes such a column-normalized mood matrix, in which the counts per bin are divided by the total number of tweets on a daily basis.

This matrix indicates, again, a lack of any significant changes in daily mood distributions which are universally characterized by a strong central tendency that remains stationary over time (continuous horizontal line of maximum sentiment at bin $s = 5.35$).

2.5 Approximating mood

To disentangle different diachronic aspects of the mood as indicated by matrix M we determine its Singular Value Decomposition (SVD), defined as

$$M = U \cdot \Sigma \cdot V^T, \tag{2.1}$$

in which Σ is a matrix that has the singular values of the mood matrix as entries on the diagonal and zero values elsewhere.

Based on the SVD, we define the approximation of the mood matrix of a dimension k as follows: we insert the k-th singular value of M into the k-th position on the diagonal of a further zero-valued matrix, denoted by Σ_k. The approximation of the k-th dimension is then defined as

$$\tilde{M}_k = U \cdot \Sigma_k \cdot V^T. \tag{2.2}$$

Note that we can easily extend this approximation to include multiple dimensions. Panels B, D and E of Chapter 2.2 show examples of approximations of the mood matrix.

Performing this analysis for our Twitter corpora, we find that almost all of the variance in the data is explained by this first component (D cohort: 97.85%, R cohort: 99.03%, and COVID-19 NYC: 99.91%). Panel C of Chapter 2.2 displays the percentage of variance that is explained by each dimension in the COVID-19 NYC data set.

2.6 Similarity of mood distribution and approximation

To asses the similarity between the actual mood matrix and the approximation, we further analyze the first dimension approximation of the mood matrix $\tilde{M}_1$. Specifically,

we quantify the degree to which the first singular vector (FSV) approximation is similar to the actual distribution of mood in the data set.

A Kolmogorov-Smirnov (KS) test to compare the approximated distribution with the original indicates that the average of the FSV approximation is almost identical to the overall distribution across the various Twitter cohorts, as the KS statistic of each comparison is 0.0006, 0.0003, and 0.0001 for the D cohort, R cohort, and COVID-19 NYC, respectively.

Effects of natural term frequencies

As we are considering lexicon matching techniques, the similarity between FSV approximation and the complete data sets is very likely to be related to the lexicon that is used. One explanation that aligns with this finding, is the fact that each word that has a score in the lexicon has a natural probability of occurring in natural language. Combining this with the fact that removing frequently used words increases subjective well-being predictions [101], we conclude that the common signal that is captured the FSV approximation is the overall distribution of mood that is caused by the product of the natural term frequencies and the scores that are assigned to these terms. Specifically, the FSV approximation captures the effects of the prevailing word frequencies in natural language.

Cohort	**JS**	p_f
D	0.128	0.075
R	0.114	0.034
Covid-19	0.118	0.015

TABLE 2.1: JS is the Jensen-Shannon Divergence between the first singular vector as a distribution and word-level happiness score distribution in the Brown corpus, p_f is the probability of finding a smaller JS, estimated from 100,000 random reshuffles of per-word happiness scores in the Brown corpus

If the FSV approximation is indeed an approximation of the distribution of sentiment due to natural word frequencies, we may expect it to be similar to the distribution of single-word sentiment in other corpora. Here we explore the Brown corpus as an alternative source of natural language word frequencies. The Brown corpus is a collection of texts across a variety of categories including news, reviews, fiction, etc. originally collected in 1961 [102]. We can measure the similarity between the FSV approximation and the single-word sentiment distribution by taking the Jensen-Shannon divergence (JS), a measure of the dissimilarity of two probability distributions, assigning values from 0 (the same) to 1 (no similarity) [103]. The results are shown in Table 2.1, where we see the largest divergence as 0.128 bits between the D Twitter cohort and the single-word sentiment distribution in the Brown corpus. We investigate how likely such a divergence would be by chance p_f through a null-model of random word frequencies, estimated by sampling 10^5 random shuffles of the observed word frequencies. We find R and $Covid-19$ cohorts both are significantly similar to the single-word Brown corpus distribution at $p < 0.05$; there was only a 3.4% and 1.5% chance respectively of finding a smaller JS divergence with random word frequencies. Despite the changes in language use over the last sixty years, we still find the that the FSV approximates the sentiment distribution in words better than we would expect by chance, except in the case of the D cohort, perhaps indicating that those with such a mental condition exhibit more significant changes in the frequency of sentimental word use.

These results are not specific to these Twitter cohorts and selection of sentiment tool. We show similar results for a 10% random sample of Tweets between 2010 and 2014 in Appendix A.1.

2.7 Artificial Toy Example

To better understand what the Eigenmood decomposition finds when applied to binned sentiment distributions over time, especially in the presence of a strong base language distribution, we apply the method to artificial data. The data is generated from a known, and controlled, mixture model, binned, and the matrix decomposition methods are applied to the data.

Data Generation

The data is generated by drawing from a mixture of Gaussian distributions over time. Data is limited to the range between 1 and 9, numbers drawn outside this range were discarded and redrawn. Numbers had a 30% to be drawn from a base distribution G_{base} with mean 5 and standard deviation 2, $G_{base} = Gaussian(\mu = 5, \sigma = 2)$ to simulate a basic language distribution. Two additional Gaussian distributions, G_{low} and G_{high} were each drawn from with a 35% chance, to represent a transient bimodal distribution. G_{low} and G_{high} both have scale 2, but a mean that changes in time according to a sine wave, moving away from 5 until the mean of G_{low} is 2 and the mean of G_{high} is 8, and moving back together. The distance between these means is graphed in Fig 2.3. These distributions are thus functions of time t, relative to the total length of simulated time t_{max}. The distributions are then $G_{low}(t) = Gaussian(\mu = 5 - 3 \cdot sin(t/t_{max} \cdot \pi), \sigma = 2)$ and $G_{high}(t) = Gaussian(\mu = 5+3 \cdot sin(t/t_{max} \cdot \pi), \sigma = 2)$. The probability of a number $1 < x < 9$ being drawn from this simulation at a given point in time is given by Equation 2.3.

$$p(x|t) = .3 \cdot G_{base}(x) + .35 \cdot G_{low}(x,t) + .35 \cdot G_{high}(x,t) \tag{2.3}$$

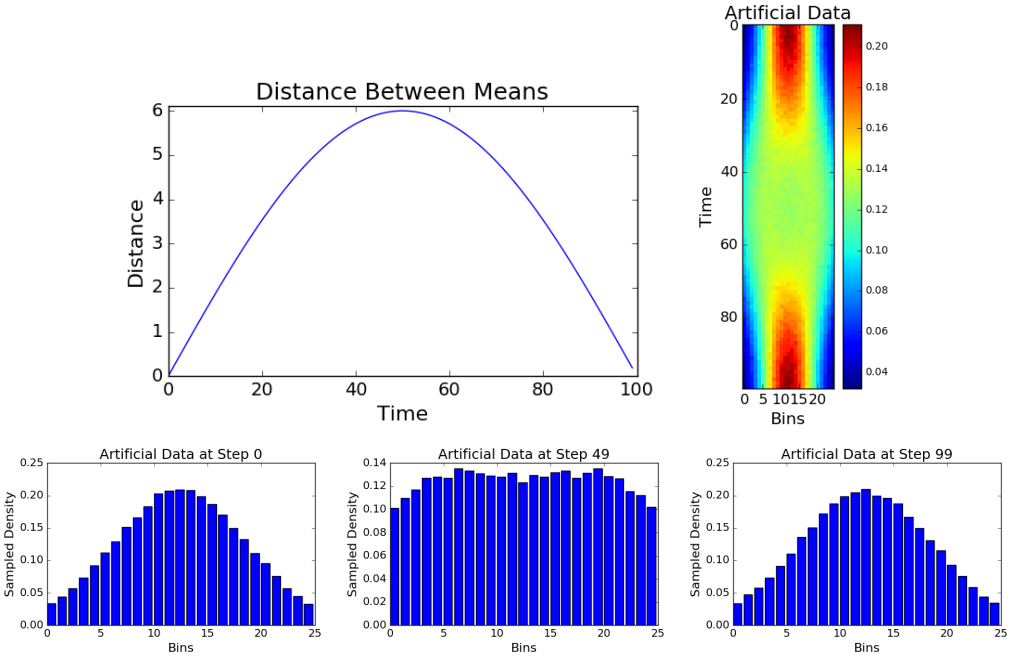


FIGURE 2.3: **Top:Left:**Distance between bimodal means over simulated time. **Right:** Heatmap of the data with columns representing bins and rows representing time **Bottom:** Left-to-Right: The histogram of simulated sentiment at time step 0, 49, and 99

10,000 samples were drawn from this simulated mixture at each of 100 time steps. These numbers were then binned into 25 equally sized bins between 1 and 9, and the counts in each bin were standardized to a probability distribution, such that the value in each bin summed to one at each time step. A heatmap of bins values over time is shown in Fig. 2.3, as well as histograms at the beginning, middle, and end.

SVD of Artificial Sentiment Over Time

We take the SVD of the artificial sentiment distributions shown in Fig. 2.3. This decomposition produces paired left and right singular vectors, the left representing characteristic time patterns, or *Eigenbins*, also interpreted as the relative weight given to each right singular vector over time. The right singular vectors in turn represent characteristic patterns in the distribution for the given time period, e.g. if the time period is divided into weeks these singular vectors could be interpreted as *Eigenweeks*. The first three components are shown in Fig. 2.4. The first singular vector is unimodal, but without the curved tails of a Gaussian; it is most emphasized at the beginning and end of the simulation. The second singular vector is similarly unimodal, but has both positive values in the middle and negative values at the tails, allowing it to distinguish differences in bins; at the beginning and end of the simulation it adds weight to the central peak while subtracting from the extreme bins, while in the middle of the simulation it subtracts weight from the central bins and adds weight to the extremes. The third component emphasizes a middle bimodal distribution that occurs most prominently at the quarter and three-quarter points in time, removing weight from both the middle and extreme bins. Two further components are shown in Fig. 2.5, but they are mostly noise. The scree plot for the SVD is shown in Fig. 2.6 Left.

The data reconstructed without the first component is shown as a heatmap in Fig. 2.6 Right. We can see that the reconstructed data captures the main changes in the distribution, the low and high Gaussian draws, removing the base distribution from the data.

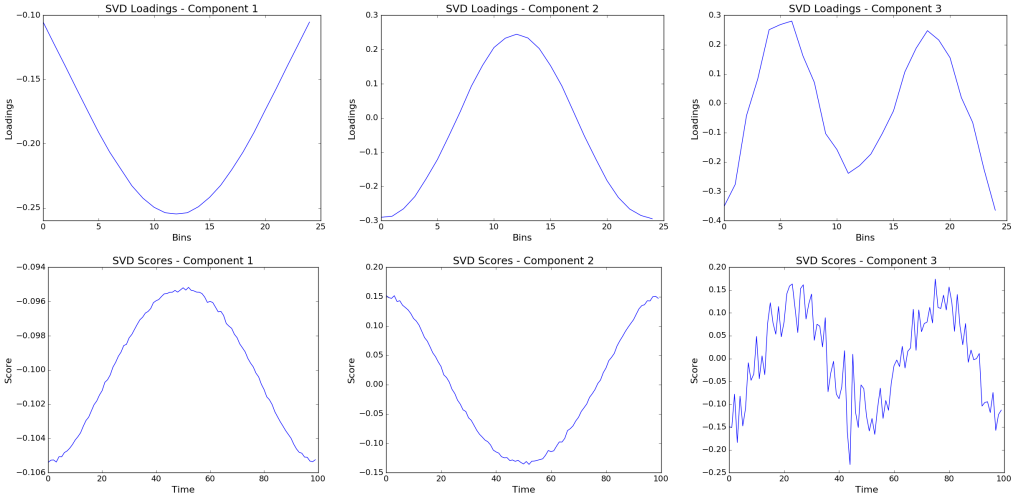


FIGURE 2.4: First 3 SVD Components. **Top:** right singular vectors, represent characteristic distribution patterns. **Bottom:** left singular vectors, represent characteristic time patterns, also interpreted as the relative weight given to each right singular vector over time.

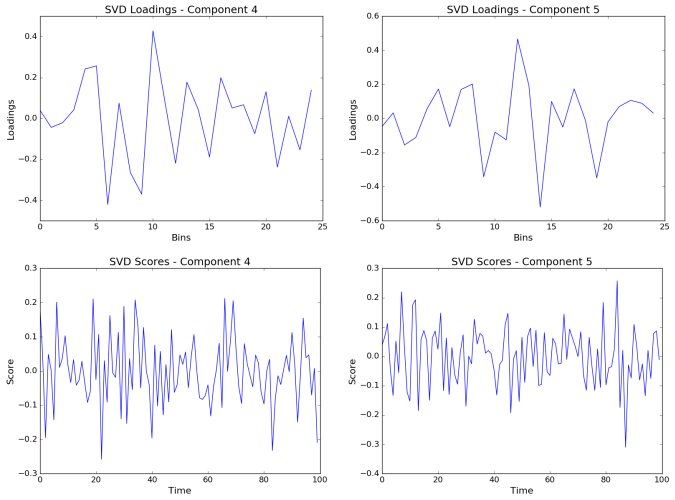


FIGURE 2.5: First 3 SVD Components. **Top:** right singular vectors, respresent characteristic distribution patterns. **Bottom:** left singular vectors, represent characteristic time patterns, also interpreted as the relative weight given to each right singular vector over time.

2.8 Sentiment Correspondence with Covid-19 Mortality

To demonstrate how Eigenmood components can aid in modeling other phenomena, we model mortality during the Covid-19 pandemic with and without sentiment and Eigenmood components as exogenous factors.

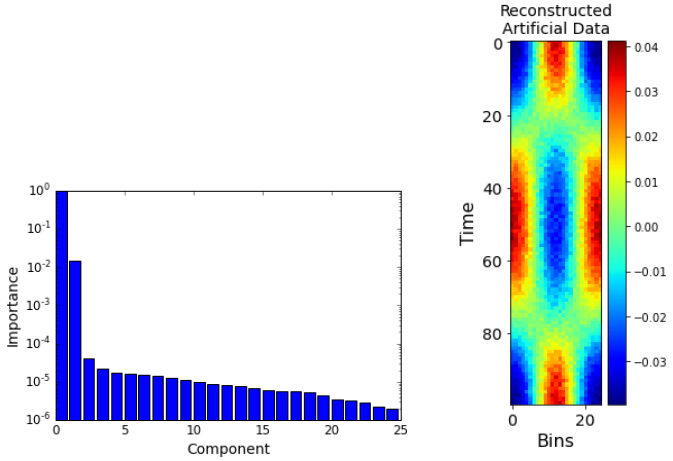


FIGURE 2.6: Left: The scree plot of relative importance for each vector. Right: The artificial data reconstructed for without the first component.

Data Collection

Data were downloaded from an open-access repository of all English-language COVID-19 related tweets published in the United States beginning January 22, 2020. Broadly, this repository aims to "contribute towards enabling informed solutions and prescribing targeted policy interventions" by making this data freely accessible for analysis [96]. For purposes of this study we sought to create two time-specific corpora using posts available through the repository. First, we collected all tweets published in this corpus between January 22, 2020 (the first day with available data and roughly one week prior to the first confirmed US COVID-19 case) through April 10, 2020. Next, we downloaded personal twitter timelines of unique social media users who contributed to the corpus and also resided within the 20 US cities with the most COVID-19 cases per 100,000 people according to their Twitter profile location, which we refer to as user-timeline data; n = 354,738 users, 380,361,628 tweets within the time period 2020-01-05 to 2022-03-26. The user-timeline data was used for more specific time-series analyses. The cities included in this list were: Atlanta, GA; Baltimore, MD; Boston, MA Charlotte, NC; Chicago, IL; Cleveland, OH; Denver, CO; Detroit, MI; Houston, TX; Indianapolis, IN; Las Vegas,

NV; Los Angeles, CA; Miami, FL; Nashville, TN; New Orleans, LA; New York, NY; Philadelphia, PA; San Francisco, CA; Seattle, WA; Washington, DC. In tables below, we label models using data from each city by the state abbreviation, except in the case of Los Angeles and San Francisco, both in California, which we label CA LA and CA SF.

Eigenmood around Covid

For each tweet in the Covid-19 user timeline data, we remove all tweets containing covid-related words to attempt to limit our analysis to the underlying mood, and not just reporting on the pandemic. We take the VADER score of each tweet, and remove from the distribution all those that did not match the VADER lexicon and were thus assigned a score of 0. The data were binned into 41 buckets of size 0.05 of VADER ranging from -1.025 to 1.025 to create a discrete distribution in sentiment scores each day. This was summed weekly, from Sunday to Saturday, and normalized to get a binned distribution each week. This matrix was decomposed using a Singular Value Decomposition as described above to obtain *eigenbins*, the contribution of each *eigenweek* component of mood in time. Mortality data was collected from the CDC website [104] to collect the total count of mortality by all causes in each state aggregated weekly from Sunday to Saturday.

ARIMA Models of All Causes Mortality Data

We focused on the period from 2020-01-05 to 2022-03-26. To model trends in mortality we trained an autoregressive integrated moving average (ARIMA) model. ARIMA models are frequently used in econometrics to model univariate time series, using a

number p of autoregressive terms looking backwards in time, an order of integration d to difference the time series for stationarity, and a number of moving average terms q directly regressing on errors in prior time steps [105][106]. ARIMA models have also been successfully used to model the spread of Covid-19 [107, 108]. Mood variables, both mean value each week and eigenbins, are treated as exogenous variables in the regression, differenced once to meet stationarity requirements. We divided the period of data into a training data set of 35% of the data (2020-01-18 to 2020-10-24, note the start is the second week of full data collection to account for differencing) a validation set of 35% of the data (2020-10-31 to 2021-07-31), and a test set of 30% of the data (2021-08-07 to 2022-03-26). The training data was used to create the initial SVD of the sentiment, in a sense learning the component eigenweeks. The data in the validation and test set were then projected onto those eigenweeks in order to continue each corresponding eigenbin into the validation and test set without allowing future data to influence the learning of the eigenweeks. The validation set was used to select the best hyperparameters of the model according to the best performing R^2 value of one-step-ahead predictions in the validation set. A grid search was performed over the hyper-parameters p, and q of the ARIMA model from 1 to 3 for p and from 0 to 3 for q, to select the best performing model without exogenous mood variables. Additionally, for models with included mood variables, a hyperparameter was selected to either use mood lagged by one week or contemporaneous mood. For models with eigenbin mood variables, we additionally selected the best 2 components of the first 11 based on performance on the validation set. Models were then retrained on the combined training and validation sets, the performance for this training is reported in Table 2.2 and the performance of the one-step-ahead forecast from these models on the test set is reported in Table 2.3

As we can see from Table 2.2, in almost all cases the selected eigenmood components

State	*NoMood*	*MeanVader*	*SelectedMoodComponents*
GA	0.905	0.908	**0.910**
MD	0.869	0.873	**0.884**
MA	0.901	0.901	**0.907**
NC	0.874	0.876	**0.880**
IL	0.915	0.916	**0.917**
OH	0.941	**0.943**	0.934
CO	0.815	0.816	**0.840**
MI	0.817	0.818	**0.847**
TX	0.913	0.914	**0.924**
IN	0.891	0.892	**0.897**
NV	0.813	0.816	**0.834**
CA LA	0.974	0.973	**0.974**
FL	0.864	0.865	**0.868**
TN	0.889	0.893	**0.901**
LA	0.809	0.806	**0.823**
NY	0.883	0.884	**0.886**
PA	0.933	0.935	**0.942**
CA SF	0.974	**0.975**	0.971
WA	**0.531**	0.510	0.518
DC	0.594	0.594	**0.600**

TABLE 2.2: R^2 values for validation selected models trained on the full training and validation set on in-sample 1-step ahead predictions for the full training and validation set. Bold values denote the best performance for a state

produce the best fit on the in-sample predictions. In two cases, for Ohio and California with mood based on tweets from San Francisco, the mean Vader score outperforms the selected eigenmood, while in only one case, Washington, the model without exogenous mood components has a better fit of the data.

As shown in Table 2.3, for 16 of the 20 cities investigated, including either mean sentiment or selected mood components as exogenous factors improves the forecast of the model on the held-out test set. However, in only 6 cases do selected components perform the best, while in 10 cases inclusion of the mean Vader sentiment has the best performance (in 8 cases selected mood components outperform no sentiment, while in 12 cases mean Vader sentiment outperforms no sentiment). More details on these models are included in Appendix A.8, including full regression tables for each model.

Mortality during the Covid pandemic is difficult to model, with large peaks that

State	*NoMood*	*MeanVader*	*SelectedMoodComponents*
GA	0.887	0.834	**0.891**
MD	0.668	**0.699**	0.416
MA	0.851	0.851	**0.857**
NC	0.567	**0.570**	0.547
IL	0.848	**0.849**	0.848
OH	0.862	0.858	**0.897**
CO	0.861	**0.862**	0.805
MI	**0.895**	0.882	0.887
TX	0.881	0.880	**0.884**
IN	0.845	0.846	**0.849**
NV	0.636	**0.650**	0.601
CA LA	0.902	0.903	**0.910**
FL	**0.933**	0.930	0.931
TN	0.714	**0.731**	0.715
LA	0.828	**0.830**	0.803
NY	0.795	**0.810**	0.801
PA	0.905	**0.905**	0.851
CA SF	0.902	**0.916**	0.875
WA	**0.632**	0.624	0.604
DC	**0.253**	0.253	0.222

TABLE 2.3: R^2 values for validation selected models trained on the full training and validation set on out-of-sample 1-step ahead predictions for the held-out test set. Bold values denote the best performance for a state

appear and disappear suddenly, in and out of training - test splits. In general, adding sentiment information provides more information for the model leading to better model fit. However, it is difficult to generalize to forecast far into the future without refitting the model. The mean sentiment allows for better generalization to unseen data, as seen in Table2.3, but in near-term forecasts we can usually find eigenmood components that allow for better model fit, as seen in Table 2.2. Additionally, the training/validation/test split disadvantages the SVD method of finding eigenmood components by limiting the training of the components themselves to the training set, such that the future test set is more distant in time from the data used to create the eigenweek loadings. Retraining the model more frequently, perhaps at every time step, as well as updating the components, may allow for better one-step-ahead forecasts using eigenmood components. We leave this for future work.

2.9 Discussion

Observing the mean sentiment in a corpus of written artifacts provides a good deal of emotional information, however, it is only part of the story. Examining how individual word frequencies affect this mean is useful to observe changes in an individual day's word frequency compared to the natural frequency in language, but this is very fine-grained. A singular value decomposition offers a more comprehensive approach, splitting a distribution into components that describe many different days. We have shown that the first component closely represents the overall distribution of sentiment in natural language and dominates distribution of sentiment each day. By removing this component we can construct clear visualizations of changes in the distribution of sentiment over time, and we have shown how the later components of the decomposition can be useful in understanding other phenomena, although these components have difficulty extending to out-of-sample data. In the next chapter, we will show how an Eigenmood decomposition can be used to characterize holidays and help to investigate long-standing hypotheses about population-level behavior.

Chapter 3

Human Sexual Cycles are Driven by Culture and Match Collective Moods

This majority of this Chapter is a reproduction of a paper published in Scientific Reports [109] and includes writing from co-authors Pedro Leal Varela, Johan Bollen, Luis M. Rocha, and Gonçalves-Sá.

3.1 Abstract

Human reproduction does not happen uniformly throughout the year and what drives human sexual cycles is a long-standing question. The literature is mixed with respect to whether biological or cultural factors best explain these cycles. The biological hypothesis proposes that human reproductive cycles are an adaptation to the seasonal (hemisphere-dependent) cycles, while the cultural hypothesis proposes that conception

dates vary mostly due to cultural factors, such as holidays. However, for many countries, common records used to investigate these hypotheses are incomplete or unavailable, biasing existing analysis towards Northern Hemisphere Christian countries. Here we show that interest in sex peaks sharply online during major cultural and religious celebrations, regardless of hemisphere location. This online interest, when shifted by nine months, corresponds to documented human births, even after adjusting for numerous factors such as language and amount of free time due to holidays. We further show that mood, measured independently on Twitter, contains distinct collective emotions associated with those cultural celebrations.

Our results provide converging evidence that the cyclic sexual and reproductive behavior of human populations is mostly driven by culture and that this interest in sex is associated with specific emotions, characteristic of major cultural and religious celebrations.

3.2 Introduction

Human reproduction shows a yearly cyclical pattern and whether this periodicity is driven primarily by cultural or by biological factors has been an open question for several decades. In Western, Northern Hemisphere countries, births tend to peak in September, corresponding to early winter conceptions [86]. These conception dates are aligned with the December solstice which has been taken as evidence for the existence of an environment-induced biological clock that drives human reproduction cycles [110, 111]. Proposed evolutionary explanations include temperature [112], libido, or the availability of food [86, 113]. However, this conception peak also coincides with religious celebrations, like Christmas, suggesting that culture drives the observed birth cycles. Culture and

biology certainly influence each other, and it is very likely that both influence sexual drive. However, whether biological or cultural factors best explain the reproduction cycle has long been debated in the literature, with biological explanations dominating the argument [86].

The biological hypothesis, proposes that human reproductive cycles are an adaptation to the seasonal cycles caused by hemisphere positioning in the yearly orbit of the Earth around the Sun. If true, reproductive periodicity should be similar among Northern Hemisphere countries, less pronounced closer to the equator, and reversed in Southern Hemisphere countries [114]. On the other hand, the cultural hypothesis proposes that conception dates vary mostly due to cultural factors, such as holidays or seasonal marriage patterns [111]. If true, we should see similar sexual cycles in similar cultures independent of hemisphere. To study these hypotheses we need to measure sexual activity on a planetary scale. Common proxies for such measurements include birth records, incidence of sexually transmitted diseases, or condom sales [115]. However, for many countries these records are inaccurate with respect to the timing of sexual activity [116, 117] and a focus on hospital records (for births or sexually transmitted diseases) would largely restrict analysis to "Western" countries, where such data tends to be most commonly available. Thus, previous indicators do not offer sufficiently accurate data from across the globe to help distinguish between the two hypotheses.

The recent availability of large-scale population data from web searches and social media now allows us to study collective social behavior on a global scale. In this work, we gauge interest in sex directly from Google searches and characterize seasonal population sentiment from the analysis of Twitter feeds. We show that analysis of this large-scale online activity can be used as proxies for real-life actions and help answer longstanding scientific questions about human behavior.

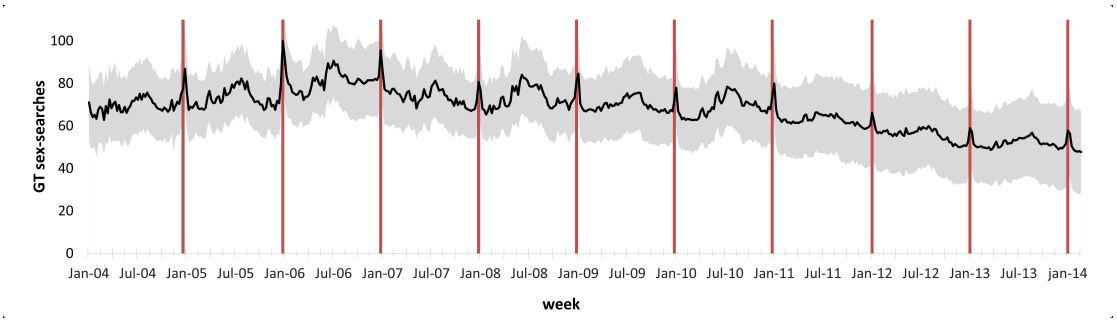


FIGURE 3.1: **Weekly queries for the term "sex" for a group of representative western Northern countries.** The black line represents the averaged queries in a 10-year period, obtained from Google Trends, which is normalized by overall search volume. These countries are: Austria, Canada, Denmark, Finland, France, Germany, Italy, Lithuania, Malta, Netherlands, Poland, Portugal, Spain, Sweden and the United States of America. Shaded grey represents the standard deviation. The red vertical line marks Christmas week.

3.3 Results

3.3.1 Worldwide Variations in Sexual Interest

To measure interest in sex, for each country, we retrieved the frequency by which people searched for the word "sex" using Google Trendstm (GT) [118] (Methods 3.5.1, 3.5.2,3.5.3); henceforth referred to as "sex-searches." Interestingly, even in countries where English is not an official language, the English term "sex" is either more searched for than the corresponding word in the local languages or they are strongly correlated (Supplementary Table S1). Moreover, the terms most associated with searches for "sex" in GT refer to direct interest in sex and pornography (Supplementary Table S1). Therefore, GT searches for the term "sex" are a good proxy for interest in sexual behavior in the countries analyzed in this study.

Fig. 3.1 depicts GT weekly sex-search data for 10 years from January 2004 to February 2014 for a set of Northern countries, which celebrate Christmas on December 25th. Yearly maximum peaks occur during Christmas week (red vertical lines), as previously observed for the USA [119]. While one may think that this increased interest in sex

results simply from more free time during the holiday season, GT data is normalized by overall search volume [118]; even in a situation of increased general online activity, the increase in sexual interest is higher. Conversely, we could expect the holiday season to lead to a decrease in overall searches, led by school vacations for instance, originating an artificial peak for sex-related interest. However, we do not observe similar increases in weekly sex-searches for other widely observed holidays, such as Thanksgiving in the USA or Easter in France (Fig. S1A and S1B). Furthermore, a putative decrease in overall searches is unlikely, as a decrease in searches for school-related material can be compensated by a strong increase in searches for "presents" or "recipes". In fact, when we control for search-volume of very common words, such as "on", "and", or "the", there is some variation around the holiday period but it is in different directions for different search terms (Fig S2A and S2B), probably resulting in an overall neutral change. Therefore, and although other dates lead to an increase in sex-searches (Fig. S1A and B), the Christmas holiday is uniquely associated with the highest peaks in sex-searches observed in these Northern countries. It is also known that, in Western Northern countries, conceptions peak around Christmas, in what some refer to as the "holiday effect" [120]. Indeed, the observed sex-search peaks match birth rate increases for this set of countries when shifted by nine months (Fig. S3A), which further confirms GT sex-searches as a good proxy for sexual activity.

Compared to the observation of sex-search peaks in Northern countries that celebrate Christmas on December 25th (and corresponding increase in September birth rates where such data is available), the two hypotheses outlined above would predict quite distinct observations for other cultures and hemisphere locations. If the biological hypothesis is correct, all Northern countries should have similar sex-search peaks around the same time, and these peaks should occur in a counter-phase pattern (six months later) in all

Southern countries—irrespective of culture. On the other hand, if the cultural hypothesis is true, these peaks should appear anywhere Christmas is celebrated—irrespective of hemisphere—and other similar celebrations in different cultures should lead to sex-search peaks in other times of the year.

To test these predictions, we extracted GT sex-search time-series data for all 129 countries for which GT offered consistent data. Countries were categorized according to hemisphere (North or South) and their predominant religion [121, 122]. Countries where at least 50% of the population self-identifies as Christian were considered culturally Christian countries, and similarly for Muslim countries. Other countries, where neither of these religions is dominant, were grouped separately; Supplementary Table S2 shows the complete list of countries and categorization.

Both Northern and Southern countries show a prominent peak in sex-searches around Christmas and we observe no counter-phase pattern corresponding to the southern hemisphere winter solstice of June 21st (see Fig. S4A, Fig. S5C, and Fig. S5D). In fact, there is a strong significant correlation ($R^2 = 0.54$, $p < 0.001$) between the mean sex-search time series of Northern and Southern countries (Supplementary Table S3). Since most Northern and Southern countries for which we have data identify as Christian (80 of 129), the observed correlation suggests that a cultural effect, rather than hemisphere location, drives the Christmas sex-search peak. Indeed, the birth data available for Christian, Southern countries peaks with Christmas sex-searches when shifted by nine months in much the same way as for Christian, Northern Countries, even though it is summer in the former and winter in the latter (Fig. S3). Notice further that there is neither a sex-searches increase in December nor a birth peak in September for Northern countries that do not celebrate Christmas on December 25th (Fig. S7). As reliable birth data is not generally available, particularly for Southern and Muslim countries,

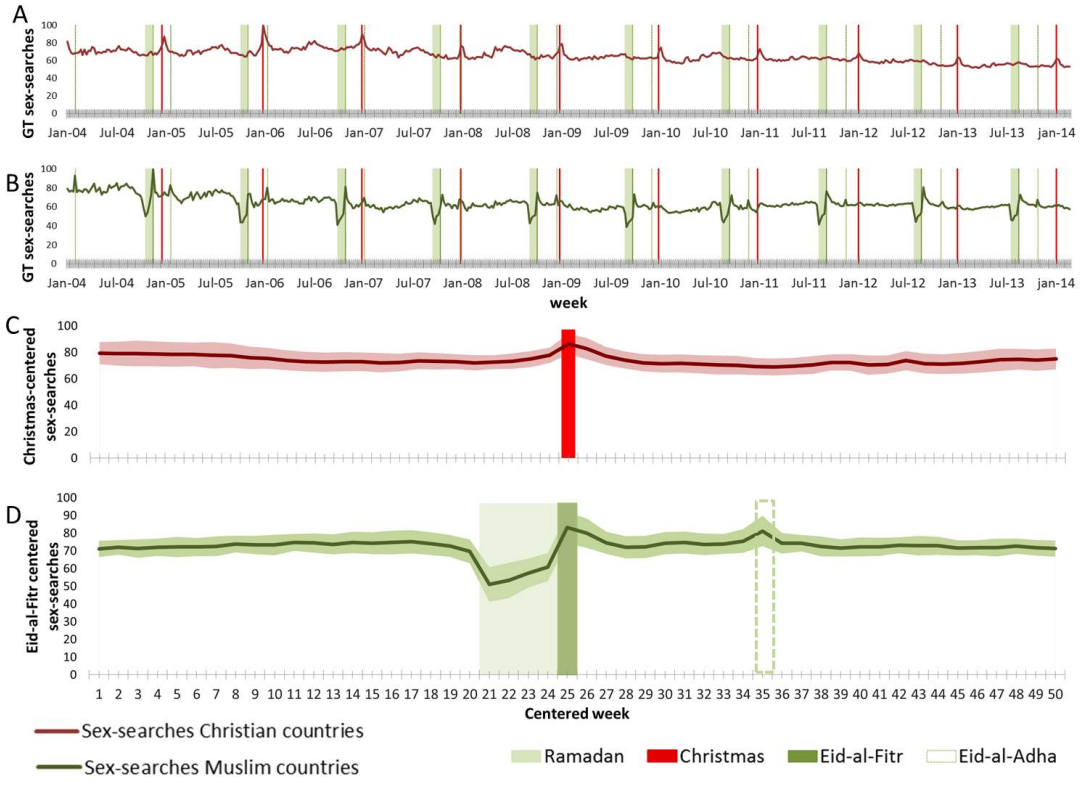


FIGURE 3.2: **Weekly queries for the term "sex" in culturally different countries.** (A) Normalized and averaged queries for all available countries identified as Christian (dark red line). (B) Normalized and averaged queries for all available countries identified as Muslim (dark green line). (C) Searches in all Christian countries centered around Christmas week (26). (D) Searches in all Muslim countries centered around Eid-al-Fitr week (25). See Supplementary Table 2 for country identification and availability on GT. The vertical red lines mark Christmas week, the shaded light green area represents Ramadan, with the darker green lines marking Eid-al-Fitr (solid) and Eid-al-Adha (dashed). Shaded areas around the lines in C and D show the standard deviation.

and is only available for four Southern countries, all of them predominantly Christian, (Methods 3.5.6, Supplementary Table S9 and Figs. S3 and S6), we use GT sex-search data instead to observe many more countries and address the two hypotheses.

Parsing all countries by religion (Fig. 3.2A&B, Fig. S4 and Supplementary Table S3), it is clear that the mean sex-search time-series are periodic but uncorrelated between Christian and Muslim countries ($R^2 = 0.19$, $p < 0.001$). The difference in sex-search behavior between these two sets of countries is further revealed in Fig. 3.2C&D, where we averaged the sex-search yearly time-series across all ten years centered on Christmas

week (for Christian countries) or centered on Eid-al-Fitr, the major family holiday that ends Ramadan (for Muslim countries). In Christian countries, the only clear peak occurs during the Christmas week. In contrast, in Muslim countries there is a peak during the week of Eid-al-Fitr and a second peak during the week of Eid-al-Adha, the other major religious and family celebration in Muslim culture; also noteworthy is a steep decrease during Ramadan, consistent with that period of general abstinence (as further discussed below). Both of these groups of countries clearly show sex-search peaks associated with distinct cultural celebrations, rather than with hemisphere. Indeed, it is worth noting that the Muslim calendar does not follow the solar calendar: every year Ramadan shifts by 10 days relative to its date during the previous Gregorian calendar year. Nevertheless, sex-searches peak during the moving week of Eid-al-Fitr (and Eid-al-Adha) in Muslim countries. The moving sex-search peaks associated with major religious events in Muslim countries further emphasizes the cultural driver behind such collective behavior.

To resolve the incompatible predictions of the biological and cultural hypotheses we made country-specific comparisons between hemisphere and culture, beyond the group-average behavior described above. We averaged the yearly sex-search time-series for each of the 129 individual countries across all years in four different ways: centered on Christmas week (fixed relative to the solar calendar), centered on Eid-al-Fitr week (moving relative to the solar calendar), and centered on each of the solstices, fixed on June 21st and December 21st (Methods 3.5.4, Supplementary Tables S4-6 and Fig. S5). We then measured the response of countries to a holiday as the sex-search z-score deviation above the mean at Christmas, Eid-al-Fitr and the two solstice weeks (Methods 3.5.5 and Supplementary Table S7). Fig. 3.3 shows a world map with color-coded countries: shades of red indicate countries whose highest sex-search deviation from mean occurs during the Christmas week, and shades of green indicate countries

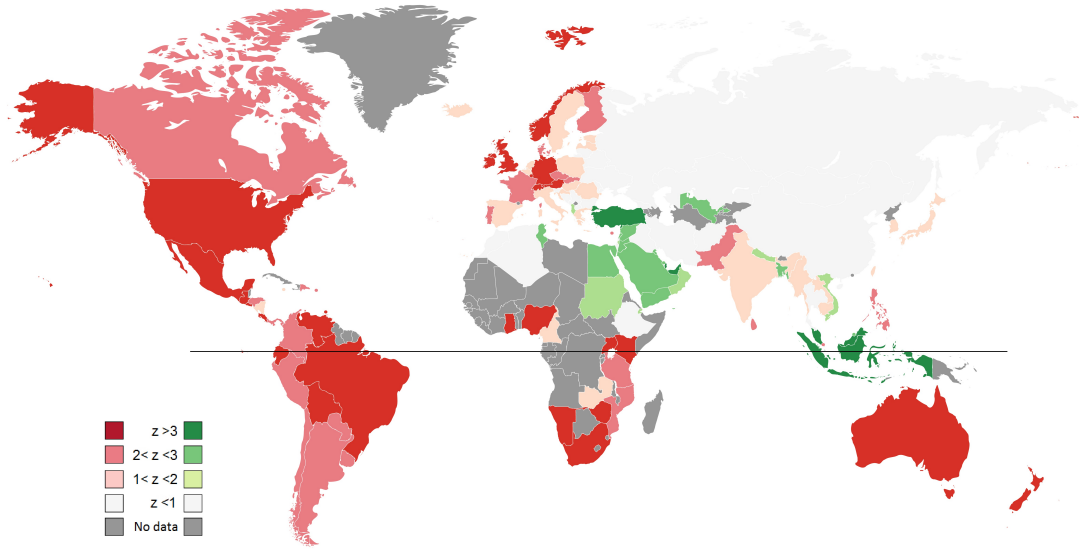


FIGURE 3.3: **World-wide sex-search profiles.** The world map is color-coded according to the z-score of each individual country's sex-search time-series. Shades of red represent a higher z-score (larger increase in searches) during Christmas week (on Christmas-centered data). Shades of green represent a higher z-score (larger increase in searches) during Eid-al-Fitr week (on Eid-al-Fitr centered data). Light grey denotes countries with no significant variation above mean in either of these weeks. Dark grey countries are those for which there is no GT data available. Black line represents the equator separating the hemispheres. Built using: https://mapchart.net/.

whose highest sex-search deviation from mean occurs during Eid-al-Fitr week (Methods 3.5.7). It is clear that this response yields a map organized according to culture rather than hemisphere.

We then compared this new country classification (according to the individual countries' sex-search profile, Supplementary Table S7 and Methods 3.5.14) with our previous identification based only on the proportion of the population that self-identified as Christian, Muslim or Other (Supplementary Table S2) (13,14). Out of the 30 countries originally identified as Muslim (14), 77% show a significant increase (z¿1) in sex-searches during the week of Eid-al-Fitr, and out of the 80 countries originally identified as Christian (13), 80% show a significant increase (z¿1) during the Christmas week, regardless of the hemisphere. It is important to note that this correspondence is even higher (91%) when we identify as "Other" the ten Christian countries that do not celebrate Christmas

on December 25th. In fact, we do not see an increase in sex searches around December 25th in any of these Northern Russian and Serbian Orthodox Christian countries, which celebrate Christmas in early January, and this further supports the cultural hypothesis (Methods 3.5.2, 3.5.14, Supplementary Figure S7). Moreover, only 14% of Southern countries showed a significant increase in sex-searches during the June solstice (Supplementary Tables S7 and S8B), demonstrating that there is no significant counter-phase sex-search peak in the southern hemisphere, contradicting the biological hypothesis.

3.3.2 Trends in Holiday Moods

The Christmas and Eid-al-Fitr holidays carry significant cultural and religious meaning, but they are not directly associated with sex. It is, in fact, very counter-intuitive to think of Christmas and Eid as the times of the year with the most online searches for sex. However, these events may trigger specific and collective moods, leading to a striking correspondence between these holidays and sexual interest. To investigate the emotional factors involved we measured changes in public sentiment on Twitter [3, 4, 8]. The analysis was performed before, during, and after Christmas and Eid-al-Fitr in a set of seven countries with sufficient Twitter traffic in our data: Australia, Argentina, Brazil, Chile, Indonesia, Turkey, and the USA (Methods 3.5.9 and Fig. S8). Although it is not possible to know whether the Google and Twitter populations are the same per country, given the large volume of Google searches and tweets, it is very likely that they provide a significant sample of the same populations.

Twitter sentiment was quantified by rating a random 10% sample of all tweets posted between September 2010 to February 2014 using the Affective Norms for English Words (ANEW) lexicon [6] (Methods 3.5.8 and 3.5.9). The ANEW lexicon consists of 1,034

English words that carry a sentiment score along three dimensions: Arousal (a), Dominance (d), and Valence (v), corresponding respectively to whether the word makes human raters feel calm vs. excited, controlled vs. in-control, and sad vs. happy. The sentiment value of a single tweet is defined as the mean ANEW score of its words. We translated the lexicon to Spanish and Portuguese to capture public sentiment in those languages as well, but did not have the ability to translate into additional languages. To avoid bias from holiday-related language, we ignored all words used in traditional greetings for all known holidays in the World (Supplementary Table S13); we also removed the word "Christmas" and "valentine" from the lexicon, which does not include other holiday names.

We first observed that the weekly volume of sex-searches significantly correlates with the mean weekly sentiment derived from the three ANEW dimensions in a multiple linear regression (Methods 3.5.15, Supplementary Table S10). In every country, valence yields a positive coefficient, while dominance a negative coefficient; thus the happier but less in-control the population mood is, the more sex-searches tend to increase in every country (Methods 3.5.10 and 3.5.15). Interestingly, while public sentiment displays a strong linear relationship with sex-search volume when all mood dimensions are considered, there is little correlation with each ANEW dimension on its own (Supplementary Table S11).However, the observed linear correlation does not allow us to characterize the population mood in the target cultural celebrations. To investigate if days that are similar in mood to Christmas in Christian Countries or to Eid-al-Fitr in Muslim Countries also tend to observe increased volume of sex-searches, we need a more nuanced characterization of the mood profile each week.

Because collective mood sentiment, as measured here, is derived from many tweets of large and diverse populations, it can contain distinct and informative components.

Thus, we employed an eigenvector-based analysis (20) to characterize the distribution of sentiment values, rather than just average sentiment. We thus obtain the components of public sentiment that explain most of the variance in the data not attributable to regular language use, hereafter referred to as "eigenmoods." Specifically, an eigenmood is a small set of components (eigenvectors) of a matrix. In this matrix, the rows denote sentiment scores in a given range or bin, and the columns denote the weeks (Methods 3.5.11 and 3.5.16), and elements are the number of tweets during a week that fall in that bin. Thus, an eigenmood is not an average sentiment value (per week in our analysis), but rather a change in the distribution of sentiment that explains a significant proportion of the variation in the time-series data [123].

We found that two components were sufficient to describe public sentiment associated with each holiday and country – a characterization that is independent of sex-search volume, and relies only on measurement of sentiment on Twitter (Methods 3.5.10, 3.5.11, 3.5.12,3.5.16,3.5.17,3.5.18 and Supplementary Fig. S10 and Fig. S11). Fig. 4 (Column A), Fig. S9 and Fig. S14 show the sentiment distribution of the selected eigenmoods that best characterize the holidays of interest, per every week of the year; redder (greener) colors represent increased (decreased) numbers of tweets falling in the respective mood dimension bins – e.g., for valence, upper bins on vertical axis denote increased happiness and lower bins denote increased sadness. The sentiment distributions of rows 1, 2, and 3 in Fig. 4 column A are centered on Christmas for USA (Northern, Christian) and Brazil (Southern, Christian), and Eid-al-Fitr for Indonesia (Southern, Muslim). While the eigenmood that describes Christmas in the USA uses only the valence dimension of ANEW, the best eigenmood for Christmas in Brazil requires valence and arousal, and for Eid-al-Fitr in Indonesia requires valence and dominance. The sentiment distribution of these eigenmoods per week clearly shows that significant and unique changes in sentiment

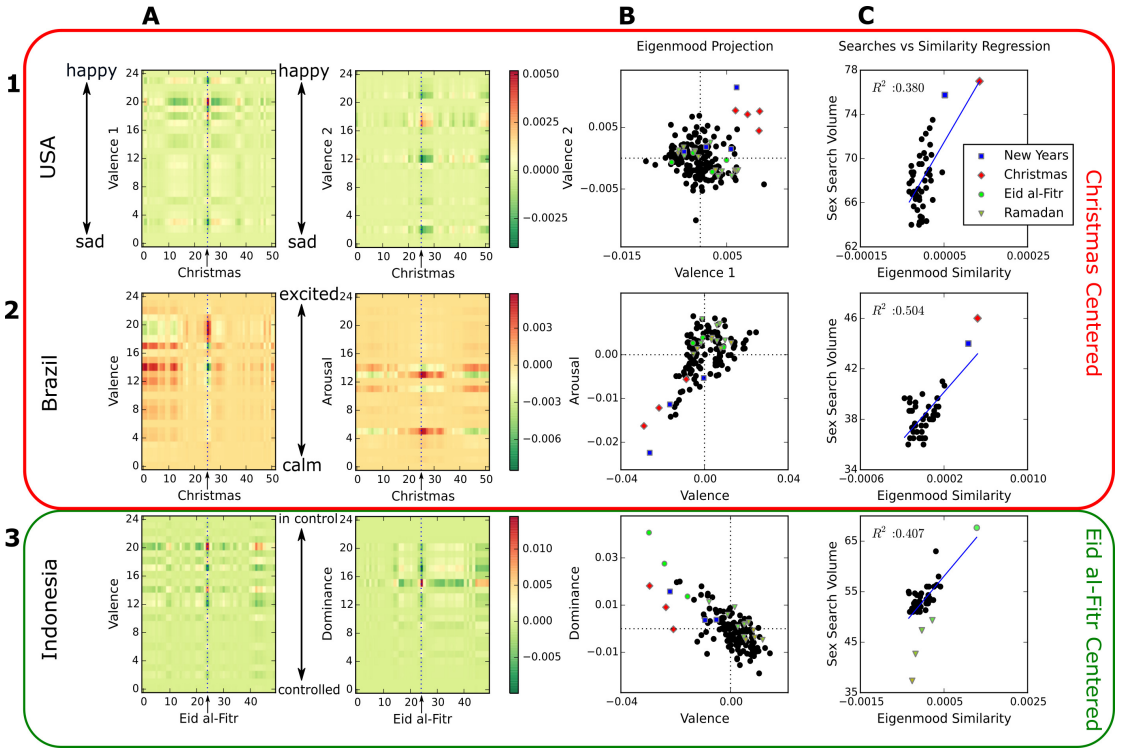


FIGURE 3.4: **Mood distributions and their correlations with sex-searches.** Rows: 1 – USA centered on Christmas, 2 – Brazil centered on Christmas, 3 – Indonesia centered on Eid-al-Fitr. Columns: A – Heatmaps of sentiment distribution reconstructed from selected eigenmoods. Vertical axis specifies the bins of the ANEW distribution for a given mood dimension, from low (bottom) to high (top) values. Eigenmood components were selected to best characterize the respective holiday and country (after removing the first component). In the case of the USA (Row 1), the two selected components both fall in the "valence" dimension and are labelled valence1 and valence2; for Brazil (Row 2) and Indonesia (Row 3) the first component also falls in the "valence" mood dimension, but the second falls in the "arousal" and "dominance" dimensions, respectively. Horizontal axis specifies the week of the centered, averaged year (52 weeks for the Gregorian calendar, 50 for the Muslim Calendar). The dotted line in the center marks the holiday of interest, on week 26 for Christmas, or week 25 for Eid-al-Fitr. Color represents the weight of the eigenmood per bin per week. B – Projections of weeks into the space formed by the selected eigenmood components. Each axis specifies the projection of week onto each component that defines the eigenmood. See text for details and supplemental materials for more information on component selection. C – Linear regressions between GT sex search volume (vertical-axis) and similarity to holiday center in the Twitter eigenmood space depicted in column B (horizontal-axis) for averaged weeks. The weeks of Ramadan are shown with increasing color intensity from more yellow to more green as they approach Eid-al-Fitr. The R^2 values for the regressions are 0.380 for Christmas in the USA, 0.504 for Christmas in Brazil, and 0.407 (0.637 without the Ramadan weeks) for Eid-al-Fitr in Indonesia.

occur during the target holidays. In all these cases, the public mood of the holiday in question generally shifts to "happy" bins (more red in higher valence) and away from "sad" bins (more green in lower valence). In Brazil, the mood also shifts to more "calm" bins during Christmas week (more red in lower arousal), and in Indonesia it also shifts to neither "in-control" nor "controlled" bins during the Eid al-Fitr week (more red in mid dominance). More detailed characterization of eigenmoods and their selection for each country is provided in Supplementary Material (Methods 3.5.16,3.5.17,3.5.18, Fig. S12-13).

Fig. 4, column B, shows all weeks in the data projected onto the selected eigenmood space of two components for each country. It is clear that in this space Christmas weeks (red diamonds) cluster together for the USA and Brazil, and Eid-al-Fitr weeks (green circles) cluster together for Indonesia, demonstrating that the eigenmoods are consistent in different years for each holiday in each country. Fig. 4 column C depicts the linear regression between sex-search volume as calculated before (vertical axis), and mood similarity to the target holiday in the respective eigenmood space (horizontal axis) for all weeks in the data set denoted by black circles in the plot (Methods 3.5.19). We observe a significant correlation for all countries studied, with $R^2 \geq 0.38$ for Christmas in all Christian Countries and $R^2 \geq 0.34$ for Eid-al-Fitr in all Muslim Countries, irrespective of hemisphere (Supplementary Table S12). Thus, in Christian countries we can say that the more the public mood of any week resembles the Christmas eigenmood, the larger the volume of observed sex-searches tends to be. Similarly, in Muslim Countries the more public mood is similar to the Eid-Al-Fitr eigenmood, the larger is the volume of sex-searches. In the case of both Muslim Countries studied (Indonesia and Turkey), there is a striking result pertaining to Ramadan: those 4 weeks (4 lowest green triangles in Fig. 4C, bottom right, for Indonesia), have the lowest sex-search volume by far in

the data, consistent with the period of abstinence that marks Ramadan (see also Fig. 2B, Fig. 2D). The public mood during these weeks of Ramadan is also quite distinct from the Eid-al-Fitr mood (horizontal axis in Fig. 4C, bottom right), but, becomes more similar the closer the week is to Eid-al-Fitr in time; and as the mood becomes closer to the Eid-al-Fitr mood as Ramadan approaches its end, the sex-search volume also increases. Naturally, due to the low, outlier sex-search volume during Ramadan weeks, the linear regression is much stronger if those weeks are removed, with $R^2 \geq 0.64$ (Supplementary Table S12).

Thus, not only there are specific moods associated with Christmas and Eid-al-Fitr, the eigenmoods that best characterize these holidays significantly correlate with increased interest in sex throughout the calendar. This is true in all countries studied, in both hemispheres and cultures. Moreover, and although these moods, occur at different times in different cultures, they seem to be similar in essence and, in general, the "happier" the mood, the more it associates with sex interest.

We have shown that components of mood corresponding to major cultural holidays also correspond to sex searches online, but did not fully investigate whether one variable causes the other. Granger Causality [124] is a widely used method developed for economic data that argues that we can perform a statistical test between two time series for "causal" relationships. This method argues that if lagged values of one series X offer a significant improvement when included in a linear model with lagged values of the other Y, we have evidence to reject a null-hypothesis that Y is not Granger-caused by X, and say that X Granger-causes Y.

	arousal mean	dominance mean	valence mean	holiday similarity
variable g-cause sex-search	0.0981	0.0136*	0.0164*	0.003 47**
sex-search g-cause variable	0.0446*	0.360	0.459	0.0480*

TABLE 3.1: Mean and Holiday Similarity Granger-Causality p-values. * indicates $p < 0.05$, ** indicates table Bonferroni corrected $p < 0.00625$

In Table 3.1 we have the p-values for granger causality tests between the time series of mean ANEW values against the time series of sex searches for the United States. Dominance and Valence sentiment means Granger-cause sex searches at $p < 0.05$, and holiday eigenmood similarity Granger-causes sex searches at $p < 0.00625$, while sex searches significantly Granger-cause arousal mean and holiday similarity at $p < 0.05$. More analysis is included in Appendix B.1.

3.4 Discussion

Taken together, our analyses provide strong converging evidence for the cultural hypothesis: human reproductive cycles are driven by culture rather than biological adaptation to seasonal cycles. Furthermore, the observed peaks of interest in sex occur around family-oriented religious holidays, across different hemispheres and cultures, and the measured collective mood on these holidays correlates with interest in sex throughout the year, beyond these holidays. This correlation suggests that the cultural driver of reproductive cycles depends on the collective mood of human societies, though establishing such causality warrants further study. It is also worth noticing that while other major holidays in each country lead to increased sex-search volume (e.g. Eid-al-Adha), not all holidays exhibit this effect (e.g. Easter and Thanksgiving), suggesting that certain holidays have unique eigenmoods which lead to increased interest in sex at the population level. Thus, specific mood states–typically happier, calmer, and neither in-control nor

controlled–are associated with interest in sex, and this collective emotion is universal and maximized during cultural celebrations such as Christmas and Eid-al-Fitr. The fact that the Muslim holidays do not follow a solar calendar, with the interest in sex varying according to the religious calendar, provides additional support for the cultural hypothesis.

It is clear from this work that culture (particularly the religious calendar) best explains the pattern of sexual interest. Naturally, it is important to stress that if collective mood states drive interest in sex at the individual level, there must ultimately be a common biological response to the cultural, contextual driver. Several hypotheses can be entertained – though not adaptation to seasonal cycles. For instance, some studies show that depressed people lose interest in sex and that "happy moods," such as those uncovered for Christmas and Eid-al-Fitr, are usually more conducive to sexual arousal [125, 126]. Increased food consumption has also been shown to have a relationship with sexual maturation and interest [127, 128], however, we do not see similar increase in sex-searches during other holidays associated with high food intake, such as Thanksgiving in the USA or Easter in France. And given the children and family focus of both Christmas and Eid-al-Fitr, it is reasonable to consider psychological and symbolic triggers to the observed behavior, but the neurological and biochemical pathways involved in such responses are as yet unknown.

That the culturally motivated surge in sexual interest can be so easily anticipated and measured has implications for public health and policy. Hospitals should be prepared for an increase in STD testing and possibly even abortions in the weeks following such holidays and when the corresponding collective mood is observed at other times of the year.

Overall, this work emphasizes the need for more world-scale studies and the importance of a better understanding of global collective behaviors at the level of individual countries. These will enable better-informed decisions and the more effective fine-tuning of policy towards the distinct needs of countries, cultures, and communities.

3.5 Methods

3.5.1 Google Trends Data

Google Trends (GT) provides a time series index of the search volume of a given Google query (10). GT allows for searches in a selected region (country, state, city, etc.) and for a selected time range starting in January 2004 for most countries. Google normalizes the resulting query index relative to the total amount of query volume for a search term in the chosen area, per week, so that the maximum query share of the time series is set to be 100. GT queries are also broad matched, meaning that queries such as "sex videos" are counted in the calculation of the query index for "sex".

3.5.2 Country Selection and Categorization

We considered all countries for which GT is available and for which a search for "sex" had a least two contributing cities and had enough time points to analyze at least four consecutive holiday seasons (Christmas and Ramadan), thus starting at least in the last week of 2009. This was the case for 129 countries in all continents. In the paper these countries are identified either by their name or by the country code, as in Supplementary Table S2.

Countries were categorized according to their major religion and geographical location (continent and Northern or Southern Hemisphere according to Wikipedia) and this categorization is referred to "identification" in the main manuscript. A country was considered "culturally Christian" when at least half of its population identified as Christian (Catholic, Protestant, Orthodox, or other) (13). A country was considered "culturally Muslim" when at least half of its population identified as Muslim (14). A country was labeled as "Other" when the majority of its population didn't identify as either Christian or Muslim. In the case of countries that have parts of their territory in both hemispheres, we used the location of the capital as the deciding criteria. Out of the countries identified as Christian, eleven have a majority that follow either the Russian or Serbian Orthodox Churches (namely: Belarus, Bosnia and Herzegovina, Bulgaria, Georgia, Macedonia, Moldova, Montenegro, Serbia, Slovenia, Russia and Ukraine). In ten of these countries (Bulgaria being the exception), Christmas is celebrated in early January (of the Gregorian Calendar) and they could have been labeled as Other for the proposes of this analysis.

3.5.3 Searches for "sex"

We downloaded the time-series corresponding to searches for "sex" for each of the available countries in GT as long as they had at least two cities contributing data, and had enough time points to analyze at least four consecutive holiday seasons (Christmas and Ramadan), thus starting at least in the last week of 2009. Supplementary Table S2 shows all countries included in the analysis. Because Google does not provide the absolute number of searches and we do not have access to the normalization algorithm, all the analyzed data is relative to the total search volume and it has been noticed by ourselves and by others that there is some variation the output GT provide, from week

to week. To limit this variation all of the analyzed data was downloaded on the same week.

For a subset of 50 countries (on all continents) we downloaded GT data for 2 search queries: (1) for the term "sex" and (2) for its translation in the local language. We compared the volume of searches between the two queries and calculated their correlation over time. Supplementary Table S1 shows the 25 countries and languages that retrieved a sufficiently significant search volume in the local language to support our analysis. We then calculated the "Search Volume Ratio", as the number of searches for "sex" divided by the number of searches for the corresponding translation. We also calculated the Correlation between the two time series ("sex" and the translated word) as the Pearson's R.

GT also provides and ranks the top words associated with the search term and these are also shown on Supplementary Table S1.

3.5.4 Centered Calendars

Data were organized into yearly "calendars" centered around the holidays of interest in order to better compare time series across cultures, and to create better summaries of averaged yearly time-series. Five "yearly calendars", or sets, were constructed:

1. The first, a "Civil Calendar" starts on the first week that includes January 1st and ends on the following December 31st.

2. The second was centered around the weeks that contain Christmas. In this paper we refer to it as the "Christian Calendar".

3. The third was centered around the weeks that contain the Eid-al-Fitr celebrations. In this paper we refer to it as the "Muslim Calendar".

4. The fourth was centered around June 21st and is referred to as the "June Solstice Calendar";

5. The fifth was centered around December 21st and is referred to as the "December Solstice Calendar".

Each week of each calendar was given an index ranging from 1 to the maximum number of weeks in that year. The first week GT indexes starts at the Jan 1 2004, so all remaining weeks will start seven days from this first index. In our centered calendars, the week containing Christmas and the solstices becomes week 26 and the week containing Eid-al-Fitr becomes week 25. This is because both the "Civil", "Solstices" and "Christmas" calendars follow the Gregorian Calendar with 52.177457 weeks per year, but the "Muslim Calendar" follows a lunar calendar with 29.53 days per month, leading to 354 or 355 days per year. Since the "Muslim Calendar" is consistently shorter than the solar year, it shifts with respect to the Gregorian calendar, necessitating the removal of these extra weeks as they contained no major event or holiday. Thus, Christmas was specified as week 26 in a 52 week calendar (starting from week 1), and Eid-al-Fitr as week 25 in a 50 week calendar. Occasional exception weeks were dropped from analysis if they did not fit into these calendars, without greatly altering the analysis; see Supplementary Tables S4-6 for the complete list. Supplementary Figure S5 shows the plot of all countries, centered around the weeks that contain Christmas, Eid-al-Fitr or January 1st, averaged according to their cultural identification (see above).

3.5.5 Country Classification from sex-searches

If sex searches correspond to countries' self-reported religions or locations (as described in the Country Selection and Categorization section), we can use sex searches as a feature to classify countries. Here we describe the process by which sex searches were used to measure a country's response to events: Eid al-Fitr, Christmas, the December Solstice, and the June Solstice. These responses were used to evaluate sex searches as a feature in a classification task. The centered time series described before were calculated for all countries in Supplementary Table S2. For each country we obtained between 4 and 9 yearly time series for all years for which data is available. These yearly time-series were averaged in five different ways per country: one following the civil Gregorian calendar, one centered on Christmas week, one centered on Eid- al-Fitr week, one centered on June 21st, representing the June solstice, and the last centered on December 21st, representing the December solstice. Average yearly time-series were created by first normalizing the data by year, such that the highest valued week each year was given a value of 1, and other weeks were expressed as a proportion of that maximum, in order to correct for bias towards years with more searches. To identify weeks with peak sex-search behavior, z-scores for each of these averaged time series were calculated as $z = (x - \mu)/\sigma$ where μ is the mean and σ is the (population) standard deviation

We then pursued a simple classification of countries according to their behavior on the Christmas and Eid-al-Fitr weeks. When the averaged Christmas-centered (Eid-al-Fitr-centered) time-series for a country yields z ¿ 1 on the Christmas (Eid-al-Fitr) week, the country was classified as a Christian Country (Muslim Country). If z ¡ 1 for both the Christmas- and Eid-al-Fitr-centered time-series, then such a country is classified as Other. If z ¿ 1 for both Christmas- and Eid-al-Fitr-centered time-series, the country was

culturally associated with largest z. Results can be seen in Supplementary Table S7. A similar procedure was followed to compare countries according to geographical location. See also Supplementary Methods S1.

3.5.6 Birth Data

There are biases and problems with birth data. This data is particularly uncommon in Muslim and Southern countries and is further confused in Muslim countries both by the fact that religious events do not follow the solar calendar and that registration dates do not accurately match actual birth dates (see Supplementary Materials Fig. 6). Nevertheless, if online sex-searches correspond to an actual increase in sexual activity, it should be possible to see an increase in births for countries where good records exist.

Monthly birth rates were collected from the United Nations Database [129] [1], See Supplementary Table S9 for data.

For each country, each month was divided by the number of days in the month (February months were divided by 28.25), then each year was normalized to its maximum value. This removes any bias towards years with more births.

To compare monthly birth rates with GT results we were restricted by the time range constraints of both data sets. We only have GT results from 2004 onwards and we rarely have birth data beyond 2012. In Supplementary Table S9 shows the availability of birth data for all countries used in this study.

There is also no increase in sex-searches or September births in Northern countries that do not celebrate Christmas on December 25th (Supplementary Figures S7). In

[1] except for South Africa,
retrieved from http://www.statssa.gov.za/publications/P0305/P03052012.pdf

addition, there is independent evidence that, even within the same country, religiously distinct populations–such as the Muslim and Jewish populations of Israel–have different conception patterns that correlate with their religious holidays [130].

3.5.7 World Map

Countries were color coded according to the z-scores presented in Supplementary Table S7. The World Map was built using the online tool: https://mapchart.net/

3.5.8 ANEW

The sentiment in tweets was quantified according to the Affective Norms for English Words (ANEW) lexicon [6, 131]. The ANEW assigns a number between 1 and 9 along three dimensions to 1034 words. These dimensions are arousal (a), dominance (d), and valence (v). The scores were determined though a survey as the mean score participants assigned each word. The valence scores correspond to whether (from 1 to 9) the word made participants feel sad to happy, arousal from calm to excited, and dominance from controlled to in-control For example, the word "laughter" has a valence score of 8.5, while "leprosy" has a score of 2.1. A basic translation to Spanish and Portuguese was performed through Google Translate and refined by speakers.

3.5.9 Twitter Data

The source of the twitter data used comes from IU's twitter garden hose feed, a 10% sample of all tweets. Geo-location data in combination with shape objects [132] allowed the country from which a tweet came to be determined for many tweets. We focus on tweets collected between September 2010, when the collection stabilized, and February

2014, when the tweet collection dropped, complicating homogeneous analysis of the data. We analyzed seven countries that yielded a sufficiently large number of tweets per week (about ten thousand): Argentina, Australia, Brazil, Chile, Indonesia, Turkey, and the USA. This includes countries in both hemispheres, both culturally Christian and Muslim, and with both English and Other official languages. Individual country's tweets are only examined after their collection had stabilized, starting in September 2010 for the US, Australia, and Chile; May 2011 for Indonesia, Brazil and; June 2011 for Argentina, and September 2011 for Turkey. Days were defined according to Greenwich Mean Time, and weeks from Sunday to midnight Saturday. The overall number of weekly collected tweets are shown in Supplementary Fig. S8, ranging from nearly a million scored tweets per week from the USA and Brazil, to only about ten thousand scored tweets from Turkey and Australia. The proportion of scored tweets to all collected tweets is usually quite small, usually below 5%.

An individual tweet's sentiment score was determined by finding all words within the tweet that matched the ANEW lexicon, and taking the average of their scores in each dimension. In the case that multiple languages were matched, the scores from the language with the most matched words were used. In case of a tie, the average scores over the tying languages were calculated. To better find the actual sentiment during the holidays without generic seasonal greetings, we don't score words if they appear in generic holiday greetings, such as "happy holidays", and we remove the ANEW words Christmas and Valentine from the lexicon entirely. The list of holidays whose greetings we removed were collected from http://www.officeholidays.com/. The complete list of phrases we removed from score calculation is included in Supplementary Table S13.

3.5.10 Mean Sentiment Correlations with Sex-Search Volume

To see if sentiment in tweets correlates with sex search volume we computed the ordinary least squares estimate of a multiple linear regression for each country, using the time series of mean tweet sentiment each week along the three ANEW dimensions as independent variables, with the weekly volume of sex searches as the dependent variable. To compute the weekly mean sentiment time series for ANEW dimension, we first calculated the mean tweet sentiment score for each day and then calculated the mean sentiment of the week such that each day has an equal weight in the weekly average.

3.5.11 Singular Value Decomposition for Eigenmood Analysis

Aggregating all sentiment in tweets into a mean value discards information in the distribution of sentiment across tweets. Therefore, we use binned distributions of sentiment across tweets in the following analysis. We focus on a 25-binned distribution of tweet sentiment between the lowest and highest possible ANEW score as a moderately-grained distribution, with fine enough resolution to capture some detailed structure while aggregating an adequate number of tweets per bin, 400 on average for a collection of 104 tweets.

We applied a singular value decomposition (SVD) (20) to the binned distribution of ANEW scores over time. Our matrix M has columns representing bins, and rows representing weeks. The left and right singular vectors then have an interpretation as the "eigenbins" and "eigenweeks" respectively. We will also refer to the singular vectors as components. The first component explains the vast majority of the variance, and is similar to the base distribution of the language, as expected from the Brown corpus, shown in (19). The second component explains a trend over time, while further

components correspond to other fluctuations, including yearly variations for holidays. For more information see also Methods 3.5.16.

3.5.12 Data Reconstruction

To analyze how sentiment varies, rather than its basic distribution in language use, we reconstructed the original data without the first component. After recalculating the relative variances, we can remove noise by also removing the components explaining the least variance. Reconstruction, then includes only those components that explain 95% of the remaining variance after the first component is removed. This leaves cyclic patterns and outlier weeks deviating strongly from the baseline sentiment distribution, which we visualize as a heatmap of the distribution over time in. We average over all full years in the data for multiple countries, centered on the week of a strong cultural holiday, to emphasize the change in these distributions, as shown in Supplementary Fig. 14. For more information see also Methods 3.5.17.

3.5.13 Eigenmood Selection

To investigate the distribution of sentiment in a country during a holiday, we selected an *eigenmood* composed of the two components that best characterized the mood distribution on the holiday. Supplementary Figure S15 and Methods 3.5.18. These two components were selected to describe a country's twitter sentiment on a holiday in the following way. First, the average projection of the holiday was found over all years of the data, as well as the standard deviation. The two eigenweeks with the highest absolute value of the holiday's projection minus its standard deviation were selected. The standard deviation is calculated over very few points, but subtracting it from the mean

allows us to know how small the magnitude of the projected vector we may expect. This way, the mood of the holiday of interest can be expected to have a strong correlation with the selected components and cluster closely together.

3.5.14 Notes on "misclassifications" for Country Classification from sex-searches

Some of the countries identified as Christian celebrate the nativity according to Julian calendar, with Christmas falling on January 7th or January 14th of the Gregorian calendar. Such is the case of the Christian countries: Belarus, Bosnia and Herzegovina, Georgia, Macedonia, Moldova, Montenegro, Serbia, Slovenia, Russia and Ukraine. Neither of these countries has a national holiday on December 25th nor shows an increase in sex-searches around December 25th. Had these countries been labeled as "Other", the percentage of countries identified as Christian for which we see a significant increase (z-score¿1) in sex-searches would have been of 91%. In addition to not celebrating the Christmas on December 25th, some of these countries also have a sizeable percentage of population that self-identifies as Muslim. Such is the case of Montenegro (29%), Macedonia (39%) and Bosnia and Herzegovina (45%). From the 30 Muslim countries, Pakistan was classified as Christian and 6 other countries didn't make the threshold. Pakistan is highly related to Christmas, probably due to the fact that there is a public holiday on 25th December, which coincidentally celebrates the birthday of Muhammad Ali Jinnah, founder of Pakistan. The other six countries also correspond to the ones for which the quality of the sex-search data was the poorest. Keeping in mind that we were looking for countries that culturally relate to a Christian or Muslim religious background, all countries that didn't make the threshold to be labelled as either are classified as Other. Unsurprisingly, there are many countries who are originally labelled

as Other and end up classified as either Christian or Muslim. European countries, such as the Czech Republic, Estonia and the Netherlands, whose majority does not identify as religious are classified as Christian, most likely due to the fact that these populations celebrate the holiday as well, even if secular.

3.5.15 Mean Sentiment Correlations with Sex-Search Volume

As shown in Supplementary Table S9A, there is a highly significant, moderate fit ($R^2 > 0.1$) across all countries, demonstrating a significant correlation between volume of sex-searches and mean sentiment as measured by the three ANEW dimensions. The coefficient of determination is generally stronger for Christian countries than Muslim Countries. Similarly to the GT data, the multiple linear regression models can be improved by averaging sentiment and sex-search volume across years using the 52-week Christmas centered calendar for the USA, Australia, Brazil, Argentina, and Chile, , and the 50-week Eid-al-Fitr centered calendar for Indonesia and Turkey. This smooths out extraordinary events that are picked up by sentiment analysis. The results of this centered-data regression are presented in Supplementary Table S9B. The fit is highly significant for all countries, and improves for all countries, ($R^2 > 0.26$). In every case, valence is yields a positive coefficient, while dominance a negative coefficient; so the happier but less dominant the sentiment expressed by a country, the more sex-searches tend to increase. As far as significance is concerned, t-tests reveal that the valence dimension is most often significant, followed by dominance, with arousal the least likely to be a significant factor. Interestingly, as shown in Supplementary Table S10, when we computed the ordinary least squares estimate of a standard linear regression on each ANEW dimension independently, we obtained very poor (but significant) goodness of fit, as measured by R^2. Therefore, the mean value of each ANEW dimension on its own

is a poor predictor of sex-search volume in all countries (with few exceptions such as Arousal in Brazil). We can thus say that mean sentiment correlates with sex-search volume (Supplementary Table S9) but the timeseries of mean weekly values of each ANEW dimension do not yield a nuanced characterization of sentiment correlated with interest in sex.

3.5.16 Singular Value Decomposition

Singular value decomposition (SVD) is a method by which a matrix can be linearly decomposed into ordered orthonormal components, each explaining as much of the linear variation as possible, after the components that came before it. The SVD of any $m \times n$ matrix M of real or complex numbers can represented as follows in Equation 3.1:

$$M = USV^T \tag{3.1}$$

Where U is an $m \times n$ matrix with orthonormal columns, V is an $n \times n$ matrix with orthonormal columns, and S is an $n \times n$ diagonal matrix. The columns of U and V are referred to as the left and right singular vectors of M respectively. These singular vectors are eigenvectors of the matrices MM^T and M^TM respectively. The diagonal entries of S, called the singular values of M, are the square roots of the eigenvalues of the matrices MM^T and M^TM. By convention, the singular values are ordered from greatest to least. The columns of U form a basis for the column space of M and the columns of V form a basis for the row space of M. The right singular vectors are also known in principal component analysis (PCA) as the loadings of the original variables (bins) onto the new coordinate system. The relative variance explained by each component can then be calculated for each component k as $s_k^2 / \sum_i s_i^2$ where s_k is the kth diagonal

component of S. It is important to note that matrices can be reconstructed with a lower rank by setting elements of S to zero. Typically only the top l singular values are kept in order to reduce noise and create the closest rank-l approximation of the original matrix [63].

3.5.17 Data Reconstruction

It can be clearly seen from the data reconstruction averages in Extended Data Fig. 8 and Supplementary Fig. S6, that the distribution of sentiment shifts towards higher bins during holidays, represented by redder high bins and greener low bins on holidays. Christmas stands out in the USA (US), Australia (AU), and Brazil (BR). Eid-al-Fitr stands out in both Turkey (TR) and Indonesia (ID), and in Turkey the beginning of Ramadan is emphasized a few weeks before. The centering performed only looks at weeks within the surrounding cultural year, such that Christmas is week 26 of a 52 week year (starting with a first week 1), while Eid-al-Fitr is week 25 of a 50 week year. Other weeks are averaged in this range according to their displacement from the holiday week (e.g., a week two weeks before the Christmas week in 2012 is averaged with weeks two weeks before Christmas in all other years). This obscures the emphasis on holidays using another calendar, such that Indonesia also has a strong signal on Christmas, but these signals are averaged over multiple weeks when the calendars are misaligned. The heatmaps for all countries centered on all holidays are included in Supplementary Fig. S6.

3.5.18 Eigenmood Selection and Characterization

The mean value of a holiday's projection on various components for different countries are shown in Supplementary Figures S2 and S3 for Christmas and Eid-al-Fitr respectively, with the two components selected for each country highlighted in red. As described, since the first component corresponds to the basic distribution of sentiment in the language and overwhelms projections because of how much it explains, and the last few components are mostly noise, we only look at the components explaining 95% of the variance after the removal of the first. The second component usually describes a variation over the whole time series of out data, thus it tends to have a large standard deviation.

To better understand how the selected components describe the mood, we define an interpretable linguistic variable [132]. The linguistic variable can take five fuzzy values, "low", "medium-low", "medium", "medium-high", and "high" with membership functions defined over the 25 bins of the original twitter sentiment distribution. These membership functions are shown in Supplementary Fig. S4 and were chosen such that each original bin's membership in all values sums to one, and the area under each membership function is the same.

The response of the linguistic variable to the holiday in each selected eigenmood is shown in Supplementary Figure S5 for the selected relevant holiday for each country. These responses were calculated by reconstructing the distribution bins with only the eigenmood selected for the country and holiday, multiplying the reconstructed bin value by its memberships, and summing over all bins for each linguistic value. These responses can be interpreted as the change from the language's base sentiment distribution on the holiday contributed by the selected eigenmood. The response characterized

by the Christmas eigenmood in the USA is an increase in medium-high happiness, with decreases in other levels of happiness, low and medium happiness in particular. How mood changes on a major holiday varies between countries but generally we see that the selected eigenmood describes increases medium-high or high valence on the holidays, with decreases in low, medium-low, and medium valence, as well as lower or more moderate dominance and arousal. The behavior of the dominance mood dimension in the week of Eid-al-Fitr in Indonesia highlights the importance of the more nuanced mood measurement that eigenmoods afford. While the ANEW mean value measurement above suggested a dominance decrease towards a less "in-control" mood, what we have at Eid-al-Fitr is a shift away from the extremes to a collective mood state that is neither very "in-control" nor very "controlled" – coherent with a happier and calmer mood scenario typically found in these holidays for all countries. In other words, during most weeks of the year, there is increased bimodal dominance activity in higher and lower bins (simultaneously high "in-control" and "controlled", respectively), but in the week of Eid-al-Fitr, the dominance mood converges to a mid-level dominance (Fig. 4 column A, row 3, dominance panel).

3.5.19 Eigenmood correlations to sex-search volume in target holidays

As a measure of mood similarity between weeks in a space defined by a selected eigenmood, we use the dot product between their coordinates in this space [63]. This measure increases between weeks with similar (positive or negative) projections onto the eigenweeks forming the space, becomes negative with opposite projections, and decreases in magnitude with weeks that are not correlated with the eigenweeks and are thus projected near the origin. Due to the properties, it is important to select an eigenmood that strongly corresponds to a week or weeks of interest, by containing high-magnitude

values in the corresponding eigenbins. The similarity can then be expressed as $w \cdot c$ where w and c are weeks projected into the eigenmood, which is equivalently the vector of corresponding weighted eigenbin values. In comparison between weeks and a holiday averaged over years, these vectors are the element-wise averages of the week's projection coordinates over the years. We report results with these averages, but these results are robust to yearly, non-averaged data, as well as different selection criteria for the eigenmoods (for example, allowing a greater number of components). The projection spaces for each eigenmood are shown in Supplementary Fig. S7.

In general, weeks close in proximity in time will be more similar in eigenmood, but certain weeks, often other holidays, more distant in time can have a high similarity in eigenmood to the selected holiday. In the USA, for example, the weeks closest in eigenmood to Christmas are, in order, the week of New Year's Day, the other weeks of December, and the weeks following July 4th, Father's Day, and Memorial Day. National Day in Chile is similar in eigenmood and sex searches to Chile's Christmas. New Year's Day and Christmas in Indonesia are similar to Eid-al-Fitr's eigenmood and high sex searches. In Turkey, weeks in late June, early July, and the week following Eid-al-Fitr are the most similar in terms of eigenmood and sex search volume to Eid-al-Fitr.

To investigate the relationship between a week's similarity in eigenmood to a holiday and the number of sex searches, we perform an ordinary least squares regression between sex searches as the dependent variable, and similarity as the independent variable. Displayed in Figure 4 and reported in Extended Data Table 2 are the results of this regression as well as Brownian distance correlation statistics, a nonlinear measure of correlation [133]. The plots of all linear regressions are included in Supplementary Fig. S7.

There is a fairly strong correspondence ($R^2 \geq .380$) between similarity in eigenmood to Christmas and sex searches in the C countries: the US, Brazil, Australia, Argentina, and Chile. The southern hemisphere Christian countries Brazil, Argentina, and Chile also have a noticeable correlation with Eid-al-Fitr, however, the slope of the regression is negative, implying that the less like the mood during the winter week of Eid-al-Fitr, the more sex searches are conducted.

In Muslim countries Turkey and Indonesia, we were limited by having less Twitter data and fewer tweets that match. However, there are significant correlations between similarity to Eid-al-Fitr and increased sex searches. The linear correlation is reduced compared to Christmas in Christian countries, since over time the weeks of Ramadan become more similar in eigenmood to Eid-al-Fitr, the festival at Ramadan's conclusion, while the cultural pressure is one of abstinence, such that these weeks have unusually low sex searches. In the case of Turkey in particular, the holiday of Eid-al-Adha, or the Sacrifice Feast, also has high sex searches, but is different in eigenmood from Eid-al-Fitr. The positive correlation between sex searches and Christmas eigenmood in Indonesia is likely caused by the sizable Christian population living there and effects due to summer.

Turkey is an interesting case, since it has a very strong negative correlation between sex searches and similarity to Christmas although the response to Eid-al-Fitr is smaller. In part, this may be due to limitations in our data gathering and method application, since our ANEW is only available in English, Spanish, and Portuguese. However, we still have a good number of tweets from Turkey, so we look more closely at its eigenmood. The projection of all weeks into its eigenmoods for Christmas and Eid-al-Fitr is shown in Supplementary Fig. S7, which happen to be same in this case. The regressions between sex searches and the similarity of averaged weeks to Christmas and Eid-al-Fitr are shown in Supplementary Fig. S7. The mood associated with Eid is also associated

with Ramadan, which emphasizes abstinence. During the weeks of Ramadan, there are much fewer sex searches than usual, although the weeks are not too far different in mood. In addition, there is a separate holiday, Eid-al-Adha, that is associated with a second peak in sex searches, but with a different mood. Perhaps due to Turkey's small Christian population and winter timing, Christmas and weeks like it in eigenmood have low sex searches and averaging over years decreases the effects of holiday traditions (like Eid-al-Fitr) due to misaligned calendars.

3.5.20 Granger Causality

Toda and Yamamoto [134] noted and corrected an issue that can occur frequently with time series - if the time series are not stationary, the results of a Granger causality test are incorrect. Their correction is to add additional lags of the variables to the VAR model, equal to the largest order of integration of any variable, but to exclude them during the final statistical test. The basic steps are: find the order of integration m of each variable, determine the number of lags for the VAR model from AIC, add lags to correct for any auto-correlation, and run a wald test on the model minus the last m lags to correct for non-stationarity [135][136].

Chapter 4

Small cohort of patients with epilepsy showed increased activity on *Facebook* before sudden unexpected death

This Chapter is a reproduction of a paper published in *Epilepsy and Behavior* [137] and includes writing from co-authors Rion Brattig Correia, Wendy R. Miller, and Luis M. Rocha.

4.1 Abstract

Sudden Unexpected Death in Epilepsy (SUDEP) remains a leading cause of death in people with epilepsy. Despite the constant risk for patients and bereavement to family members, to date the physiological mechanisms of SUDEP remain unknown. Here we

explore the potential to identify putative predictive signals of SUDEP from online digital behavioral data using text and sentiment analysis. Specifically, we analyze *Facebook* timelines of six epilepsy patients deceased due to SUDEP, donated by surviving family members. We find preliminary evidence for behavioral changes detectable by text and sentiment analysis tools. Namely, in the months preceding their SUDEP event patient social media timelines show: i) increase in verbosity; ii) increased use of functional words; and iii) sentiment shifts as measured by different sentiment analysis tools. Combined, these results suggest that social media engagement, as well as its sentiment, may serve as possible early-warning signals for SUDEP in people with epilepsy. While the small sample of patient timelines analyzed in this study prevents generalization, our preliminary investigation demonstrates the potential of social media data as complementary data in larger studies of SUDEP and epilepsy.

4.2 Introduction

Sudden Unexpected Death in Epilepsy (SUDEP) remains a leading cause of death for people with epilepsy (PWE), and includes all epilepsy-related deaths not due to trauma, drowning, status epilepticus, or other identifiable causes. The incidence of SUDEP is about 0.35 cases per 1,000 person-years [138]. While research into the physiological mechanisms underlying SUDEP continue to be thoroughly studied, and new SUDEP-related guidelines for clinicians treating PWE have been published in order to minimize SUDEP risk, SUDEP incidence remains steady [87, 139]. To date, the most espoused preventive strategy for SUDEP remains seizure control via appropriate self-management [140], and especially medication adherence, since a clear risk factor for SUDEP is a higher frequency of seizures [104]. While these risk factors have been disseminated

broadly, including to the public, SUDEP remains a leading cause of death for PWE, leading organizations such as The Institute of Medicine, American Epilepsy Society, and Epilepsy Foundation to call for increased study into SUDEP.

Apart from research related to the ways in which providers, patients, and their families discuss SUDEP [87, 141], very little behavioral research has been conducted to reveal potential behavioral or social attributes that may precede SUDEP. Should such specific attributes exist, they would provide an area of preventive intervention for SUDEP. In this study, we utilize digital behavioral data and investigate its potential for uncovering behavioral signatures preceding SUDEP that could be leveraged as early-warning signals to inform self-management interventions in PWE. As patients are known to not fully recall important events or even display recognizable behavior change during clinical consultations, digital behavioral data, such as social media data, can offer a complementary view of patient behavior of clinical significance [94]. Specifically, we use text and sentiment analysis to evaluate temporal changes in emotional states and communication patterns of the subjects in the study. The methodology gives us the unique opportunity to examine longitudinally the emotional states of a cohort of PWE with a known outcome of SUDEP. Our preliminary results show that social media may reveal behavioral experiences leading up to SUDEP, and thus guide areas for SUDEP-preventing interventions. This study also demonstrates the successful use of alternative, real-world data sources in studying SUDEP [94, 142].

Psychological stress is known to increase the risk of certain diseases, like the common cold [143]. Directly related to PWE, stress and major life events are known to increase the risk of seizures, which in turn can increase the risk of SUDEP [144, 145]. However, direct physiological measurements of stress involves expensive and invasive tools. A

compelling alternative is to measure stress and other cognitive states indirectly in self-reported digital behavioral data, such as in social media posting on *Facebook*. This is one of the focuses of the interdisciplinary field of *affective computing*, which has developed methods to measure human emotion (including stress) via linguistic and other computer-based features, such as keystroke dynamics [146]. For instance, Pennebaker [147] found a correspondence between textual features and physiological signals of stress. Similarly, Vizer, Zhou, & Sears [148] found that increased lexical complexity (diversity of words) tends to correspond with increased physical or cognitive stress. However, such studies are often conducted in controlled laboratory conditions, asking participants to write essays with particular prompts. This is not the case with social media, where users write posts spontaneously without being prompted in laboratory settings. Our assumption is that stress and other mood states influence whether and how a social media post is written, and can thus be measured via textual analysis of those posts. A substantial body of literature already reports that social media data enables quantitative measurement and prediction of various behavioral processes of biomedical relevance, i.e. a real-world data source to study "humans as their own model organism" [94]. Indeed, social media data has already been shown to be useful, alone or in combination with other data sources, for a variety of other biomedical problems. For instance, data from Twitter and Instagram helps in the detection of health conditions including the spread of flu pandemics [149], warning signals of drug adverse reactions [150], human reproduction [109], and even depression [151]. Social media users who self-reported their diagnosis of depression have been shown to exhibit distorted modes of thinking (cognitive distortions) in their writing, an early warning that can lower the burden of this underdiagnosed condition and leading cause of disability worldwide [92]. A long list of successful applications using social media data for biomedical and health-related problems is discussed in our recent

review [94].

To infer relevant cognitive states in our cohort of deceased SUDEP subjects we use textual and sentiment analysis of their social media posts. These methods were originally developed to determine the positive or negative feelings expressed in natural language texts towards specific product ratings, often used for marketing purposes [152, 153]. However, a number of sentiment analysis tools have been developed from psychological experiments, and can be used to model the emotional states of authors based on their written text [94]. In fact, sentiment analysis has been very useful to track various individual and cohort specific behaviors of relevance to biomedicine, especially mental health [92, 94, 109]. Similarly to other domains, these computational methods are likely to be useful to characterize the behavior of SUDEP cohorts, including any possible stress markers hidden in their social media discourse that can be leveraged to inform interventions aimed at improving self-management, a key predictor of epilepsy-related outcomes. Next, we detail the data gathering, textual methods, and three different sentiment analysis tools we apply to our SUDEP cohort.

4.3 Materials and methods

We began by eliciting families from which a member was known to have died of SUDEP. To do so, we advertised our research goals on the bulletin boards of the Epilepsy Foundation website and epilepsy-related *Facebook* groups. We also distributed information about our study to the Epilepsy Foundation's SUDEP Institute, which passed on the information to members of SUDEP bereavement groups within the Institute. The Epilepsy

Foundation website is one of the most popular sites for people with epilepsy. All procedures were evaluated by the Indiana University Institutional Review Board, who ultimately deemed that the study was exempt/not human subjects research. Family members self-referred to our study via email, and were given information about the study, its goals, and were also informed that participation was voluntary. We received about 20 inquiries from families who wanted to donate social media content from their deceased family members. From these, a majority of users had *Facebook* accounts, and only a few had *Twitter* or *Instagram* accounts. Due to data availability we decided to focus our analysis solely on *Facebook* timelines. This yielded a small cohort of $n = 12$ *Facebook* timelines (four males and eight females) from which we had timelines to collect data from. For six subjects we obtained full login information, and for the remaining we had varying viewing access to timeline posts, as listed in Table 4.1.

Data collection for subjects with login information was conducted through an in-house developed application using *Facebook*'s official application programming interface (API). Family members logged into the deceased *Facebook* account and accessed a specific app webpage. The app then collected all of the subject's timeline posts, including text, meta-data (e.g., date, posting device, etc), and the number of likes, comments and shares. Similarly, when only viewing access to the subject's timeline was available, family members (or a researcher when family was unable/unavailable) were instructed to scroll the deceased timeline, thus loading all posts, and export the subject's timeline content as an HTML file. A script developed in-house was used to process the HTML file, collecting text, available meta-data, and number of likes, comments, and shares from posts. Importantly, unlike the app-collected timelines that made use of subject's login information, timelines collected via the HTML-scraping script may not contain all subject posts, as privacy settings putatively put in place by the subject may have blocked

the person collecting the data from viewing them in the first place. In addition, in 2009 *Facebook* made a significant change to their interface: the prompt to the post box changed from "Update your Status", followed by "<Subject name> is..." to "What's on your mind?". Naturally, we believe this interface change may elicit a different response from the user. To avoid any possible interface bias in our analysis, we only consider subject posts that occurred after 2009, when the change took place. All collected data were securely stored within our servers for further analysis. For each subject Table 4.1 lists basic demographic, subject posting time range, and any notable life event discussed by the subject on their *Facebook* timeline in the month preceding their SUDEP, which was manually annotated by the researchers.

The number of posts collected for each subject varies widely, from only 4 posts written by Subject 12, all the way to 2,271 posts written by Subject 2 (see Table 4.1). The average number of posts per subject is 726. In total, we collected and processed 8,717 posts with text that were written after 2009, when considering all 12 subjects. However, because some subjects had very little number of posts—as is the case of Subject 12—we opted to limit our analysis to subjects with more than 500 posts that contained text and were written after 2009. In other words, next we only present results on subjects 1-3, 6, 8, and 10, a cohort of $n = 6$ subjects. These subjects are highlighted in Table 4.1.

Textual content of individual posts were processed using the dictionaries of three sentiment analysis tools: *Affective Norms for English Words* (ANEW) [154], *Valence Aware Dictionary for sEntiment Reasoning* (VADER) [155], and *Linguistic Inquiry and Word Count* (LIWC) [156]. These three tools are widely used in the sentiment analysis literature. In fact, VADER and LIWC were consistently among the best tools for 3-class polarity classification (negative, neutral, or positive emotion) across a number of corpora in a benchmark comparison study [23].

Subj.	Collection	Sex	Age	Posts	Window of posts*	Notable life event before SUDEP
1	App	F	23	1,410	2,526	New apartment, job, and city
2	App	M	20	2,271	2,157	Releasing DVD copies of new movie
3	App	F	18	844	2,071	Lonely as new college freshman
4	App	F	24	273	1,865	Graduating a Master's program
5	App	M	14	51	911	Birthday
6	App	F	15	473	843	Return from Europe trip
7	FoF	F	29	62	2,334	n/a
8	Public	F	n/a	2,201	2,315	n/a
9	Public	F	n/a	10	52	Party and writing paper
10	Friend	M	24	984	2,373	Recent concussion and recovery
11	FoF	M	28	134	1,524	Hospitalization
12	Friend	F	16	4	413	Braces Removed

TABLE 4.1: Demographics and data collection details for study subjects. Six subject timeline posts were collected via a custom-built app accessed using subject's login and password information. Six subject timelines were collected via HTML scraping of pages as visible to the public, to *Facebook* friends, or to friends of friends (FoF), as noted. The number of posts column tallied only posts with written text after 2009 (due to a significant *Facebook* interface change). * Column "window of posts" denote the number of days between a subject's first and last post.

Dictionaries were used to match against single words in subject posts. Matched words were then scored over several sentiment and textual dimensions per post. For instance, ANEW includes ratings from 1 to 9 in a dictionary of 1,034 words along three dimensions: *valence*, from unhappy to happy; *arousal* from calm to excited; and *dominance* from controlled to in-control. These ratings were originally collected from surveys given to undergraduates in a psychology class using a 9-point Likert-like scale [154]. We used ANEW to find the mean sentiment along these three dimensions for each post by averaging the sentiments of each word, while neglecting words absent from the dictionary. VADER [155] is a tool for measuring the intensity of positive or negative affect through lexical scores modified by syntactical rules, and is readily available as part of the Natural Language Toolkit for python [157]. In addition to dictionary-based sentiment scores, VADER looks at nearby words and modifies sentiment scores based on 5 simple rules: the presence of exclamations, capitalization, adverbs, negations, and contrasting conjunctions. Using this tool, we computed normalized scores describing the intensity of positive, neutral, and negative emotion present in each subject post. LIWC (pronounced

Luke) is the third dictionary-based tool used. It was developed with a well-documented procedure of consistent categorization between a majority of human judges. The latest version of the software, LIWC2015, has dictionaries containing nearly 6,400 words and evaluates text across nearly 90 linguistic and sentiment variables, including summary variables, pronouns, articles, cognitive processes, time focus, personal concerns, and informal language categories [156, 158].

4.4 Results

Assuming some type of stressor prior to SUDEP, which in turn could manifest as a change in the subject's digital verbosity, first we characterize the number of words per subject *Facebook* post (word count) with a simple negative binomial regression. The binomial regression tests whether there was a significant difference in the amount of words per post when comparing posts written in two different epochs of the subject's digital behavior. More specifically, we compare the average number of words per post in the two months (56 days) preceding the subject's SUDEP against the average number of words per post in the rest of the available timeline. We choose the last two months as a conservative time range for a subject behavioral change that at the same time holds enough examples (posts) for a robust statistical analysis—as a 10 samples minimum is a frequently recommended heuristic for an accurate estimation of model parameters [159]. However we note that posting behavior varies between subjects and we do not know whether, or when, stressors proceedings SUDEP may appear for each subject. We also tested different epochs, ranging from one to twelve weeks prior to SUDEP. Results are consistent for subjects with sufficient data in the last period being considered, and are shown in Fig. C.1. From our six analysed subjects, subjects 1, 2, 6, and 10 had

significantly higher word count in the two months preceding their SUDEP. Subject 8 also had a higher word count in the last two months, albeit not significant at $p < 0.05$. Conversely, subject 3 had a significantly lower word count in the last two months. Results are shown in Table 4.2 and Figure 4.1 shows the average word count for each subject timeline. Two regressions are fitted to the data highlighting the slope of the increase (or decrease) in subject verbosity: one considering the complete subject timeline (dotted line) and one only considering the last two months of posts (solid line).

subject	n_{early}	n_{last}	μ_{early}	μ_{last}	$time_p$
2	2,162	109	12.431	**34.413**	**1.197e-32**
1	1,547	54	9.592	**17.889**	**4.146e-06**
8	2,185	16	12.070	**18.375**	0.081
6	717	23	5.252	**7.304**	**0.021**
10	1147	7	13.983	**23.571**	**0.048**
3	834	10	**11.125**	4.100	**0.001**

TABLE 4.2: Significance tests for differences in word counts in posts during the last two months preceding SUDEP compared to other posts. The mean word count for the posts written during the last two months (μ_{last} with n_{last} samples) are compared to the mean word count of all other posts written by the subject before this period (μ_{early} with n_{early} samples). Significance is estimated from a negative binomial regression, with $p < 0.05$ highlighted in bold. Subjects are ordered according to the rank-product of the number of samples during the last month and the number prior to the last month.

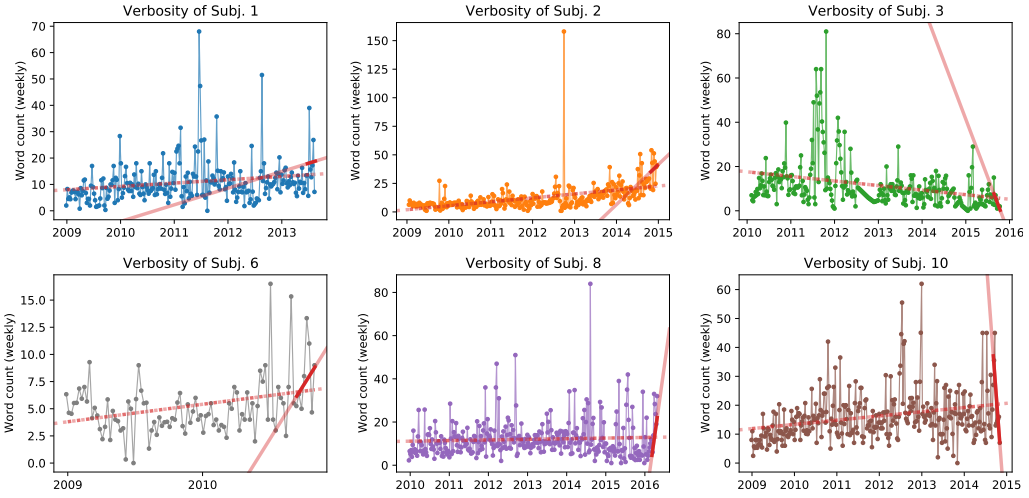


FIGURE 4.1: **Subject verbosity measured by word count**. Values are shown as weekly average to improve readability. Dashed red line shows the trend over the entire range of subject's posts. Solid red line is the trend over the last two months of data with darker color denoting the period length.

Since digital behavioral changes may be reflected not only in post length but also in

how frequent posts are made, next we use a zero-inflated negative binomial regression to examine whether the observed verbosity (word count) and frequency of posting prior to SUDEP was significantly different from subject's previous epochs. A zero-inflated negative binomial regression is an extension of the binomial regression where there is an assumption that a different process governs the likelihood that a subject makes no posts in a day (zero word count), which is then modeled by a logistic regression. Results are consisted and are presented in Table C.2; different epochs considered are shown in Figure C.2. In general we see that both subjects 1 and 2 were more likely to post in the two months preceding SUDEP, as well as writing longer posts. Perhaps due to increased model complexity, changes in subject 6's posting behavior are less significant, being less likely to post in weeks preceding SUDEP with little difference in the number of words written per day. Subject 8 and 10 were significantly less likely to post in the final weeks before SUDEP, with a non-significant increase in words per day when they did. Lastly, subject 3 did not have a significant change in the number of days with a post, but did write significantly fewer words.

Having analyzed subject verbosity, we now turn to the sentiment of the text they wrote. We remember each sentiment dimension is calculated by averaging per-word sentiment scores calculated for ANEW, LIWC, and VADER, three independent sentiment tools. In the following Figures 4.2-4.4, line plots denote the average of a specific sentiment dimension measure over all posts each week. Some particular sentiment trends can be observed in these figures. For instance, four of the six subject show an over time increase in happiness sentiment, as measured by ANEW's valence dimension (see dotted lines in Fig. 4.2). Only two subjects, 3 and 10, show a decrease in happiness in the last two months (solid line). Importantly, Subject 3 has an overall happiness increase but the a sharp sentiment shift in the last two months, reflected by her described feelings of

loneliness of being a college freshmen. On the other hand, subject 6 has an over time happiness decrease, but a sharp happiness increase in the last two months, reflecting a sentiment shift due to her European travels. Overall, despite some subjects having reversed valence sentiment, when their complete timeline sentiment is compared to the sentiment in the last two months of posting, they all have something in common: a significant sentiment shift, as measured by the difference in slope of the two regressions.

To show this phenomena is not simply an effect of the sentiment tool of choice, Figures 4.3 & 4.4 show subject use of emotion-neutral words and functional words, measured by VADER and LIWC, respectively. Functional words includes a broad category of words such as pronouns ('him', 'she'), articles ('the', 'a'), conjunctions ('and', 'but'), interjections ('oh', 'ah'), pro-sentences ('yes', 'no', 'okay'), and others. We observe an over time increase in the average number of such words used per post for 5 of the 6 subjects (see Fig. 4.4). In addition, for 4 subjects the amount of functional words used increases substantially in the last two months of posting. In regards to emotion-neutral words, five of the six subjects show an increase use of emotion-neutral words—a sentiment dimension that other tools, such as ANEW, ignores (see Fig. 4.3). However, similarly to subject verbosity, all subjects have a drastic shift in the analyzed sentiment categories when their complete timeline is compared to the last two months, again as measured by the regression slope (see red lines in aforementioned plots).

4.5 Discussion

First, we would like to emphasize that we cannot claim SUDEP causation, or the predictive accuracy of these tools applied to the social media posts of living individuals. However, the noticeable increase in functional words and the overall verbosity preceding

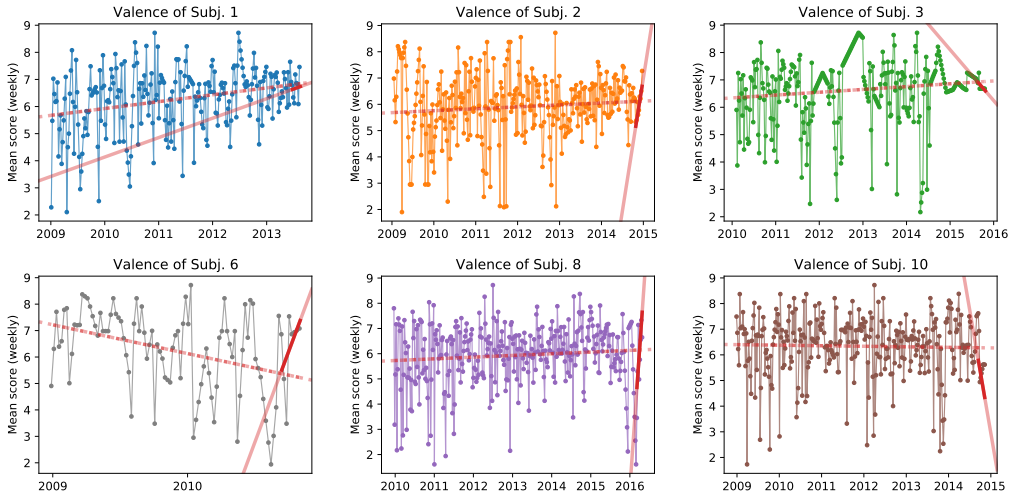


FIGURE 4.2: **Subject happiness measured by ANEW's Valence score**. Values are shown as weekly average to improve readability. Dashed red line shows the trend over the entire range of subject's posts. Solid red line is the trend over the last two months of data with darker color denoting the period length.

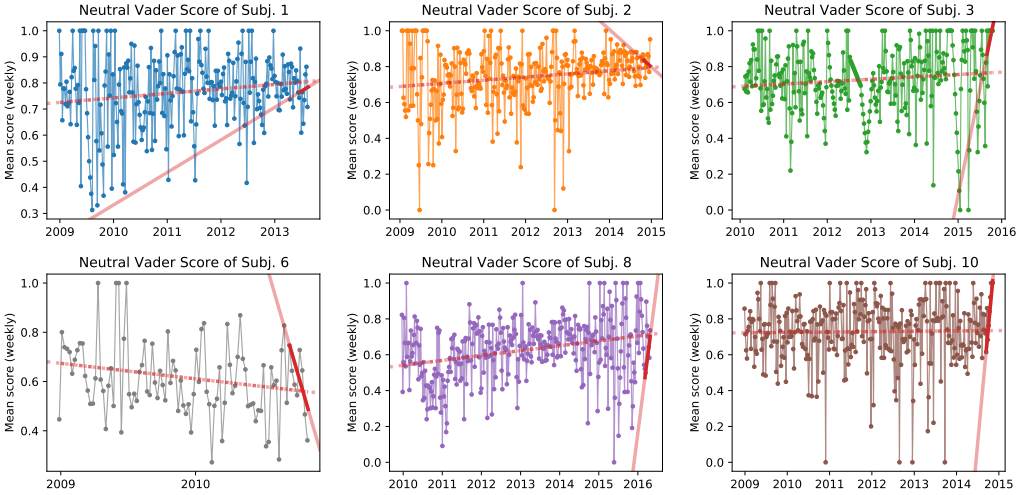


FIGURE 4.3: **Subject use of neutral words measured by VADER's Neutral score**. Values are shown as weekly average to improve readability. Dashed red line shows the trend over the entire range of subject's posts. Solid red line is the trend over the last two months of data with darker color denoting the period length.

SUDEP for a number of subjects is particularly suggestive of some detectable changes in the digital behavior of subjects, and that may serve as early-warning signals correlating with SUDEP. It is known that stress and major life events are likely to increase the risk of epilepsy [144, 145], and that in turn may increase the risk of SUDEP. Several of our subjects had major life changes in the weeks preceding their death, from concussions, moving to another city, returning from an overseas trip, or feeling lonely as a new college

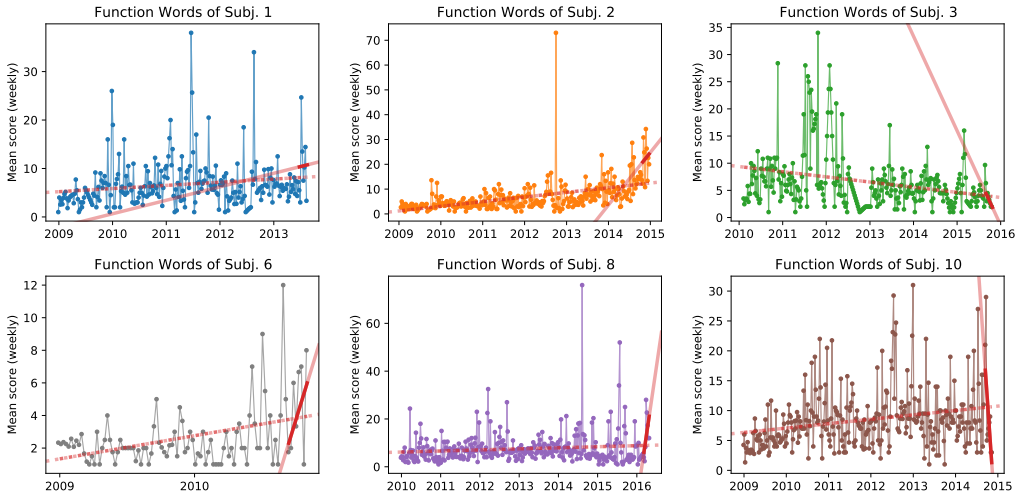


FIGURE 4.4: **Subject use of functional words measured by LIWC**. Values are shown as weekly average to improve readability. The dashed red line shows the trend over the entire range of a subject's posts, while the solid red line is the trend over the last two months of data.

freshmen. In addition, the misuse of functional words has been associated with Aphasia, a language impairment attributed to the Wernicke's area, a brain area in the left (dominant) temporal-parietal region characterized by EEG abnormalities in epilepsy patients [160–162]. Unlike impairment to Broca's area where patients speak slow, in hesitating ways, and phrases are devoid of functional words, impairment to the Wenicke's area cause patients to speak warmly and fluidly but using functional words with no content at all [160]. We manually checked sentence construction in the last two months of posting for our subjects and found no trace of functional words misuse aside from their increased occurrence. Nonetheless, if an increase in verbosity or changes in functional word use is indicative of stress or major life changes, the use of textual and sentiment tools may allow for a predictive, quantitative measure in larger studies, complementing current qualitative analyses. But we do stress that the lack of appropriate sample size and a rigorous case-control in our current study hinders generalization of our findings at this point. Nonetheless, our preliminary results serve to invite additional research into this problem, especially to encourage attention to social media and other digital

behavior data, thus contributing to better prediction of warning signals of SUDEP.

One possible avenue to evaluate the potential of sentiment analysis for predicting SUDEP is to employ statistical machine learning models using the text and sentiment analysis tools we described above. We attempted to build such models to predict changes in the last day or week of posts in a subject's timeline—instead of the last two months of posts shown in regressions above. However, we encountered two common machine-learning problems, especially in shorter window scenarios. The first was over-fitting and the subsequent false positive prediction. Since sentiment tools possess many sentiment variables (dimensions), it is easy to perfectly fit posts used in training the algorithm. Yet, the resulting prediction/classification models do not generalize to predicting subject posts left out for testing. Stricter model regularization and dimensionality reduction methods can help, but in the end, using shorter prediction windows results in a classification scenario with a very large class imbalance with very few positive instances (i.e., posts preceding SUDEP) which does not allow automatic machine learning classification. This is because most posts occur when subjects are deemed healthy, and only very few instances can be safely set as being SUDEP related—those that happened right before death. Given this problem of class imbalance, classifiers for automatic prediction are not possible with our current dataset.

The second problem pertains to the labeling of posts as SUDEP-relevant. Assuming that only the last posts before SUDEP are relevant, may miss prior days and posts (positive instances) that may have been close calls for SUDEP. Without the proper labeling of these instances, our algorithms are potentially missing several learning opportunities. The two-month window prior to SUDEP we used in the regression analysis is reasonable for the observed cohort, allowing a reasonable amount of positive posts for most subjects (see Table C.2). But the regression serves as an observation tool more

than an automatic predictor. Indeed, at the current stage, social media analysis can only enhance and provide a different perspective to other health data, such as electronic health records, personal diaries, epilepsy warning devices, service animals, etc. A more systemic and complete picture of SUDEP may emerge by combining these seemingly heterogenous data sources.

Going forward, our goal is to combine clinical (e.g., physician notes, laboratory exams, genetic profiling, questionnaire responses, electronic health records) with non-clinical digital behavioral data (e.g. electronic diaries, discussion boards, email exchange, phone usage patterns, social media posting and consumption) into research design. This is planned via recruitment of epilepsy patients who consent the to the collection of their digital behavioral data, such as social media IDs [94]. Our own work with focus groups of epilepsy patients and their caretakers have demonstrated willingness to donate digital behavioral data for studies. Indeed, as shown in the work we report here, this can be even done postmortem to avoid an observer bias—patients changing their behavior by knowing they are being observed. With enough subjects to account for the increase in variables, the next step is to validate the predictive power of social media signals in case-control experiments. We intend to focus on specific questions such as: why are subjects writing or using certain words more often prior to their death? Can this be statistically correlated with an increased risk of SUDEP? Can we pinpoint a behavioral phase shift to inform self- and caretaker-management as an early warning? The preliminary results we now report demonstrate the feasibility of extracting such signals. As we recruit additional subjects in planned larger studies, it will be possible to answer these questions more quantitatively and conclusively.

To compile additional digital behavioral data sources, our team is currently developing myAura [163], a personalized web service for epilepsy management. MyAura will

include self-reported patient diaries, such as seizure tracking, food and water intake, medication adherence, physician encounters, among others. One of its goals is to test a variety of clinical and non-clinical temporal variables that may be proven useful in epilepsy management. The use of patient donated social media timelines, as we have shown here, can prove to be the next frontier in informing our understanding of SUDEP and other epilepsy outcomes. MyAura will include the option for users to donate their social media timelines, thus allowing the recruitment of larger patient cohorts. Findings from analysis of the data of larger cohorts is likely to inform self-management recommendations for PWE, including allowing for SUDEP-predicting behaviors to be identified. For instance, patients with epilepsy could be monitored for an increased risk for SUDEP. In addition, our text and sentiment analysis could be used to inform individualized self-management interventions based on patient's posts and behaviors. At the same time behavioral results can help direct physiologic studies, as cellular-level or biomarker changes can, for example, ultimately be correlated with behavioral experiences (e.g. cortisol and physiologic or psychological stress).

As a small pilot, our study has demonstrated the feasibility of mining social media data for SUDEP (and other epilepsy-related) research, as well as very preliminary findings regarding increased social media activity preceding SUDEP. While the sample size of this study is too small to render generalizations in terms of SUDEP prediction, our work here demonstrates the feasibility of a novel way of investigating epilepsy-related phenomena, including SUDEP. This work also demonstrates the value in the interdisciplinary collaboration between clinical/behavioral epilepsy researchers and informatics/complex systems scientists.

Chapter 5

Conclusion

The sentiment expressed in our writing allows a window into our internal state, both at an individual level and collectively. Social media has allowed the mass expression of this sentiment, expressing otherwise unseen emotion and being in turn influenced by external forces. Collective mood requires tools for understanding the parts that compose it, while individual mood can reveal indications and warnings about our physical well-being.

I explored the first research question in Chapter 2: *Can meaningful components of collective mood states be extracted from time-series analysis of the sentiment of entire populations measured on social media?* I developed the *Eigenmood* methodology of decomposing a binned distribution of sentiment over time through a Singular Value Decomposition, and illustrated it through a toy example. Although mean sentiment is frequently used as a summary of collective mood, it is dominated by the frequency of words within natural language. Through an *Eigenmood* decomposition we can find a first singular vector that accounts for the majority of the variance in the data and corresponds to the frequency of sentiment in natural language, as verified against single-word sentiment frequencies in external corpora. The reconstruction of the original matrix without

this first component provides visualizations of underlying phenomena, like decreasing sentiment during the Covid-19 pandemic.

In Chapter 2 I also began the exploration of the second research question: *Are components of collective mood predictive of the future collective behavior of populations?* I also demonstrated that the inclusion of these mood components can aid in the modeling of other population-level phenomena, such as mortality during the Covid-19 pandemic. By using eigenmood components as exogenous variables in an ARIMA models for twenty populous cities we can usually improve model fit on in-sample data beyond mean sentiment, however this method of decomposition has difficulty extending into the future, finding that models with mean sentiment better generalize to out-of-sample data. In nearly every case, however, including information of collective sentiment improves the performance of the ARIMA model.

I continued the exploration of this question in Chapter 3. By taking advantage of Google sex searches as a proxy of human reproductive interest, we were able to investigate a long-standing debate between the biological and the cultural hypotheses of human reproduction. By centering searches on major cultural holidays, distinct patterns emerge dependent only on culture and not hemisphere or climate. These patterns find a peak in sex searches on the major cultural holiday across countries, corresponding to peaks in the best available birth data nine months later. The eigenmood methodology allowed us to further understand this behavior. By finding an eigenmood representation of each holiday for each country, we were able to show that throughout the year, the more similar the collective mood was to the holiday eigenmood, the more sex searches occur.

In Chapter 4, I investigated the final research question: *Can characteristic temporal*

patterns associated with individuals or small cohorts could be used to predict specific medical conditions? While previous chapters show how collective sentiment can be used to understanding and predict aggregate statistics, in this chapter I investigated how text and sentiment analysis of individual timelines can be used to understand individual health outcomes. By examining the Facebook timelines of a small cohort of individuals deceased due to SUDEP, we found preliminary evidence of changes in social media posting behavior in the months preceding SUDEP. In particular for 5 of the 6 most prolific posters, we found increases in verbosity during the two months preceding SUDEP and for the remaining subject found a significant drop. Previous studies found that stress leads to increased verbosity in laboratory settings; this increase in verbosity on social media then lends some quantitative credence to otherwise anecdotal evidence that SUDEP follows periods of significant stress. While we also found large changes in various sentiment measures preceding SUDEP, we did not find consistent changes across our subjects. The small sample size of our study prevents the drawing of broader conclusions, but we hope it can serve as a pilot study to build upon with the MyAura project. With a larger group of subjects, journalled data around seizures and corresponding social media timelines the team will be able to continue this research towards a better understanding of SUDEP.

The actions of large groups of people are often poorly understood and difficult to predict, with conflicting narratives of cause and effect proposed by competing interests. Collective action is often taken even when a rational basis for the collective good is difficult to find. Perhaps a less rational basis exists for collective action. Perhaps instead it is feeling that drives a collective forward, a collective mood state as an emergent property that drives group decisions. It is my hope that the study of such collective moods may allow us to better understand how the groups we are a part of will behave,

and allow us to better see how we may make collective decisions.

Bibliography

[1] James W Pennebaker, R Boyd, K Jordan, and K Blackburn. The development and psychometric properties of LIWC2015. (SEPTEMBER 2015):1–22, 2015. doi: 10.15781/T29G6Z. URL http://www.liwc.net/LIWC2007LanguageManual.pdf.

[2] Bing Liu. *Sentiment Analysis and Opinion Mining*, volume 5. Morgan & Claypool Publishers, 2012.

[3] Scott A. Golder and Michael W. Macy. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 333:1878–1881, 2011.

[4] Peter Sheridan Dodds and Christopher M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010. ISSN 13894978. doi: 10.1007/s10902-009-9150-9.

[5] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135, 2008. doi: 10.1561/1500000001.

[6] Margaret Mm Bradley and Pj Peter J Lang. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. *Psychology*, Technical(C-1): 0, 1999. ISSN 10897801. doi: 10.1109/MIC.2008.114. URL http://dionysus.

psych.wisc.edu/methods/Stim/ANEW/ANEW.pdf$\delimiter"026E30F$nhttp://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Affective+Norms+for+English+Words+(+ANEW+):+Instruction+Manual+and+Affective+Ratings{#}0$\delimiter"026E30F$nhttp://scholar.google.com/scholar?hl=en{&}btnG=S.

[7] Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112 (8):2389–2394, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1411678112. URL http://www.pnas.org/content/112/8/2389.abstract.

[8] Peter Sheridan Dodds, Kameroncker Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12), 2011. ISSN 19326203. doi: 10.1371/journal.pone.0026752.

[9] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Www2010*, pages 17–21, 2009. URL http://arxiv.org/abs/0911.1583.

[10] Douglas M McNair, Maurice Lorr, Leo F Droppleman, et al. *Profile of mood states*. Educational and Industrial Testing Service San Diego, CA, 1981.

[11] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3): 165–210, 2005. ISSN 1574020X. doi: 10.1007/s10579-005-7880-9.

[12] Rosalind W Picard. Affective Computing. *Pattern Recognition*, 73(321):304, 1997. ISSN 14337541. doi: 10.1007/BF01238028. URL http://vismod.media.mit.edu/tech-reports/TR-321.pdf.

[13] Philip J. Stone, Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as a Unit of Information. *Behavioral science*, 1966.

[14] Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. The Spanish adaptation of ANEW (affective norms for English words). *Behavior research methods*, 39(3):600–605, 2007. ISSN 1554-351X. doi: 10.3758/BF03193031.

[15] Ana Paula Soares, Montserrat Comesaña, Ana P Pinheiro, Alberto Simões, and Carla Sofia Frade. The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, 44(1):256–69, 2012. ISSN 15543528. doi: 10.3758/s13428-011-0131-7. URL http://www.ncbi.nlm.nih.gov/pubmed/21751068.

[16] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45 VN - r(4):1191–1207, 2013. ISSN 1554-3528. doi: 10.3758/s13428-012-0314-x. URL http://www.ncbi.nlm.nih.gov/pubmed/23404613$\delimiter"026E30F$nhttp://dx.doi.org/10.3758/s13428-012-0314-x.

[17] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011. ISSN 18777503. doi: 10.1016/j.jocs.2010.12.007. URL http://dx.doi.org/10.1016/j.jocs.2010.12.007.

[18] Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422, 2006. ISSN 09255273. doi: 10.1.1.61.7217. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.7217.

[19] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3 . 0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining SentiWordNet. *Analysis*, 0:1–12, 2010.

[20] CJ J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International AAAI Conference on Weblogs and ...*, pages 216–225, 2014. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109$\delimiter"026E30F$nhttp://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf.

[21] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder : A system for subjectivity analysis. *October*, (October):34–35, 2005. doi: 10.3115/1225733.1225751. URL http://portal.acm.org/citation.cfm?id=1225751{&}dl=GUIDE,$\delimiter"026E30F$nhttp://dl.acm.org/citation.cfm?id=1225751.

[22] Andrew Reagan, Brian Tivnan, Jake Ryland Williams, Christopher M. Danforth, and Peter Sheridan Dodds. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. page 34, 2015. URL http://arxiv.org/abs/1512.00531.

[23] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–23, 2016. doi: 10.1140/epjds/s13688-016-0085-1.

[24] How the general inquirer is used and a comparison of general inquirer with other text-analysis procedures. URL http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html.

[25] Y. R. Tausczik and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010. ISSN 0261-927X. doi: 10.1177/0261927X09351676.

[26] George A Miller. More Than 166,000 Word Form and Sense Pairs. 38(11):39–41, 1995.

[27] T Wilson, J Wiebe, and P Hoffman. Recognizing contextual polarity in phrase level sentiment analysis. *Acl*, 7(5):12–21, 2005. ISSN 0891-2017. doi: 10.3115/1220575.1220619.

[28] Karo Moilanen and Stephen Pulman. Sentiment Composition. *Proceedings of the Fourth International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, (M):378–382, 2007. ISSN 13138502. URL http://www.clg.ox.ac.uk/{_}media/people:karo:sentcompranlp07final.pdf.

[29] Hassan Saif, Yulan He, and Harith Alani. Semantic Sentiment Analysis of Twitter. *CEUR Workshop Proceedings*, 917:56–66, 2012. ISSN 16130073. doi: 10.1007/978-3-642-35176-1_32. URL http://dx.doi.org/10.1007/978-3-642-35176-1{_}32.

[30] Adnan Duric and Fei Song. Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems*, 53(4):704–711, 2012. ISSN 01679236. doi: 10.1016/j.dss.2012.05.023. URL http://dx.doi.org/10.1016/j.dss.2012.05.023.

[31] Taku Kudo and Yuji Matsumoto. A Boosting Algorithm for Classification of Semi-Structured Text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17–24, 2004.

[32] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002. ISSN 1554-0669. doi: 10.3115/1118693.1118704. URL http://portal.acm.org/citation.cfm?id=1118693.1118704.

[33] R. L. Robinson, R. Navea, and W. Ickes. Predicting Final Course Performance From Students' Written Self-Introductions: A LIWC Analysis. *Journal of Language and Social Psychology*, 32(4):469–479, 2013. ISSN 0261-927X. doi: 10.1177/0261927X13476869.

[34] D. Nadeau, C. Sabourin, J. De Koninck, S. Matwin, and P. Turney. Automatic Dream Sentiment Analysis. 2000. doi: 10.1023/B.

[35] John Pestian, John Pestian, Pawel Matykiewicz, Brett South, Ozlem Uzuner, and John Hurdle. Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, 5:3, 2012. ISSN 1178-2226. doi: 10.4137/BII.S9042.

[36] Mitja D Back, Albrecht C P Küfner, and Boris Egloff. The emotional timeline of September 11, 2001. *Psychological science : a journal of the American Psychological Society / APS*, 21(10):1417–1419, 2010. ISSN 0956-7976. doi: 10.1177/0956797610382124.

[37] M D Back, A C P Kufner, and B Egloff. "Automatic or the People?": Anger on September 11, 2001, and Lessons Learned for the Analysis of Large Digital Data Sets. *Psychological Science*, 22(6):837–838, 2011. ISSN 0956-7976. doi: 10.1177/0956797611409592.

[38] Onur Varol, Emilio Ferrara, Christine L Ogan, Filippo Menczer, and Alessandro Flammini. Evolution of online user behavior during a social upheaval. In *Proceedings of the 2014 ACM conference on Web science*, pages 81–90. ACM, 2014.

[39] Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. Online popularity and topical interests through the lens of instagram. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 24–34. ACM, 2014.

[40] Carleen Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2): 361–368, 2009.

[41] Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 468–479. World Scientific, 2016.

[42] Henry Kautz. Data mining social media for public health applications. In *23rd Int. Joint Conf. on Artificial Intelligence (IJCAI 2013),(AAAI Press, 2013)*, 2013.

[43] Emily K Seltzer, NS Jean, Emily Kramer-Golinkoff, David A Asch, and RM Merchant. The content of social media's shared images about ebola: a retrospective study. *Public health*, 129(9):1273–1277, 2015.

[44] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

[45] Adam Sadilek, Henry A Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *ICWSM*, pages 322–329, 2012.

[46] Emily H Chan, Vikram Sahai, Corrie Conrad, and John S Brownstein. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS neglected tropical diseases*, 5(5):e1206, 2011.

[47] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[48] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[49] Andrea López, Alissa Detz, Neda Ratanawongsa, and Urmimala Sarkar. What patients say about their doctors online: a qualitative content analysis. *Journal of general internal medicine*, 27(6):685–692, 2012.

[50] Jeffrey Segal, Michael Sacopulos, Virgil Sheets, Irish Thurston, Kendra Brooks, and Ryan Puccia. Online doctor reviews: do they track surgeon volume, a proxy for quality of care? *Journal of medical Internet research*, 14(2), 2012.

[51] Patricia A Cavazos-Rehg, Melissa Krauss, Sherri L Fisher, Patricia Salyer, Richard A Grucza, and Laura Jean Bierut. Twitter chatter about marijuana. *Journal of Adolescent Health*, 56(2):139–145, 2015.

[52] Leah Thompson, Frederick P Rivara, and Jennifer M Whitehill. Prevalence of marijuana-related traffic on twitter, 2012–2013: a content analysis. *Cyberpsychology, Behavior, and Social Networking*, 18(6):311–319, 2015.

[53] Marcel Salathé, Duy Q Vu, Shashank Khandelwal, and David R Hunter. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science*, 2(1):4, 2013.

[54] Marcel Salathé and Shashank Khandelwal. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology*, 7(10):e1002199, 2011.

[55] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7: 45141, 2017.

[56] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.

[57] Ingrid A van de Leemput, Marieke Wichers, Angélique OJ Cramer, Denny Borsboom, Francis Tuerlinckx, Peter Kuppens, Egbert H van Nes, Wolfgang Viechtbauer, Erik J Giltay, Steven H Aggen, et al. Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, 111(1):87–92, 2014.

[58] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[59] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[60] Andrea Lancichinetti, M Irmak Sirer, Jane X Wang, Daniel Acuna, Konrad Körding, and Luís A Nunes Amaral. High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, 5(1):011007, 2015.

[61] Michael W Berry, Susan T Dumais, and Gavin W O'Brien. Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595, 1995.

[62] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[63] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[64] Robert Cudeck and Robert C MacCallum. *Factor analysis at 100: Historical developments and future directions*. Routledge, 2012.

[65] Morrison Donald. *Multivariate statistical methods*. McGraw Hill, New York, 1990.

[66] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[67] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[68] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[69] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

[70] Steffen Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010.

[71] Steffen Rendle. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57, 2012.

[72] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[73] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

[74] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[75] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM, 2016.

[76] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4): 1253–1278, 2000.

[77] Richard A Harshman. Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis. 1970.

[78] Mónica Bécue-Bertaut and Jérôme Pagès. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis*, 52(6):3255–3268, 2008.

[79] Hervé Abdi, Lynne J Williams, and Domininique Valentin. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary reviews: computational statistics*, 5(2):149–179, 2013.

[80] John Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983.

[81] Peter Filzmoser, Karel Hron, and Clemens Reimann. Principal component analysis for compositional data with outliers. *Environmetrics*, 20(6):621–632, 2009.

[82] Eugene Kaciak and Sheahan Jerome N. Market segmentation: an alternative principal components approach. *Proceedings of the Annual Conference of the Administrative Sciences Association of Canada - Marketing Division*, 8:139–148, 1988.

[83] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

[84] Clayton A Davis, Giovanni Luca Ciampaglia, Luca Maria Aiello, Keychul Chung, Michael D Conover, Emilio Ferrara, Alessandro Flammini, Geoffrey C Fox, Xiaoming Gao, Bruno Gonçalves, et al. Osome: the iuni observatory on social media. *PeerJ Computer Science*, 2:e87, 2016.

[85] Division of Public Affairs Office of the Associate Director for Communication, Digital Media Branch. Weekly u.s. influenza surveillance report, 2018. URL https://www.cdc.gov/flu/weekly/.

[86] Wendy Macdowall, Kaye Wellings, Judith Stephenson, and Anna Glasier. Summer nights: A review of the evidence of seasonal variations in sexual health indicators among young people. *Health Education*, 108(1):40–53, 2007.

[87] Wendy R. Miller, Neicole Young, Daniel Friedman, Janice M. Buelow, and Orrin Devinsky. Discussing sudden unexpected death in epilepsy (sudep) with patients: Practices of health-care providers. *Epilepsy & Behavior*, 32(Supplement C):38–41, 2014. ISSN 1525-5050. doi: 10.1016/j.yebeh.2013.12.020.

[88] Marijn Ten Thij, Ian B. Wood, Luis M. Rocha, and Johan Bollen. Decomposition of online sentiment reveals societal eigenmoods. *In Preparation*, 2023.

[89] Yoshihiko Suhara, Yinzhan Xu, and Alex 'Sandy' Pentland. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, WWW

'17, pages 715–724, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4913-0. doi: 10.1145/3038912.3052676.

[90] Johan Bollen, Bruno Gonçalves, Guangchen Ruan, and Huina Mao. Happiness is assortative in online social networks. *Artificial life*, 17(3):237–251, 2011.

[91] Rui Fan, Onur Varol, Ali Varamesh, Alexander Barron, Ingrid A van de Leemput, Marten Scheffer, and Johan Bollen. The minute-scale dynamics of online emotions reveal the effects of affect labeling. *Nature Human Behaviour*, 3(1):92–100, 2019.

[92] Krishna C. Bathina, Marijn ten Thij, Lorenzo Lorenzo-Luaces, Lauren A. Rutter, and Johan Bollen. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, (5):458–466, 2021. doi: 10.1038/s41562-021-01050-7.

[93] Danny Valdez, Marijn ten Thij, Krishna Bathina, Lauren A Rutter, and Johan Bollen. Social media insights into us mental health during the covid-19 pandemic: Longitudinal analysis of twitter data. *J Med Internet Res*, 22(12):e21418, Dec 2020. ISSN 1438-8871. doi: 10.2196/21418. URL http://www.jmir.org/2020/12/e21418/.

[94] Rion Brattig Correia, Ian B Wood, Johan Bollen, and Luis M Rocha. Mining social media data for biomedical signals and health-related behavior. *Annual Review of Biomedical Data Science*, 3:433–458, 2020. doi: 10.1146/annurev-biodatasci-030320-040844.

[95] George K. Zipf. *Human behavior and the principle of least effort.* 1949.

[96] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*, 6(2):e19273, May 2020. ISSN 2369-2960. doi: 10.2196/19273. URL http://publichealth.jmir.org/2020/2/e19273/.

[97] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1961.10482090.

[98] Isabel M. Kloumann, Christopher M. Danforth, Kameron Decker Harris, Catherine A. Bliss, and Peter Sheridan Dodds. Positivity of the English language. *PLoS ONE*, 7(1):e29484–e29484, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0029484.

[99] David Garcia, Antonios Garas, and Frank Schweitzer. Positive words carry less information than negative words. *EPJ Data Science*, 1(1):3, 2012. ISSN 2193-1127. doi: 10.1140/epjds3.

[100] Geoff Cumming and S. Finch. Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist*, 60:170–180, 2005.

[101] Kokil Jaidka, Salvatore Giorgi, H. Andrew Schwartz, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences of the United States of America*, 117(19):10165–10171, may 2020. ISSN 10916490. doi: 10.1073/pnas.1906364117.

[102] W Nelson Francis and Henry Kucera. Brown corpus manual. *Letters to the Editor*, 5(2):7, 1979.

[103] Simon DeDeo, Robert XD Hawkins, Sara Klingenstein, and Tim Hitchcock. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276, 2013.

[104] National Center for Health Statistics. Weekly provisional counts of deaths by state and select causes, 2020-2022. https://data.cdc.gov/d/muzy-jte6, 2022. Accessed: 6-June-2022.

[105] George EP Box, Gwilym M Jenkins, and Gregory C Reinsel. *Time series analysis: forecasting and control*. John Wiley & Sons, fourth edition, 2008.

[106] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[107] Haneen Alabdulrazzaq, Mohammed N Alenezi, Yasmeen Rawajfih, Bareeq A Alghannam, Abeer A Al-Hassan, and Fawaz S Al-Anzi. On the accuracy of arima based prediction of covid-19 spread. *Results in Physics*, 27:104509, 2021.

[108] Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, and Massimo Ciccozzi. Application of the arima model on the covid-2019 epidemic dataset. *Data in brief*, 29:105340, 2020.

[109] Ian B Wood, Pedro L Varela, Johan Bollen, Luis M Rocha, and Joana Gonçalves-Sá. Human sexual cycles are driven by culture and match collective moods. *Scientific reports*, 7(1):17973, 2017.

[110] David R Cummings. Human birth seasonality and sunshine. *American Journal of Human Biology*, 22(3):316–324, 2010.

[111] Franklin H Bronson. Seasonal variation in human reproduction: environmental factors. *The Quarterly Review of Biology*, 70(2):141–164, 1995.

[112] Till Roenneberg and Jürgen Aschoff. Annual rhythm of human reproduction: Ii. environmental correlations. *Journal of Biological Rhythms*, 5(3):217–239, 1990.

[113] K Anand, G Kumar, S Kant, SK Kapoor, et al. Seasonality of births and possible factors influencing it in a rural area of haryana, india. *Indian pediatrics*, 37(3): 306–311, 2000.

[114] Ursula M Cowgill. Season of birth in man. contemporary situation with special reference to europe and the southern hemisphere. *Ecology*, 47(4):614–623, 1966.

[115] Kaye Wellings, W Macdowall, M Catchpole, and J Goodrich. Seasonal variations in sexual activity and their implications for sexual health promotion. *Journal of the Royal Society of Medicine*, 92(2):60–64, 1999.

[116] Completeness of birth registration. http://data.worldbank.org/indicator/SP.REG.BRTH.ZS, 2015. Accessed: 31-March-2015.

[117] Stacy Beck, Daniel Wojdyla, Lale Say, Ana Pilar Betran, Mario Merialdi, Jennifer Harris Requejo, Craig Rubens, Ramkumar Menon, and Paul FA Van Look. The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity. *Bulletin of the World Health Organization*, 88:31–38, 2010.

[118] Google trends. http://www.google.com/trends/, 2014. Accessed: 4-November-2014 for sex-related data, and 28-August-2017 for Figure S2.

[119] Patrick M Markey and Charlotte N Markey. Seasonal variation in internet keyword searches: a proxy assessment of sex mating behaviors. *Archives of sexual behavior*, 42(4):515–521, 2013.

[120] Martin L Levin, Xiaohe Xu, and John P Bartkowski. Seasonality of sexual debut. *Journal of Marriage and Family*, 64(4):871–884, 2002.

[121] Wikipedia, christianity by country — wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Christianity_by_country, 2014. Accessed: 4-November-2014.

[122] Wikipedia, islam by country — wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Islam_by_country, 2014. Accessed: 4-November-2014.

[123] Ian B Wood, Joana Gonçalves-Sá, Johan Bollen, and Luis M Rocha. Eigenmood twitter analysis: Measuring collective mood variation. In Preparation.

[124] Clive Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, 1969.

[125] Moniek M Ter Kuile, Stephanie Both, and Janneke Van Uden. The effects of experimentally-induced sad and happy mood on sexual arousal in sexually healthy women. *The journal of sexual medicine*, 7(3):1177–1184, 2010.

[126] Guy Bodenmann and Thomas Ledermann. Depressed mood and sexual functioning. *International Journal of Sexual Health*, 19(4):63–73, 2008.

[127] Stergios Moschos, Jean L Chan, and Christos S Mantzoros. Leptin and reproduction: a review. *Fertility and sterility*, 77(3):433–444, 2002.

[128] AA Ammar, F Sederholm, TR Saito, AJW Scheurink, AE Johnson, and P Sodersten. Npy-leptin: opposing effects on appetitive and consummatory ingestive behavior and sexual behavior. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 278(6):R1627–R1633, 2000.

[129] United nations, worldwide births. Online, 2014. Accessed: 4-November- 2014.

[130] Michael Friger, Ilana Shoham-Vardi, and Kathleen Abu-Saad. Trends and seasonality in birth frequency: a comparison of muslim and jewish populations in

southern israel: daily time series analysis of 200 009 births, 1988–2005. *Human reproduction*, 24(6):1492–1500, 2009.

[131] T. Lee. Generating the affective norms for english words (anew) dataset. https://tomlee.wtf/search/ANEW, 2010. Accessed: 18-June 2010.

[132] Made with natural earth. free vector and raster map data @ naturalearthdata.com. http://www.naturalearthdata.com, 2014.

[133] Gábor J Székely, Maria L Rizzo, et al. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

[134] Hiro Y Toda and Taku Yamamoto. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, 66(1):225–250, 1995.

[135] Dave Giles. Testing for granger causality, 2011. URL http://davegiles.blogspot.de/2011/04/testing-for-granger-causality.html.

[136] Christoph Pfeiffer. Toda-yamamoto implementation in r, 2012. URL https://christophpfeiffer.org/2012/11/07/toda-yamamoto-implementation-in-r/.

[137] Ian B Wood, Rion Brattig Correia, Wendy R Miller, and Luis M Rocha. Small cohort of patients with epilepsy showed increased activity on facebook before sudden unexpected death. *Epilepsy & Behavior*, 128:108580, 2022.

[138] Richard D Bagnall, Douglas E Crompton, and Christopher Semsarian. Genetic basis of sudden unexpected death in epilepsy. *Frontiers in neurology*, 8:348, 2017.

[139] Cynthia Harden, Torbjörn Tomson, David Gloss, Jeffrey Buchhalter, J Helen Cross, Elizabeth Donner, Jacqueline A French, Anthony Gil-Nagel, Dale C Hesdorffer, W Henry Smithson, et al. Practice guideline summary: sudden unexpected

death in epilepsy incidence rates and risk factors: report of the guideline development, dissemination, and implementation subcommittee of the american academy of neurology and the american epilepsy society. *Neurology*, 88(17):1674–1680, 2017.

[140] W Henry Smithson, Brigitte Colwell, and Jane Hanna. Sudden unexpected death in epilepsy: addressing the challenges. *Current neurology and neuroscience reports*, 14(12):502, 2014.

[141] Mark J Stevenson and Thomas F Stanton. Knowing the risk of sudep: two family's perspectives and the danny did foundation. *Epilepsia*, 55(10):1495–1500, 2014.

[142] Sreeram V. Ramagopalan, Alex Simpson, and Cormac Sammon. Can real-world data really replace randomised clinical trials? *BMC Medicine*, 18(1):13, 12 2020. doi: 10.1186/s12916-019-1481-8.

[143] Sheldon Cohen, David AJ Tyrrell, and Andrew P Smith. Psychological stress and susceptibility to the common cold. *New England journal of medicine*, 325(9): 606–612, 1991.

[144] H. McConnell, J. Valeriano, and J. Brillman. Prenuptial seizures: a report of five cases. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 7(1):72–75, 1995. doi: 10.1176/jnp.7.1.72.

[145] Heather R. McKee and Michael D. Privitera. Stress as a seizure precipitant: Identification, associated factors, and treatment options. *Seizure*, 44:21–26, 2017. doi: 10.1016/j.seizure.2016.12.009.

[146] Jing Zhai and Armando Barreto. Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In *Engineering in*

Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, pages 1355–1358. IEEE, 2006.

[147] James W Pennebaker. Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy*, 31(6):539–548, 1993.

[148] Lisa M Vizer, Lina Zhou, and Andrew Sears. Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies*, 67(10):870–886, 2009.

[149] Nicholas A. Christakis and James H. Fowler. Social network sensors for early detection of contagious outbreaks. *PLOS ONE*, 5:e12948, 2010. doi: 10.1371/journal.pone.0012948.

[150] Rion Brattig Correia, Lang Li, and Luis M. Rocha. Monitoring potential drug interactions and reactions via network analysis of instagram user timelines. In *Pacific Symposium on Biocomputing*, volume 21, pages 492–503. 2016.

[151] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proc. 5th Annual ACM Web Science Conf.*, WebSci'13, pages 47–56. ACM, 2013.

[152] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135, 2008.

[153] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[154] Margaret M. Bradley and Peter J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida., Gainesville, FL, 1999.

[155] C.J. Hutto and E.E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, pages 216–225, Ann Arbor, MI, June 2014.

[156] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[157] Edward Loper Bird, Steven and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[158] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Technical report, Austin, TX: University of Texas at Austin, 2015.

[159] Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[160] Cindy Chung and James Pennebaker. The psychological functions of funcion words. In Klaus Fiedler, editor, *Social Communication*, pages 343–359. Psychology Press, New York, 2007.

[161] Aninda B. Acharya and Michael Wroten. Wernicke aphasia. In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island, FL, 2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK441951/.

[162] Epilepsy Foundation of America. Types of language problems in epilepsy. Online, 2021. Accessed on Aug 18, 2021.

[163] Luis M. Rocha, Katy Börner, and Wendy R. Miller. myaura: personalized web service for epilepsy management. Available from https://hsrproject.nlm.nih.gov/view_hsrproj_record/20191123, 2019. Accessed Nov 29.

[164] Fred Morstatter, Jürgen Pfeffer, and Huan Liu. When is it biased?: assessing the representativeness of twitter's streaming api. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 555–556. International World Wide Web Conferences Steering Committee, 2014.

[165] Lotfi Asker Zadeh. The concept of a linguistic variable and its application to approximate reasoning—i. *Information sciences*, 8(3):199–249, 1975.

[166] Zero-inflated negative binomial regression r data analysis examples, Aug 2018. URL https://stats.idre.ucla.edu/r/dae/zinb/.

[167] Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in R. *Journal of Statistical Software*, 27(8), 2008. URL http://www.jstatsoft.org/v27/i08/.

Appendix A

Chapter 2 Appendix

A.1 Eigenmood of ANEW

To demonstrate that the Eigenmood analysis is robust to selection of sentiment tool and dataset, we also run it on an ANEW-scored sample of Tweets from 2010 to 2014. We use a 10% random sample of Tweets from Twitter's garden hose[1] from June 1, 2010 to Februrary 13, 2014 as our source of written sentiment. Each tweet was scored according to the Affective Norms for English Words (ANEW) as described in [4]. The ANEW is a lexical tool based on survey results that assigns 1034 English words a value from 1 to 9 along three dimensions: *arousal*, from calm to excited; *dominance*, from controlled to in-control; and *valence*, from sad to happy [6]. Each tweet containing words in the ANEW was scored with the average sentiment values of those words. We limit our analysis to tweets geo-located in the USA, yielding about ten-thousand tweets per day, and a 25-bin distribution of tweet scores along each dimension. Some choice of binning is required for the singular value decomposition, and a 25 bin distribution was found to be a good medium between resolution and sampling coverage.

[1]Data system supported by NSF Award No. IIS-0811994

Our initial focus was on the Latin American countries, so we included a translation of the ANEW to Spanish and Portuguese as well. To translate, each ANEW word was passed through Google translate and verified by native speakers of the languages on the team. If a tweet contained words matching multiple languages, the language with the most matches was used to score the tweet, or the average of the language's scores were taken in the case of ties. Other ANEW translations are available from separate experiments using native speakers in Spanish [14] and European Portuguese [15], however, these studies as well as Warriner et al.'s analysis of these and other studies [16] and Dodds et al.'s experiments with Amazon Mechanical Turk [8] show that generally there is a good agreement between the affective norms assigned to english glosses, and those assignment by ANEW, especially in the case of valence. For computational ease we thus use the same ANEW scores for each language translation.

A.1.1 Singular Value Decomposition

A singular value decomposition splits a matrix of data into orthonormal bases for its row and column spaces composed of *left* and *right singular vectors* respectively, along with relative variance explained by each, their *singular values*. This is related to a principal component analysis, where the right singular values are the principal component loadings and the left singular vectors multiplied by the corresponding singular values are the scores [63]. For each ANEW dimension, we create a matrix with rows corresponding to days, and columns corresponding to the bins of the distribution. The resulting left and right singular vectors can then be interpreted as *eigenbins* and *eigendays* respectively. The eigendays are vectors with 25 elements that describe patterns in a day's binned distribution; when all eigendays are summed together with appropriate coefficients, they

reproduce the distribution of any day's data. Similarly, the eigenbins describe patterns in a bin's value over time.

A.1.2 Statistics vs Mean

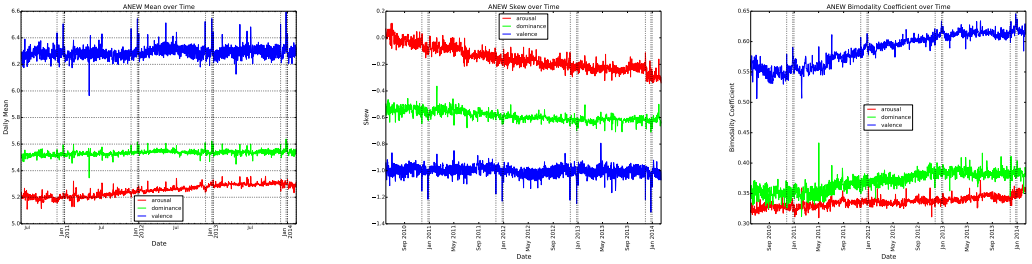


FIGURE A.1: Mean, Skew, and Bimodality coefficient over time for scored tweets from the United States

Mean values are used as a measure of central tendency in order to summarize sentiment distributions in a useful manner. However, as seen in figure A.1, the sentiment distribution is highly skewed, especially for valence. This negative skew is noted by [16], while a positive skew appears for the most frequently used words in [7]. Due to this skew, the median value is often chosen as a summary statistic since it is less influenced by outliers. However, even the median cannot accurately report the fluctuations in the distribution over time. Valence, the component that receives the most interest, also has a substantial bimodality coefficient; the negative skew corresponds to a large secondary peak in negative sentiment. To account for the various shapes the distribution assumes, we can turn to the eigendays and eigenbins found through a singular value

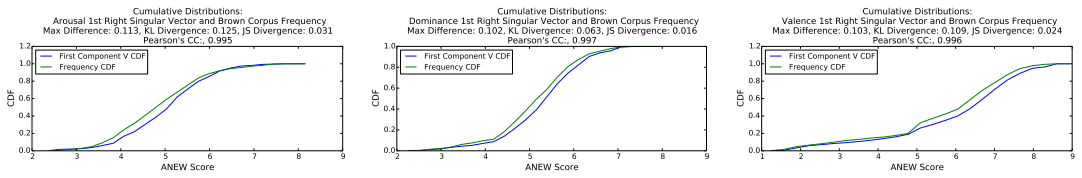


FIGURE A.2: Cumulative Distribution Comparison between First Right Singular Vectors and Brown Corpus, for each ANEW Dimension

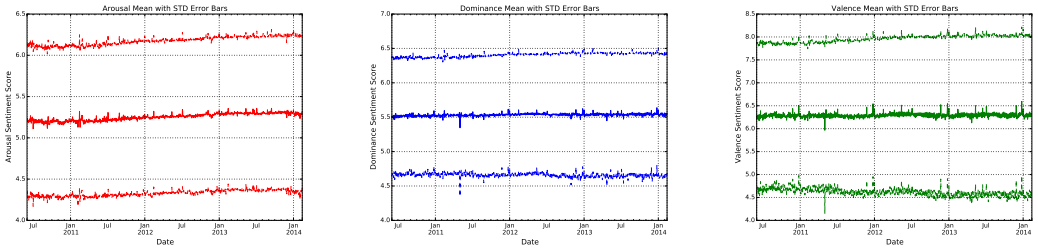


FIGURE A.3: Mean with standard deviation error bars over time for scored tweets from the United States

decomposition. Each eigenbin contains a time series of how much a particular change in distribution shape, the corresponding eigenday, contributes to each day's distribution.

The first eigenday corresponds closely to the overall distribution of sentiment in written language, according to word frequencies in the Brown corpus, included with the ANEW data [6]. This similarity is shown between their cumulative distribution functions in Figure A.2. To measure how close this eigenday is to natural language frequency, we normalized the eigenday to sum to 1 and found the Jensen-Shannon divergence (JS), a measure of the dissimilarity of two probability distributions, assigning values from 0 (the same) to 1 (no similarity) [103]. The divergence is quite low, the highest being 0.031 bits for arousal as shown in Table 1. In fact, the chance that a random reshuffling of word frequencies would result in a distribution so similar, estimated from 10^5 such reshuffles, is negligible for all but arousal, which has a 0.050 chance. In addition, this first component accounts for almost all of the variance in the data: 99% for arousal and dominance, and more for valence. This suggests that the distribution of sentiment is overwhelmed by the natural frequency of words. To really observe how it changes we need to look at further components.

Dimension	**JS**	p_f
Arousal	0.031	0.050
Dominance	0.016	0.000
Valence	0.024	0.000

TABLE A.1: JS is the Jensen-Shannon Divergence, p_f is the probability of finding a smaller JS, estimated from 100,000 random reshuffles of word frequencies

A.2 Additional Results

A.2.1 Event Detection

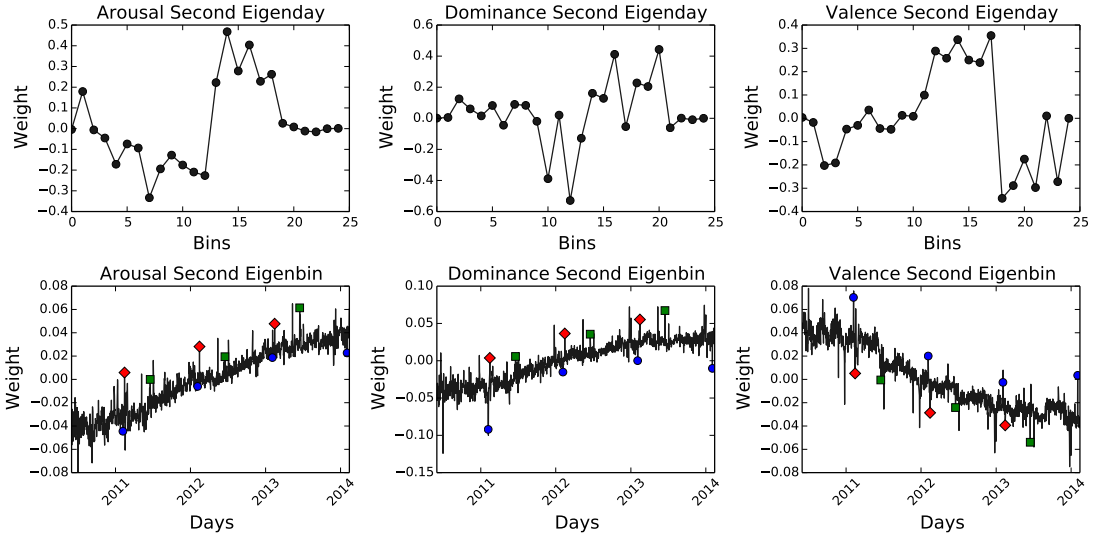


FIGURE A.4: Second Eigenday and Eigenbin for each ANEW dimension. Blue circles mark Super Bowls, red diamonds mark Valentine's Day, and green squares mark Father's Day

To describe changes in mood over time, we look at Z-scores for statistics like standard deviation and bimodality coefficient, as well as the mean, calculated over a trailing window of the previous 90 days. Significant changes are defined as Z-scores with absolute values greater than a threshold of 3. The emotional state of a number of days can be described as exhibiting more interesting behavior than the mean sentiment alone would suggest. For example, consider the Super Bowl, a major national American football event. The mean for dominance only significantly changes the day after the 2011 Super

Bowl, falling with a Z-score of -3.968. However, its bimodality coefficient plummets every year, with an average Z-score of -4.163, suggesting that expressions of dominance are more similar than usual when the same major sporting event is watched by most users. On Valentine's Day, the mean arousal peaks with an average Z-score of 5.465, however, for 2012 and 2013 the arousal bimodality coefficient also peaks with an average Z-score of 3.603. While most people are excited by Valentine's Day, a significant number respond oppositely. Father's Day only sees a significant change in mean sentiment values during 2013. However, the bimodality coefficient for valence increases during 2012 and 2013 with a Z-score of 3.106, while the overall distribution of valence changes in a similar way every year, as will be discussed later.

The second eigenday corresponds to long term changes over the course of the data, but is also highly relevant to certain holidays, as shown in Figure A.4. Since major holidays like Christmas, New Year's Day, and the USA's Independence Day are well described by changes to the mean, we continue to focus on the Super Bowl, Father's Day, and Valentine's Day. During the Super Bowl, Dominance reverses its trend over time and emphasizes the middle bins of its distribution, accounting for the lessened bimodality. This is surprising since words like "win" have a high dominance, and "humiliate" a low dominance, suggesting that the overall feeling is dominated by more neutral dominance words like "game" and "party". The increase in a bimodal arousal sentiment during Valentine's Day is largely explained by the contribution of the second arousal eigenday, emphasizing both higher bins and a very low bins which includes words like "sleep" and "bored". Similarly, the increase in valence bimodality on Father's Day is described by the second valence eigenday, emphasizing bins with happy words like "home" and "family" as well as low bins including the words "lonely" and "hate". We can also see that although the change isn't significant according to our rolling Z-score threshold,

Valentine's Day also emphasizes the same valence bins.

A.3 Twitter Data Collection

To measure and estimate collective mood states we analyzed the text of a large number of tweets. Our data source is Indiana University's connection to the Twitter garden-hose[2], which gives us access to a 10% random sample of the total volume of tweets Twitter receives. While this random sample is known to have an occasional bias [164] we do not detect or correct for the bias in this work. As shown in Fig A.5 The number of tweets collected range from on the order of 10^6 tweets at the beginning of the collection, to $5 \cdot 10^7$ towards the end, while only about 10^7 contain words in the ANEW lexicon, . We focus on tweets collected between September 2010, when the collection stabilized, and Februrary 2014, when the tweet collection suddenly dropped, complicating homogeneous analysis of the data.

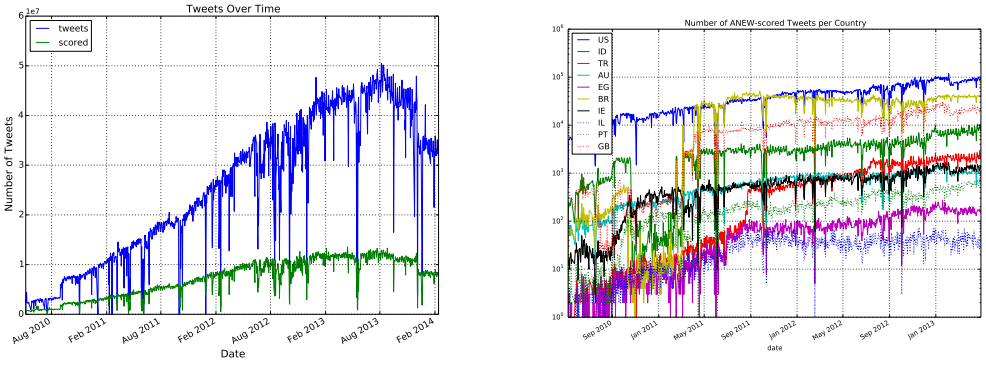


FIGURE A.5: Top: The number of tweets collected each day over time. "tweets" denotes the raw number of tweets, while "scored" denotes the number of tweets containing a word matching the ANEW lexicon, Bottom: The number of tweets collected each day over time for various countries.

[2]Data system supported by NSF Award No. IIS-0811994

Subsets of this tweet collection were divided into days based on Greenwich Mean Time (GMT), defining the beginning of a new day and end of a previous day at midnight GMT. Weeks were defined from Sunday to Sunday. These divisions are imperfect, given differences in time zones across countries, but simplifies the analysis and is somewhat ameliorated when examining weekly data. Geo-located tweets were divided into countries based on shape files[3]. As seen in Figure A.5, the number of scored tweets collected for each country varies across orders of magnitude, with the US leading in quantity with about 10^4 scored tweets collected per day. Each country's tweets are only examined after the collection has visually stabilized, starting in September 2010 for the US and Australia, May 2011 for Indonesia, Brazil and Portugal, and September 2011 for Turkey and Egypt.

The feed from twitter is naturally noisy, with a changing number of tweets received every day and a trend of increasing volume over the time period examined. Tweet collection rarely fails for an entire day, collecting a partial amount of the days tweets. These two properties make it difficult to distinguish a day in which tweet collection fails from a day's normal variation. To identify true failures in the tweet collection we use a rolling Z-Score defined in Equation A.1:

$$Z(X, t, w) = \frac{X_t - mean_w(X)}{\sigma_w(X)} \tag{A.1}$$

Where X is the time-series of tweets, w is the trailing window, t is the index of given day in X, $mean_w$ is the mean and σ_w is the standard deviation of the trailing w days (from X_{t-w} to X_{t-1})

[3]Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.

We have two time series describing the number of tweets collected on a given day, the raw numbers of tweets X^{tweets}, and the number of scored tweets containing a word matching the ANEW lexicon X^{scored}. *Collection failures* are defined as days in which $Z(X^{tweets}, t, 90) < -2$ or $Z(X^{scored}, t, 90) < -2$. These are determined in an sequential fashion for all t greater than 90, removing each collection failure as it is found (so that the window is more precisely defined over the last 90 non-collection failure days). To simplify analysis, weekly data ignores collection failures, since weekly aggregation smooths over daily failures.

A.4 Scoring Tweets

To calculate numerical values for sentiment in a twitter feeds, we matched words in tweets to words in ANEW. On each day, for each country, between the start of stabilized collection until 2014-02-13 all words in all tweets were matched against the ANEW lexicon. Each tweet containing words in the ANEW received scores along the three ANEW dimensions equal to the mean of the corresponding scores of the matching words. For weekly analysis, each bin receives the average of the daily bin probabilities each week, giving equal weight to each day.

Since the original focus of the project was on Latin American countries, we performed a basic expansion of the ANEW to match Spanish and Portuguese words. The translations were found by initially running each word through Google Translate, and then refining the translations by hand. For basic language detection, we find the language with the greatest number matches and assign the tweet a score based on that language. In case of a tie, the average scores over the tying languages are calculated. To find the actual sentiment during the holidays without generic seasonal greetings, we don't score words

if they appear in generic holiday greetings, such as "happy holidays", and we remove the ANEW words Christmas and Valentine from the lexicon entirely. The list of holidays whose greetings we removed were collected from http://www.officeholidays.com/.

Aggregating all sentiment in tweets into a mean value discards information in the distribution of sentiment across tweets. Therefore, we use binned distributions of sentiment across tweets in the following analysis. We focus on a 25-binned distribution between the lowest and highest possible ANEW scored as a moderately-grained distribution, with fine enough resolution to capture some detailed structure while aggregating an adequate number of tweets per bin, 40 on average for a collection of 10^3 tweets.

We also examine an alternative distribution that lends itself to easier interpretation. We can create a linguistic variable for each ANEW dimension, a variable whose values are linguistic values [165], in this case: low, medium-low, medium, medium-high, and high. Each value is a fuzzy set, defined as a membership function over the 100-binned distribution of tweet ANEW scores as shown in A.6. These membership functions were defined such that each function has the same area under its curve, while the memberships of each original bin across linguistic values sum to one. By multiplying the probability in each bin by its membership in linguistic values and summing the distribution of memberships for each linguistic value, we can produce a probability distribution over the values of the linguistic variable.

During analysis of the daily tweet sentiment, we also correct for data collection failures. Simply throwing away collection failures would interfere with models of the time series dependent upon previous days. To reduce the effect of these failures, we replaced collection failures with a trailing mean. First, to ensure that it is necessary to replace the data, we perform an additional identification check. For each $\mathbf{X}$ representing

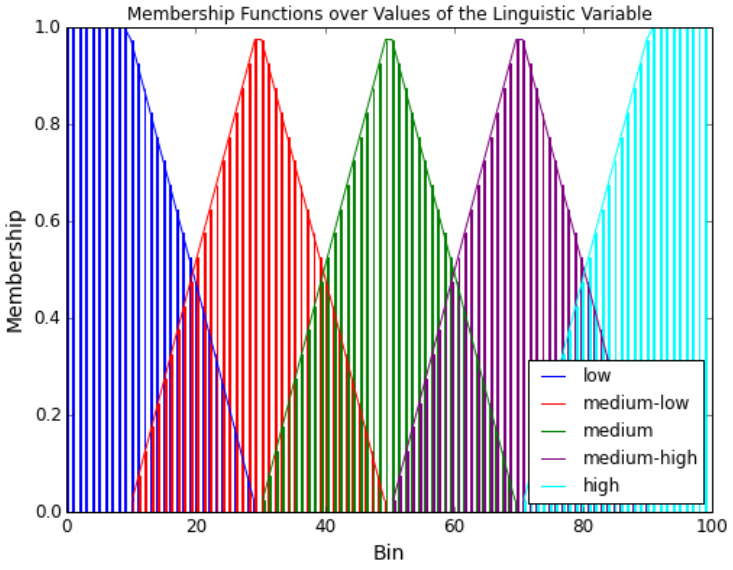


FIGURE A.6: The membership functions for the values the linguistic variable can take

the timeseries matrix of histogram vectors for an ANEW dimension, a singular value decomposition is performed. The first component explains the most variation in the data, therefore the first *eigenbin* b (explained below) can be used to determine how well each day fits the first component. We consider it only necessary to replace the data of a collection failure if there is an abrupt change in how well that collection failure fits the first component compared to previous days, given by $|Z(\Delta b, t, 90))| > 2$ where Δb denotes the one-day differences in the eigenbin. In these cases, the data for $\mathbf{X_t}$ is replaced by the trailing mean (for each element of the vector) of all occurrences of same day of the week as t in the past 90 days.

From these scored tweets we can examine statistics regarding collective moods over time, such as the mean value for each ANEW dimension estimated from 25-bin distributions, as shown in Figure A.7. The means vary little over time, nor do the distributions of ANEW scores visibly change when plotted as a heatmap in Figure A.8. As we will argue below, this is due to the natural frequency of ANEW words in the language overwhelming the distribution, washing out structure in sentiment.

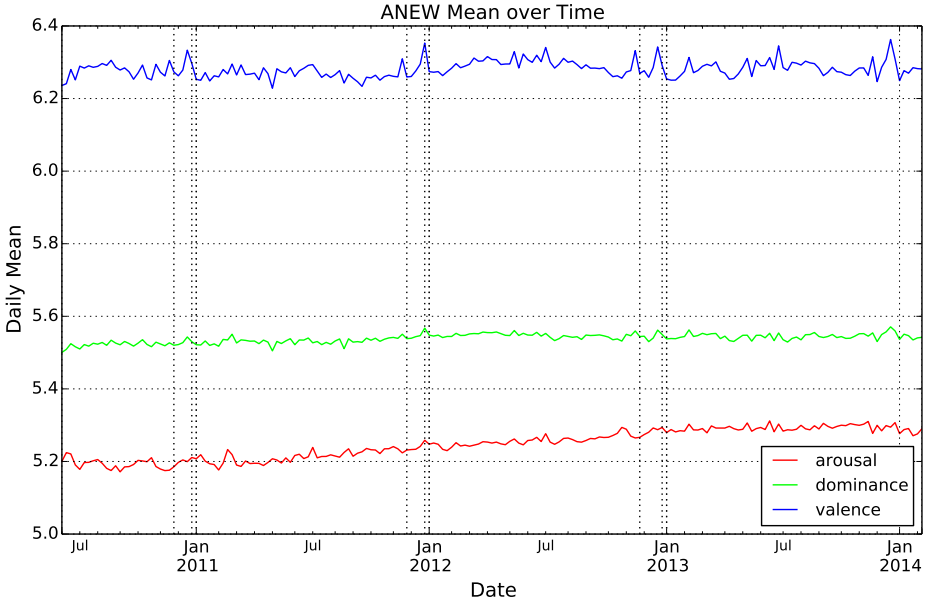


FIGURE A.7: Mean weekly values for the ANEW dimensions estimated from 25 bins. The vertical lines mark the holidays Thanksgiving, Christmas, and New Year's Day each year. Offsets from peaks are due to the difference between the holiday and the Sunday marking the start of the week

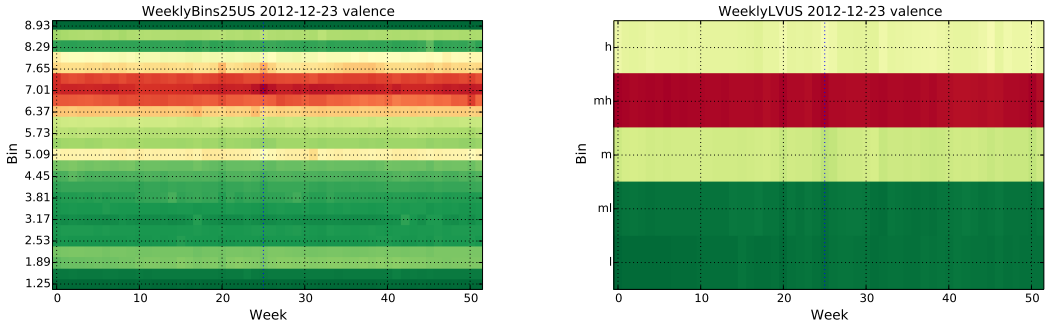


FIGURE A.8: Top: Heatmap of the 25-bin weekly distribution of tweet valence for the US around Christmas 2012, Bottom: Heatmap of the linguistic variable weekly distribution of tweet valence for the US around Christmas 2012

A.5 Singular Value Decomposition on binned ANEW distribution

Singular value decomposition (SVD) is a method by which a matrix can be linearly decomposed into ordered orthonormal components, each explaining as much of the linear variation as possible, after the components that came before it. The SVD of any $m \times n$

matrix M of real or complex numbers can represent M as follows in Equation A.2:

$$M = USV^T \tag{A.2}$$

Where U is $m \times n$ matrix with orthonormal columns, V is $n \times n$ matrix with orthonormal columns, and S is a diagonal matrix. The columns of U and V are referred to as the left and right singular vectors of M respectively. These singular vectors are eigenvectors of the matrices MM^T and M^TM respectively. The diagonal entries of S, called the singular values of M, are the square roots of the eigenvalues of the matrices MM^T and M^TM. By convention, the singular values are ordered from greatest to least. The columns of U form a basis for the column space of M and the columns of V form a basis for the row space of M. The right singular vectors are also known in *principal component analysis* (PCA) as the loadings of the original variables (bins) onto the new coordinate system. It is important to note that matrices can be reconstructed with a lower rank by setting elements of S to zero. Typically only the top l singular values are kept in order to create the closest rank-l approximation of the original matrix [63].

We applied the SVD to the binned distribution of ANEW scores over time. Our matrix M has columns representing bins, and rows representing days (weeks). The left and right singular vectors then have an interpretation as the "eigenbins" and "eigendays" ("eigenweeks") respectively. We will also refer to the singular vectors as components.

The relative variance explained by each component can then be calculated for each component k as $s_k^2/(\Sigma_i s_i^2)$ where s_k is the kth diagonal component of S. The variance explained by each component for the weekly 25-bin US distribution over each ANEW dimension is displayed in Figure A.9. The first component explains the vast majority of the variance.

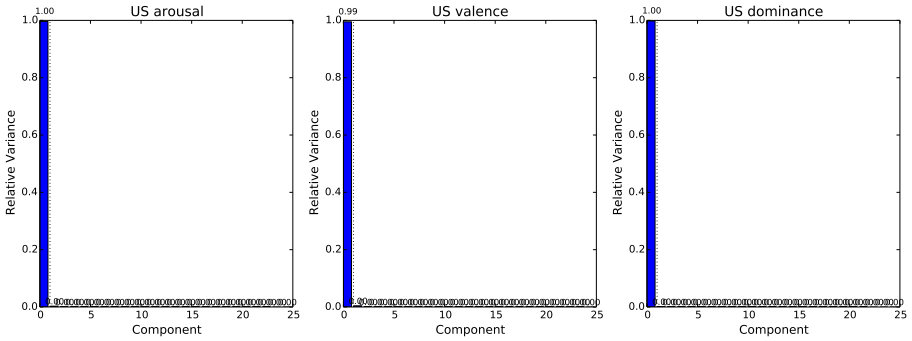


FIGURE A.9: Relative variance explained by each component

Each element of an eigenbin is the original day's (week's) coordinate along the axis formed by the corresponding eigenday. This is equal to the cosine of the angle between the original row and the eigenday. The multiplication of an element of an eigenbin by its singular vector gives the projection of the original row onto the corresponding eigenday, and in PCA this referred to as a principal component score. The reverse relationship holds between eigendays and eigenbins as well. The first three components for each ANEW dimension are shown in Figure A.10. Lines have been added to show that Eid Al-Fitr, Thanksgiving, Christmas, and New Years deviate from normal days to either correlate more or less with corresponding eigendays. We will suggest an interpretation for the first three components. The first component captures the usual distribution of ANEW sentiment in tweets given their natural frequency in the language, the second component the general change in the distribution over time, and the third component the cyclical yearly changes in the distribution.

While the first component explains the majority of the variation, it only explains the usual frequency of ANEW words in written language. In the ANEW study, the frequency of each term in the Brown corpus is included [6]. If we consider the first eigenday of the 100-bin daily ANEW scores as a distribution itself, we can see its similarity to the distribution of ANEW scores in the Brown corpus. The plots of the cumulative

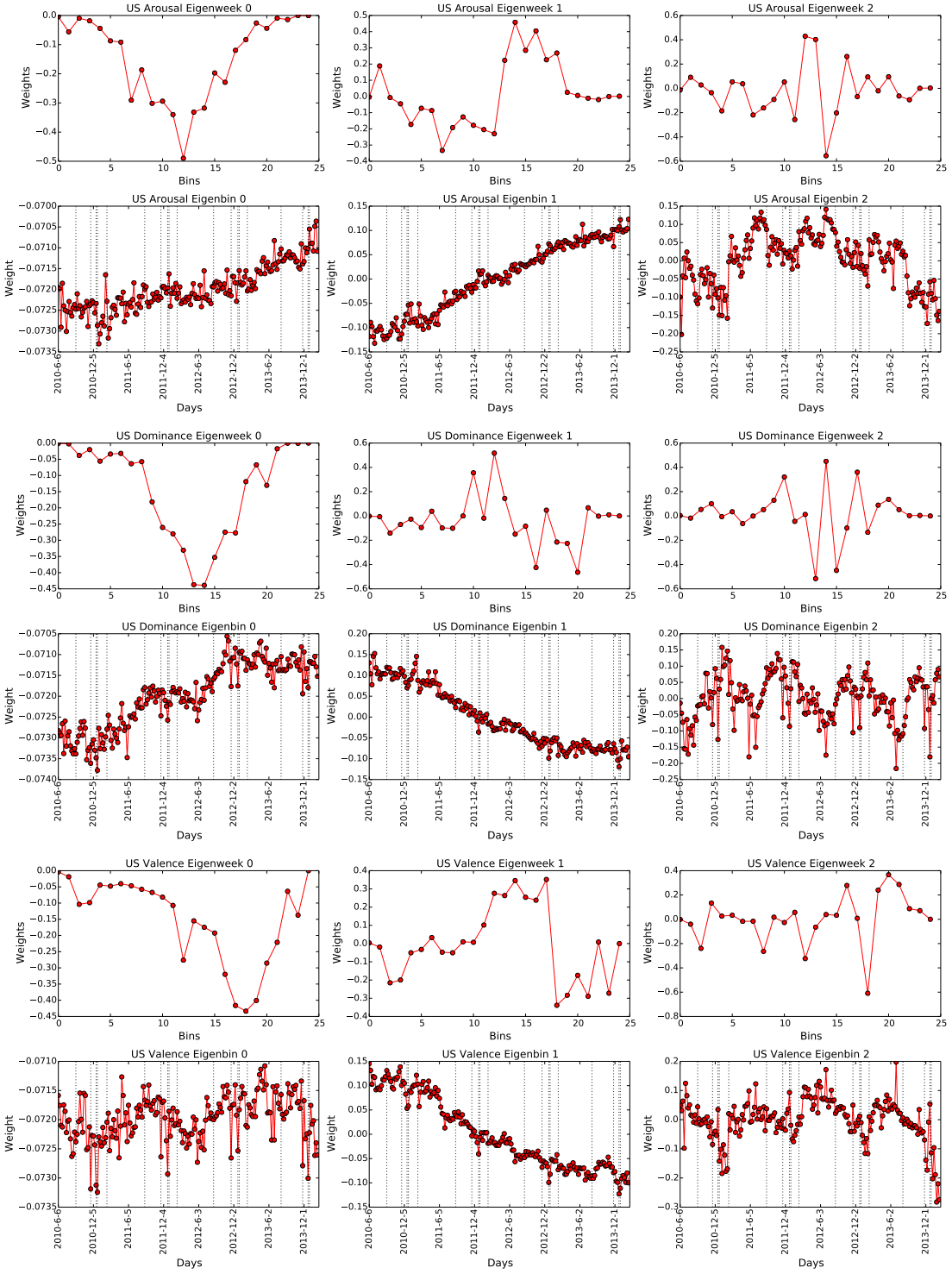


FIGURE A.10: Top to Bottom: Arousal, Dominance, and Valence components. Vertical lines are holidays: Eid al-Fitr, Thanksgiving, Christmas, New Year's Day, and Valentine's Day

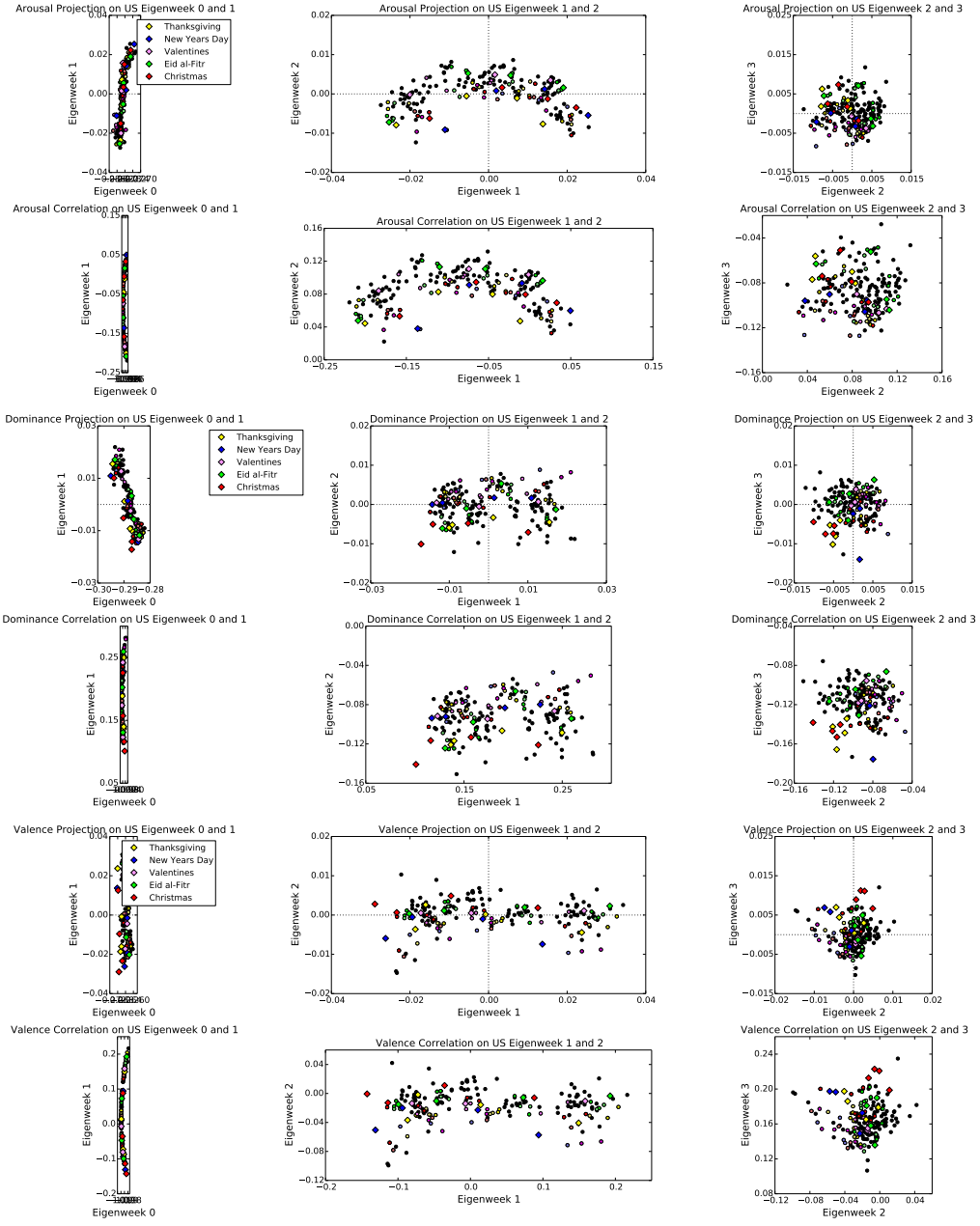


FIGURE A.11: Top to Bottom: arousal, dominance, and valence, projection and correlation onto components

distribution for both the first eigendays and the Brown corpus are compared in Figure A.2.

We also plot the projection and correlation of the original data onto eigenbins in Figure A.11, and see once again that Holidays are clustered outliers, especially for valence components.

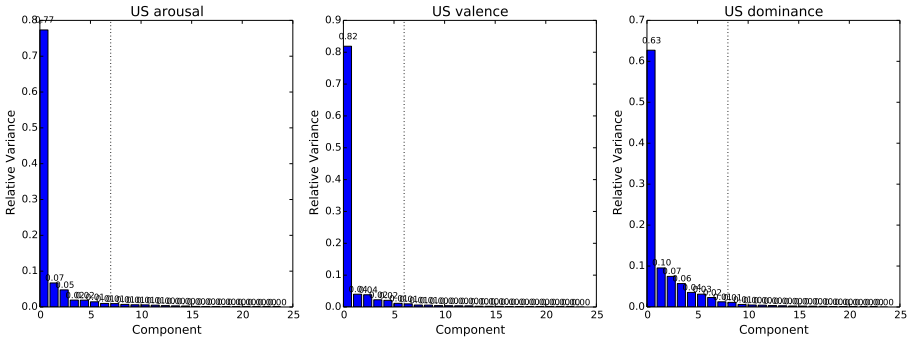


FIGURE A.12: Relative variance explained by each component, vertical line indicates where at least 95% of the variance has been explained by the components to the left

A.6 Reconstruction

Since we are more interested in how sentiment varies, rather than its basic distribution in language use, we reconstruct the original data without the first component. After recomputing the relative variances explained after removing the first component, we can remove noise by also removing the components explaining the least variance. Reconstruction, then includes only those components that explain 95 % of the remaining variance after the first component is removed.

This leaves cyclic patterns and outlier days deviating strongly from the baseline sentiment distribution, which we visualize as a heatmap of the distribution over time. The reconstructed heatmap for the US centered around Christmas 2012 is shown in Figure A.13

We can average over all full years in the data for multiple countries, centered on the week of a strong cultural holiday, to emphasize the change in these distributions, as shown in Figure A.14. It can be clearly seen from these averages that the distribution of sentiment shifts towards higher bins during holidays, represented by redder high bins and greener low bins on holidays. Christmas stands out in the USA (US), Australia (AU), Brazil (BR), and Portugal and Spain (PT and ES). Portugal and Spain are combined

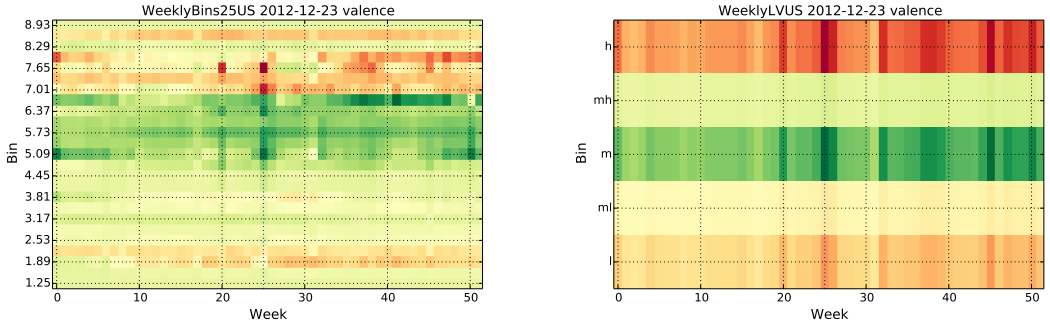


FIGURE A.13: Top: Reconstructed heatmap of the 25-bin weekly distribution of tweet valence for the US around Christmas 2012, Bottom: Reconstructed heatmap of the linguistic variable weekly distribution of tweet valence for the US around Christmas 2012

in this figure due to low tweet counts. Eid al-Fitr stands out in both Turkey (TR) and Indonesia (ID), and in Turkey the beginning of Ramadan is emphasized a few weeks before. The centering performed only looks at weeks within the surrounding cultural year, such that Christmas is week 26 of a 52 week year (starting with a first week 1), while Eid al-Fitr is week 25 of a 50 week year. Other weeks are averaged in this range according to their displacement from the holiday week (e.g., a week two weeks before the Christmas week in 2012 is averaged with weeks two weeks before Christmas in all other years). This obscures the emphasis on holidays using another calendar, such that Indonesia also has a strong signal on Christmas and Portugal and Spain have a strong signal on Eid al-Fitr, but these signals are averaged over multiple weeks when the calendars are misaligned.

A.7 Descriptive Statistics

Once we have scored tweets, we can compute descriptive statistics for the distribution of ANEW scores on each day. Below are plots of the mean, standard deviation, skewness, excess kurtosis, and bimodality coefficient for the distribution of ANEW scores over US tweets. Three vertical lines are placed during each year to mark Thanksgiving,

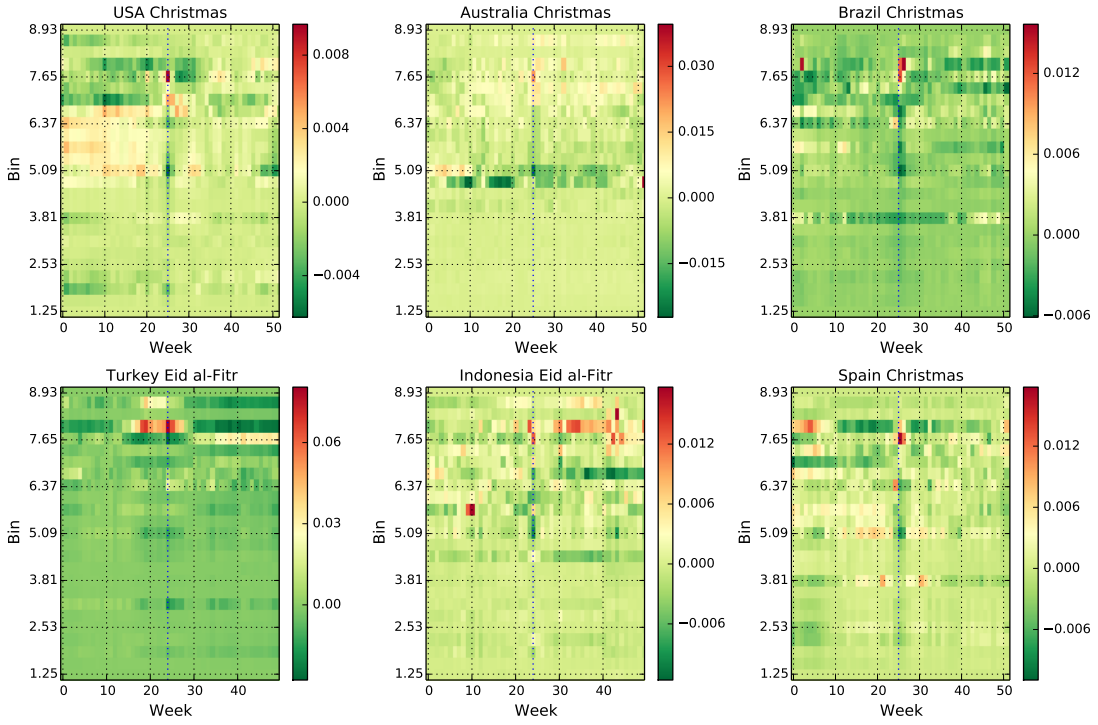


FIGURE A.14: Reconstructed Heatmaps for multiple countries, centered on cultural holidays

Christmas, and New Year's Day. For computational expediency, all descriptive statistics were approximated from a 100-bin distribution of scores between the minimum and maximum scores possible.

For a random variable T over the set of tweets on a given day, and E the expectation operator, the descriptive statistics are defined as:

$$mean = \mu = E[T] \tag{A.3}$$

$$standarddeviation(STD) = \sigma = \sqrt{E[(T-\mu)^2]} \tag{A.4}$$

$$skewness = \gamma = E[(\frac{T-\mu}{\sigma})^3] \tag{A.5}$$

$$kurtosis = \kappa = E[(\frac{T-\mu}{\sigma})^4] \tag{A.6}$$

$$excesskurtosis = \kappa - 3 \tag{A.7}$$

$$bimodalitycoefficient = \frac{\gamma^2 + 1}{\kappa} \tag{A.8}$$

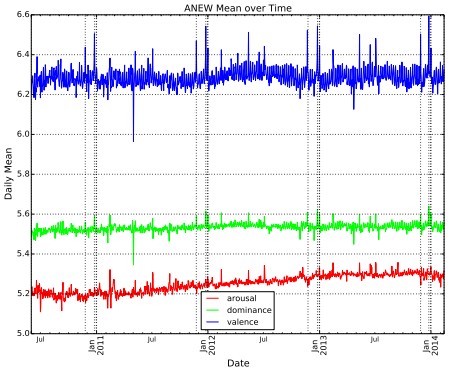


FIGURE A.15: Mean over time for scored tweets from the United States

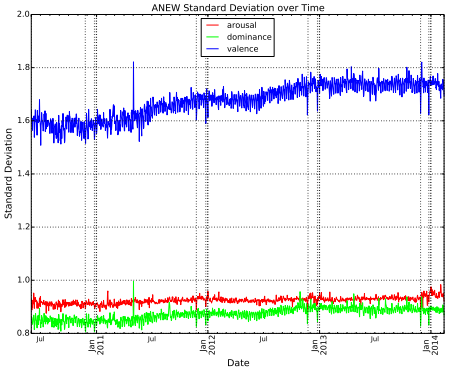


FIGURE A.16: Standard deviation over time for scored tweets from the United States

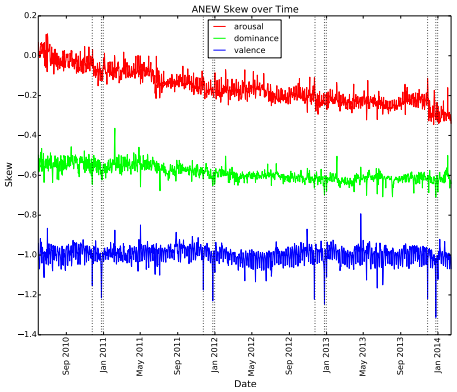


FIGURE A.17: Skewness over time for scored tweets from the United States

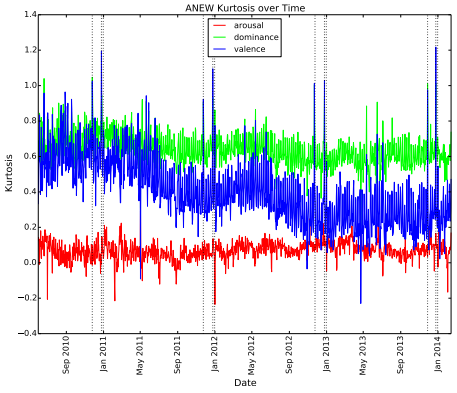


FIGURE A.18: Kurtosis over time for scored tweets from the United States

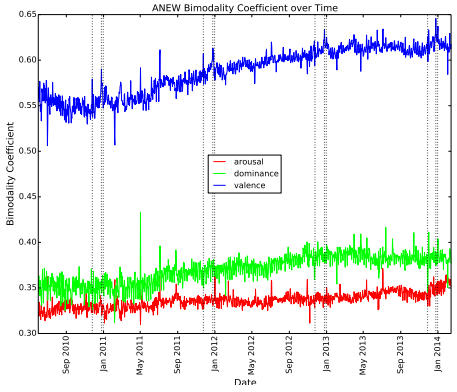


FIGURE A.19: Bimodality coefficient over time for scored tweets from the United States

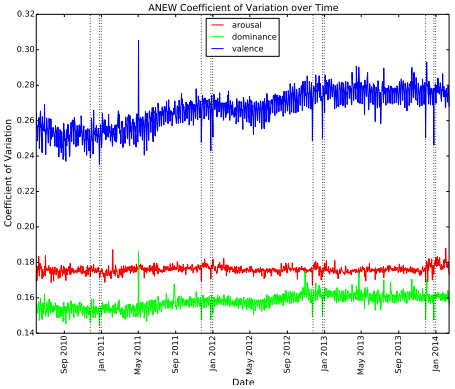


FIGURE A.20: Bimodality coefficient over time for scored tweets from the United States

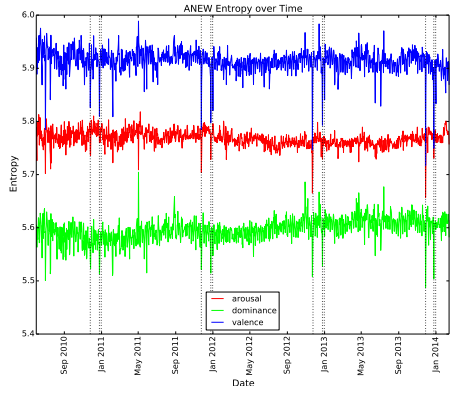


FIGURE A.21: Bimodality coefficient over time for scored tweets from the United States

A.8 Arima Models for Covid Mortality

Full performance statistics on all datasets for validation selected models are listed below. The *fullTrain* dataset is the combination of training and validation datasets. Order refers to the ARIMA parameters, the first number is the number of previous lags used as variables, the second the order of integration (differencing), the third the number of moving average components. This summary of overall performance is followed by visualizations of each selected model and its predictions as well as its full regression table.

A.8.1 Selected Models and Performance Stats

state	$trainR^2$	$validationR^2$	$fullTrainR^2$	$testR^2$	$order$
GA	0.798	0.911	0.905	0.887	(3, 1, 3)
MD	0.864	0.859	0.869	0.668	(1, 1, 2)
MA	0.914	0.801	0.901	0.851	(1, 1, 0)
NC	0.158	0.863	0.874	0.567	(1, 1, 0)
IL	0.866	0.935	0.915	0.848	(1, 1, 3)
OH	0.396	0.922	0.941	0.862	(3, 1, 3)
CO	0.597	0.872	0.815	0.861	(2, 1, 2)
MI	0.821	0.785	0.817	0.895	(1, 1, 0)
TX	0.863	0.920	0.913	0.881	(3, 1, 1)
IN	0.536	0.910	0.891	0.845	(2, 1, 0)
NV	0.506	0.811	0.813	0.636	(1, 1, 1)
CA LA	0.627	0.940	0.974	0.902	(2, 1, 3)
FL	0.911	0.800	0.864	0.933	(2, 1, 3)
TN	0.679	0.872	0.889	0.714	(1, 1, 0)
LA	0.769	0.846	0.809	0.828	(1, 1, 3)
NY	0.876	0.898	0.883	0.795	(1, 1, 0)
PA	0.847	0.949	0.933	0.905	(2, 1, 0)
CA SF	0.627	0.940	0.974	0.902	(2, 1, 3)
WA	0.521	0.182	0.531	0.632	(3, 1, 3)
DC	0.716	0.224	0.594	0.253	(1, 1, 1)

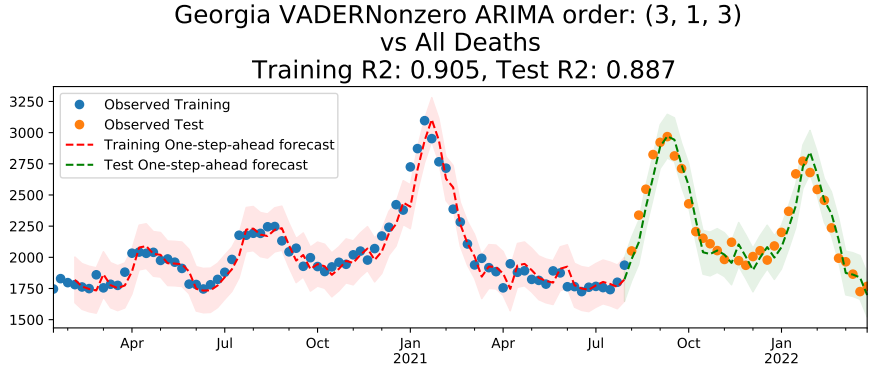
TABLE A.2: Training and Test performance statistics for ARIMA models without sentiment factors, order selected by validation set performance

state	$trainR^2$	$validationR^2$	$fullTrainR^2$	$testR^2$	$mood_{shift}$	*order*	*totalcount*
GA	0.796	0.913	0.908	0.834	0	(3, 1, 3)	30676524
MD	0.873	0.862	0.873	0.699	0	(2, 1, 3)	10203927
MA	0.914	0.801	0.901	0.851	0	(1, 1, 0)	17543538
NC	0.185	0.866	0.876	0.570	1	(1, 1, 0)	8527596
IL	0.867	0.935	0.916	0.849	1	(1, 1, 3)	36741068
OH	0.390	0.919	0.943	0.858	1	(3, 1, 3)	6969260
CO	0.597	0.872	0.816	0.862	0	(2, 1, 2)	9704942
MI	0.829	0.770	0.818	0.882	1	(1, 1, 0)	7898692
TX	0.863	0.920	0.914	0.880	0	(3, 1, 1)	27588633
IN	0.539	0.910	0.892	0.846	0	(2, 1, 0)	5710193
NV	0.510	0.814	0.816	0.650	0	(1, 1, 1)	12978313
CA LA	0.655	0.947	0.973	0.903	0	(3, 1, 3)	56629277
FL	0.911	0.799	0.865	0.930	0	(2, 1, 3)	11555324
TN	0.686	0.876	0.893	0.731	1	(1, 1, 0)	8584755
LA	0.764	0.837	0.806	0.830	1	(1, 1, 3)	8081540
NY	0.877	0.905	0.884	0.810	0	(1, 1, 0)	38762121
PA	0.857	0.950	0.935	0.905	1	(1, 1, 2)	18005902
CA SF	0.655	0.946	0.975	0.916	0	(3, 1, 3)	14887659
WA	0.513	0.181	0.510	0.624	1	(2, 1, 3)	18671561
DC	0.716	0.220	0.594	0.253	0	(1, 1, 1)	30640803

TABLE A.3: Training and Test performance statistics for ARIMA models with mean Vader sentiment score as exogneous factor, ARIMA order and mood lag selected by validation set performance

state	$trainR^2$	$validationR^2$	$fullTrainR^2$	$testR^2$	$mood_{shift}$	*order*	*components*
GA	0.825	0.918	0.910	0.891	0	(3, 1, 3)	(4, 5)
MD	0.883	0.867	0.884	0.416	1	(3, 1, 3)	(0, 7)
MA	0.917	0.832	0.907	0.857	0	(2, 1, 1)	(8, 10)
NC	0.196	0.876	0.880	0.547	0	(1, 1, 0)	(3, 10)
IL	0.868	0.937	0.917	0.848	0	(2, 1, 2)	(6, 10)
OH	0.369	0.931	0.934	0.897	1	(1, 1, 1)	(3, 9)
CO	0.601	0.891	0.840	0.805	0	(1, 1, 2)	(2, 3)
MI	0.855	0.798	0.847	0.887	1	(2, 1, 3)	(1, 5)
TX	0.874	0.930	0.924	0.884	1	(3, 1, 1)	(1, 5)
IN	0.554	0.913	0.897	0.849	0	(2, 1, 0)	(1, 2)
NV	0.542	0.833	0.834	0.601	1	(1, 1, 1)	(5, 8)
CA LA	0.705	0.950	0.974	0.910	1	(3, 1, 3)	(4, 7)
FL	0.913	0.808	0.868	0.931	0	(2, 1, 3)	(1, 3)
TN	0.720	0.891	0.901	0.715	1	(1, 1, 1)	(1, 5)
LA	0.779	0.862	0.823	0.803	1	(1, 1, 3)	(4, 7)
NY	0.878	0.909	0.886	0.801	0	(1, 1, 0)	(0, 5)
PA	0.885	0.953	0.942	0.851	0	(1, 1, 2)	(9, 10)
CA SF	0.722	0.955	0.971	0.875	0	(1, 1, 2)	(2, 7)
WA	0.449	0.266	0.518	0.604	0	(2, 1, 3)	(3, 5)
DC	0.716	0.235	0.600	0.222	0	(1, 1, 1)	(3, 6)

TABLE A.4: Training and Test performance statistics for ARIMA models with two Vader eigenbin components as exogneous factors, ARIMA order, component selection, and mood lag selected by validation set performance



SARIMAX Results

Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(3, 1, 3)	Log Likelihood	-474.280
Date:	Mon, 17 Apr 2023	AIC	962.560
Time:	11:30:54	BIC	979.234
Sample:	01-18-2020	HQIC	969.245
	- 07-31-2021		
Covariance Type:	opg		

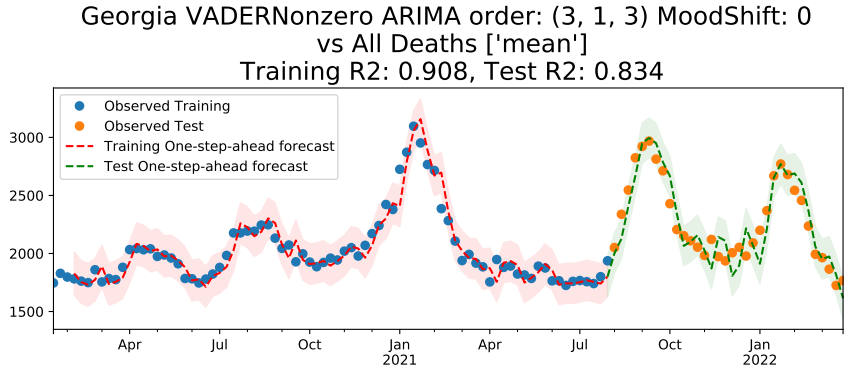
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.2646	0.363	3.487	0.000	0.554	1.975
ar.L2	-0.0190	0.614	-0.031	0.975	-1.223	1.185
ar.L3	-0.3290	0.319	-1.032	0.302	-0.954	0.296
ma.L1	-1.2238	0.884	-1.384	0.166	-2.957	0.509
ma.L2	0.2345	0.649	0.362	0.718	-1.037	1.506
ma.L3	-0.0098	0.304	-0.032	0.974	-0.607	0.587
sigma2	7975.7515	5891.637	1.354	0.176	-3571.645	1.95e+04

Ljung-Box (Q):	42.42	Jarque-Bera (JB):	4.88
Prob(Q):	0.37	Prob(JB):	0.09
Heteroskedasticity (H):	1.71	Skew:	0.48
Prob(H) (two-sided):	0.17	Kurtosis:	3.73

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.22: Atlanta GA Selected Model, No Sentiment

A.8.2 Predictions and Regression Tables



SARIMAX Results

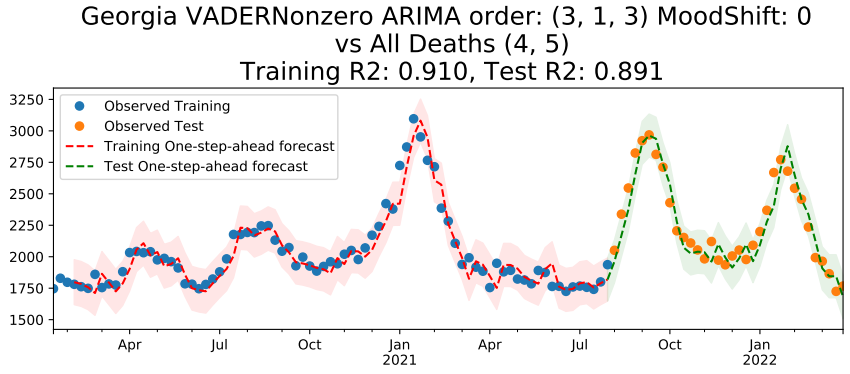
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(3, 1, 3)	Log Likelihood	-473.292
Date:	Mon, 17 Apr 2023	AIC	962.584
Time:	11:32:27	BIC	981.640
Sample:	01-18-2020	HQIC	970.224
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	418.2121	310.248	1.348	0.178	-189.863	1026.287
ar.L1	0.5022	0.569	0.882	0.378	-0.614	1.618
ar.L2	-0.5493	0.199	-2.756	0.006	-0.940	-0.159
ar.L3	0.4071	0.301	1.351	0.177	-0.184	0.998
ma.L1	-0.2696	9.196	-0.029	0.977	-18.293	17.754
ma.L2	0.9091	88.562	0.010	0.992	-172.669	174.487
ma.L3	-0.4650	42.999	-0.011	0.991	-84.742	83.812
sigma2	7597.8757	7.05e+05	0.011	0.991	-1.37e+06	1.39e+06

Ljung-Box (Q):	42.13	Jarque-Bera (JB):	21.09
Prob(Q):	0.38	Prob(JB):	0.00
Heteroskedasticity (H):	1.99	Skew:	-0.10
Prob(H) (two-sided):	0.08	Kurtosis:	5.51

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.23: Atlanta GA Selected Model, Mean Vader Sentiment



SARIMAX Results

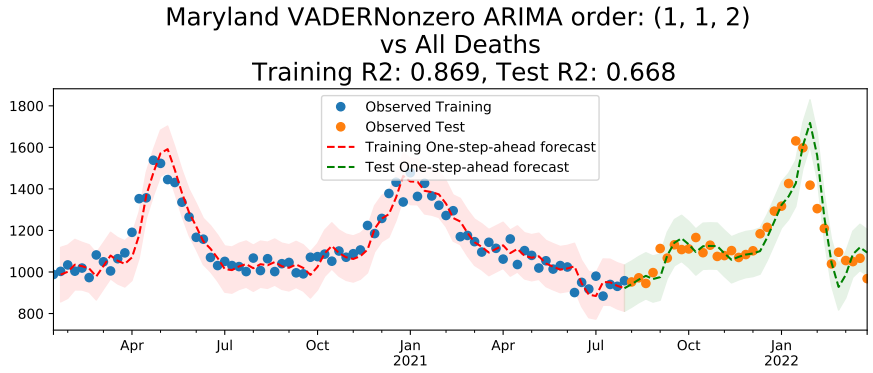
Dep. Variable:	All Cause	No. Observations:	81			
Model:	ARIMA(3, 1, 3)	Log Likelihood	-472.025			
Date:	Mon, 17 Apr 2023	AIC	962.050			
Time:	11:30:51	BIC	983.488			
Sample:	01-18-2020 - 07-31-2021	HQIC	970.645			
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
0	-7871.2966	7840.183	-1.004	0.315	-2.32e+04	7495.179
1	9384.5634	7344.264	1.278	0.201	-5009.929	2.38e+04
ar.L1	1.2170	0.399	3.053	0.002	0.436	1.998
ar.L2	0.0471	0.683	0.069	0.945	-1.292	1.386
ar.L3	-0.3505	0.348	-1.007	0.314	-1.033	0.332
ma.L1	-1.1420	0.986	-1.158	0.247	-3.075	0.791
ma.L2	0.1429	0.650	0.220	0.826	-1.131	1.417
ma.L3	4.865e-05	0.312	0.000	1.000	-0.611	0.611
sigma2	7540.5147	6585.331	1.145	0.252	-5366.498	2.04e+04

Ljung-Box (Q):	39.83	Jarque-Bera (JB):	4.57
Prob(Q):	0.48	Prob(JB):	0.10
Heteroskedasticity (H):	1.71	Skew:	0.42
Prob(H) (two-sided):	0.17	Kurtosis:	3.82

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.24: Atlanta GA Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

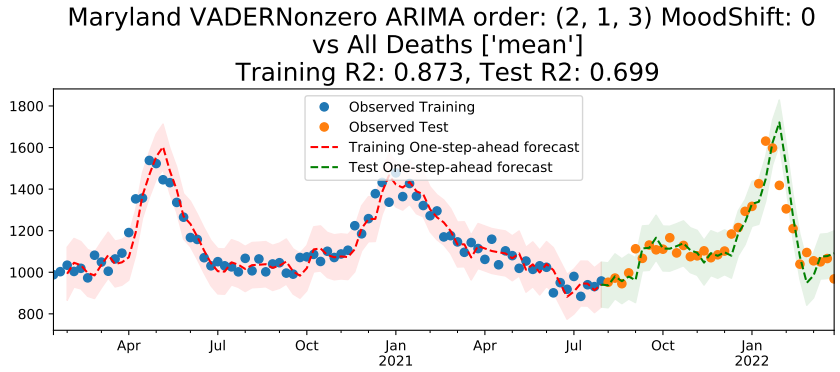
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 2)	Log Likelihood	-437.197
Date:	Mon, 17 Apr 2023	AIC	882.395
Time:	11:31:00	BIC	891.923
Sample:	01-18-2020	HQIC	886.215
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4883	0.136	3.591	0.000	0.222	0.755
ma.L1	-0.7890	0.111	-7.107	0.000	-1.007	-0.571
ma.L2	0.6252	0.104	6.020	0.000	0.422	0.829
sigma2	3225.8740	498.968	6.465	0.000	2247.914	4203.834

Ljung-Box (Q):	36.55	Jarque-Bera (JB):	1.87
Prob(Q):	0.63	Prob(JB):	0.39
Heteroskedasticity (H):	0.60	Skew:	0.20
Prob(H) (two-sided):	0.20	Kurtosis:	3.63

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.25: Baltimore MD Selected Model, No Sentiment



SARIMAX Results

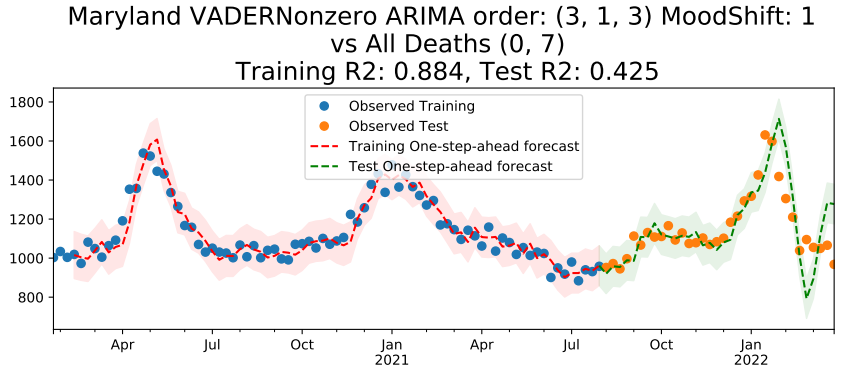
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 3)	Log Likelihood	-435.411
Date:	Mon, 17 Apr 2023	AIC	884.823
Time:	11:32:29	BIC	901.497
Sample:	01-18-2020	HQIC	891.508
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	357.4083	445.605	0.802	0.423	-515.962	1230.778
ar.L1	-0.4238	0.161	-2.637	0.008	-0.739	-0.109
ar.L2	0.5759	0.149	3.863	0.000	0.284	0.868
ma.L1	0.2143	0.489	0.439	0.661	-0.743	1.172
ma.L2	-0.2567	0.379	-0.678	0.498	-0.999	0.485
ma.L3	0.5157	0.279	1.849	0.065	-0.031	1.062
sigma2	3029.7857	1337.003	2.266	0.023	409.308	5650.263

Ljung-Box (Q):	26.63	Jarque-Bera (JB):	1.79
Prob(Q):	0.95	Prob(JB):	0.41
Heteroskedasticity (H):	0.55	Skew:	0.16
Prob(H) (two-sided):	0.12	Kurtosis:	3.66

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.26: Baltimore MD Selected Model, Mean Vader Sentiment



SARIMAX Results

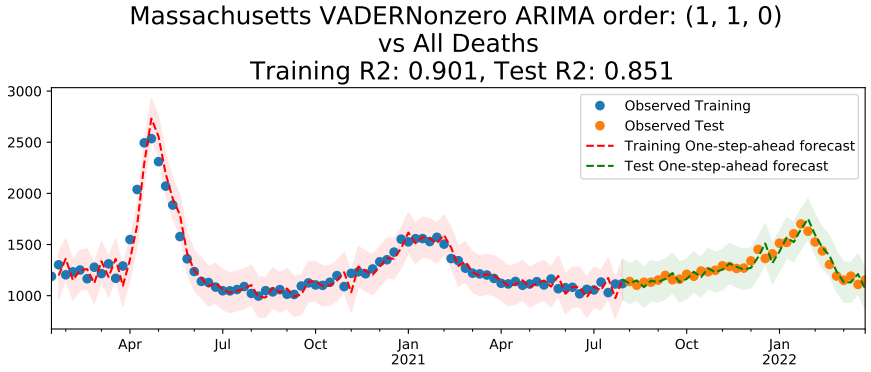
Dep. Variable:	All Cause	No. Observations:	80			
Model:	ARIMA(3, 1, 3)	Log Likelihood	-426.216			
Date:	Mon, 17 Apr 2023	AIC	870.432			
Time:	11:30:58	BIC	891.757			
Sample:	01-25-2020	HQIC	878.975			
	- 07-31-2021					
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
0	-918.0995	846.042	-1.085	0.278	-2576.311	740.112
1	2554.8151	2656.043	0.962	0.336	-2650.934	7760.564
ar.L1	-0.0762	0.152	-0.503	0.615	-0.374	0.221
ar.L2	0.5138	0.119	4.300	0.000	0.280	0.748
ar.L3	-0.4080	0.132	-3.084	0.002	-0.667	-0.149
ma.L1	-0.1592	0.655	-0.243	0.808	-1.443	1.125
ma.L2	-0.1341	0.510	-0.263	0.793	-1.134	0.866
ma.L3	0.9818	1.150	0.853	0.393	-1.273	3.236
sigma2	2641.5047	3023.055	0.874	0.382	-3283.575	8566.584

Ljung-Box (Q):	29.99	Jarque-Bera (JB):	3.61
Prob(Q):	0.88	Prob(JB):	0.16
Heteroskedasticity (H):	0.50	Skew:	0.23
Prob(H) (two-sided):	0.08	Kurtosis:	3.94

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.27: Baltimore MD Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

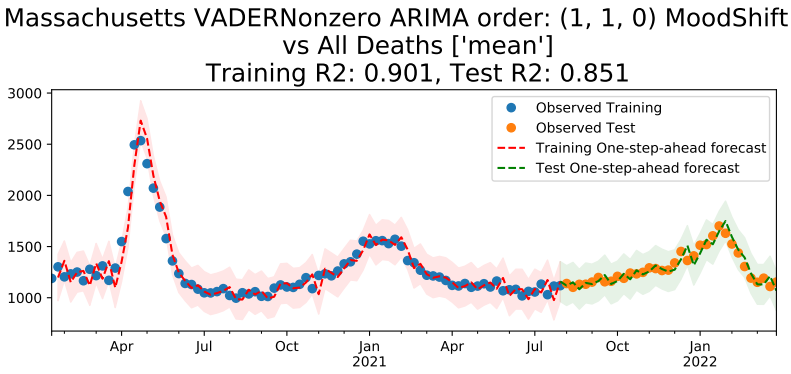
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-481.542
Date:	Mon, 17 Apr 2023	AIC	967.085
Time:	11:31:05	BIC	971.849
Sample:	01-18-2020	HQIC	968.995
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5203	0.058	8.938	0.000	0.406	0.634
sigma2	9888.5386	1284.599	7.698	0.000	7370.771	1.24e+04

Ljung-Box (Q):	19.51	Jarque-Bera (JB):	8.41
Prob(Q):	1.00	Prob(JB):	0.01
Heteroskedasticity (H):	0.22	Skew:	0.38
Prob(H) (two-sided):	0.00	Kurtosis:	4.40

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.28: Boston MA Selected Model, No Sentiment



SARIMAX Results

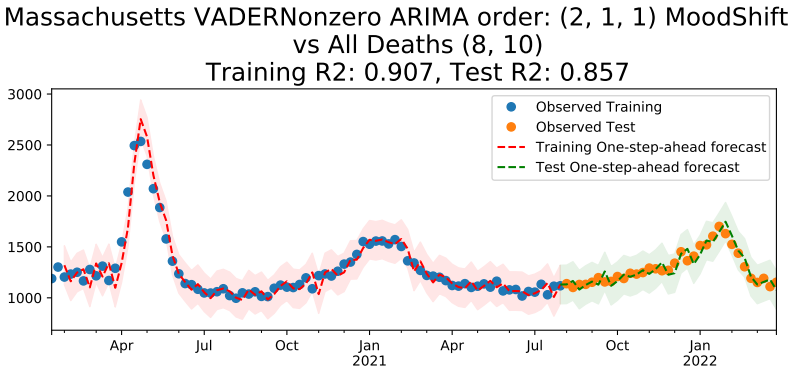
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-481.479
Date:	Mon, 17 Apr 2023	AIC	968.958
Time:	11:32:31	BIC	976.104
Sample:	01-18-2020	HQIC	971.823
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-44.3671	435.763	-0.102	0.919	-898.447	809.713
ar.L1	0.5212	0.059	8.842	0.000	0.406	0.637
sigma2	9873.4340	1291.474	7.645	0.000	7342.191	1.24e+04

Ljung-Box (Q):	19.53	Jarque-Bera (JB):	8.38
Prob(Q):	1.00	Prob(JB):	0.02
Heteroskedasticity (H):	0.22	Skew:	0.37
Prob(H) (two-sided):	0.00	Kurtosis:	4.40

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.29: Boston MA Selected Model, Mean Vader Sentiment



SARIMAX Results

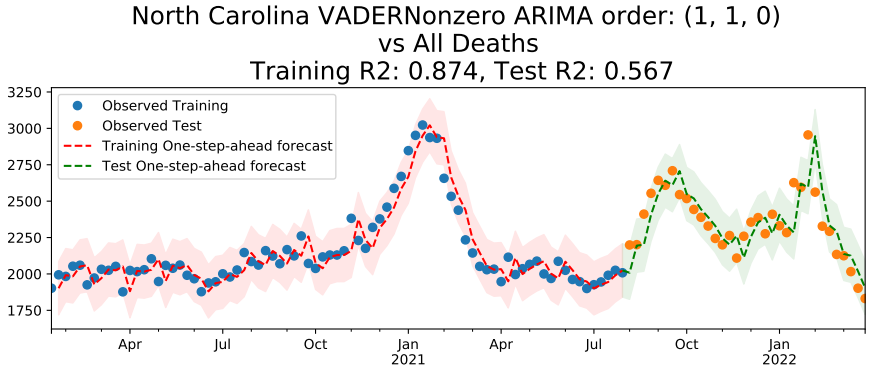
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 1)	Log Likelihood	-479.133
Date:	Mon, 17 Apr 2023	AIC	970.266
Time:	11:31:03	BIC	984.558
Sample:	01-18-2020	HQIC	975.996
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	-5423.1286	5464.189	-0.992	0.321	-1.61e+04	5286.484
1	-8670.3232	8176.360	-1.060	0.289	-2.47e+04	7355.048
ar.L1	-0.0812	0.334	-0.243	0.808	-0.735	0.573
ar.L2	0.4127	0.162	2.542	0.011	0.094	0.731
ma.L1	0.5579	0.345	1.618	0.106	-0.118	1.234
sigma2	9287.9625	1304.131	7.122	0.000	6731.913	1.18e+04

Ljung-Box (Q):	14.07	Jarque-Bera (JB):	13.31
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.18	Skew:	0.34
Prob(H) (two-sided):	0.00	Kurtosis:	4.88

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.30: Boston MA Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

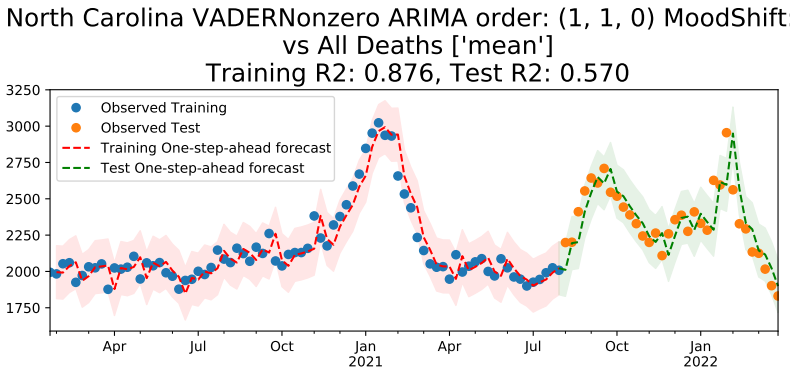
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-476.416
Date:	Mon, 17 Apr 2023	AIC	956.833
Time:	11:31:09	BIC	961.597
Sample:	01-18-2020	HQIC	958.743
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0232	0.102	-0.227	0.820	-0.224	0.177
sigma2	8714.8461	1335.685	6.525	0.000	6096.952	1.13e+04

Ljung-Box (Q):	45.63	Jarque-Bera (JB):	1.26
Prob(Q):	0.25	Prob(JB):	0.53
Heteroskedasticity (H):	1.50	Skew:	-0.30
Prob(H) (two-sided):	0.30	Kurtosis:	3.15

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.31: Charlotte NC Selected Model, No Sentiment



SARIMAX Results

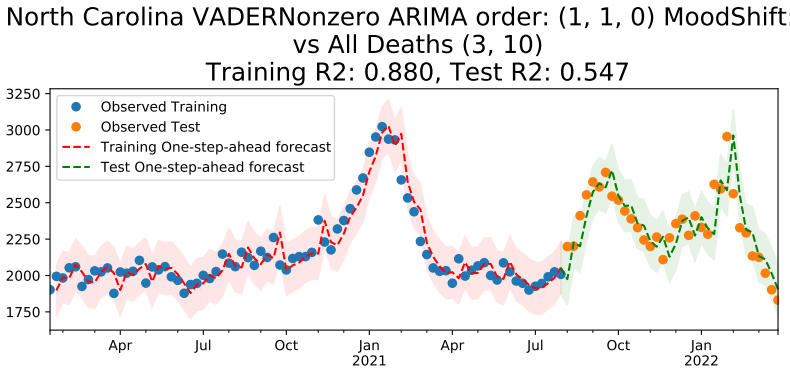
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 0)	Log Likelihood	-470.195
Date:	Mon, 17 Apr 2023	AIC	946.391
Time:	11:32:32	BIC	953.499
Sample:	01-25-2020	HQIC	949.239
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-294.0182	455.240	-0.646	0.518	-1186.272	598.235
ar.L1	-0.0156	0.104	-0.150	0.880	-0.219	0.188
sigma2	8654.1795	1371.487	6.310	0.000	5966.115	1.13e+04

Ljung-Box (Q):	41.52	Jarque-Bera (JB):	1.53
Prob(Q):	0.40	Prob(JB):	0.47
Heteroskedasticity (H):	1.58	Skew:	-0.29
Prob(H) (two-sided):	0.25	Kurtosis:	3.35

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.32: Charlotte NC Selected Model, Mean Vader Sentiment



SARIMAX Results

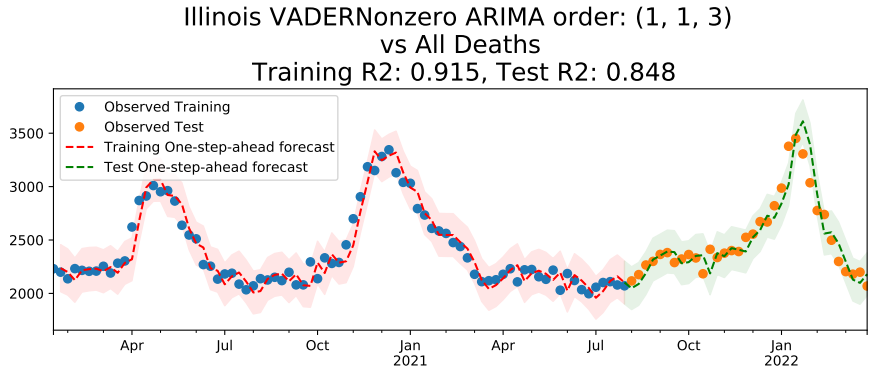
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-474.431
Date:	Mon, 17 Apr 2023	AIC	956.861
Time:	11:31:07	BIC	966.389
Sample:	01-18-2020	HQIC	960.681
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	4888.6244	3946.259	1.239	0.215	-2845.900	1.26e+04
1	6215.4010	3727.783	1.667	0.095	-1090.920	1.35e+04
ar.L1	0.0334	0.106	0.314	0.754	-0.175	0.242
sigma2	8278.8004	1286.667	6.434	0.000	5756.980	1.08e+04

Ljung-Box (Q):	39.32	Jarque-Bera (JB):	8.80
Prob(Q):	0.50	Prob(JB):	0.01
Heteroskedasticity (H):	1.93	Skew:	-0.56
Prob(H) (two-sided):	0.09	Kurtosis:	4.17

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.33: Charlotte NC Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

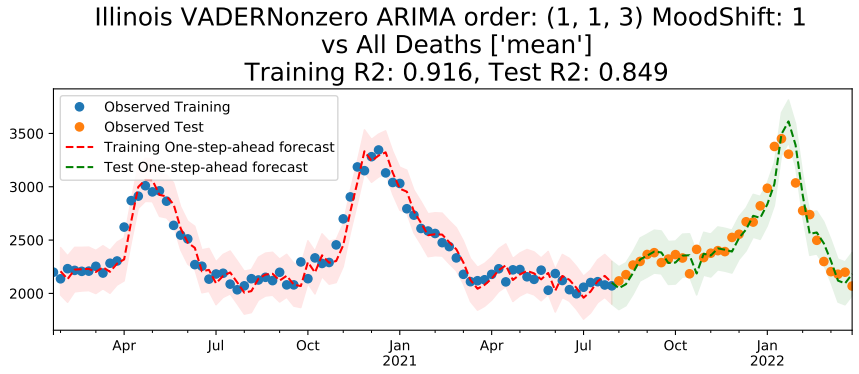
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 3)	Log Likelihood	-484.534
Date:	Mon, 17 Apr 2023	AIC	979.069
Time:	11:31:12	BIC	990.979
Sample:	01-18-2020	HQIC	983.844
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2384	0.360	0.663	0.507	-0.466	0.943
ma.L1	-0.1556	0.376	-0.414	0.679	-0.892	0.581
ma.L2	0.2923	0.114	2.560	0.010	0.068	0.516
ma.L3	0.2383	0.162	1.473	0.141	-0.079	0.555
sigma2	1.048e+04	1713.994	6.116	0.000	7123.913	1.38e+04

Ljung-Box (Q):	41.38	Jarque-Bera (JB):	2.80
Prob(Q):	0.41	Prob(JB):	0.25
Heteroskedasticity (H):	0.62	Skew:	0.46
Prob(H) (two-sided):	0.22	Kurtosis:	3.09

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.34: Chicago IL Selected Model, No Sentiment



SARIMAX Results

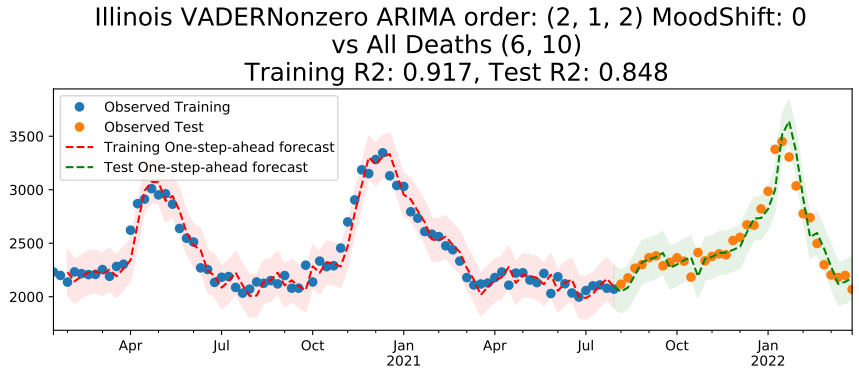
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 3)	Log Likelihood	-478.472
Date:	Mon, 17 Apr 2023	AIC	968.945
Time:	11:32:34	BIC	983.161
Sample:	01-25-2020	HQIC	974.640
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-201.7926	604.245	-0.334	0.738	-1386.091	982.506
ar.L1	0.2568	0.367	0.699	0.484	-0.463	0.977
ma.L1	-0.1779	0.385	-0.462	0.644	-0.933	0.577
ma.L2	0.3097	0.120	2.588	0.010	0.075	0.544
ma.L3	0.2244	0.174	1.290	0.197	-0.116	0.565
sigma2	1.061e+04	1741.342	6.091	0.000	7194.303	1.4e+04

Ljung-Box (Q):	40.23	Jarque-Bera (JB):	2.56
Prob(Q):	0.46	Prob(JB):	0.28
Heteroskedasticity (H):	0.62	Skew:	0.44
Prob(H) (two-sided):	0.22	Kurtosis:	3.12

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.35: Chicago IL Selected Model, Mean Vader Sentiment



SARIMAX Results

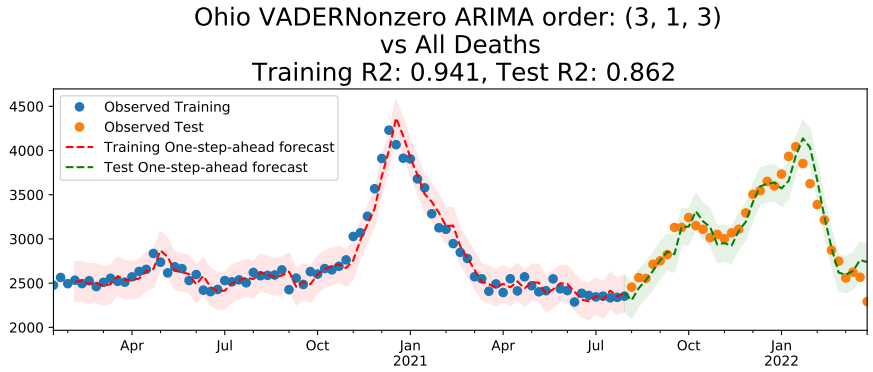
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 2)	Log Likelihood	-483.417
Date:	Mon, 17 Apr 2023	AIC	980.834
Time:	11:31:10	BIC	997.508
Sample:	01-18-2020	HQIC	987.519
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	-4361.7635	8985.339	-0.485	0.627	-2.2e+04	1.32e+04
1	1.575e+04	1.14e+04	1.377	0.169	-6669.506	3.82e+04
ar.L1	0.7459	0.320	2.328	0.020	0.118	1.374
ar.L2	-0.2109	0.303	-0.696	0.486	-0.805	0.383
ma.L1	-0.6541	0.301	-2.172	0.030	-1.244	-0.064
ma.L2	0.5036	0.239	2.108	0.035	0.035	0.972
sigma2	1.031e+04	2003.931	5.145	0.000	6381.783	1.42e+04

Ljung-Box (Q):	47.03	Jarque-Bera (JB):	1.78
Prob(Q):	0.21	Prob(JB):	0.41
Heteroskedasticity (H):	0.64	Skew:	0.36
Prob(H) (two-sided):	0.25	Kurtosis:	2.86

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.36: Chicago IL Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

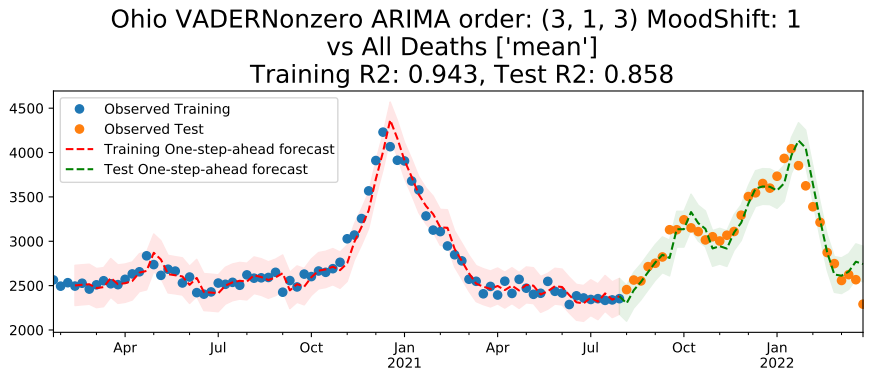
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(3, 1, 3)	Log Likelihood	-487.353
Date:	Mon, 17 Apr 2023	AIC	988.706
Time:	11:31:18	BIC	1005.380
Sample:	01-18-2020	HQIC	995.391
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2407	0.182	1.326	0.185	-0.115	0.596
ar.L2	0.5283	0.098	5.400	0.000	0.337	0.720
ar.L3	-0.6991	0.088	-7.960	0.000	-0.871	-0.527
ma.L1	-0.2650	0.229	-1.155	0.248	-0.715	0.185
ma.L2	-0.2475	0.233	-1.063	0.288	-0.704	0.209
ma.L3	0.9392	0.197	4.771	0.000	0.553	1.325
sigma2	1.078e+04	1744.481	6.180	0.000	7361.088	1.42e+04

Ljung-Box (Q):	36.11	Jarque-Bera (JB):	0.99
Prob(Q):	0.65	Prob(JB):	0.61
Heteroskedasticity (H):	1.23	Skew:	-0.12
Prob(H) (two-sided):	0.60	Kurtosis:	3.49

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.37: Cleveland OH Selected Model, No Sentiment



SARIMAX Results

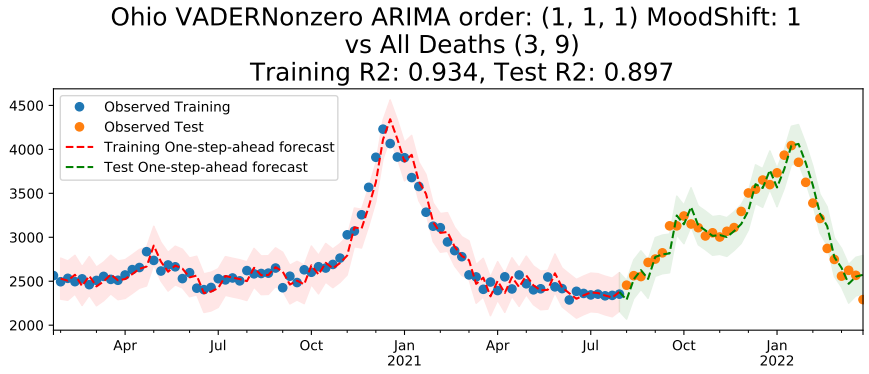
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(3, 1, 3)	Log Likelihood	-480.450
Date:	Mon, 17 Apr 2023	AIC	976.900
Time:	11:32:36	BIC	995.855
Sample:	01-25-2020	HQIC	984.494
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-281.6948	519.112	-0.543	0.587	-1299.135	735.745
ar.L1	0.2370	0.117	2.025	0.043	0.008	0.466
ar.L2	0.5253	0.060	8.789	0.000	0.408	0.642
ar.L3	-0.7115	0.089	-7.970	0.000	-0.886	-0.537
ma.L1	-0.2500	0.548	-0.456	0.648	-1.325	0.825
ma.L2	-0.2600	0.652	-0.399	0.690	-1.537	1.017
ma.L3	0.9767	0.546	1.789	0.074	-0.094	2.047
sigma2	1.045e+04	5387.430	1.940	0.052	-106.343	2.1e+04

Ljung-Box (Q):	38.11	Jarque-Bera (JB):	1.61
Prob(Q):	0.56	Prob(JB):	0.45
Heteroskedasticity (H):	1.07	Skew:	-0.10
Prob(H) (two-sided):	0.86	Kurtosis:	3.67

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.38: Cleveland OH Selected Model, Mean Vader Sentiment



SARIMAX Results

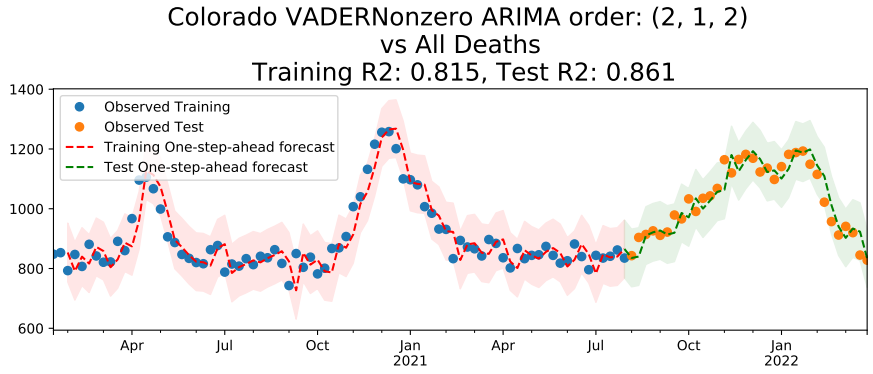
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 1)	Log Likelihood	-485.351
Date:	Mon, 17 Apr 2023	AIC	980.702
Time:	11:31:14	BIC	992.550
Sample:	01-25-2020	HQIC	985.449
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	6248.3446	3987.834	1.567	0.117	-1567.666	1.41e+04
1	7131.4160	4271.115	1.670	0.095	-1239.816	1.55e+04
ar.L1	0.7578	0.159	4.763	0.000	0.446	1.070
ma.L1	-0.5246	0.241	-2.177	0.029	-0.997	-0.052
sigma2	1.267e+04	2279.393	5.558	0.000	8200.390	1.71e+04

Ljung-Box (Q):	78.39	Jarque-Bera (JB):	0.01
Prob(Q):	0.00	Prob(JB):	0.99
Heteroskedasticity (H):	1.50	Skew:	-0.02
Prob(H) (two-sided):	0.31	Kurtosis:	2.95

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.39: Cleveland OH Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

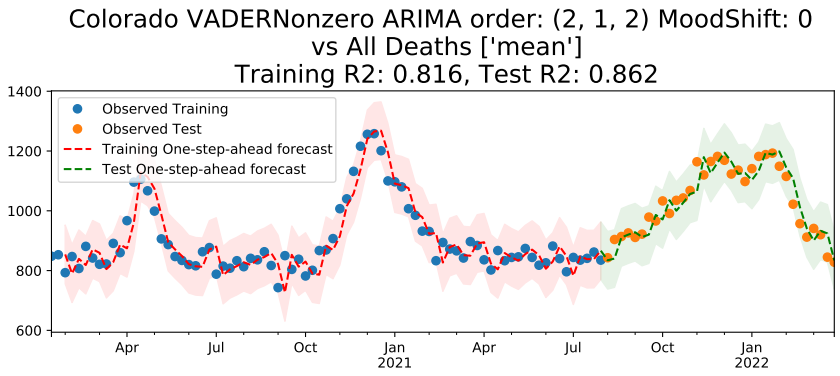
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 2)	Log Likelihood	-425.382
Date:	Mon, 17 Apr 2023	AIC	860.765
Time:	11:31:24	BIC	872.675
Sample:	01-18-2020	HQIC	865.540
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3419	0.503	-0.680	0.497	-1.328	0.644
ar.L2	0.4775	0.500	0.954	0.340	-0.503	1.458
ma.L1	0.3990	0.534	0.746	0.455	-0.649	1.447
ma.L2	-0.3134	0.537	-0.584	0.559	-1.366	0.739
sigma2	2428.1309	422.839	5.742	0.000	1599.382	3256.880

Ljung-Box (Q):	37.62	Jarque-Bera (JB):	1.66
Prob(Q):	0.58	Prob(JB):	0.44
Heteroskedasticity (H):	0.56	Skew:	0.35
Prob(H) (two-sided):	0.14	Kurtosis:	2.97

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.40: Denver CO Selected Model, No Sentiment



SARIMAX Results

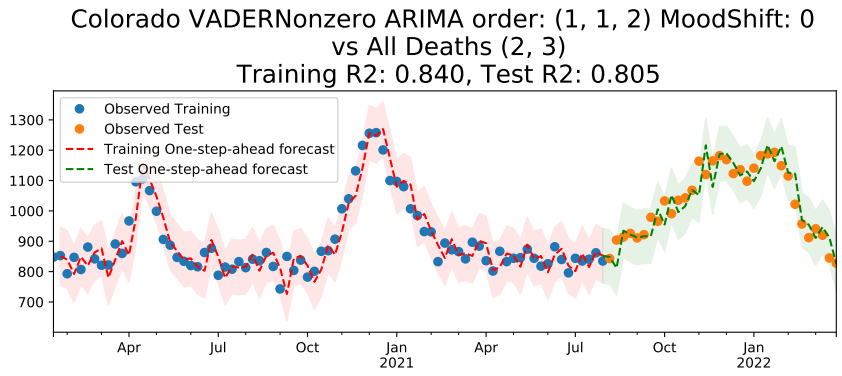
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 2)	Log Likelihood	-425.225
Date:	Mon, 17 Apr 2023	AIC	862.450
Time:	11:32:38	BIC	876.742
Sample:	01-18-2020	HQIC	868.180
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-47.7518	177.854	-0.268	0.788	-396.339	300.836
ar.L1	-0.3490	0.497	-0.703	0.482	-1.322	0.624
ar.L2	0.4755	0.490	0.971	0.332	-0.485	1.436
ma.L1	0.4078	0.529	0.771	0.441	-0.628	1.444
ma.L2	-0.3056	0.527	-0.579	0.562	-1.339	0.728
sigma2	2418.0437	421.504	5.737	0.000	1591.912	3244.175

Ljung-Box (Q):	35.62	Jarque-Bera (JB):	1.77
Prob(Q):	0.67	Prob(JB):	0.41
Heteroskedasticity (H):	0.57	Skew:	0.36
Prob(H) (two-sided):	0.15	Kurtosis:	2.98

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.41: Denver CO Selected Model, Mean Vader Sentiment



SARIMAX Results

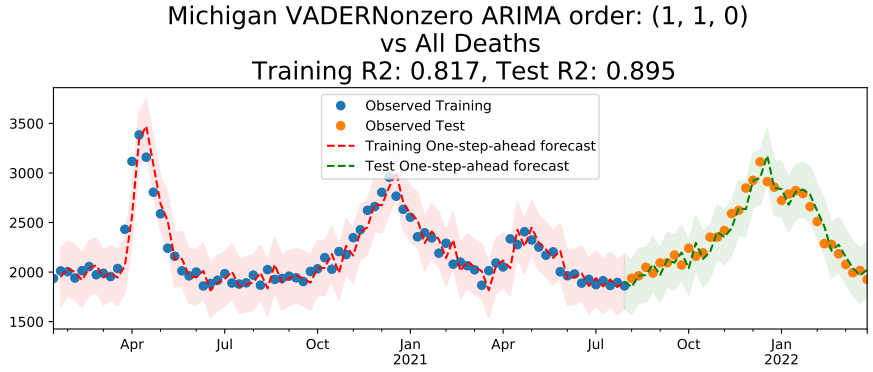
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 2)	Log Likelihood	-419.713
Date:	Mon, 17 Apr 2023	AIC	851.426
Time:	11:31:20	BIC	865.718
Sample:	01-18-2020	HQIC	857.156
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	2516.7164	1182.279	2.129	0.033	199.493	4833.940
1	-2682.6247	1499.280	-1.789	0.074	-5621.159	255.910
ar.L1	-0.9252	0.081	-11.493	0.000	-1.083	-0.767
ma.L1	1.1655	0.125	9.316	0.000	0.920	1.411
ma.L2	0.3388	0.107	3.160	0.002	0.129	0.549
sigma2	2096.6925	355.172	5.903	0.000	1400.569	2792.816

Ljung-Box (Q):	50.50	Jarque-Bera (JB):	4.57
Prob(Q):	0.12	Prob(JB):	0.10
Heteroskedasticity (H):	0.42	Skew:	0.58
Prob(H) (two-sided):	0.03	Kurtosis:	2.96

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.42: Denver CO Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

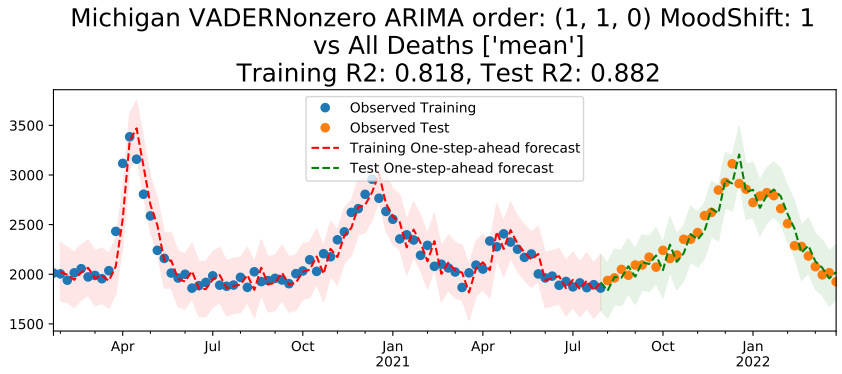
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-509.974
Date:	Mon, 17 Apr 2023	AIC	1023.949
Time:	11:31:28	BIC	1028.713
Sample:	01-18-2020	HQIC	1025.859
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3242	0.093	3.493	0.000	0.142	0.506
sigma2	2.015e+04	2829.713	7.122	0.000	1.46e+04	2.57e+04

Ljung-Box (Q):	20.25	Jarque-Bera (JB):	22.12
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.50	Skew:	0.69
Prob(H) (two-sided):	0.08	Kurtosis:	5.17

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.43: Detroit MI Selected Model, No Sentiment



SARIMAX Results

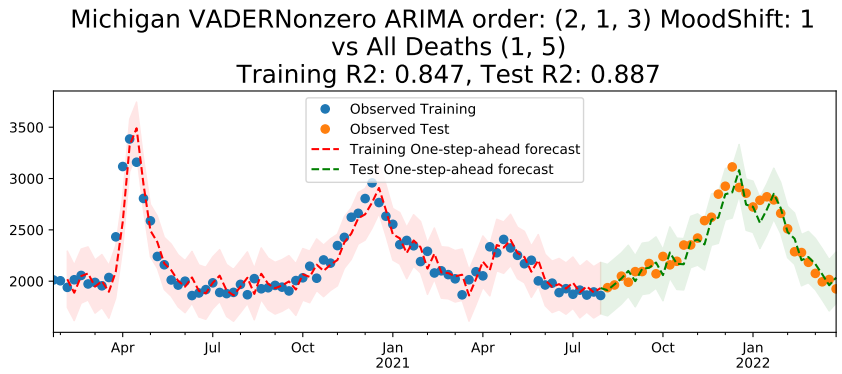
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 0)	Log Likelihood	-503.610
Date:	Mon, 17 Apr 2023	AIC	1013.220
Time:	11:32:40	BIC	1020.329
Sample:	01-25-2020	HQIC	1016.068
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	519.4541	713.107	0.728	0.466	-878.210	1917.118
ar.L1	0.3263	0.093	3.511	0.000	0.144	0.509
sigma2	2.013e+04	2827.508	7.119	0.000	1.46e+04	2.57e+04

Ljung-Box (Q):	20.21	Jarque-Bera (JB):	21.40
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.51	Skew:	0.69
Prob(H) (two-sided):	0.09	Kurtosis:	5.15

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.44: Detroit MI Selected Model, Mean Vader Sentiment



SARIMAX Results

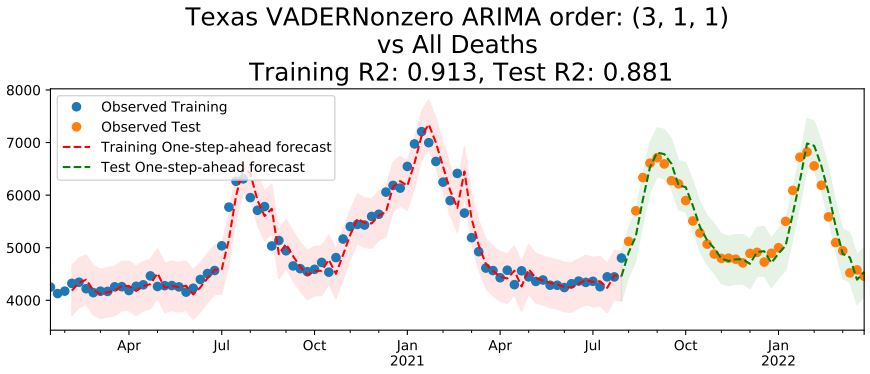
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(2, 1, 3)	Log Likelihood	-496.407
Date:	Mon, 17 Apr 2023	AIC	1008.814
Time:	11:31:26	BIC	1027.770
Sample:	01-25-2020	HQIC	1016.409
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	1876.4309	4674.762	0.401	0.688	-7285.934	1.1e+04
1	-7613.0411	7176.475	-1.061	0.289	-2.17e+04	6452.592
ar.L1	1.1669	0.340	3.435	0.001	0.501	1.833
ar.L2	-0.3559	0.325	-1.094	0.274	-0.994	0.282
ma.L1	-0.9938	36.266	-0.027	0.978	-72.074	70.087
ma.L2	0.3570	0.367	0.972	0.331	-0.363	1.077
ma.L3	-0.3631	13.210	-0.027	0.978	-26.254	25.528
sigma2	1.615e+04	5.87e+05	0.028	0.978	-1.13e+06	1.17e+06

Ljung-Box (Q):	14.24	Jarque-Bera (JB):	32.75
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.47	Skew:	0.86
Prob(H) (two-sided):	0.06	Kurtosis:	5.64

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.45: Detroit MI Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

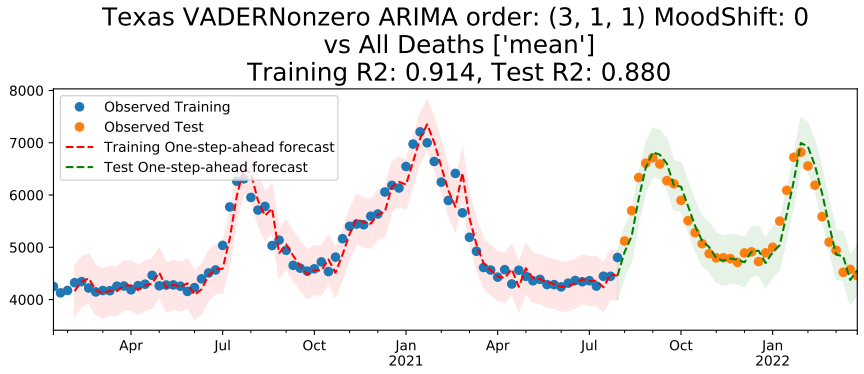
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(3, 1, 1)	Log Likelihood	-552.042
Date:	Mon, 17 Apr 2023	AIC	1114.084
Time:	11:31:32	BIC	1125.994
Sample:	01-18-2020	HQIC	1118.859
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5753	0.395	-1.455	0.146	-1.350	0.200
ar.L2	0.2999	0.172	1.745	0.081	-0.037	0.637
ar.L3	0.1948	0.128	1.523	0.128	-0.056	0.446
ma.L1	0.7884	0.386	2.044	0.041	0.032	1.545
sigma2	5.738e+04	7429.146	7.724	0.000	4.28e+04	7.19e+04

Ljung-Box (Q):	27.01	Jarque-Bera (JB):	11.22
Prob(Q):	0.94	Prob(JB):	0.00
Heteroskedasticity (H):	1.96	Skew:	-0.29
Prob(H) (two-sided):	0.09	Kurtosis:	4.74

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.46: Houston TX Selected Model, No Sentiment



SARIMAX Results

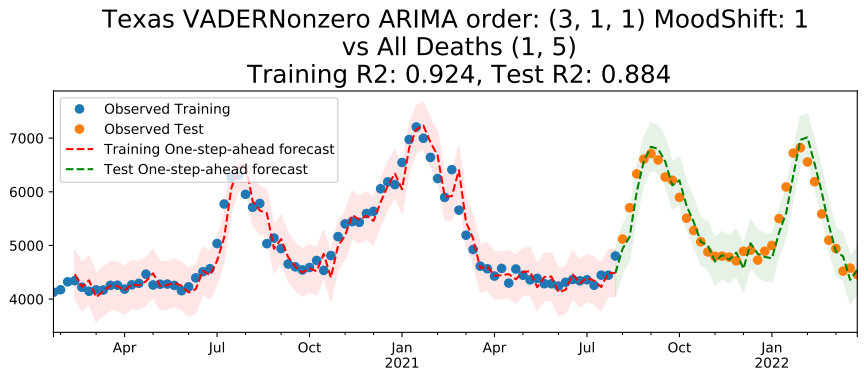
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(3, 1, 1)	Log Likelihood	-551.690
Date:	Mon, 17 Apr 2023	AIC	1115.381
Time:	11:32:42	BIC	1129.673
Sample:	01-18-2020	HQIC	1121.111
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-531.7725	1188.814	-0.447	0.655	-2861.806	1798.261
ar.L1	-0.5831	0.302	-1.930	0.054	-1.175	0.009
ar.L2	0.2934	0.158	1.855	0.064	-0.017	0.603
ar.L3	0.1975	0.123	1.602	0.109	-0.044	0.439
ma.L1	0.8114	0.303	2.676	0.007	0.217	1.406
sigma2	5.684e+04	8634.952	6.582	0.000	3.99e+04	7.38e+04

Ljung-Box (Q):	27.63	Jarque-Bera (JB):	8.95
Prob(Q):	0.93	Prob(JB):	0.01
Heteroskedasticity (H):	1.86	Skew:	-0.30
Prob(H) (two-sided):	0.11	Kurtosis:	4.53

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.47: Houston TX Selected Model, Mean Vader Sentiment



SARIMAX Results

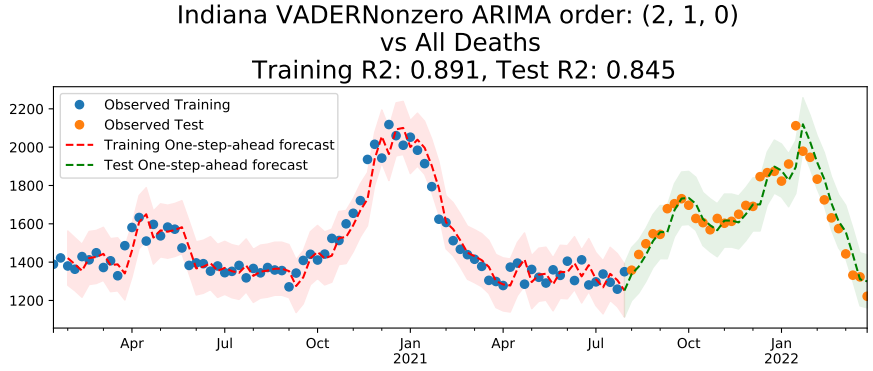
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(3, 1, 1)	Log Likelihood	-539.985
Date:	Mon, 17 Apr 2023	AIC	1093.970
Time:	11:31:30	BIC	1110.556
Sample:	01-25-2020	HQIC	1100.615
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	1.286e+04	8942.853	1.438	0.151	-4670.358	3.04e+04
1	-2.54e+04	9047.271	-2.807	0.005	-4.31e+04	-7663.911
ar.L1	-0.7037	0.104	-6.777	0.000	-0.907	-0.500
ar.L2	0.3882	0.141	2.753	0.006	0.112	0.665
ar.L3	0.2434	0.096	2.548	0.011	0.056	0.431
ma.L1	0.9999	0.133	7.533	0.000	0.740	1.260
sigma2	4.942e+04	51.215	964.971	0.000	4.93e+04	4.95e+04

Ljung-Box (Q):	33.47	Jarque-Bera (JB):	11.42
Prob(Q):	0.76	Prob(JB):	0.00
Heteroskedasticity (H):	1.91	Skew:	-0.37
Prob(H) (two-sided):	0.10	Kurtosis:	4.71

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.04e+21. Standard

FIGURE A.48: Houston TX Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

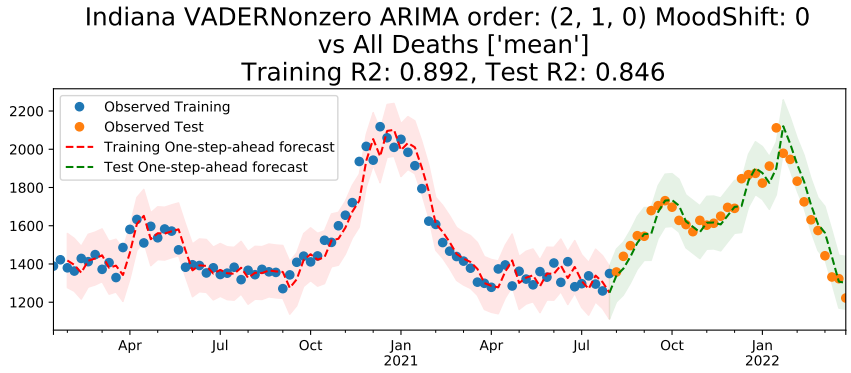
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 0)	Log Likelihood	-454.295
Date:	Mon, 17 Apr 2023	AIC	914.589
Time:	11:31:37	BIC	921.735
Sample:	01-18-2020	HQIC	917.454
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0566	0.103	-0.552	0.581	-0.258	0.144
ar.L2	0.2120	0.101	2.100	0.036	0.014	0.410
sigma2	5003.5181	806.449	6.204	0.000	3422.908	6584.128

Ljung-Box (Q):	43.70	Jarque-Bera (JB):	0.50
Prob(Q):	0.32	Prob(JB):	0.78
Heteroskedasticity (H):	1.14	Skew:	0.19
Prob(H) (two-sided):	0.73	Kurtosis:	3.01

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.49: Indianapolis IN Selected Model, No Sentiment



SARIMAX Results

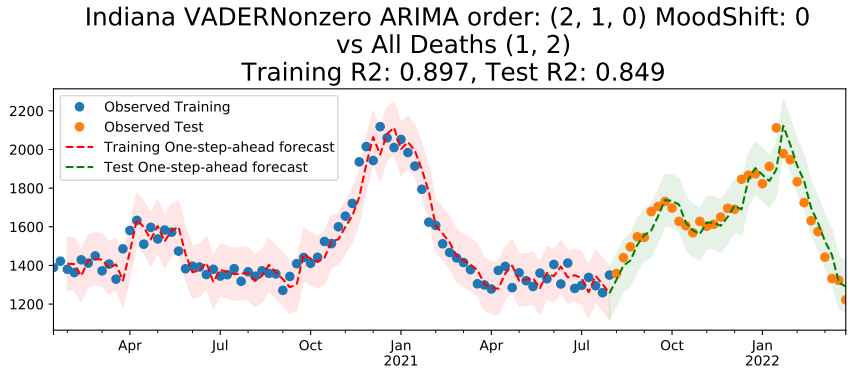
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 0)	Log Likelihood	-454.152
Date:	Mon, 17 Apr 2023	AIC	916.304
Time:	11:32:43	BIC	925.832
Sample:	01-18-2020	HQIC	920.124
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	123.8708	302.213	0.410	0.682	-468.456	716.198
ar.L1	-0.0561	0.103	-0.542	0.588	-0.259	0.147
ar.L2	0.2150	0.103	2.080	0.037	0.012	0.418
sigma2	4985.5790	807.922	6.171	0.000	3402.081	6569.077

Ljung-Box (Q):	46.00	Jarque-Bera (JB):	0.46
Prob(Q):	0.24	Prob(JB):	0.79
Heteroskedasticity (H):	1.14	Skew:	0.19
Prob(H) (two-sided):	0.73	Kurtosis:	2.98

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.50: Indianapolis IN Selected Model, Mean Vader Sentiment



SARIMAX Results

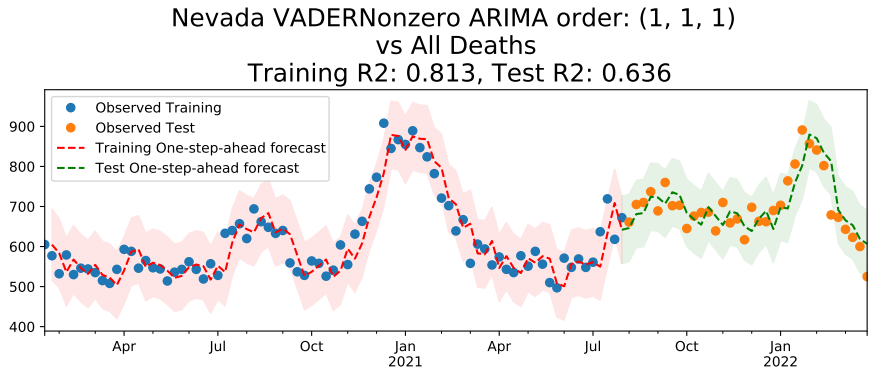
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 0)	Log Likelihood	-451.995
Date:	Mon, 17 Apr 2023	AIC	913.991
Time:	11:31:35	BIC	925.901
Sample:	01-18-2020	HQIC	918.766
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	-2624.8898	1644.811	-1.596	0.111	-5848.660	598.880
1	-3359.0538	2695.652	-1.246	0.213	-8642.435	1924.327
ar.L1	-0.0302	0.098	-0.308	0.758	-0.222	0.162
ar.L2	0.2198	0.099	2.221	0.026	0.026	0.414
sigma2	4724.3778	817.004	5.783	0.000	3123.079	6325.677

Ljung-Box (Q):	42.83	Jarque-Bera (JB):	1.06
Prob(Q):	0.35	Prob(JB):	0.59
Heteroskedasticity (H):	1.00	Skew:	0.26
Prob(H) (two-sided):	0.99	Kurtosis:	2.79

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.51: Indianapolis IN Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

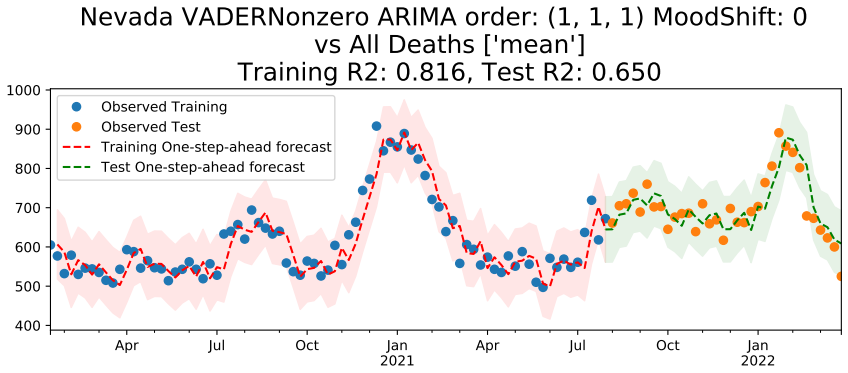
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 1)	Log Likelihood	-414.596
Date:	Mon, 17 Apr 2023	AIC	835.193
Time:	11:31:43	BIC	842.339
Sample:	01-18-2020	HQIC	838.058
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8718	0.136	-6.389	0.000	-1.139	-0.604
ma.L1	0.7085	0.186	3.815	0.000	0.345	1.072
sigma2	1852.6568	288.316	6.426	0.000	1287.567	2417.746

Ljung-Box (Q):	44.93	Jarque-Bera (JB):	1.40
Prob(Q):	0.27	Prob(JB):	0.50
Heteroskedasticity (H):	2.20	Skew:	0.31
Prob(H) (two-sided):	0.05	Kurtosis:	3.17

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.52: Las Vegas NV Selected Model, No Sentiment



SARIMAX Results

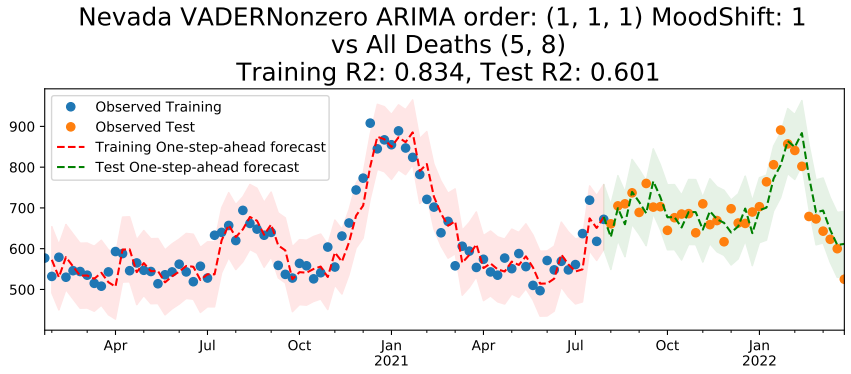
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 1)	Log Likelihood	-413.759
Date:	Mon, 17 Apr 2023	AIC	835.519
Time:	11:32:45	BIC	845.047
Sample:	01-18-2020	HQIC	839.339
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-226.1283	272.327	-0.830	0.406	-759.879	307.623
ar.L1	-0.8846	0.125	-7.081	0.000	-1.129	-0.640
ma.L1	0.7206	0.179	4.016	0.000	0.369	1.072
sigma2	1813.6910	305.238	5.942	0.000	1215.436	2411.946

Ljung-Box (Q):	48.36	Jarque-Bera (JB):	1.30
Prob(Q):	0.17	Prob(JB):	0.52
Heteroskedasticity (H):	2.10	Skew:	0.31
Prob(H) (two-sided):	0.06	Kurtosis:	3.08

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.53: Las Vegas NV Selected Model, Mean Vader Sentiment



SARIMAX Results

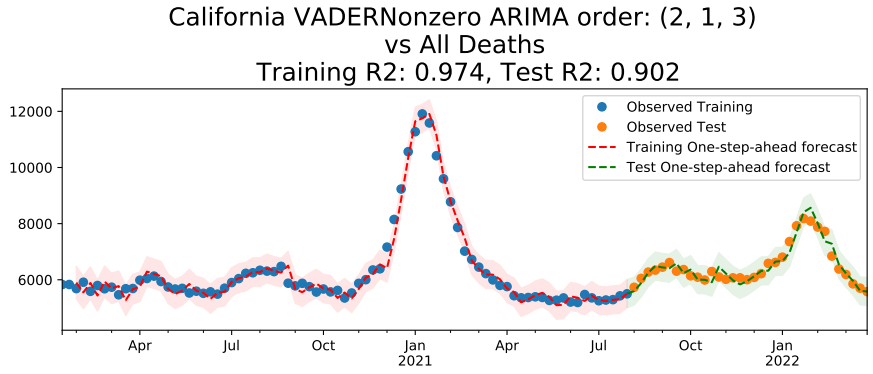
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 1)	Log Likelihood	-404.989
Date:	Mon, 17 Apr 2023	AIC	819.978
Time:	11:31:40	BIC	831.826
Sample:	01-25-2020	HQIC	824.725
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	5256.1917	2192.830	2.397	0.017	958.323	9554.060
1	5104.9294	2201.339	2.319	0.020	790.384	9419.475
ar.L1	-0.7762	0.332	-2.339	0.019	-1.427	-0.126
ma.L1	0.6535	0.389	1.679	0.093	-0.109	1.416
sigma2	1659.5425	259.808	6.388	0.000	1150.329	2168.756

Ljung-Box (Q):	42.42	Jarque-Bera (JB):	1.35
Prob(Q):	0.37	Prob(JB):	0.51
Heteroskedasticity (H):	1.44	Skew:	0.29
Prob(H) (two-sided):	0.35	Kurtosis:	3.29

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.54: Las Vegas NV Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

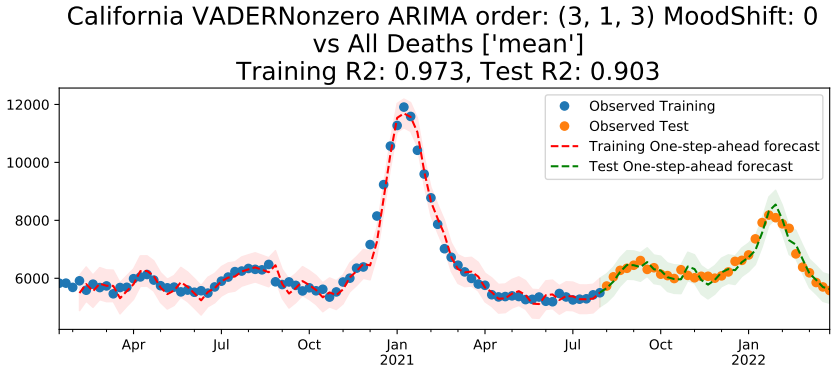
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 3)	Log Likelihood	-555.144
Date:	Mon, 17 Apr 2023	AIC	1122.289
Time:	11:31:48	BIC	1136.581
Sample:	01-18-2020	HQIC	1128.019
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8312	0.236	3.524	0.000	0.369	1.294
ar.L2	-0.4065	0.208	-1.950	0.051	-0.815	0.002
ma.L1	-0.3849	0.582	-0.662	0.508	-1.525	0.755
ma.L2	0.7246	1.338	0.542	0.588	-1.897	3.346
ma.L3	0.3606	0.611	0.590	0.555	-0.838	1.559
sigma2	6.36e+04	9.43e+04	0.675	0.500	-1.21e+05	2.48e+05

Ljung-Box (Q):	29.08	Jarque-Bera (JB):	6.74
Prob(Q):	0.90	Prob(JB):	0.03
Heteroskedasticity (H):	0.66	Skew:	0.06
Prob(H) (two-sided):	0.29	Kurtosis:	4.42

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.55: Los Angeles CA LA Selected Model, No Sentiment



SARIMAX Results

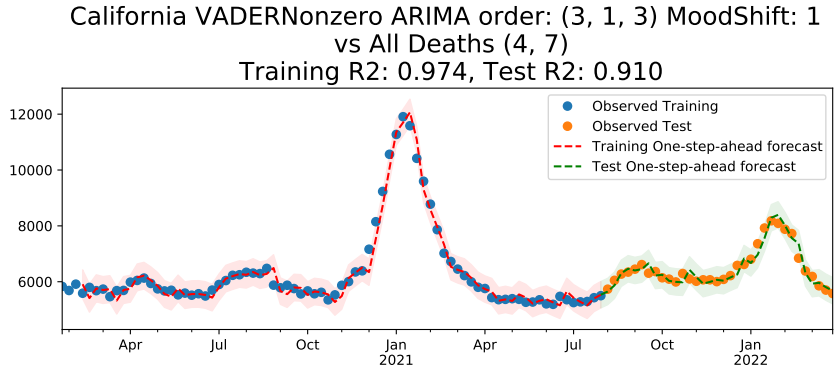
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(3, 1, 3)	Log Likelihood	-555.926
Date:	Mon, 17 Apr 2023	AIC	1127.852
Time:	11:32:48	BIC	1146.908
Sample:	01-18-2020 - 07-31-2021	HQIC	1135.492
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-380.9485	631.705	-0.603	0.546	-1619.067	857.170
ar.L1	1.9679	0.128	15.369	0.000	1.717	2.219
ar.L2	-1.4542	0.252	-5.776	0.000	-1.948	-0.961
ar.L3	0.4486	0.154	2.905	0.004	0.146	0.751
ma.L1	-1.7272	0.158	-10.931	0.000	-2.037	-1.418
ma.L2	1.6944	0.241	7.017	0.000	1.221	2.168
ma.L3	-0.9670	0.170	-5.688	0.000	-1.300	-0.634
sigma2	6.281e+04	0.065	9.62e+05	0.000	6.28e+04	6.28e+04

Ljung-Box (Q):	36.21	Jarque-Bera (JB):	16.23
Prob(Q):	0.64	Prob(JB):	0.00
Heteroskedasticity (H):	0.86	Skew:	0.51
Prob(H) (two-sided):	0.70	Kurtosis:	4.96

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.4e+23. Standard er

FIGURE A.56: Los Angeles CA LA Selected Model, Mean Vader Sentiment



SARIMAX Results

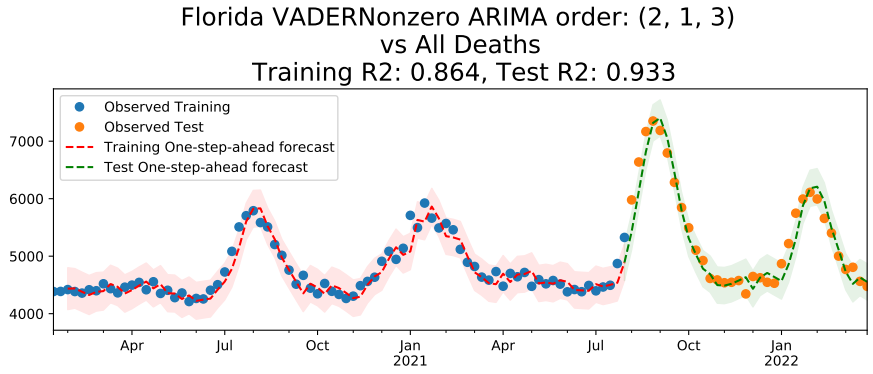
Dep. Variable:	All Cause	No. Observations:	80			
Model:	ARIMA(3, 1, 3)	Log Likelihood	-548.384			
Date:	Mon, 17 Apr 2023	AIC	1114.768			
Time:	11:31:46	BIC	1136.093			
Sample:	01-25-2020	HQIC	1123.311			
	- 07-31-2021					
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
0	1.472e+04	1.74e+04	0.847	0.397	-1.93e+04	4.88e+04
1	2.107e+04	1.38e+04	1.526	0.127	-5983.709	4.81e+04
ar.L1	2.1001	0.342	6.132	0.000	1.429	2.771
ar.L2	-1.5268	0.617	-2.475	0.013	-2.736	-0.317
ar.L3	0.3902	0.301	1.296	0.195	-0.200	0.980
ma.L1	-1.6623	0.316	-5.253	0.000	-2.282	-1.042
ma.L2	1.1519	0.388	2.972	0.003	0.392	1.912
ma.L3	-0.4896	0.146	-3.361	0.001	-0.775	-0.204
sigma2	5.742e+04	1.897	3.03e+04	0.000	5.74e+04	5.74e+04

Ljung-Box (Q):	31.03	Jarque-Bera (JB):	6.83
Prob(Q):	0.84	Prob(JB):	0.03
Heteroskedasticity (H):	0.67	Skew:	0.26
Prob(H) (two-sided):	0.31	Kurtosis:	4.34

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 2.8e+23. Standard er

FIGURE A.57: Los Angeles CA LA Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

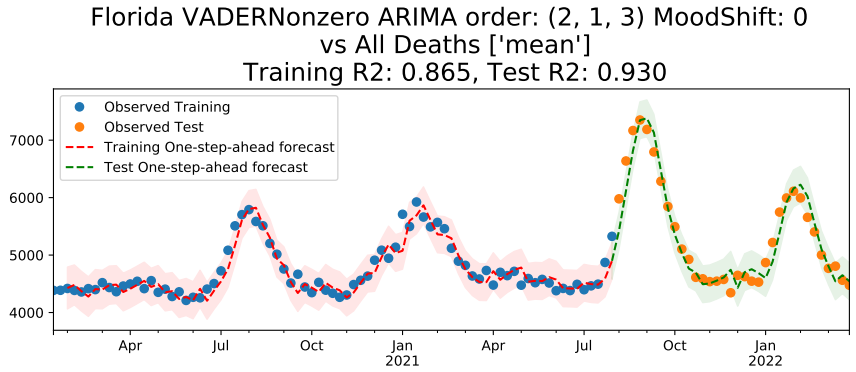
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 3)	Log Likelihood	-521.667
Date:	Mon, 17 Apr 2023	AIC	1055.334
Time:	11:31:57	BIC	1069.626
Sample:	01-18-2020	HQIC	1061.064
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.7210	0.122	14.050	0.000	1.481	1.961
ar.L2	-0.8015	0.126	-6.374	0.000	-1.048	-0.555
ma.L1	-1.8028	0.314	-5.740	0.000	-2.418	-1.187
ma.L2	1.1785	0.320	3.682	0.000	0.551	1.806
ma.L3	-0.3739	0.169	-2.206	0.027	-0.706	-0.042
sigma2	2.571e+04	7338.779	3.504	0.000	1.13e+04	4.01e+04

Ljung-Box (Q):	30.53	Jarque-Bera (JB):	18.57
Prob(Q):	0.86	Prob(JB):	0.00
Heteroskedasticity (H):	1.62	Skew:	0.90
Prob(H) (two-sided):	0.22	Kurtosis:	4.54

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.58: Miami FL Selected Model, No Sentiment



SARIMAX Results

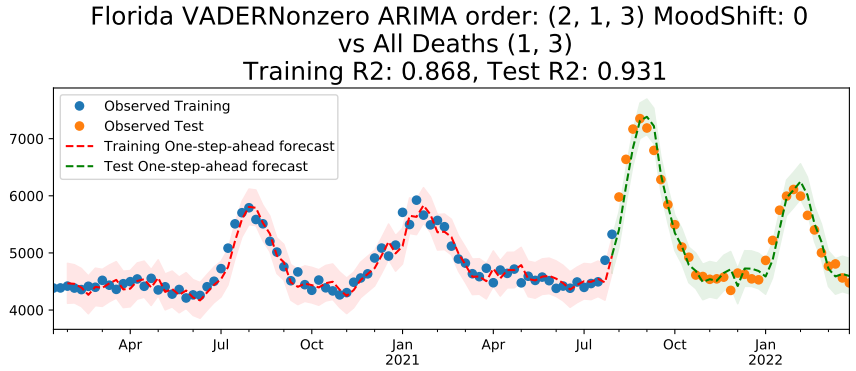
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 3)	Log Likelihood	-521.338
Date:	Mon, 17 Apr 2023	AIC	1056.676
Time:	11:32:50	BIC	1073.350
Sample:	01-18-2020	HQIC	1063.361
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	1268.7429	772.909	1.642	0.101	-246.131	2783.617
ar.L1	1.7673	0.108	16.385	0.000	1.556	1.979
ar.L2	-0.8497	0.108	-7.876	0.000	-1.061	-0.638
ma.L1	-1.8055	0.166	-10.847	0.000	-2.132	-1.479
ma.L2	1.1246	0.286	3.939	0.000	0.565	1.684
ma.L3	-0.2988	0.167	-1.788	0.074	-0.626	0.029
sigma2	2.672e+04	3831.810	6.974	0.000	1.92e+04	3.42e+04

Ljung-Box (Q):	37.43	Jarque-Bera (JB):	18.79
Prob(Q):	0.59	Prob(JB):	0.00
Heteroskedasticity (H):	1.34	Skew:	0.85
Prob(H) (two-sided):	0.46	Kurtosis:	4.67

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.59: Miami FL Selected Model, Mean Vader Sentiment



SARIMAX Results

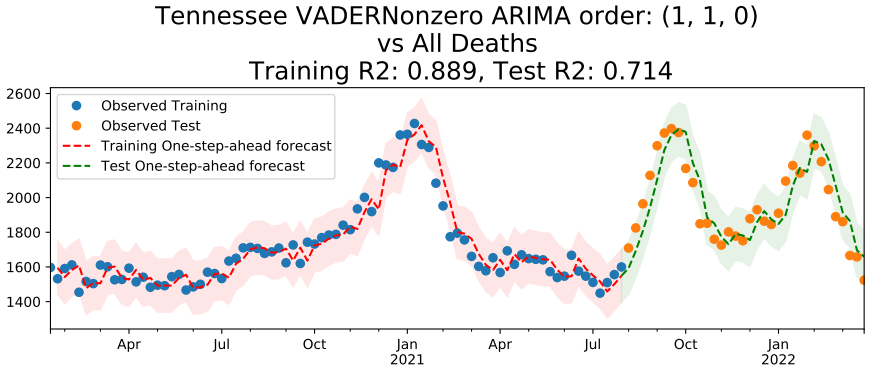
Dep. Variable:	All Cause	No. Observations:	81			
Model:	ARIMA(2, 1, 3)	Log Likelihood	-520.338			
Date:	Mon, 17 Apr 2023	AIC	1056.676			
Time:	11:31:53	BIC	1075.732			
Sample:	01-18-2020	HQIC	1064.316			
	- 07-31-2021					
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
0	-1151.4115	4788.292	-0.240	0.810	-1.05e+04	8233.467
1	-1.413e+04	7584.490	-1.864	0.062	-2.9e+04	731.144
ar.L1	1.7533	0.131	13.350	0.000	1.496	2.011
ar.L2	-0.8338	0.130	-6.435	0.000	-1.088	-0.580
ma.L1	-1.7577	0.183	-9.596	0.000	-2.117	-1.399
ma.L2	1.0523	0.242	4.343	0.000	0.577	1.527
ma.L3	-0.2728	0.139	-1.969	0.049	-0.544	-0.001
sigma2	2.538e+04	3827.255	6.631	0.000	1.79e+04	3.29e+04

Ljung-Box (Q):	29.21	Jarque-Bera (JB):	10.71
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	1.43	Skew:	0.69
Prob(H) (two-sided):	0.36	Kurtosis:	4.13

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.60: Miami FL Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

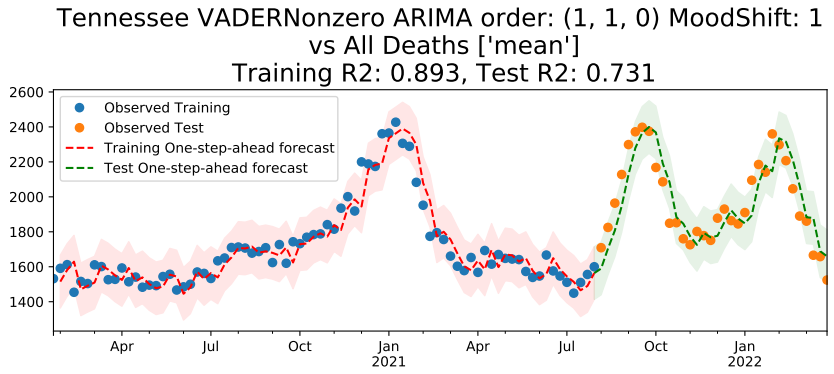
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-461.922
Date:	Mon, 17 Apr 2023	AIC	927.843
Time:	11:32:01	BIC	932.607
Sample:	01-18-2020	HQIC	929.753
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.1521	0.105	-1.454	0.146	-0.357	0.053
sigma2	6089.2577	742.602	8.200	0.000	4633.785	7544.730

Ljung-Box (Q):	31.50	Jarque-Bera (JB):	6.70
Prob(Q):	0.83	Prob(JB):	0.04
Heteroskedasticity (H):	2.13	Skew:	0.12
Prob(H) (two-sided):	0.06	Kurtosis:	4.40

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.61: Nashville TN Selected Model, No Sentiment



SARIMAX Results

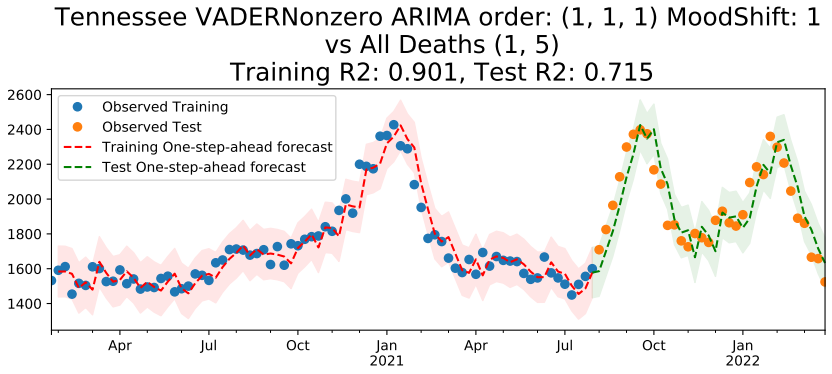
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 0)	Log Likelihood	-455.041
Date:	Mon, 17 Apr 2023	AIC	916.082
Time:	11:32:52	BIC	923.191
Sample:	01-25-2020	HQIC	918.930
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	578.1290	339.760	1.702	0.089	-87.788	1244.046
ar.L1	-0.0932	0.114	-0.814	0.416	-0.318	0.131
sigma2	5898.1265	736.928	8.004	0.000	4453.774	7342.479

Ljung-Box (Q):	33.43	Jarque-Bera (JB):	5.79
Prob(Q):	0.76	Prob(JB):	0.06
Heteroskedasticity (H):	1.55	Skew:	-0.03
Prob(H) (two-sided):	0.27	Kurtosis:	4.32

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.62: Nashville TN Selected Model, Mean Vader Sentiment



SARIMAX Results

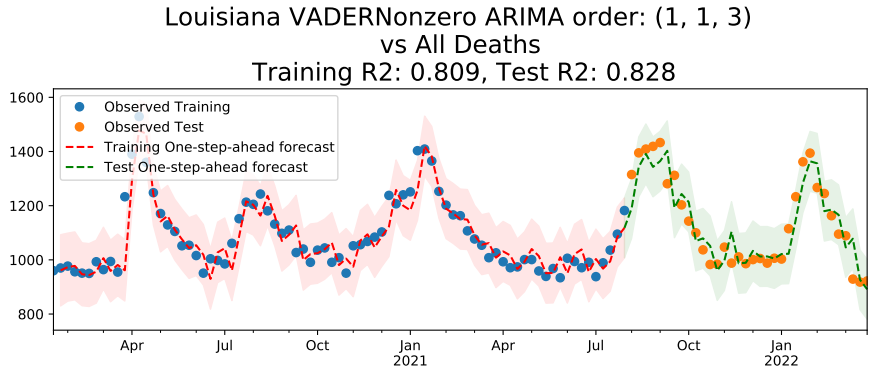
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 1)	Log Likelihood	-451.908
Date:	Mon, 17 Apr 2023	AIC	913.816
Time:	11:31:59	BIC	925.663
Sample:	01-25-2020	HQIC	918.563
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	-2315.7670	2519.812	-0.919	0.358	-7254.508	2622.973
1	-1.005e+04	3867.092	-2.598	0.009	-1.76e+04	-2466.686
ar.L1	-0.7450	0.466	-1.599	0.110	-1.658	0.168
ma.L1	0.6501	0.521	1.248	0.212	-0.371	1.671
sigma2	5446.4605	710.362	7.667	0.000	4054.177	6838.744

Ljung-Box (Q):	35.26	Jarque-Bera (JB):	6.72
Prob(Q):	0.68	Prob(JB):	0.03
Heteroskedasticity (H):	1.73	Skew:	0.20
Prob(H) (two-sided):	0.17	Kurtosis:	4.37

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.63: Nashville TN Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

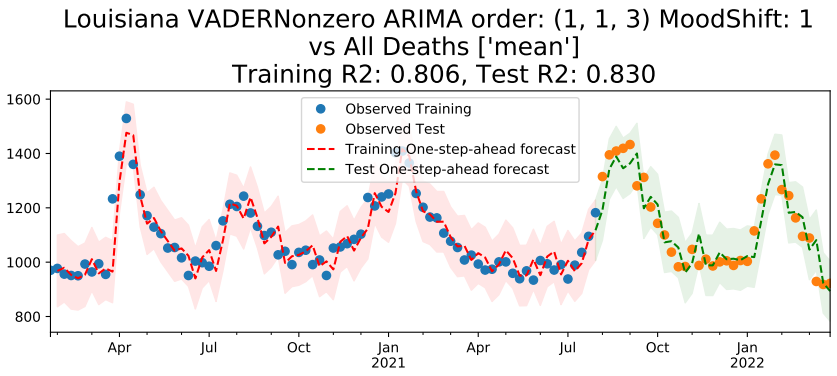
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 3)	Log Likelihood	-437.241
Date:	Mon, 17 Apr 2023	AIC	884.481
Time:	11:32:07	BIC	896.392
Sample:	01-18-2020	HQIC	889.257
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6670	0.110	6.047	0.000	0.451	0.883
ma.L1	-0.4494	16.594	-0.027	0.978	-32.973	32.074
ma.L2	-0.0059	9.141	-0.001	0.999	-17.922	17.910
ma.L3	-0.5444	9.040	-0.060	0.952	-18.263	17.174
sigma2	3144.5929	5.21e+04	0.060	0.952	-9.89e+04	1.05e+05

Ljung-Box (Q):	34.02	Jarque-Bera (JB):	90.32
Prob(Q):	0.74	Prob(JB):	0.00
Heteroskedasticity (H):	0.31	Skew:	1.36
Prob(H) (two-sided):	0.00	Kurtosis:	7.43

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.64: New Orleans LA Selected Model, No Sentiment



SARIMAX Results

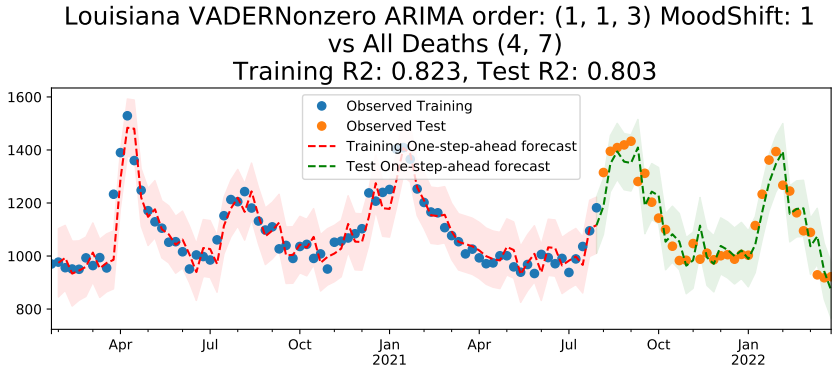
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 3)	Log Likelihood	-432.208
Date:	Mon, 17 Apr 2023	AIC	876.415
Time:	11:32:54	BIC	890.632
Sample:	01-25-2020	HQIC	882.111
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	89.2223	240.326	0.371	0.710	-381.808	560.252
ar.L1	0.6659	0.114	5.864	0.000	0.443	0.888
ma.L1	-0.4499	0.897	-0.502	0.616	-2.208	1.308
ma.L2	-0.0074	0.496	-0.015	0.988	-0.980	0.965
ma.L3	-0.5375	0.488	-1.100	0.271	-1.495	0.420
sigma2	3189.1752	2655.542	1.201	0.230	-2015.592	8393.943

Ljung-Box (Q):	34.30	Jarque-Bera (JB):	76.19
Prob(Q):	0.72	Prob(JB):	0.00
Heteroskedasticity (H):	0.33	Skew:	1.29
Prob(H) (two-sided):	0.01	Kurtosis:	7.06

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.65: New Orleans LA Selected Model, Mean Vader Sentiment



SARIMAX Results

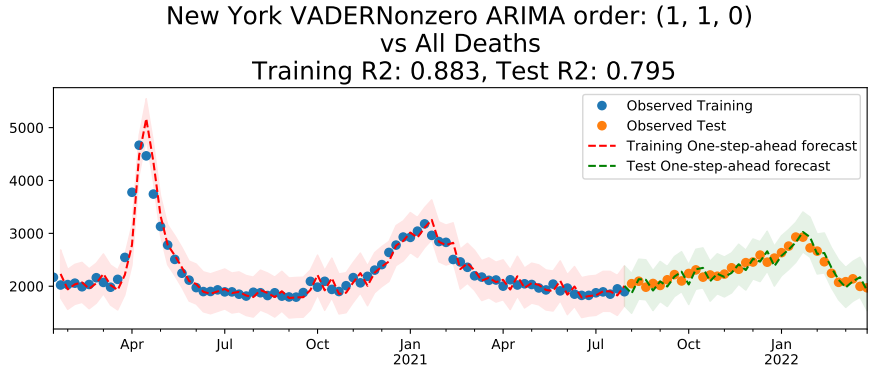
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 3)	Log Likelihood	-428.629
Date:	Mon, 17 Apr 2023	AIC	871.257
Time:	11:32:05	BIC	887.843
Sample:	01-25-2020	HQIC	877.902
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	1370.5262	2059.888	0.665	0.506	-2666.780	5407.833
1	4188.0983	2117.324	1.978	0.048	38.220	8337.976
ar.L1	0.6667	0.104	6.414	0.000	0.463	0.870
ma.L1	-0.4261	3.993	-0.107	0.915	-8.253	7.401
ma.L2	0.0274	2.312	0.012	0.991	-4.504	4.558
ma.L3	-0.5999	2.419	-0.248	0.804	-5.341	4.141
sigma2	2897.1253	1.16e+04	0.250	0.803	-1.98e+04	2.56e+04

Ljung-Box (Q):	26.45	Jarque-Bera (JB):	48.05
Prob(Q):	0.95	Prob(JB):	0.00
Heteroskedasticity (H):	0.34	Skew:	1.02
Prob(H) (two-sided):	0.01	Kurtosis:	6.23

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.66: New Orleans LA Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

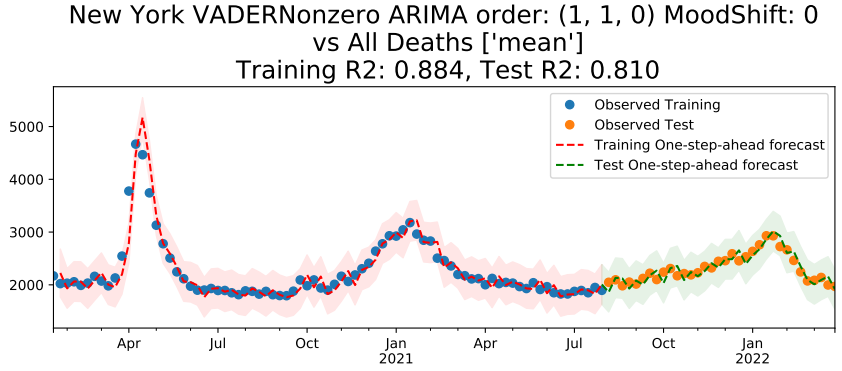
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-535.103
Date:	Mon, 17 Apr 2023	AIC	1074.206
Time:	11:32:09	BIC	1078.970
Sample:	01-18-2020	HQIC	1076.116
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5646	0.046	12.240	0.000	0.474	0.655
sigma2	3.739e+04	2474.208	15.113	0.000	3.25e+04	4.22e+04

Ljung-Box (Q):	23.49	Jarque-Bera (JB):	315.97
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	0.16	Skew:	0.70
Prob(H) (two-sided):	0.00	Kurtosis:	12.63

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.67: New York NY Selected Model, No Sentiment



SARIMAX Results

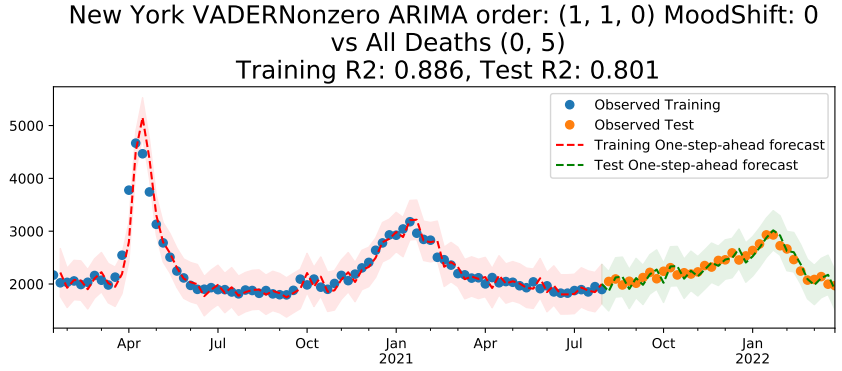
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-534.508
Date:	Mon, 17 Apr 2023	AIC	1075.015
Time:	11:32:55	BIC	1082.161
Sample:	01-18-2020	HQIC	1077.880
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	628.4477	965.023	0.651	0.515	-1262.963	2519.858
ar.L1	0.5697	0.046	12.450	0.000	0.480	0.659
sigma2	3.684e+04	2463.145	14.958	0.000	3.2e+04	4.17e+04

Ljung-Box (Q):	26.01	Jarque-Bera (JB):	325.01
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	0.15	Skew:	0.68
Prob(H) (two-sided):	0.00	Kurtosis:	12.78

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.68: New York NY Selected Model, Mean Vader Sentiment



SARIMAX Results

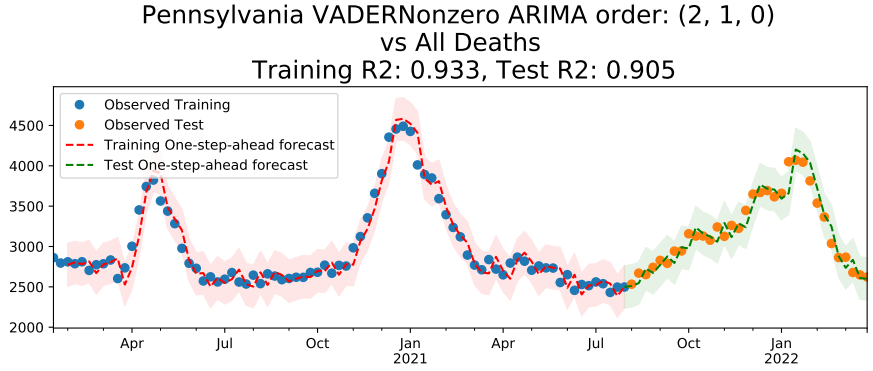
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 0)	Log Likelihood	-533.825
Date:	Mon, 17 Apr 2023	AIC	1075.650
Time:	11:32:08	BIC	1085.178
Sample:	01-18-2020	HQIC	1079.470
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	2105.9504	2922.614	0.721	0.471	-3622.267	7834.168
1	-1.06e+04	1.72e+04	-0.616	0.538	-4.44e+04	2.32e+04
ar.L1	0.5772	0.056	10.291	0.000	0.467	0.687
sigma2	3.622e+04	2563.182	14.132	0.000	3.12e+04	4.12e+04

Ljung-Box (Q):	26.56	Jarque-Bera (JB):	333.27
Prob(Q):	0.95	Prob(JB):	0.00
Heteroskedasticity (H):	0.15	Skew:	0.70
Prob(H) (two-sided):	0.00	Kurtosis:	12.90

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.69: New York NY Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

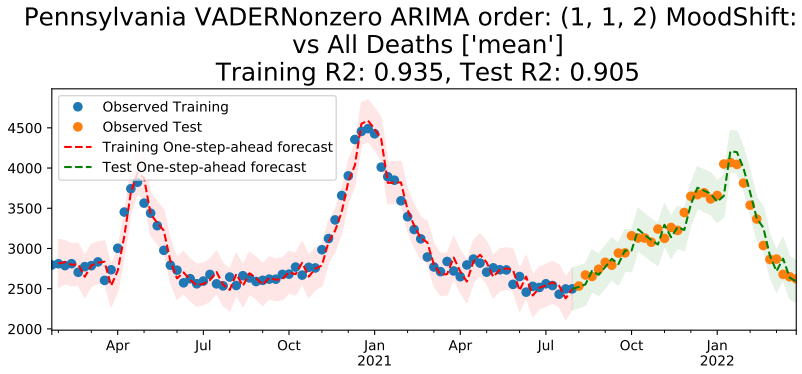
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 0)	Log Likelihood	-505.300
Date:	Mon, 17 Apr 2023	AIC	1016.600
Time:	11:32:12	BIC	1023.746
Sample:	01-18-2020	HQIC	1019.465
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3666	0.122	3.005	0.003	0.128	0.606
ar.L2	0.1977	0.105	1.875	0.061	-0.009	0.404
sigma2	1.782e+04	2735.946	6.514	0.000	1.25e+04	2.32e+04

Ljung-Box (Q):	36.17	Jarque-Bera (JB):	1.12
Prob(Q):	0.64	Prob(JB):	0.57
Heteroskedasticity (H):	0.61	Skew:	-0.16
Prob(H) (two-sided):	0.20	Kurtosis:	3.48

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.70: Philadelphia PA Selected Model, No Sentiment



SARIMAX Results

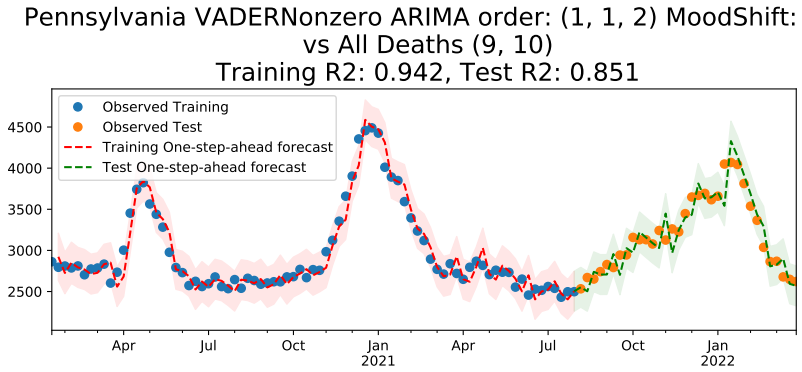
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(1, 1, 2)	Log Likelihood	-498.113
Date:	Mon, 17 Apr 2023	AIC	1006.227
Time:	11:32:57	BIC	1018.074
Sample:	01-25-2020	HQIC	1010.973
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	562.3590	676.617	0.831	0.406	-763.787	1888.505
ar.L1	0.4285	0.217	1.975	0.048	0.003	0.854
ma.L1	-0.0439	0.190	-0.231	0.817	-0.417	0.329
ma.L2	0.2340	0.135	1.736	0.083	-0.030	0.498
sigma2	1.746e+04	3062.006	5.703	0.000	1.15e+04	2.35e+04

Ljung-Box (Q):	34.87	Jarque-Bera (JB):	0.11
Prob(Q):	0.70	Prob(JB):	0.95
Heteroskedasticity (H):	0.62	Skew:	-0.03
Prob(H) (two-sided):	0.24	Kurtosis:	3.17

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.71: Philadelphia PA Selected Model, Mean Vader Sentiment



SARIMAX Results

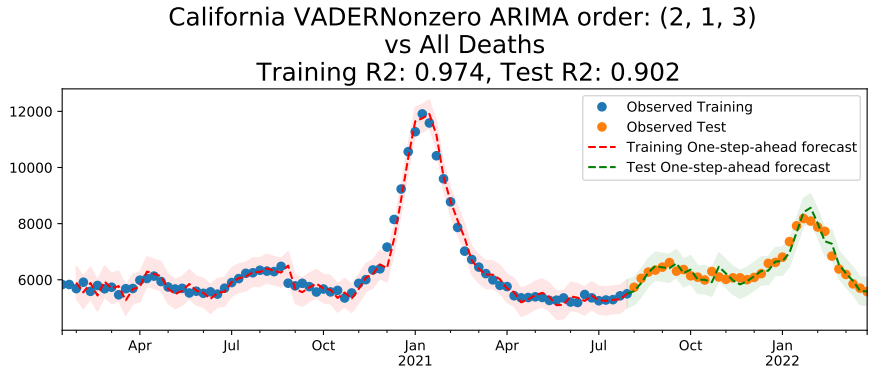
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 2)	Log Likelihood	-499.172
Date:	Mon, 17 Apr 2023	AIC	1010.344
Time:	11:32:11	BIC	1024.637
Sample:	01-18-2020	HQIC	1016.075
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	8679.7619	8478.020	1.024	0.306	-7936.853	2.53e+04
1	2.747e+04	8975.402	3.061	0.002	9878.945	4.51e+04
ar.L1	0.0384	0.308	0.124	0.901	-0.566	0.642
ma.L1	0.4330	0.282	1.537	0.124	-0.119	0.985
ma.L2	0.4257	0.136	3.124	0.002	0.159	0.693
sigma2	1.524e+04	2363.803	6.449	0.000	1.06e+04	1.99e+04

Ljung-Box (Q):	35.63	Jarque-Bera (JB):	1.10
Prob(Q):	0.67	Prob(JB):	0.58
Heteroskedasticity (H):	0.85	Skew:	0.16
Prob(H) (two-sided):	0.67	Kurtosis:	3.48

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.72: Philadelphia PA Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

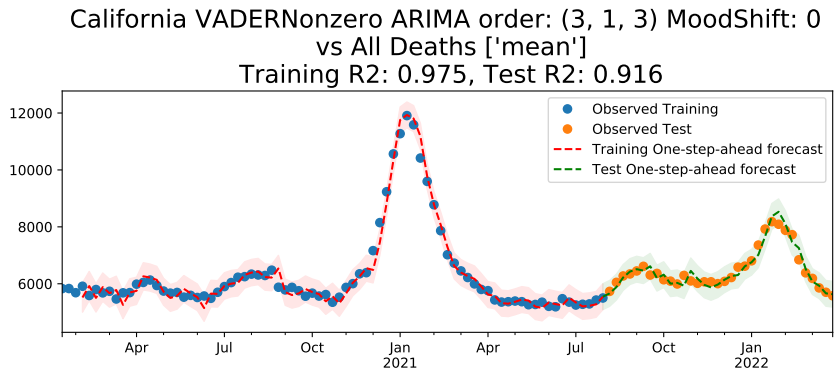
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 3)	Log Likelihood	-555.144
Date:	Mon, 17 Apr 2023	AIC	1122.289
Time:	11:32:16	BIC	1136.581
Sample:	01-18-2020	HQIC	1128.019
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8312	0.236	3.524	0.000	0.369	1.294
ar.L2	-0.4065	0.208	-1.950	0.051	-0.815	0.002
ma.L1	-0.3849	0.582	-0.662	0.508	-1.525	0.755
ma.L2	0.7246	1.338	0.542	0.588	-1.897	3.346
ma.L3	0.3606	0.611	0.590	0.555	-0.838	1.559
sigma2	6.36e+04	9.43e+04	0.675	0.500	-1.21e+05	2.48e+05

Ljung-Box (Q):	29.08	Jarque-Bera (JB):	6.74
Prob(Q):	0.90	Prob(JB):	0.03
Heteroskedasticity (H):	0.66	Skew:	0.06
Prob(H) (two-sided):	0.29	Kurtosis:	4.42

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.73: San Francisco CA SF Selected Model, No Sentiment



SARIMAX Results

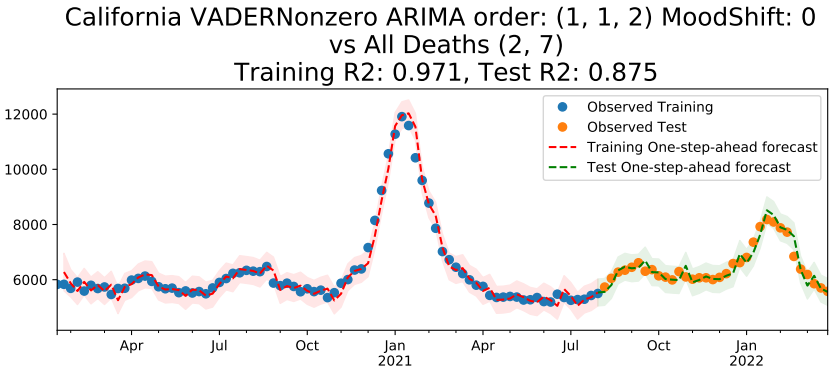
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(3, 1, 3)	Log Likelihood	-553.631
Date:	Mon, 17 Apr 2023	AIC	1123.262
Time:	11:32:59	BIC	1142.318
Sample:	01-18-2020	HQIC	1130.902
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-1090.2997	543.431	-2.006	0.045	-2155.404	-25.195
ar.L1	1.0303	0.470	2.191	0.028	0.109	1.952
ar.L2	-0.6506	0.509	-1.279	0.201	-1.648	0.347
ar.L3	0.2028	0.239	0.847	0.397	-0.266	0.672
ma.L1	-0.5815	0.604	-0.963	0.336	-1.766	0.603
ma.L2	0.9070	0.950	0.955	0.340	-0.955	2.769
ma.L3	0.1210	0.396	0.306	0.760	-0.656	0.898
sigma2	5.278e+04	4.07e+04	1.297	0.195	-2.7e+04	1.33e+05

Ljung-Box (Q):	28.77	Jarque-Bera (JB):	8.01
Prob(Q):	0.91	Prob(JB):	0.02
Heteroskedasticity (H):	0.40	Skew:	0.01
Prob(H) (two-sided):	0.02	Kurtosis:	4.55

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.74: San Francisco CA SF Selected Model, Mean Vader Sentiment



SARIMAX Results

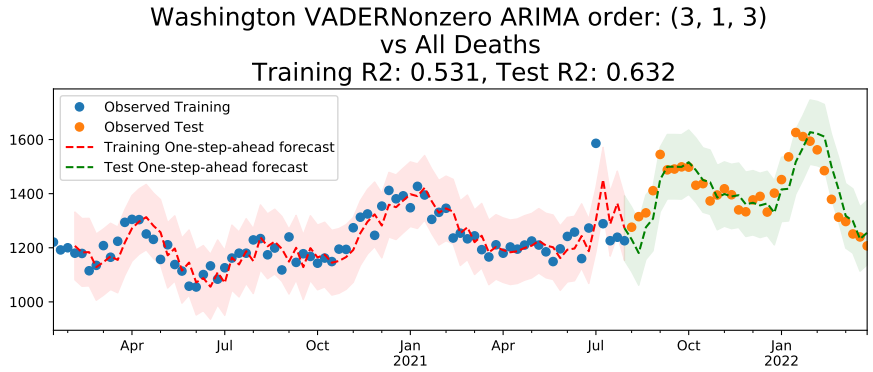
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 2)	Log Likelihood	-556.881
Date:	Mon, 17 Apr 2023	AIC	1125.761
Time:	11:32:14	BIC	1140.053
Sample:	01-18-2020	HQIC	1131.491
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	1.551e+04	5328.717	2.911	0.004	5066.116	2.6e+04
1	1.757e+04	1e+04	1.749	0.080	-2114.593	3.73e+04
ar.L1	0.6748	0.100	6.731	0.000	0.478	0.871
ma.L1	-0.0583	0.139	-0.420	0.674	-0.330	0.214
ma.L2	0.1990	0.150	1.331	0.183	-0.094	0.492
sigma2	6.169e+04	7661.369	8.052	0.000	4.67e+04	7.67e+04

Ljung-Box (Q):	33.58	Jarque-Bera (JB):	52.12
Prob(Q):	0.75	Prob(JB):	0.00
Heteroskedasticity (H):	0.94	Skew:	-0.77
Prob(H) (two-sided):	0.87	Kurtosis:	6.64

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.75: San Francisco CA SF Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

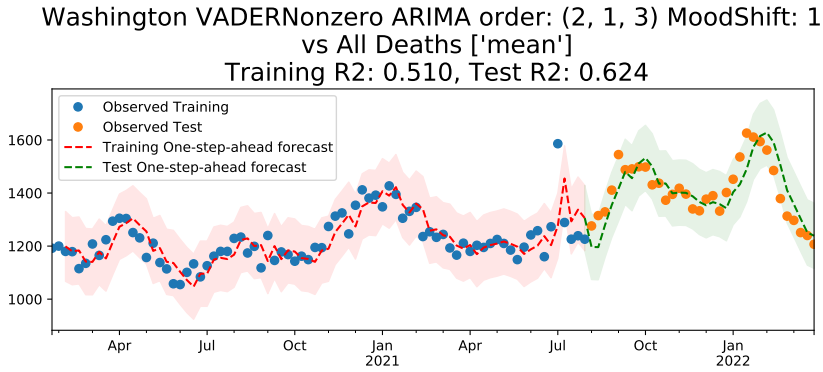
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(3, 1, 3)	Log Likelihood	-443.014
Date:	Mon, 17 Apr 2023	AIC	900.028
Time:	11:32:21	BIC	916.703
Sample:	01-18-2020	HQIC	906.714
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4139	0.445	-0.930	0.353	-1.286	0.459
ar.L2	-0.1897	0.391	-0.486	0.627	-0.955	0.576
ar.L3	-0.4862	0.307	-1.581	0.114	-1.089	0.116
ma.L1	0.1485	4.396	0.034	0.973	-8.467	8.764
ma.L2	-0.1680	3.489	-0.048	0.962	-7.007	6.671
ma.L3	0.6819	2.921	0.233	0.815	-5.042	6.406
sigma2	3663.6171	1.54e+04	0.237	0.812	-2.66e+04	3.39e+04

Ljung-Box (Q):	14.86	Jarque-Bera (JB):	79.38
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	2.80	Skew:	0.89
Prob(H) (two-sided):	0.01	Kurtosis:	7.54

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.76: Seattle WA Selected Model, No Sentiment



SARIMAX Results

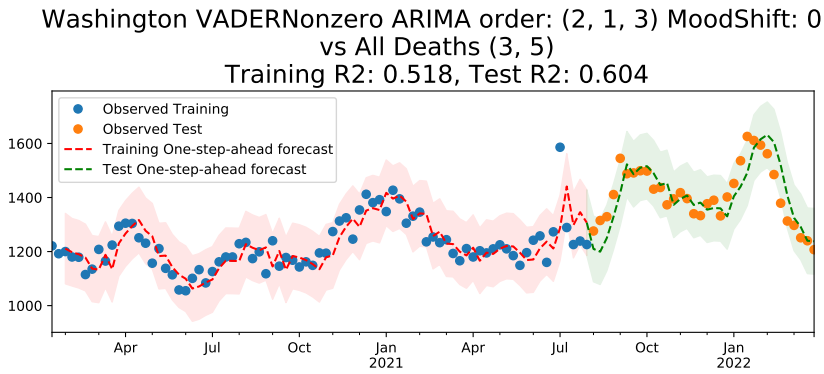
Dep. Variable:	All Cause	No. Observations:	80
Model:	ARIMA(2, 1, 3)	Log Likelihood	-439.417
Date:	Mon, 17 Apr 2023	AIC	892.834
Time:	11:33:02	BIC	909.420
Sample:	01-25-2020	HQIC	899.479
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	-178.0691	268.987	-0.662	0.508	-705.274	349.136
ar.L1	0.2675	0.397	0.674	0.500	-0.510	1.045
ar.L2	-0.7443	0.385	-1.934	0.053	-1.498	0.010
ma.L1	-0.6119	0.455	-1.346	0.178	-1.503	0.279
ma.L2	0.7260	0.412	1.762	0.078	-0.082	1.534
ma.L3	-0.0575	0.207	-0.278	0.781	-0.463	0.348
sigma2	3943.0305	435.771	9.048	0.000	3088.935	4797.126

Ljung-Box (Q):	16.22	Jarque-Bera (JB):	135.02
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	3.00	Skew:	1.19
Prob(H) (two-sided):	0.01	Kurtosis:	8.94

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.77: Seattle WA Selected Model, Mean Vader Sentiment



SARIMAX Results

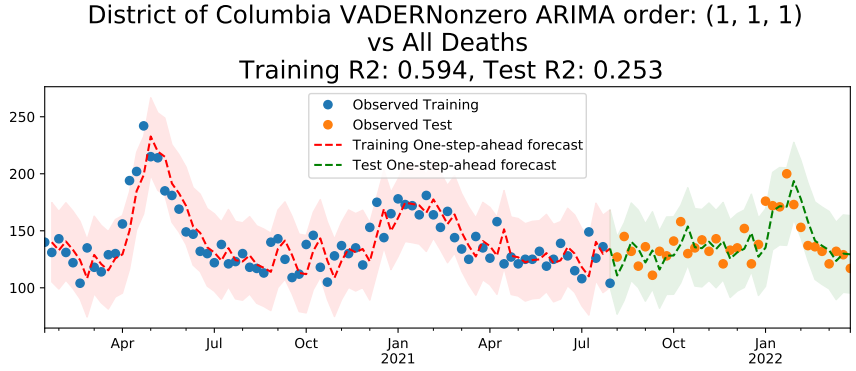
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(2, 1, 3)	Log Likelihood	-443.951
Date:	Mon, 17 Apr 2023	AIC	903.902
Time:	11:32:19	BIC	922.959
Sample:	01-18-2020	HQIC	911.543
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	-1017.6282	5280.235	-0.193	0.847	-1.14e+04	9331.441
1	4635.9759	5723.653	0.810	0.418	-6582.178	1.59e+04
ar.L1	0.3626	0.347	1.045	0.296	-0.318	1.043
ar.L2	-0.5400	0.322	-1.679	0.093	-1.170	0.090
ma.L1	-0.7036	0.401	-1.755	0.079	-1.489	0.082
ma.L2	0.5379	0.437	1.231	0.218	-0.318	1.394
ma.L3	0.0821	0.166	0.495	0.620	-0.243	0.407
sigma2	3844.1461	499.531	7.696	0.000	2865.083	4823.209

Ljung-Box (Q):	20.17	Jarque-Bera (JB):	106.58
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	3.07	Skew:	1.10
Prob(H) (two-sided):	0.00	Kurtosis:	8.21

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.78: Seattle WA Selected Model, Selected Vader Eigenmood Components



SARIMAX Results

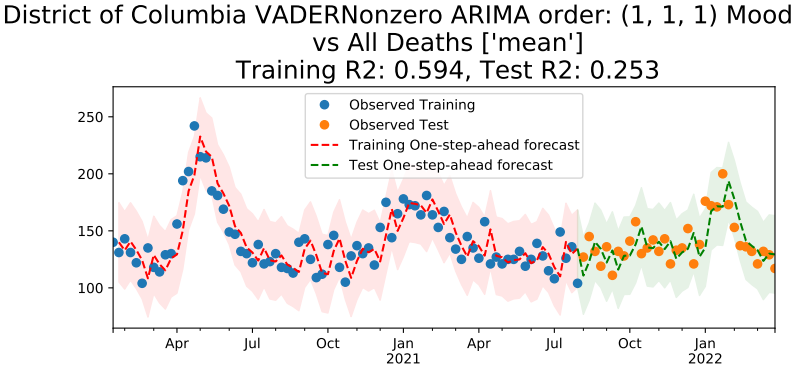
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 1)	Log Likelihood	-341.771
Date:	Mon, 17 Apr 2023	AIC	689.542
Time:	11:32:24	BIC	696.688
Sample:	01-18-2020	HQIC	692.407
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0658	0.515	-0.128	0.898	-1.074	0.943
ma.L1	-0.1535	0.480	-0.320	0.749	-1.094	0.787
sigma2	300.6063	53.616	5.607	0.000	195.520	405.692

Ljung-Box (Q):	31.30	Jarque-Bera (JB):	4.81
Prob(Q):	0.84	Prob(JB):	0.09
Heteroskedasticity (H):	0.76	Skew:	0.59
Prob(H) (two-sided):	0.49	Kurtosis:	2.74

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.79: Washington DC Selected Model, No Sentiment



SARIMAX Results

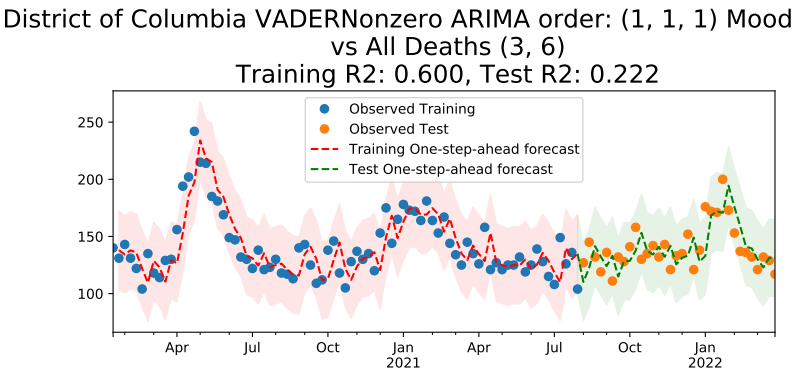
Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 1)	Log Likelihood	-341.777
Date:	Mon, 17 Apr 2023	AIC	691.554
Time:	11:33:03	BIC	701.082
Sample:	01-18-2020	HQIC	695.374
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
mean	1.4784	93.598	0.016	0.987	-181.970	184.927
ar.L1	-0.0630	0.540	-0.117	0.907	-1.121	0.995
ma.L1	-0.1559	0.504	-0.309	0.757	-1.143	0.832
sigma2	300.6409	53.746	5.594	0.000	195.301	405.981

Ljung-Box (Q):	31.34	Jarque-Bera (JB):	4.82
Prob(Q):	0.83	Prob(JB):	0.09
Heteroskedasticity (H):	0.76	Skew:	0.59
Prob(H) (two-sided):	0.49	Kurtosis:	2.73

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.80: Washington DC Selected Model, Mean Vader Sentiment



SARIMAX Results

Dep. Variable:	All Cause	No. Observations:	81
Model:	ARIMA(1, 1, 1)	Log Likelihood	-341.231
Date:	Mon, 17 Apr 2023	AIC	692.462
Time:	11:32:23	BIC	704.372
Sample:	01-18-2020	HQIC	697.237
	- 07-31-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
0	-722.6047	1197.304	-0.604	0.546	-3069.277	1624.068
1	688.4540	1432.395	0.481	0.631	-2118.988	3495.896
ar.L1	-0.0614	0.591	-0.104	0.917	-1.219	1.096
ma.L1	-0.1565	0.538	-0.291	0.771	-1.211	0.898
sigma2	296.5775	53.369	5.557	0.000	191.977	401.178

Ljung-Box (Q):	29.98	Jarque-Bera (JB):	4.27
Prob(Q):	0.88	Prob(JB):	0.12
Heteroskedasticity (H):	0.75	Skew:	0.56
Prob(H) (two-sided):	0.46	Kurtosis:	2.80

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURE A.81: Washington DC Selected Model, Selected Vader Eigenmood Components

Appendix B

Chapter 3 Appendix

Supplementary Methods

S1. Notes on "misclassifications" for Country Classification from sex-searches

Some of the countries identified as Christian celebrate the nativity according to Julian calendar, with Christmas falling on January 7th or January 14th of the Gregorian calendar. Such is the case of the Christian countries: Belarus, Bosnia and Herzegovina, Georgia, Macedonia, Moldova, Montenegro, Serbia, Slovenia, Russia and Ukraine. Neither of these countries has a national holiday on December 25th nor shows an increase in sex-searches around December 25th. Had these countries been labeled as "Other", the percentage of countries identified as Christian for which we see a significant increase (z-score>1) in sex-searches would have been of 91%. In addition to not celebrating the Christmas on December 25th, some of these countries also have a sizeable percentage of population that self-identifies as Muslim. Such is the case of Montenegro (29%), Macedonia (39%) and Bosnia and Herzegovina (45%).

From the 30 Muslim countries, Pakistan was classified as Christian and 6 other countries didn't make the threshold. Pakistan is highly related to Christmas, probably due to the fact that there is a public holiday on 25th December, which coincidentally celebrates the birthday of Muhammad Ali Jinnah, founder of Pakistan. The other six countries also correspond to the ones for which the quality of the sex-search data was the poorest.

Keeping in mind that we were looking for countries that culturally relate to a Christian or Muslim religious background, all countries that didn't make the threshold to be labelled as either are classified as Other. Unsurprisingly, there are many countries who are originally labelled as Other and end up classified as either Christian or Muslim. European countries, such as the Czech Republic, Estonia and the Netherlands, whose majority does not identify as religious are classified as Christian, most likely due to the fact that these populations celebrate the holiday as well, even if secularly.

S2. Mean Sentiment Correlations with Sex-Search Volume

As shown in Supplementary Table S9A, there is a highly significant, moderate fit ($R^2 > 0.1$) across all countries, demonstrating a significant correlation between volume of sex-searches and mean sentiment as measured by the three ANEW dimensions. The coefficient of determination is generally stronger for Christian countries than Muslim Countries. Similarly to the GT data, the multiple linear regression models can be improved by averaging sentiment and sex-search volume across years using the 52-week Christmas centered calendar for the USA, Australia, Brazil, Argentina, and Chile, , and the 50-week Eid-al-Fitr centered calendar for Indonesia and Turkey. This smooths out extraordinary events that are picked up by sentiment analysis. The results of this centered-data regression are presented in Supplementary Table S9B. The fit is highly significant for all countries, and improves for all countries, ($R^2 > 0.26$). In every case, valence is yields a positive coefficient, while dominance a negative coefficient; so the happier but less dominant the sentiment expressed by a country, the more sex-searches tend to increase. As far as significance is concerned, t-tests reveal that the valence dimension is most often significant, followed by dominance, with arousal the least likely to be a significant factor.

Interestingly, as shown in Supplementary Table S10, when we computed the ordinary least squares estimate of a standard linear regression on each ANEW dimension independently, we obtained very poor (but significant) goodness of fit, as measured by R^2. Therefore, the mean value of each ANEW dimension on its own is a poor predictor of sex-search volume in all countries (with few exceptions such as Arousal in Brazil). We can thus say that mean sentiment correlates with sex-search volume (Supplementary Table S9) but the timeseries of mean weekly values of each ANEW dimension do not yield a nuanced characterization of sentiment correlated with interest in sex.

S3. Singular Value Decomposition

Singular value decomposition (SVD) is a method by which a matrix can be linearly decomposed into ordered orthonormal components, each explaining as much of the linear variation as possible, after the

components that came before it. The SVD of any m × n matrix M of real or complex numbers can represented as follows in Equation 2:

$$M=USV^T$$

Where U is an m× n matrix with orthonormal columns, V is an n× n matrix with orthonormal columns, and S is an n× n diagonal matrix. The columns of U and V are referred to as the left and right singular vectors of M respectively. These singular vectors are eigenvectors of the matrices MM^T and M^TM respectively. The diagonal entries of S, called the singular values of M, are the square roots of the eigenvalues of the matrices MM^T and M^TM. By convention, the singular values are ordered from greatest to least. The columns of U form a basis for the column space of M and the columns of V form a basis for the row space of M. The right singular vectors are also known in principal component analysis (PCA) as the loadings of the original variables (bins) onto the new coordinate system. The relative variance explained by each component can then be calculated for each component k as $s_k^2/\sum_i(s_i^2)$ where s_k is the kth diagonal component of S. It is important to note that matrices can be reconstructed with a lower rank by setting elements of S to zero. Typically only the top l singular values are kept in order to reduce noise and create the closest rank-l approximation of the original matrix[19].

S4. Data Reconstruction

It can be clearly seen from the data reconstruction averages in Extended Data Fig. 8 and Supplementary Fig. S6, that the distribution of sentiment shifts towards higher bins during holidays, represented by redder high bins and greener low bins on holidays. Christmas stands out in the USA (US), Australia (AU), and Brazil (BR). Eid-al-Fitr stands out in both Turkey (TR) and Indonesia (ID), and in Turkey the beginning of Ramadan is emphasized a few weeks before. The centering performed only looks at weeks within the surrounding cultural year, such that Christmas is week 26 of a 52 week year (starting with a first week 1), while Eid-al-Fitr is week 25 of a 50 week year. Other weeks are averaged in this range according to their displacement from the holiday week (e.g., a week two weeks before the Christmas week in 2012 is averaged with weeks two weeks before Christmas in all other years). This obscures the emphasis on holidays using another calendar, such that Indonesia also has a strong signal on Christmas, but these signals are averaged over multiple weeks when the calendars are misaligned. The heatmaps for all countries centered on all holidays are included in Supplementary Fig. S6.

S5. Eigenmood Selection and Characterization

The mean value of a holiday's projection on various components for different countries are shown in Supplementary Figures S2 and S3 for Christmas and Eid-al-Fitr respectively, with the two components selected for each country highlighted in red. As described, since the first component corresponds to the basic distribution of sentiment in the language and overwhelms projections because of how much it explains, and the last few components are mostly noise, we only look at the components explaining 95% of the variance after the removal of the first. The second component usually describes a variation over the whole time series of out data, thus it tends to have a large standard deviation.

To better understand how the selected components describe the mood, we define an interpretable linguistic variable[29]. The linguistic variable can take five fuzzy values, "low", "medium-low", "medium", "medium-high", and "high" with membership functions defined over the 25 bins of the original twitter sentiment distribution. These membership functions are shown in Supplementary Fig. S4 and were chosen such that each original bin's membership in all values sums to one, and the area under each membership function is the same.

The response of the linguistic variable to the holiday in each selected eigenmood is shown in Supplementary Figure S5 for the selected relevant holiday for each country. These responses were calculated by reconstructing the distribution bins with only the eigenmood selected for the country and holiday, multiplying the reconstructed bin value by its memberships, and summing over all bins for each linguistic value. These responses can be interpreted as the change from the language's base sentiment distribution on the holiday contributed by the selected eigenmood. The response characterized by the Christmas eigenmood in the USA is an increase in medium-high happiness, with decreases in other levels of happiness, low and medium happiness in particular.

How mood changes on a major holiday varies between countries but generally we see that the selected eigenmood describes increases medium-high or high valence on the holidays, with decreases in low, medium-low, and medium valence, as well as lower or more moderate dominance and arousal. The behavior of the dominance mood dimension in the week of Eid-al-Fitr in Indonesia highlights the importance of the more nuanced mood measurement that eigenmoods afford. While the ANEW mean value measurement above suggested a dominance decrease towards a less "in-control" mood, what we have at Eid-al-Fitr is a shift away from the extremes to a collective mood state that is neither very "in-control" nor very "controlled" – coherent with a happier and calmer mood scenario typically found in these holidays for all countries. In other words, during most weeks of the year, there is increased bimodal dominance activity in higher and lower bins (simultaneously high "in-control" and "controlled", respectively), but in the week of Eid-al-Fitr, the dominance mood converges to a mid-level dominance (Figure 4 column A, row 3, dominance panel).

S6. Eigenmood correlations to Sex-search volume in target Holidays

As a measure of mood similarity between weeks in a space defined by a selected eigenmood, we use the dot product between their coordinates in this space[20]. This measure increases between weeks with similar (positive or negative) projections onto the eigenweeks forming the space, becomes negative with opposite projections, and decreases in magnitude with weeks that are not correlated with the eigenweeks and are thus projected near the origin. Due to the properties, it is important to select an eigenmood that strongly corresponds to a week or weeks of interest, by containing high-magnitude values in the corresponding eigenbins. The similarity can then be expressed as $w \cdot c$ where w and c are weeks projected into the eigenmood, which is equivalently the vector of corresponding weighted eigenbin values. In comparison between weeks and a holiday averaged over years, these vectors are the element-wise averages of the week's projection coordinates over the years. We report results with these averages, but these results are robust to yearly, non-averaged data, as well as different selection criteria for the eigenmoods (for example, allowing a greater number of components). The projection spaces for each eigenmood are shown in Supplementary Fig. S7.

In general, weeks close in proximity in time will be more similar in eigenmood, but certain weeks, often other holidays, more distant in time can have a high similarity in eigenmood to the selected holiday. In the USA, for example, the weeks closest in eigenmood to Christmas are, in order, the week of New Year's Day, the other weeks of December, and the weeks following July 4th, Father's Day, and Memorial Day. National Day in Chile is similar in eigenmood and sex searches to Chile's Christmas. New Year's Day and Christmas in Indonesia are similar to Eid-al-Fitr's eigenmood and high sex searches. In Turkey, weeks in late June, early July, and the week following Eid-al-Fitr are the most similar in terms of eigenmood and sex search volume to Eid-al-Fitr.

To investigate the relationship between a week's similarity in eigenmood to a holiday and the number of sex searches, we perform an ordinary least squares regression between sex searches as the dependent variable, and similarity as the independent variable. Displayed in Figure 4 and reported in Extended Data Table 2 are the results of this regression as well as Brownian distance correlation statistics, a nonlinear measure of correlation[30]. The plots of all linear regressions are included in Supplementary Fig. S7.

There is a fairly strong correspondence ($R^2 \geq .380$) between similarity in eigenmood to Christmas and sex searches in the C countries: the US, Brazil, Australia, Argentina, and Chile. The southern hemisphere Christian countries Brazil, Argentina, and Chile also have a noticeable correlation with Eid-al-Fitr, however, the slope of the regression is negative, implying that the less like the mood during the winter week of Eid-al-Fitr, the more sex searches are conducted.

In Muslim countries Turkey and Indonesia, we were limited by having less Twitter data and fewer tweets that match. However, there are significant correlations between similarity to Eid-al-Fitr and increased sex searches. The linear correlation is reduced compared to Christmas in Christian countries, since over time the weeks of Ramadan become more similar in eigenmood to Eid-al- Fitr, the festival at Ramadan's conclusion, while the cultural pressure is one of abstinence, such that these weeks have unusually low sex searches. In the case of Turkey in particular, the holiday of Eid-al-Adha, or the Sacrifice Feast, also has high sex searches, but is different

in eigenmood from Eid-al-Fitr. The positive correlation between sex searches and Christmas eigenmood in Indonesia is likely caused by the sizable Christian population living there and effects due to summer.

Turkey is an interesting case, since it has a very strong negative correlation between sex searches and similarity to Christmas although the response to Eid-al-Fitr is smaller. In part, this may be due to limitations in our data gathering and method application, since our ANEW is only available in English, Spanish, and Portuguese. However, we still have a good number of tweets from Turkey, so we look more closely at its eigenmood. The projection of all weeks into its eigenmoods for Christmas and Eid-al-Fitr is shown in Supplementary Fig. S7, which happen to be same in this case. The regressions between sex searches and the similarity of averaged weeks to Christmas and Eid-al-Fitr are shown in Supplementary Fig. S7. The mood associated with Eid is also associated with Ramadan, which emphasizes abstinence. During the weeks of Ramadan, there are much fewer sex searches than usual, although the weeks are not too far different in mood. In addition, there is a separate holiday, Eid-al-Adha, that is associated with a second peak in sex searches, but with a different mood. Perhaps due to Turkey's small Christian population and winter timing, Christmas and weeks like it in eigenmood have low sex searches and averaging over years decreases the effects of holiday traditions (like Eid-al-Fitr) due to misaligned calendars.

Supplementary Figures

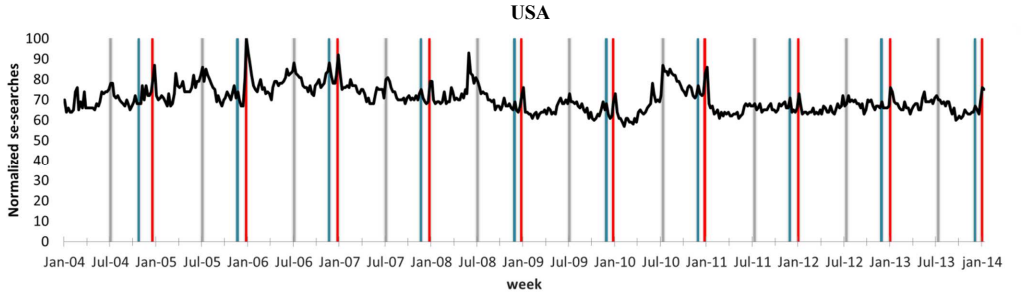


Fig. S1A**. GT query [sex] results for the USA.** The weeks containing Thanksgiving day, Christmas and the 4th of July are highlighted in blue, red and grey, respectively.

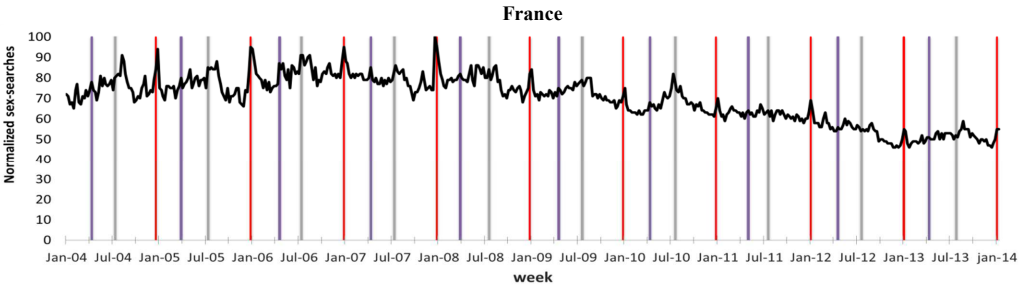


Fig. S1B**. GT query [sex] results for France.** The weeks containing Easter Sunday, July 14th and Christmas are highlighted in purple, grey and red, respectively.

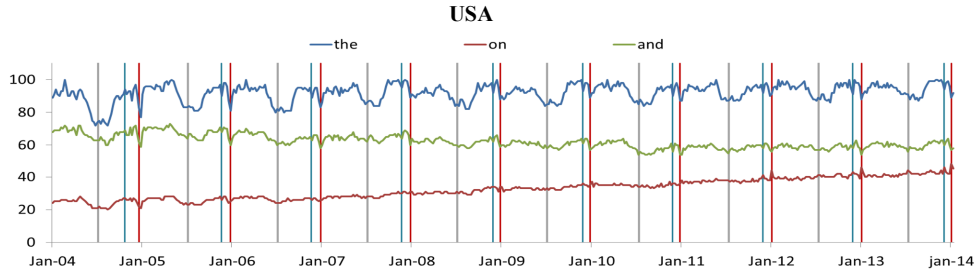


Fig. S2A**. GT queries for "the", "on" and "and", in the USA.** The weeks containing Thanksgiving day, Christmas and the 4th of July are highlighted in blue, red and grey, respectively.

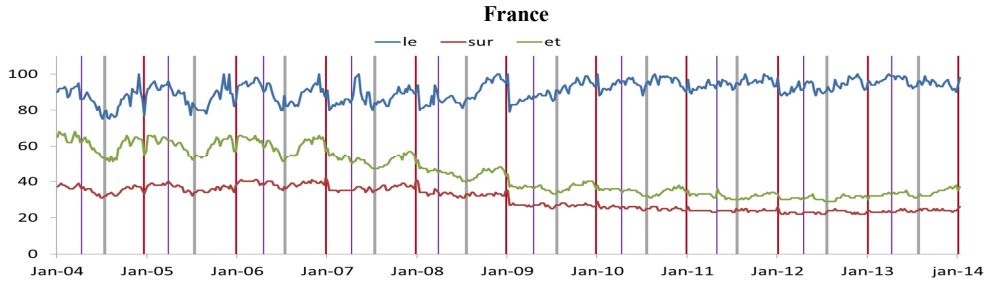


Fig. S2B**. GT queries for "le", "sur" and "et", in France.** The weeks containing Easter Sunday, July 14th and Christmas are highlighted in purple, grey and red, respectively.

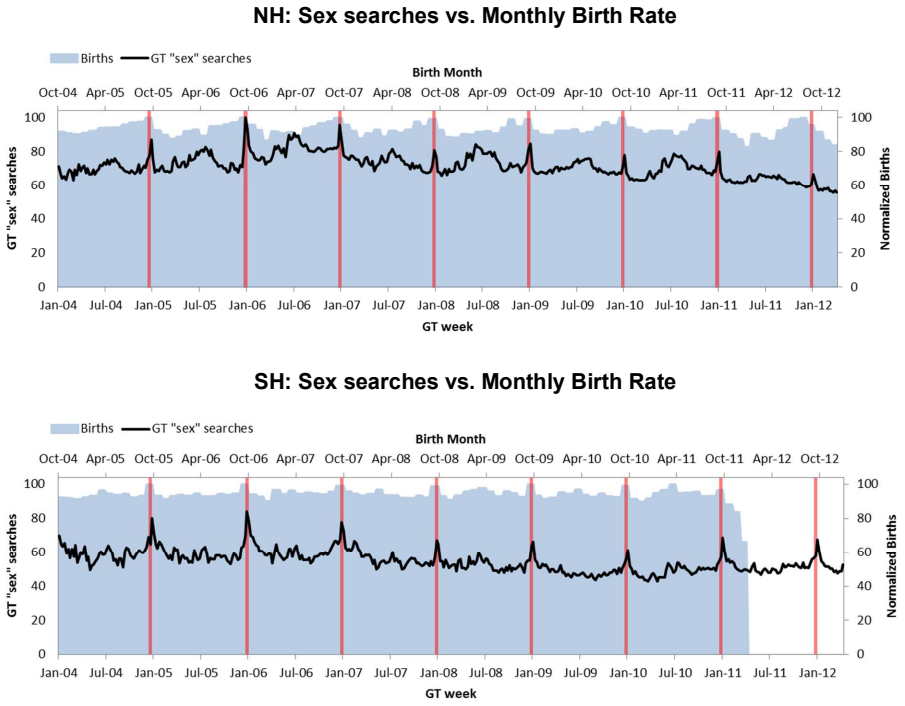


Fig. S3. Monthly birth data shifted by nine months (blue shaded area, top and right axis) and weekly averaged Google Trends results for "sex-searches" (black line, bottom and left axis) plotted for:
A) All Western Northern countries for which both birth and GT data exist (Austria, Canada, Denmark, Finland, France, Germany, Italy, Lithuania, Malta, Netherlands, Poland, Portugal, Spain, Sweden and United States of America), also represented in Fig. 1 in the main paper. Births in September are higher than the yearly average in all countries but Lithuania and Sweden, with an average variation of 6%).

B) All Southern countries for which both birth and GT data exist (Australia, New Zealand, Chile and South Africa). Births in September are higher than yearly average in all countries (average variation 5.5%, with the difference being as high at 10% in South Africa and New Zealand.)

Births were shifted nine months to match probable conception month. The red line marks Christmas week.

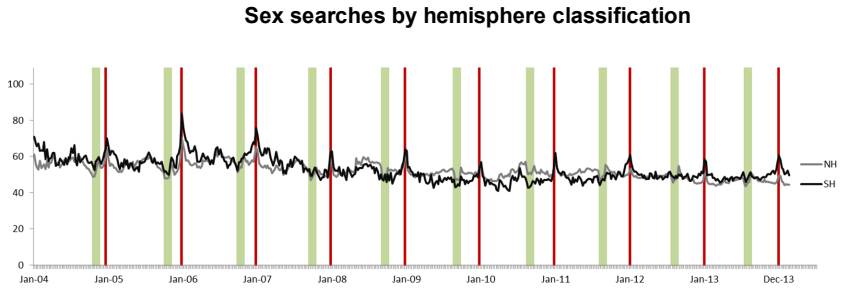


Fig. S4A. Averaged sex-searches for Northern and Southern countries. R^2 is 0.54 with a p-value of 2E-41. The weeks containing Ramadan and Christmas Day are highlighted in green and red, respectively.

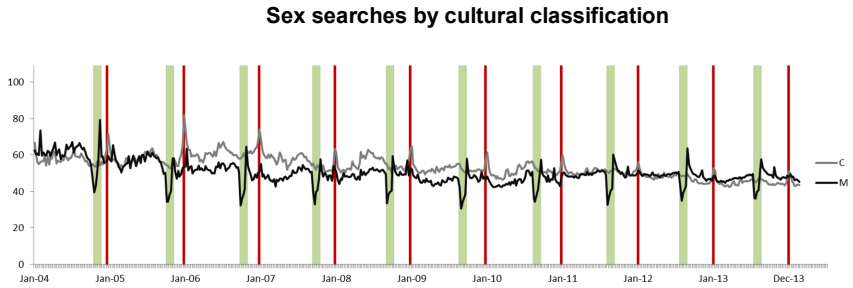


Fig. S4B. Averaged sex-searches for all Christian and Muslim countries. R^2 is 0.19 with a p-value of 3E-26. The weeks containing Ramadan and Christmas Day are highlighted in green and red, respectively.

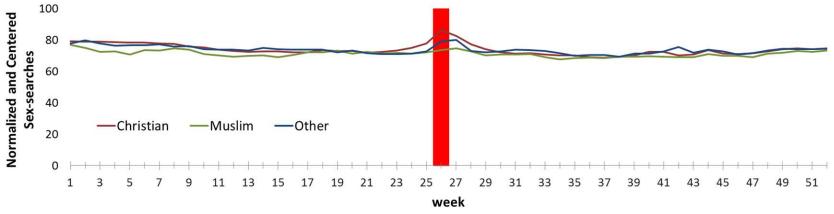


Fig. S5A. Averaged Christmas-centered results for the Christian (red), Muslim (green) and Other (dark blue) country sets. The red vertical bar represents the Christmas week, centered on week 26.

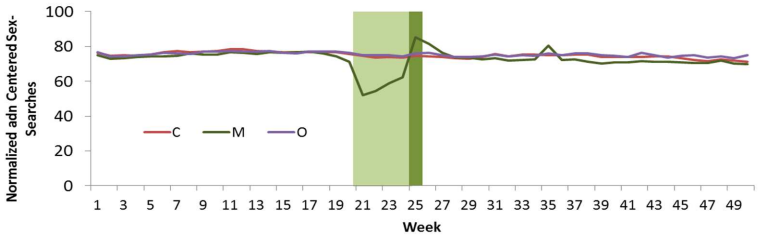


Fig. S5B. Averaged Eid-al-Fitr-centered results for the Christian (red), Muslim (green) and Other (dark blue) country sets. The darker green vertical bar represents the Eid-al-Fitr week, centered on week 25. The light green area represents the remaining Ramadan weeks.

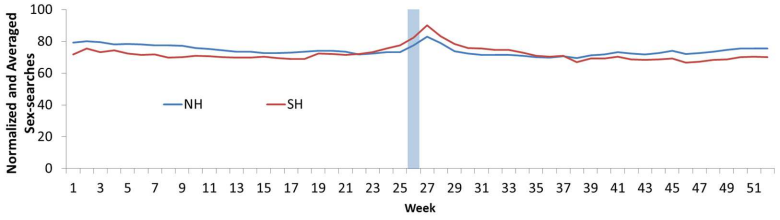


Fig. 5C. Averaged December Solstice-centered results for the Northern Hemisphere (blue) and Southern Hemisphere (red) country sets. Light blue vertical bar represents the week of the December-Solstice.

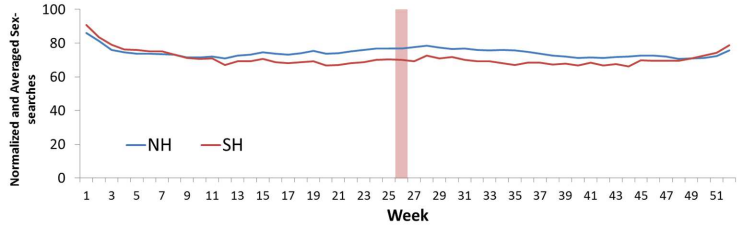


Fig. S5D. Averaged June Solstice-centered results for the Northern Hemisphere (blue) and Southern Hemisphere (red) country sets. Light pink vertical bar represents the week of the June-Solstice.

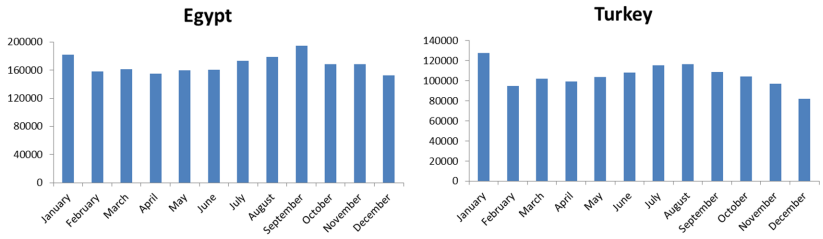


Fig. S6A. Averaged monthly births (for all available years) for Turkey and Egypt. In some Muslim countries, as in these examples, birth records are artificially at their lowest in December (in the case of Turkey, 22% below average) and peak in January (in the case of Turkey, 202% above average) , as parents prefer to have their children registered in the New Year.

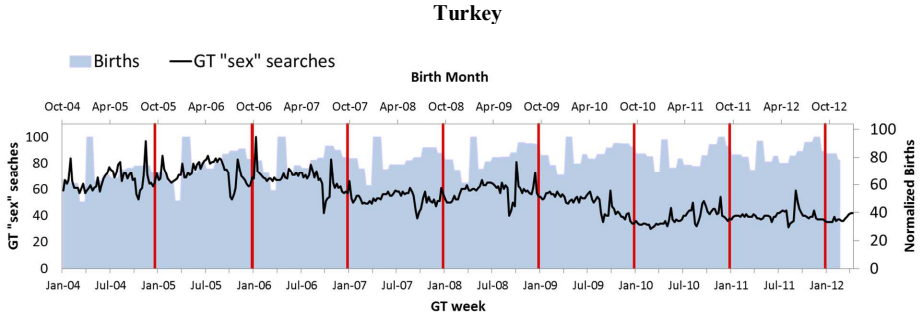


Fig. S6B Normalized monthly birth data (shaded blue, top and right axis) and Google Trends results of "sex"-searches (black line, left and bottom axis) for Turkey. Births were normalized so that each year's maximum becomes 100 and shifted nine months to match with probable conception month. The red line represents Christmas week, which was very close to Eid-al-Ada in 2005, 2006 and 2007. (It is obvious that the major registration peak happens in January of each year and it's not matched by an increase in sex-searches).

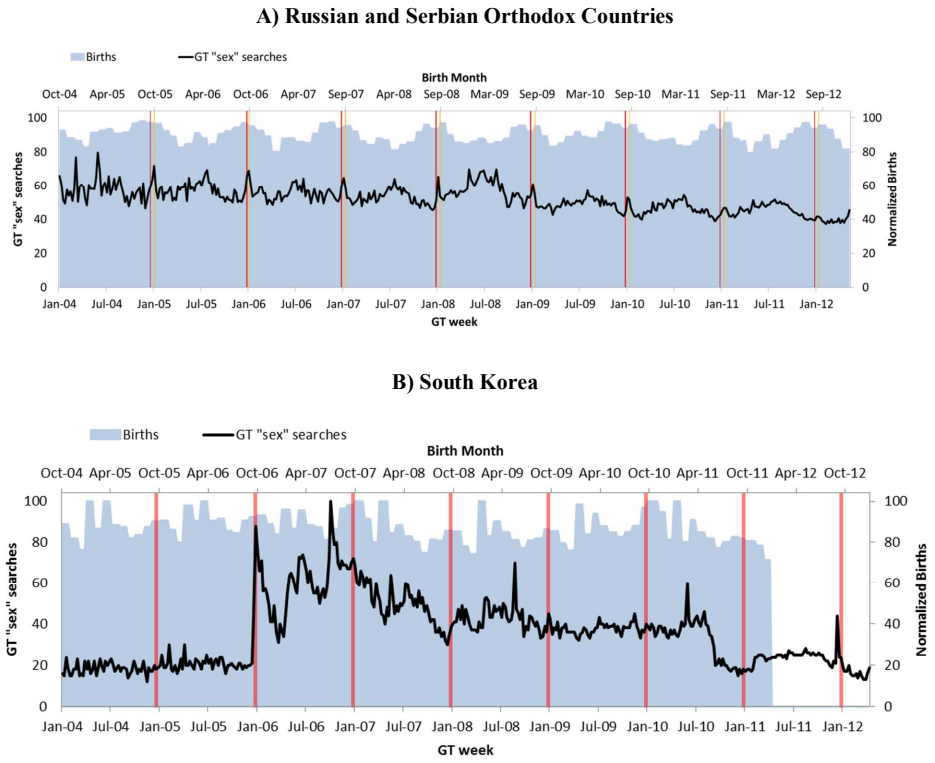


Fig. S7. Normalized monthly birth data (shaded blue, top and right axis) and Google Trends results of ”sex”-searches (black line, right and bottom axis) for
A) All Northern and Christian countries for which both birth and GT data exist that Celebrate Christmas on January 6th (Macedonia, Moldova, Serbia, Slovenia, Russia and Ukraine).
B) South Korea, as an example of a Northern Other country, for which both birth and GT data exists.

Births were shifted nine months to match with probable conception month. Vertical lines represent Christmas week with red marking the week of December 25th and orange marking the week of January 6th.

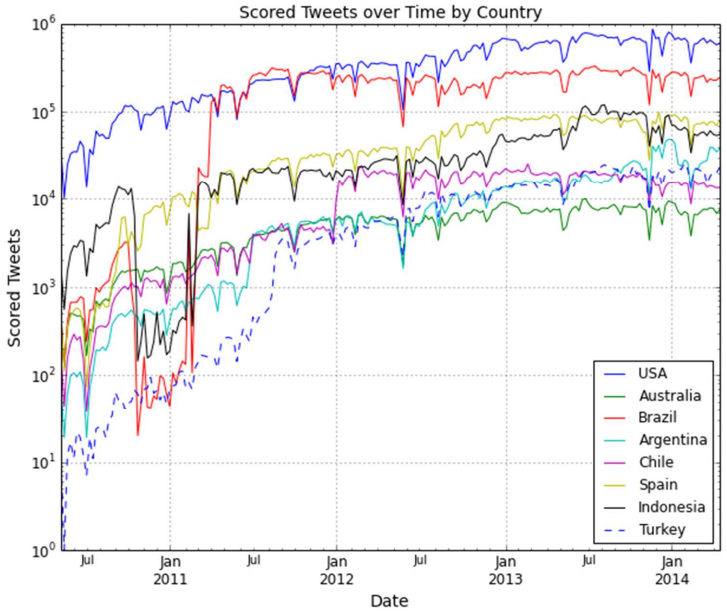


Fig. S8. Total number of weekly geolocated tweets matching ANEW for countries selected for *Eigenmood* analysis.

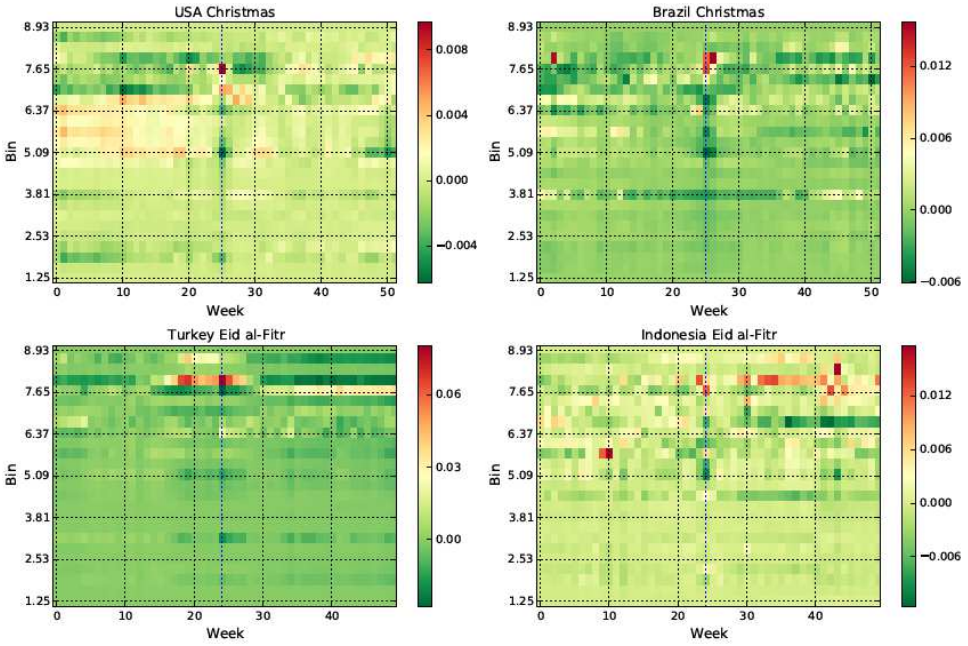


Fig. S9. Reconstructed valence heatmaps for multiple countries, centered on cultural holidays. Probability distributions of tweet valence were arranged in 25 bins (y-axis) each week (x-axis) for each country. Years were centered on a chosen holiday, marked by a central, vertical line. These data were averaged over all years, so each cell contains the average probability of a tweet's valence falling into a bin during a week. The data were reconstructed by removing the first component and components explaining less than 95% of the remaining variance.

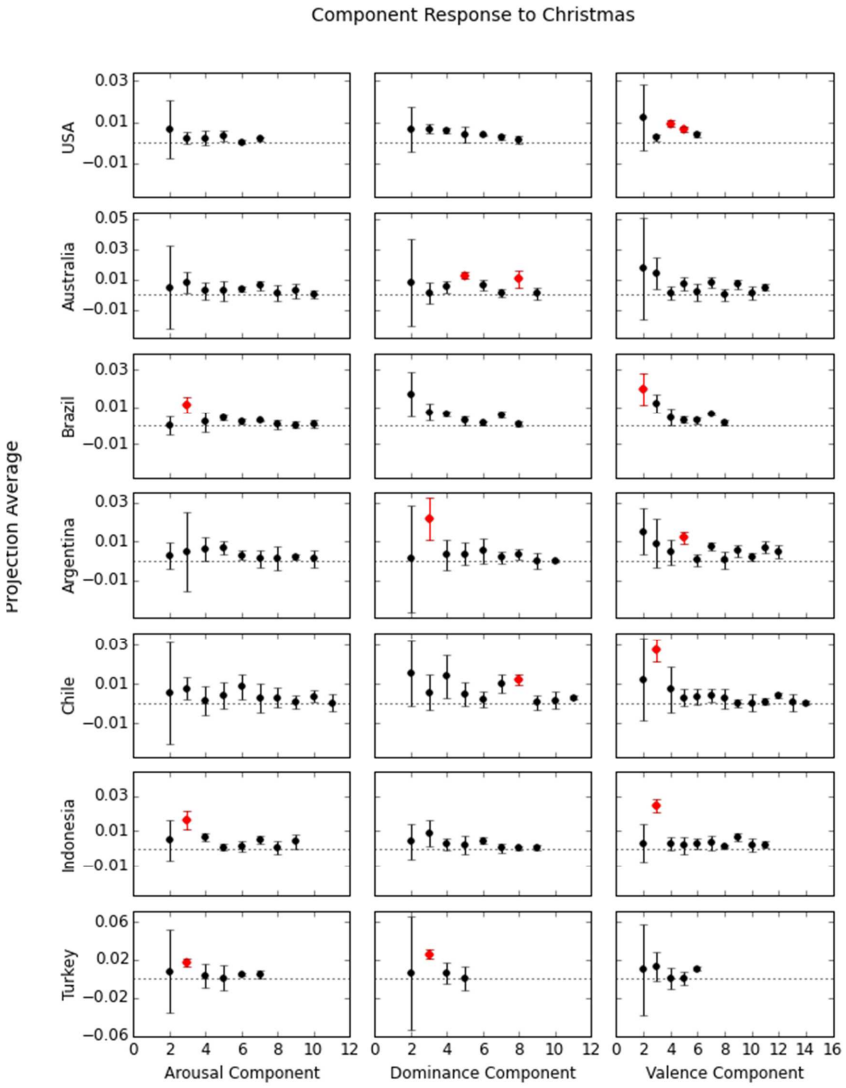


Fig. S10. ANEW component response to Christmas by country. Selected components highlighted in red.

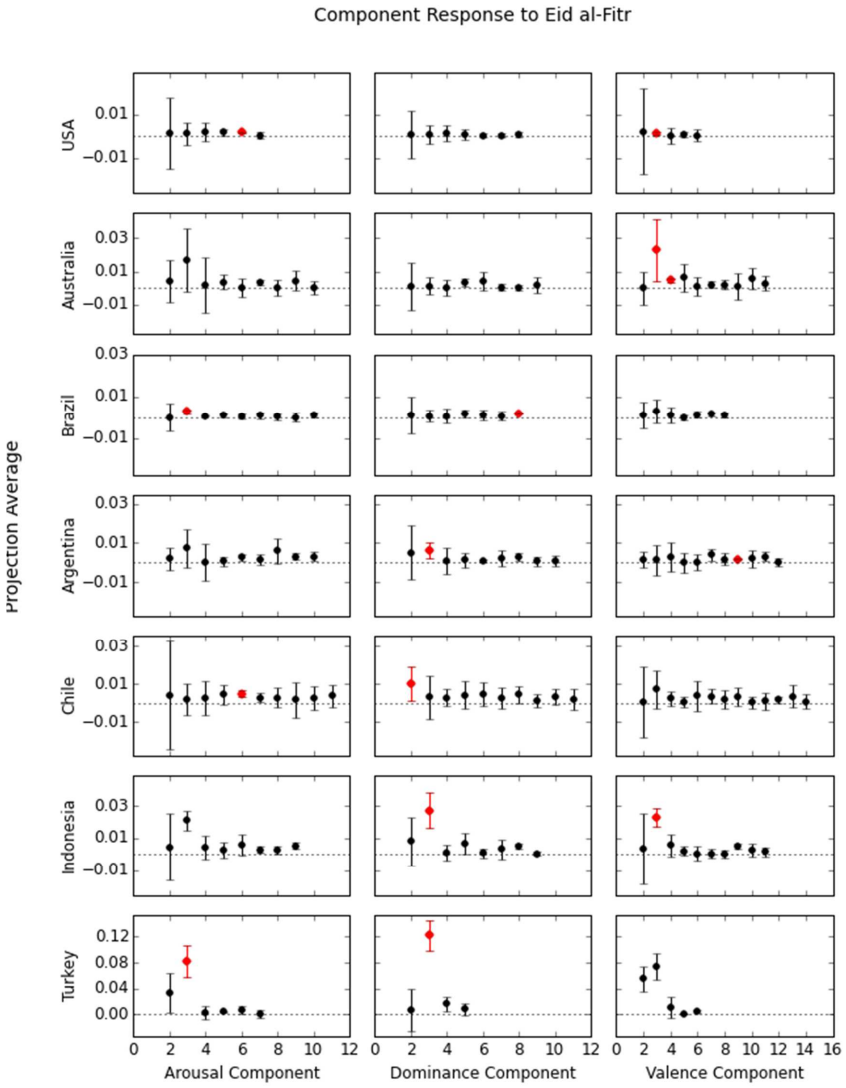


Fig. S11. ANEW component response to Eid-al-Fitr by country. Selected components highlighted in red.

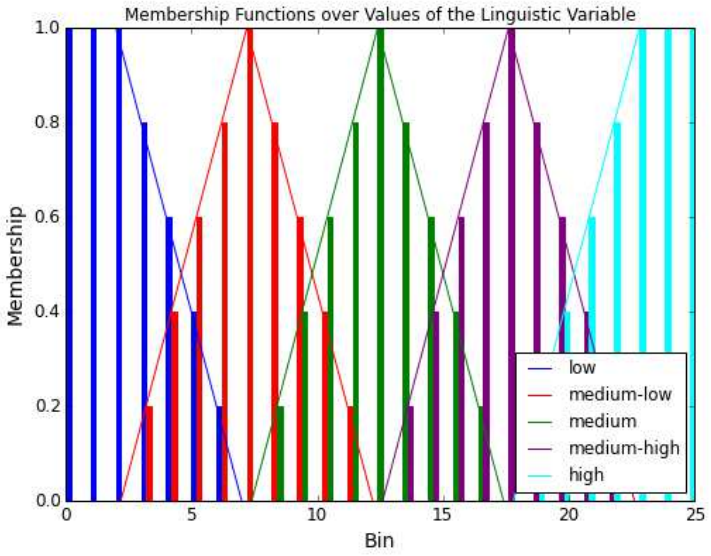


Fig. S12. Linguistic Variable value membership functions over 25 bins. The original bins belong to the values of the linguistic variable ("low", "medium-low", "medium", "medium-high", "high") to different extents. The membership functions are mappings from the original bins to a value between 0 and 1, representing membership fuzzy value of that the linguistic variable can take. The membership functions were chosen such that the sum of a bin's membership across all functions is 1, and the area under each membership function's curve is equal.

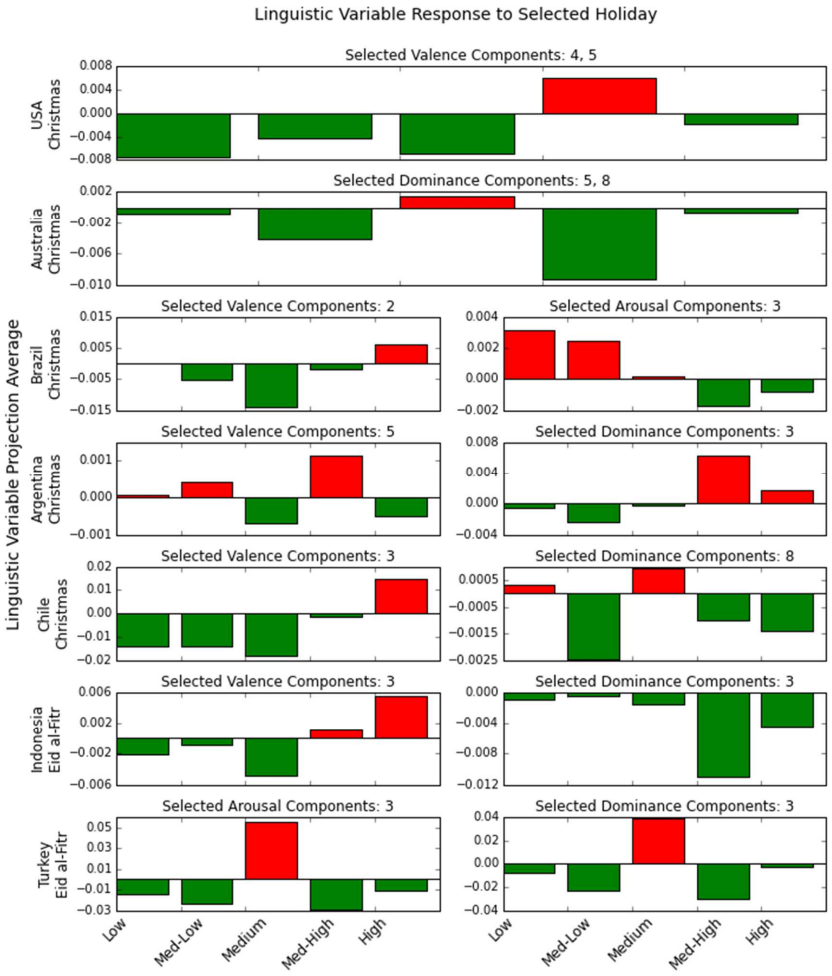
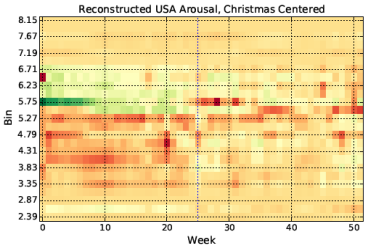


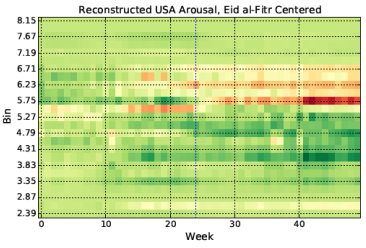
Fig. S13. Linguistic Variable Response to relevant holidays selected for each country, as an aid to interpret the effect of chosen *eigendays* during the holidays. A positive value (in red) means that the members of that value of the linguistic variable had increased weight on the holiday, while negative (in green) means they had decreased weight on the holiday.

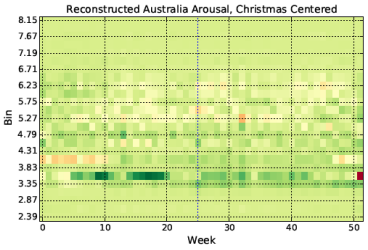
Fig. S14. Average year reconstructed heatmaps. Reconstructed valence heatmaps for each country's average year centered on different holidays. Distributions over time are reconstructed from the components that explain 95% of the variance in the data after the first component is removed. Green represents a decrease in the bin

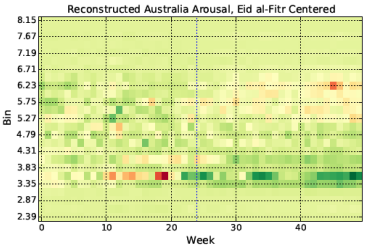
compared to the full distribution, red represents an increase, and yellow represents no change. Center dotted line is the holiday of interest. Left: Christmas, Right: Eid-al-Fitr. Countries top to bottom: USA, Australia, Brazil, Argentina, Chile, Indonesia, Turkey

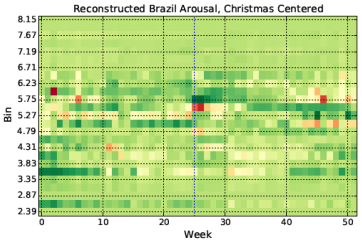
Arousal

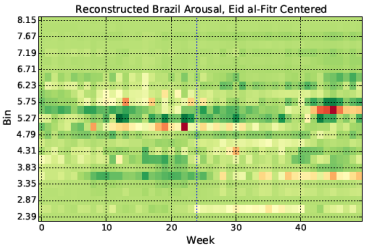


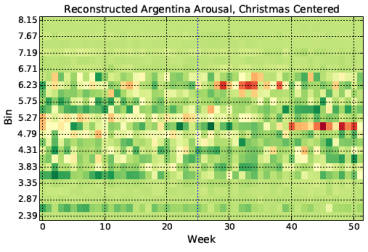


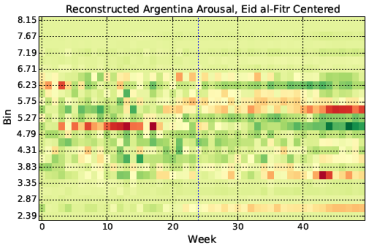


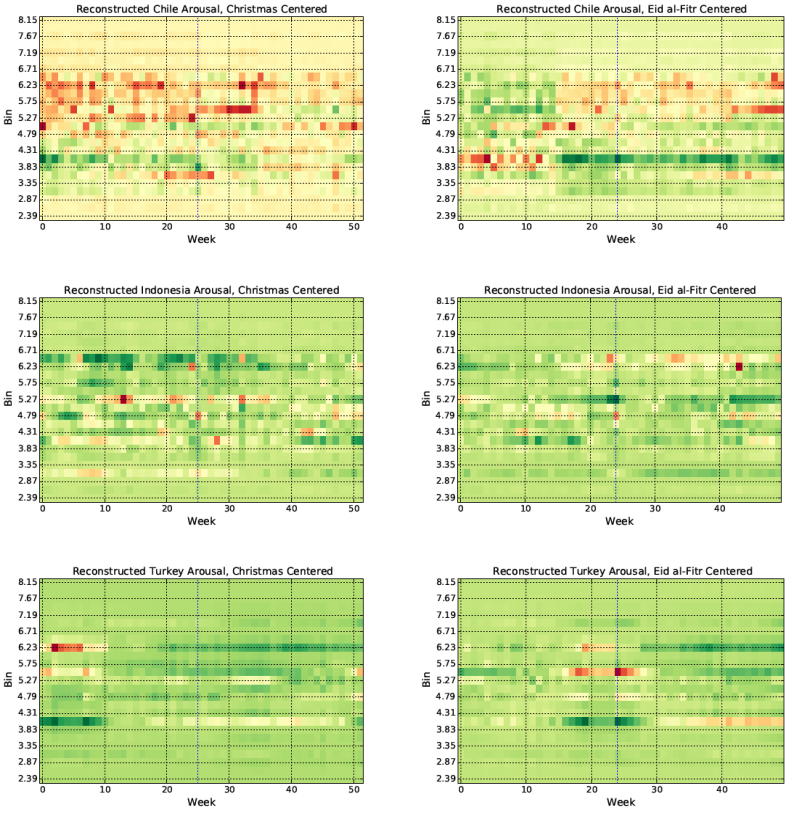




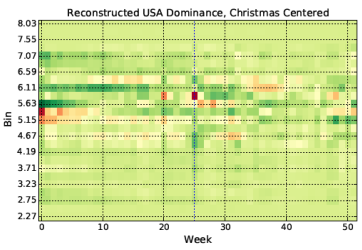


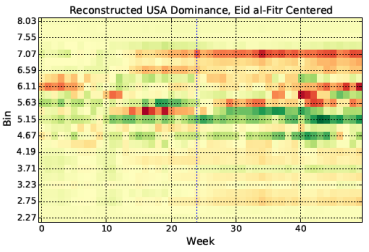


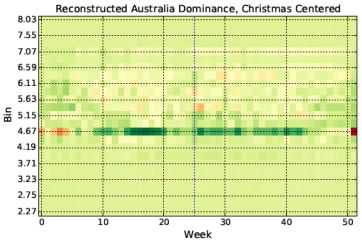


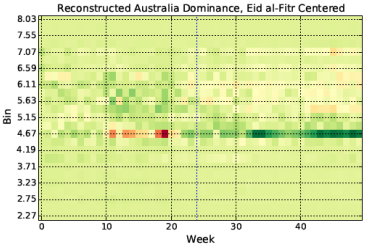

Reconstructed Chile Arousal, Christmas Centered
Reconstructed Chile Arousal, Eid al-Fitr Centered
Reconstructed Indonesia Arousal, Christmas Centered
Reconstructed Indonesia Arousal, Eid al-Fitr Centered
Reconstructed Turkey Arousal, Christmas Centered
Reconstructed Turkey Arousal, Eid al-Fitr Centered
Bin
Week
8.15
7.67
7.19
6.71
6.23
5.75
5.27
4.79
4.31
3.83
3.35
2.87
2.39
0
10
20
30
40
50

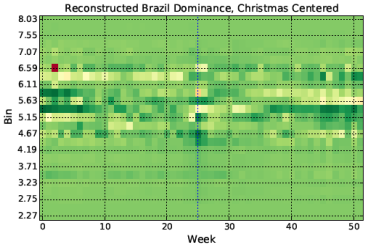
Dominance

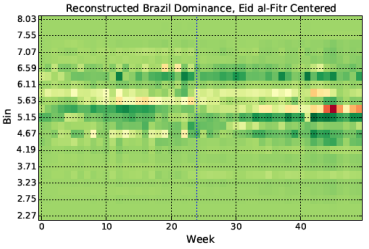

Reconstructed USA Dominance, Christmas Centered
Bin
Week

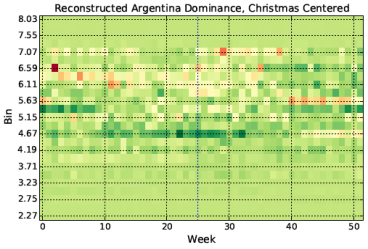

Reconstructed USA Dominance, Eid al-Fitr Centered
Bin
Week

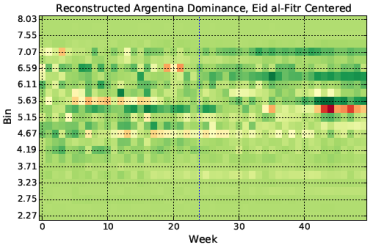

Reconstructed Australia Dominance, Christmas Centered
Bin
Week

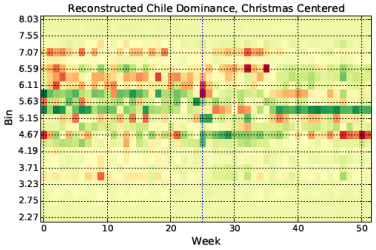

Reconstructed Australia Dominance, Eid al-Fitr Centered
Bin
Week

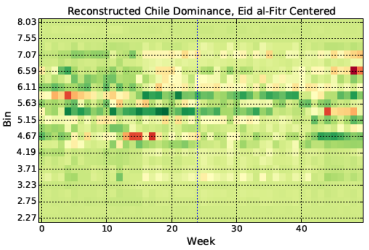

Reconstructed Brazil Dominance, Christmas Centered
Bin
Week

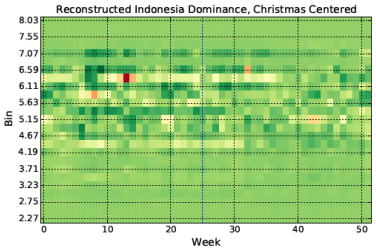

Reconstructed Brazil Dominance, Eid al-Fitr Centered
Bin
Week

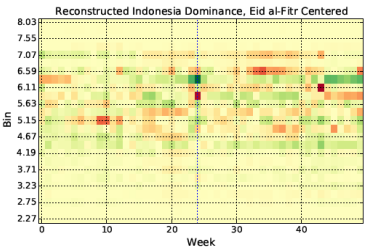

Reconstructed Argentina Dominance, Christmas Centered
Bin
Week

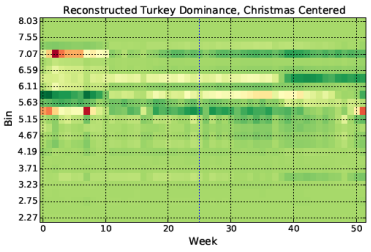

Reconstructed Argentina Dominance, Eid al-Fitr Centered
Bin
Week

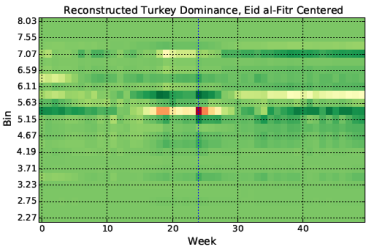

Reconstructed Chile Dominance, Christmas Centered
Bin
Week


Reconstructed Chile Dominance, Eid al-Fitr Centered
Bin
Week

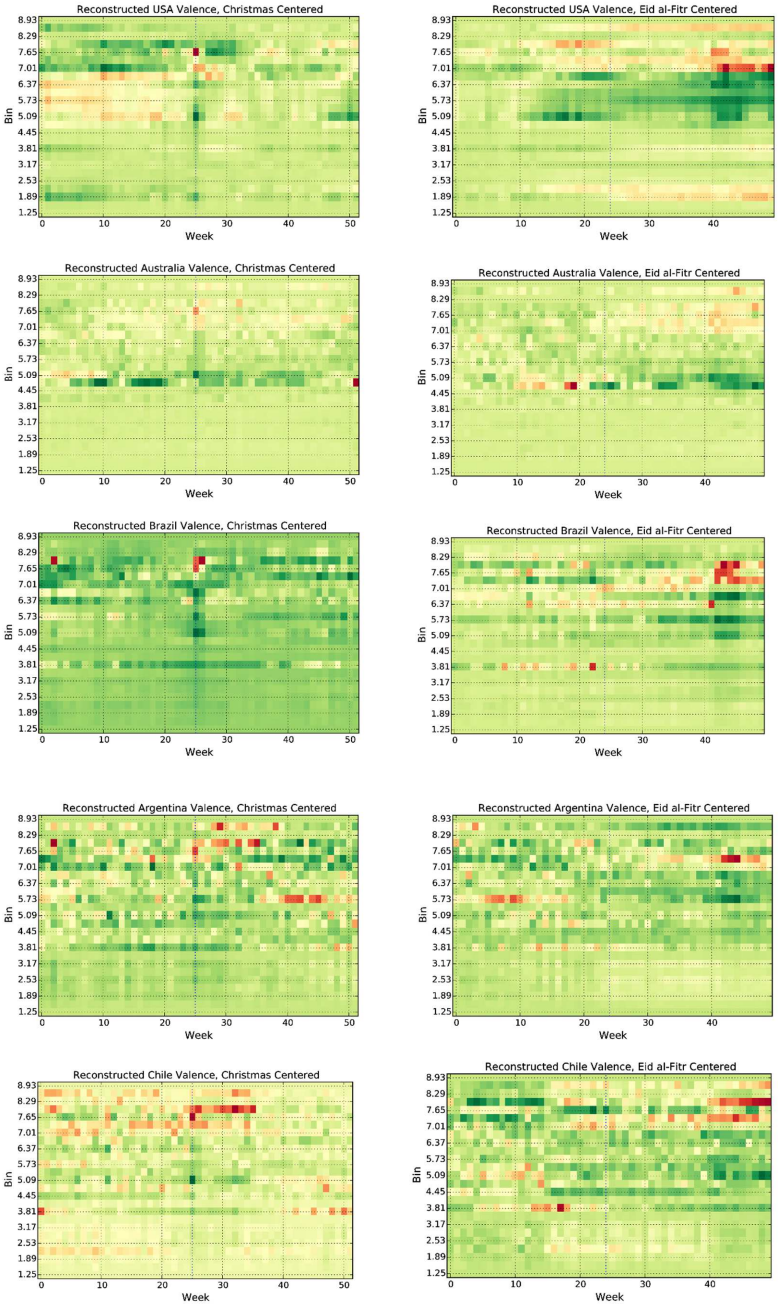

Reconstructed Indonesia Dominance, Christmas Centered
Bin
Week


Reconstructed Indonesia Dominance, Eid al-Fitr Centered
Bin
Week


Reconstructed Turkey Dominance, Christmas Centered
Bin
Week


Reconstructed Turkey Dominance, Eid al-Fitr Centered
Bin
Week

Valence


Reconstructed USA Valence, Christmas Centered
Reconstructed USA Valence, Eid al-Fitr Centered
Reconstructed Australia Valence, Christmas Centered
Reconstructed Australia Valence, Eid al-Fitr Centered
Reconstructed Brazil Valence, Christmas Centered
Reconstructed Brazil Valence, Eid al-Fitr Centered
Reconstructed Argentina Valence, Christmas Centered
Reconstructed Argentina Valence, Eid al-Fitr Centered
Reconstructed Chile Valence, Christmas Centered
Reconstructed Chile Valence, Eid al-Fitr Centered
Bin
Week

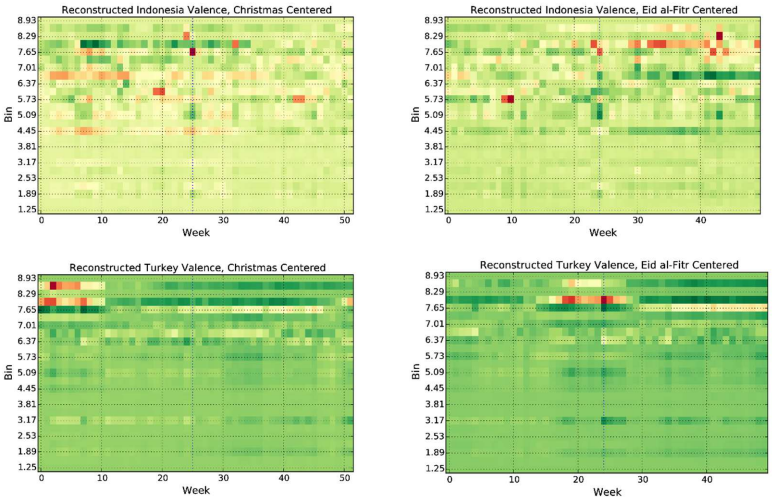
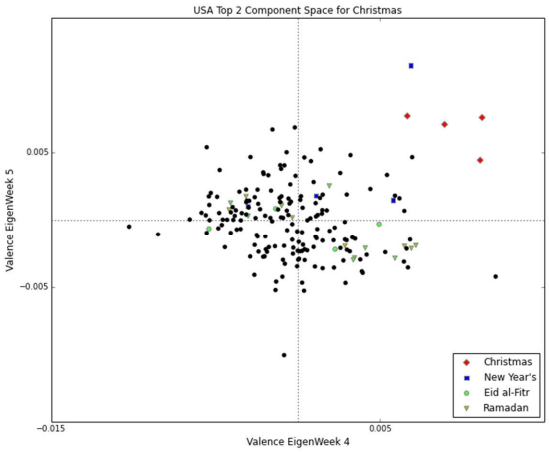
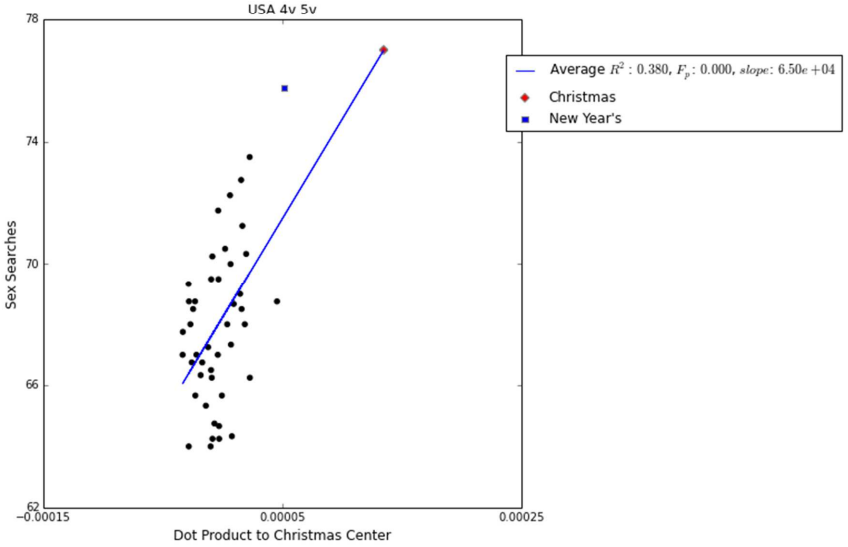

Reconstructed Indonesia Valence, Christmas Centered
Reconstructed Indonesia Valence, Eid al-Fitr Centered
Reconstructed Turkey Valence, Christmas Centered
Reconstructed Turkey Valence, Eid al-Fitr Centered
Bin
Week
8.93
8.29
7.65
7.01
6.37
5.73
5.09
4.45
3.81
3.17
2.53
1.89
1.25
0
10
20
30
40
50

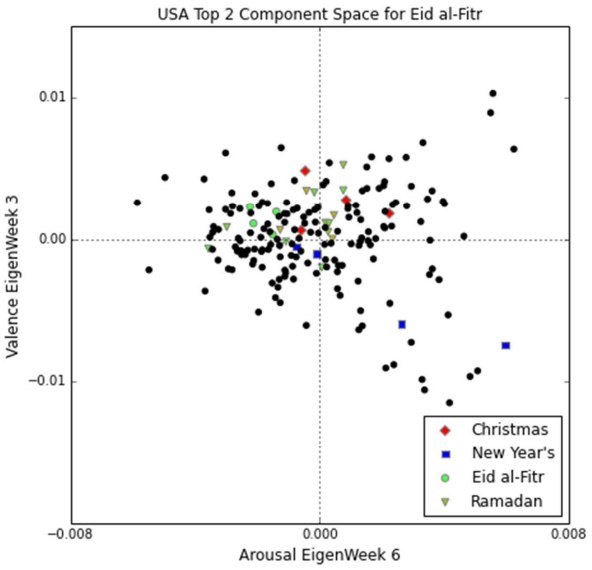
Fig. S15. Eigenmood projections and regressions. Projections show all yearly data points, projected into the space formed by the selected eigenweeks; regressions show the average year's sex searches and similarity to the holiday center.

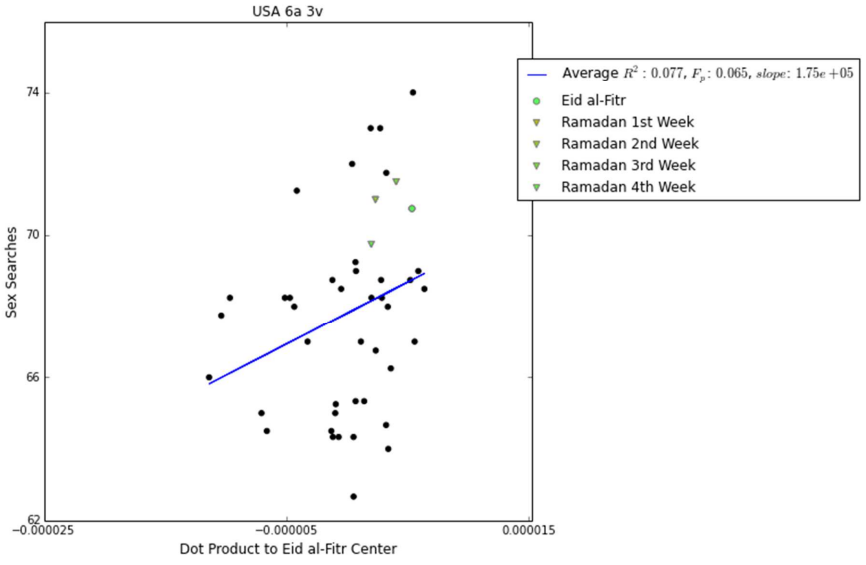
USA Christmas



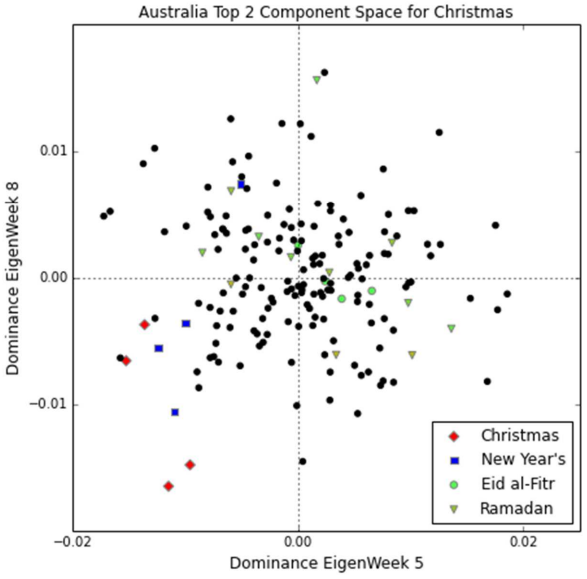


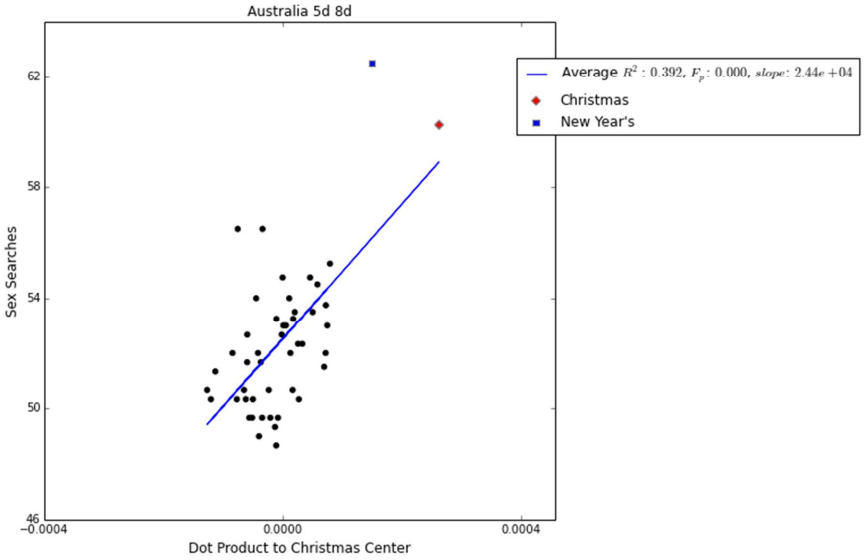
USA Eid-al-Fitr



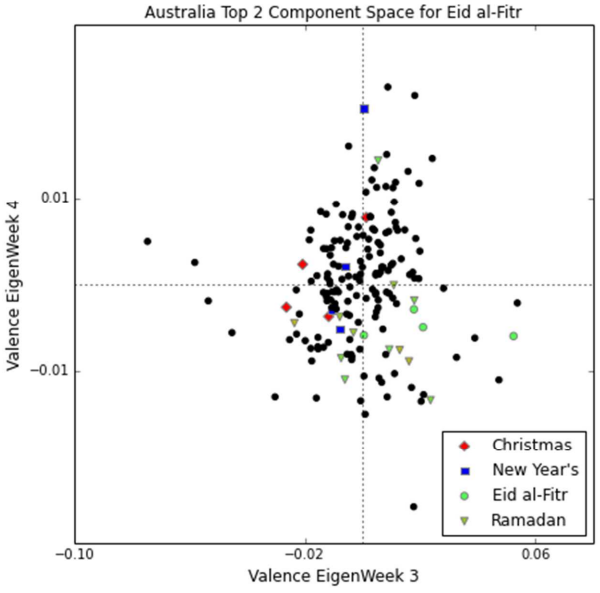


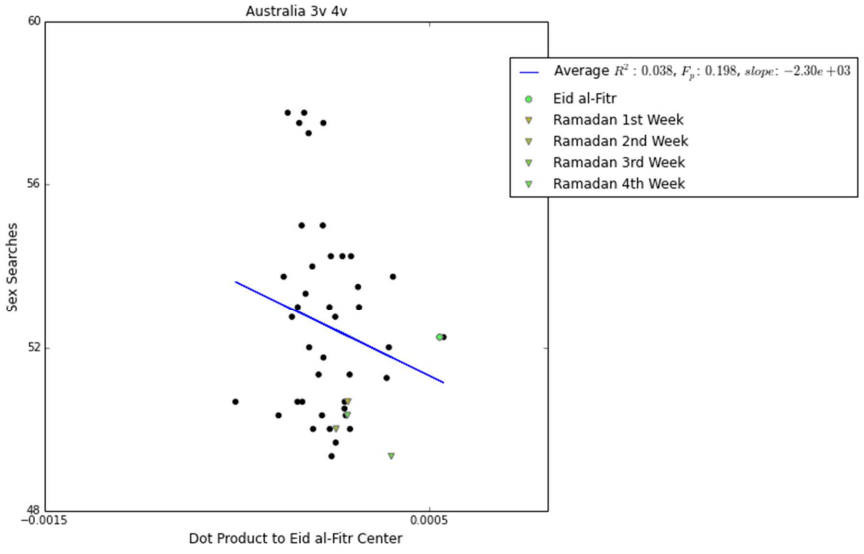
Australia Christmas



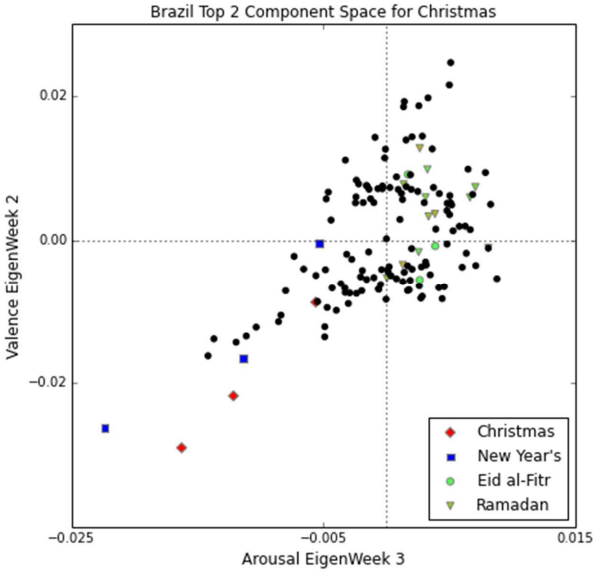


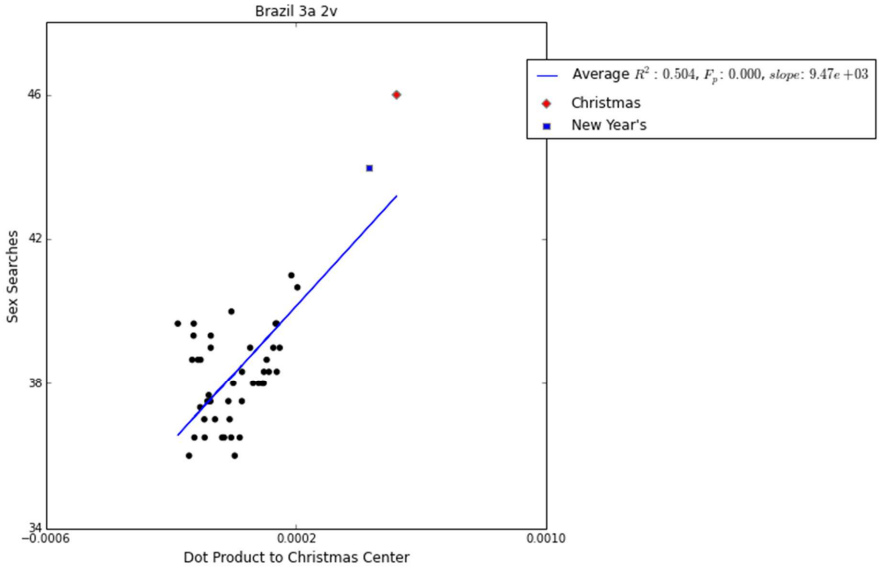
Australia Eid-al-Fitr



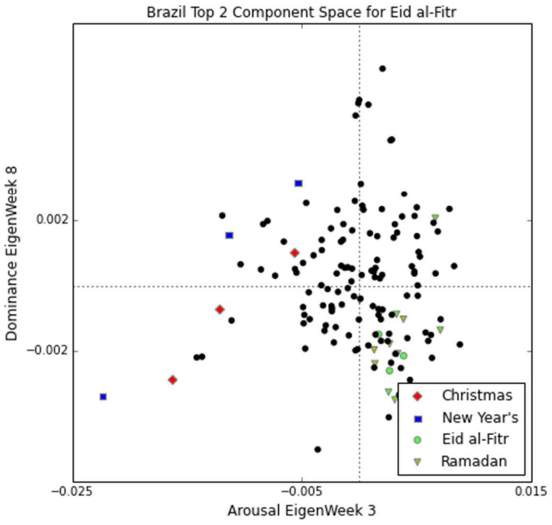


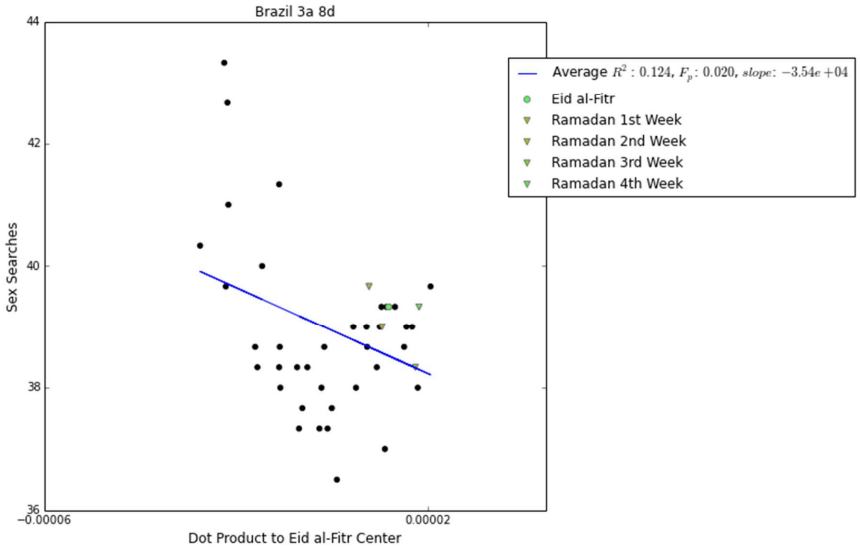
Brazil Christmas



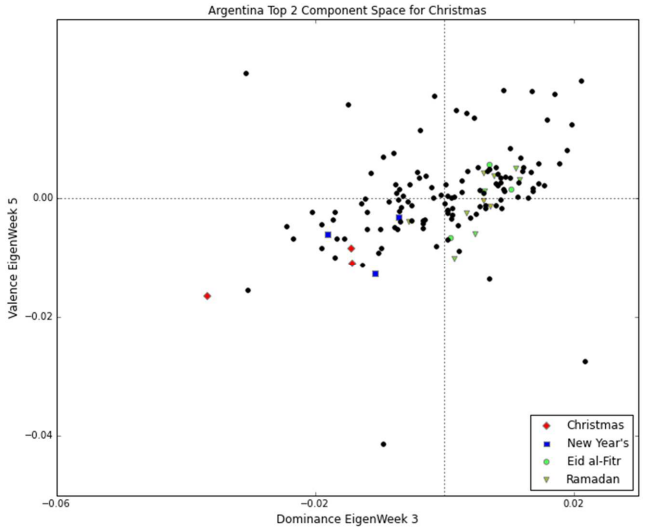


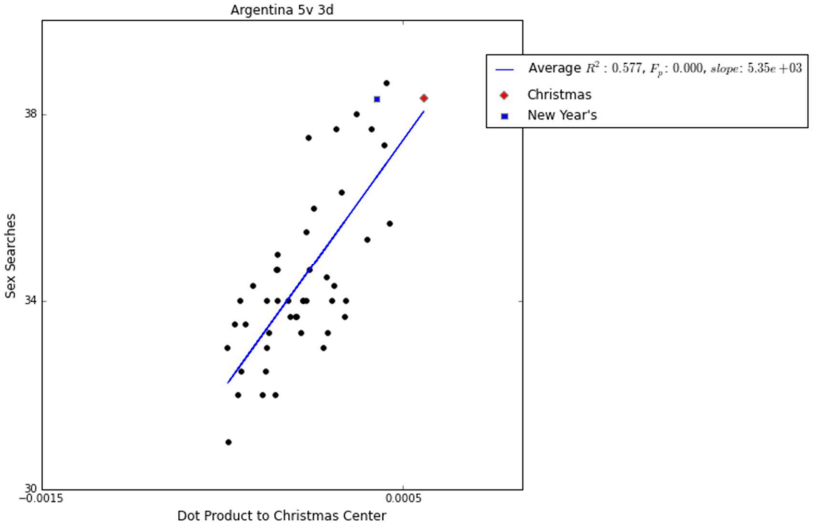
Brazil Eid-al-Fitr


Brazil Top 2 Component Space for Eid al-Fitr
Dominance EigenWeek 8
0.002
−0.002
−0.025
−0.005
0.015
Arousal EigenWeek 3
Christmas
New Year's
Eid al-Fitr
Ramadan

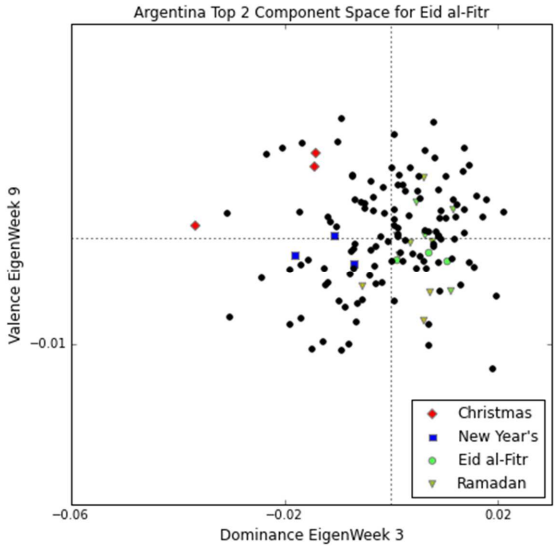

Brazil 3a 8d
Sex Searches
44
42
40
38
36
−0.00006
0.00002
Dot Product to Eid al-Fitr Center
Average R^2 : 0.124, F_p: 0.020, $slope$: −3.54e+04
Eid al-Fitr
Ramadan 1st Week
Ramadan 2nd Week
Ramadan 3rd Week
Ramadan 4th Week

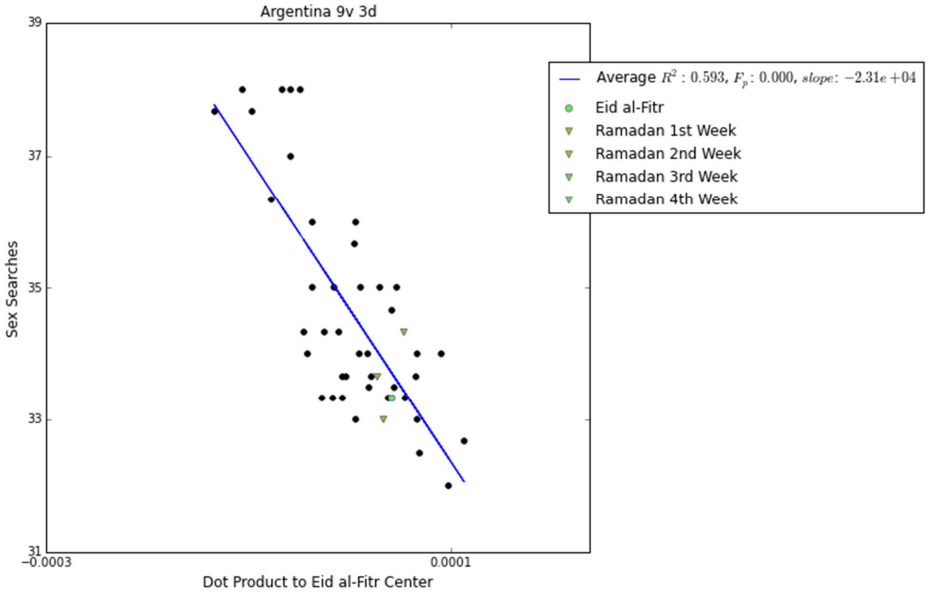
Argentina Christmas



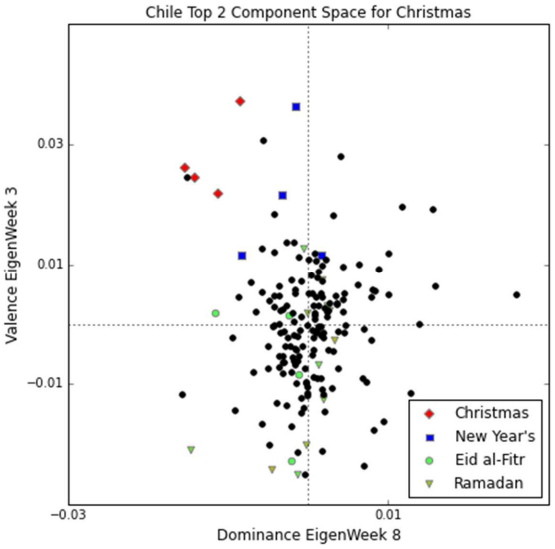


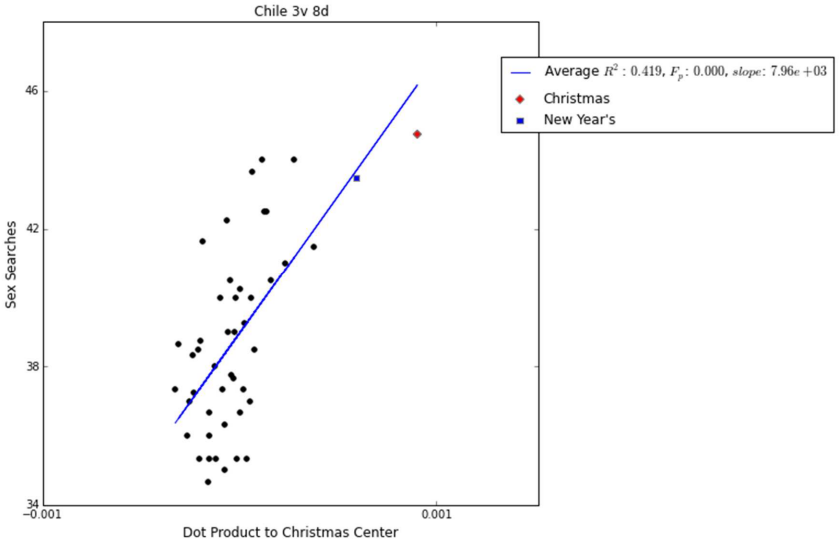
Argentina Eid-al-Fitr



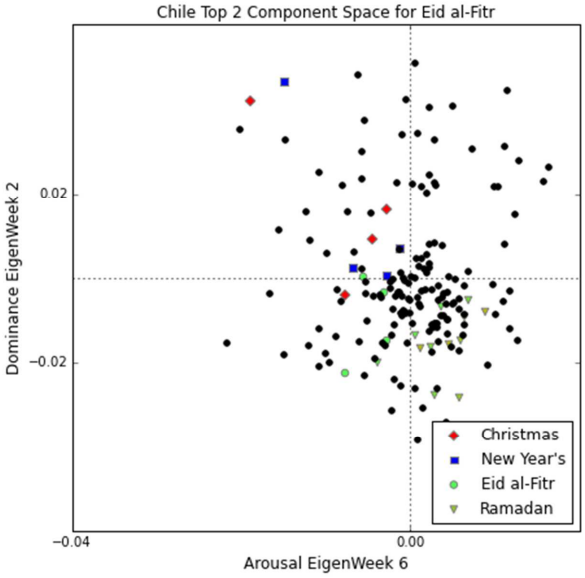


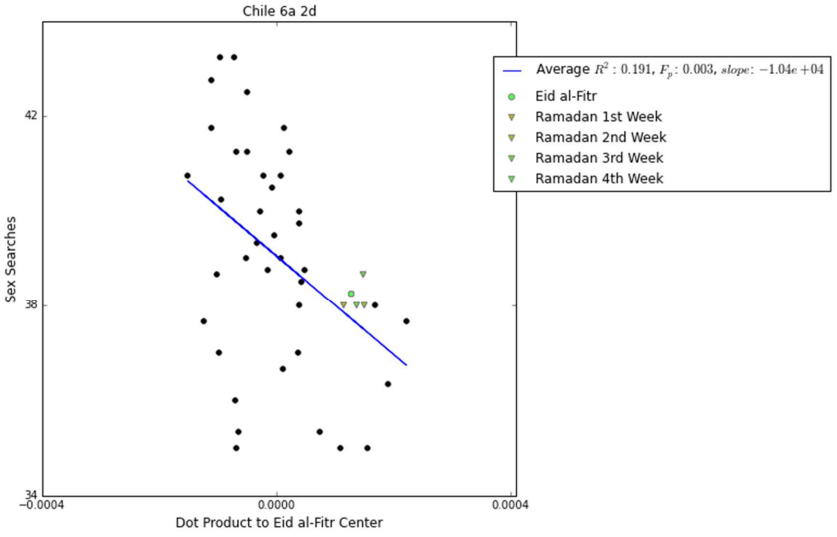
Chile Christmas



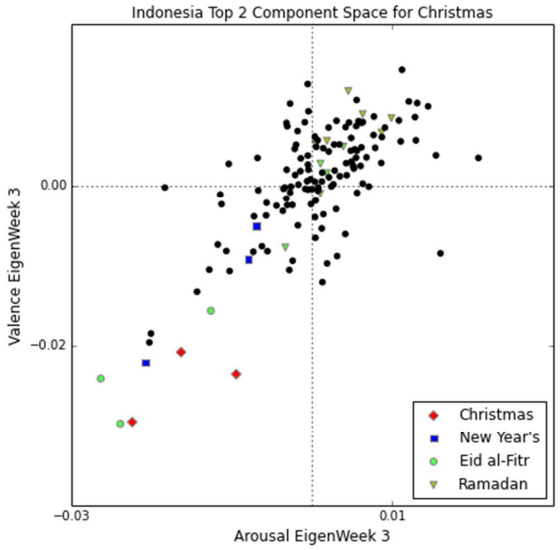


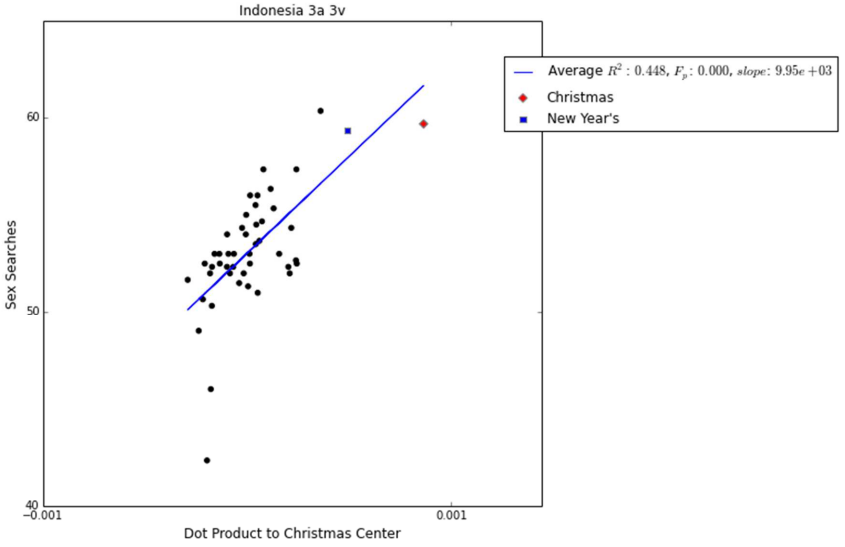
Chile Eid-al-Fitr



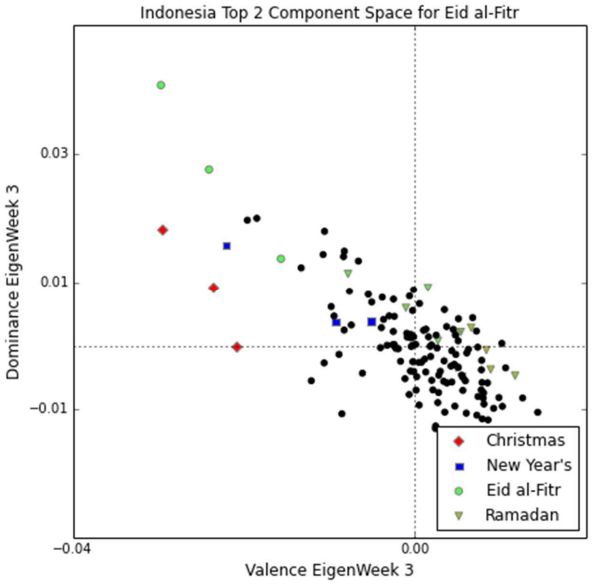


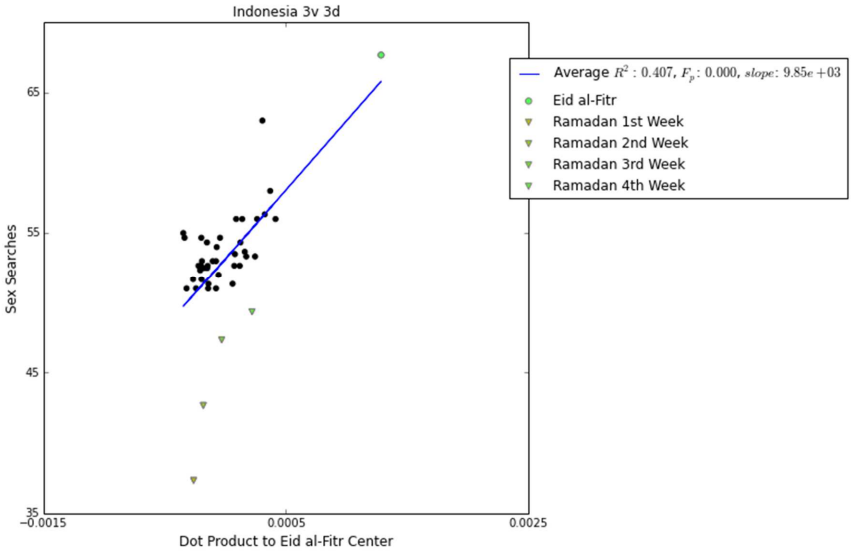
Indonesia Christmas



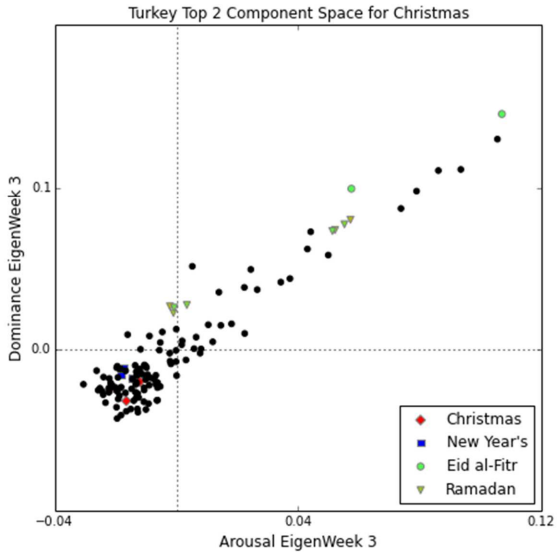


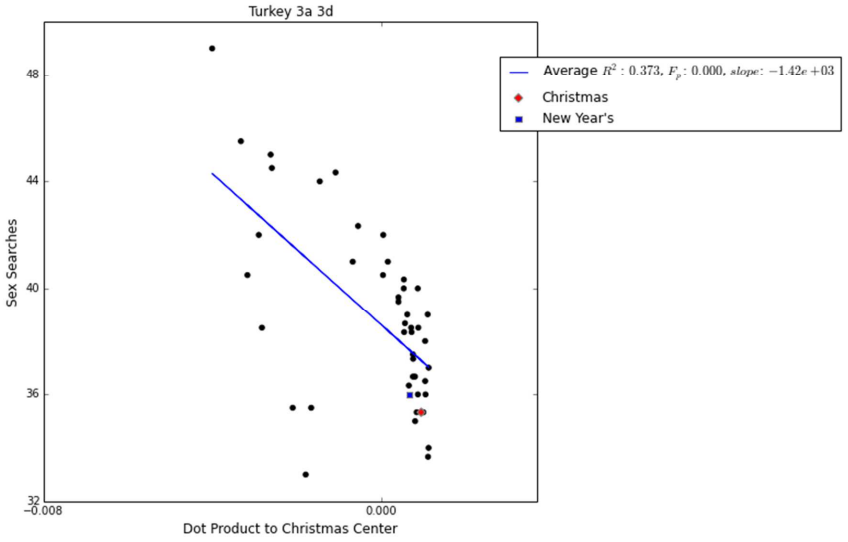
Indonesia Eid-al-Fitr



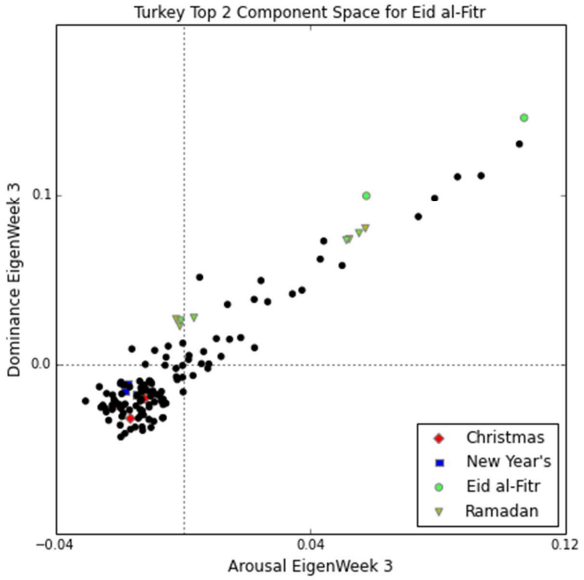


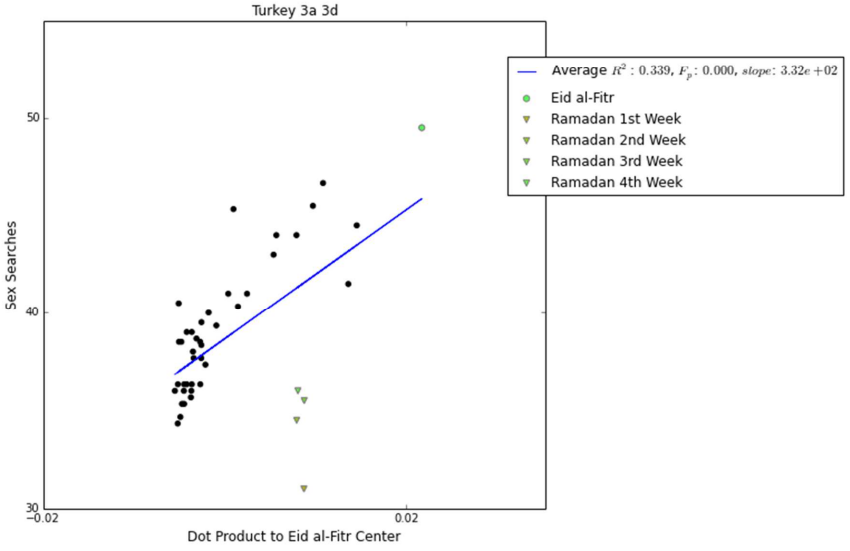
Turkey Christmas





Turkey Eid-al-Fitr





Supplementary Tables

Table S1. Searches for "sex" in select countries. Search queries for "sex" are issued in select countries, representing sexual interest in different cultures, hemispheres, and languages. Google Trends ™ allows the retrieval of search volume time series for multiple search terms. We downloaded GT data for 2 search queries: (1) for the term "sex" and (2) for its translation in the local language as detailed in Supplementary Methods. Table S1 shows the 25 countries and languages that retrieved a sufficiently significant search volume in the local language to support our analysis. From left to right, columns show the: "Countries" for which the analysis was performed; "Search term" in GT; the "Top 5 words associated with the search term", provided and ranked by Google Trends; the "Search Volume Ratio", calculated as the number of searches for "sex" divided by the number of searches for the corresponding translation; and the "Correlation between the two time series ("sex" and the translated word).
The English word "sex" is either more searched for than the corresponding word in the local language (blue to red in the 4th column) or there is a strong correlation between the search terms (red in the 5th column). This is consistent with the fact that the top 5 broad searches most associated with "sex" returned by GT refer to interest in sexual content and pornography in every country (3rd column) and that sexual materials and pornography are widely available in English. The two exceptions are Russia and Israel and neither of these countries is relevant to our analysis.

Country	Search Term	Top 5 Words associated with search term	Search volume ratio	Correlation
Argentina	sex	sex free, sex videos, porn sex, porn, video sex		
	sexo	videos, **sexo** videos, sexo gratis, videos de sexo, porno	0.27	0.81
Australia	sex	free sex, porn sex, porn, sex stories, sex videos		
Brazil	sex	videos sex, sexo, videos, sex video, sex shop		
	sexo	videos sexo, videos, videos de sexo, sexo video, sexo gratis	0.17	0.71
Bulgaria	sex	porno, sex porno, sex free, sex video, sex bg		
	секс	**секс** порно (sex porn), порно (porn), секс игри (sex game), секс клипове (sex videos), клипове (videos)	5.11	-0.08
Chile	sex	sex free, sex videos, video sex, sex porn, sexo		
	sexo	videos, **sexo** videos, sexo gratis, videos de sexo, porno	0.23	0.81
Egypt	sex	sex free, sex arab, sex tube, sex movies, hot sex		
	جنس	افلام (sex movies), افلام جنس (sex sex), سكس جنس (sex), سكس (films), قصص جنس (sex stories)	8.04	0.83
	سكس	صور (films), افلام سكس (sex videos), سكس سكس (sex sex), صور سكس (picture), (photo)	1.15	-0.56
France	sex	sex video, sex free, free, porn sex, porn		
	sexe	video **sexe,** video, sexe gratuit, sexe amateur, amateur	0.98	0.82
Germany	sex	sex free, video sex, sex videos, porn, sex porn		
Greece	sex	sex free, video sex, sex videos, porn, sex porn		
	σεξ	**σεξ** βιντεο (sex video), βιντεο (video), σεξ πορνο (sex porn), πορνο (porn), ιστοριες σεξ (sex stories)	15.60	0.43
India	sex	indian sex, sex videos, free sex, sex stories, hot sex		
Indonesia	sex	cerita sex, cerita, video sex, video, foto sex		
	seks	cerita (story), cerita **seks,** video seks, video, foto seks	7.18	0.20
Israel	sex	sex free, , sex video, sex porn, porn		
	סקס	, , (free sex) סקס חינם, (sex videos) סקס סרטי, (sex) סקס (free sex movies) סרטי סקס חינם	0.72	0.24
Italy	sex	video, sex video, free sex, porno, porno sex		
	sesso	video **sesso,** video, porno, sesso porno, sesso gratis	1.22	0.54
Japan	sex	sex xxx, xxx, sex, sex, sex		
	セックス	セックス動画 (sex video), セックス画像 (sex image), エロ (hello), 無料セックス (free sex)	1.70	-0.30
New Zealand	sex	sex free, porn, sex porn, sex stories, sex videos		
Portugal	sex	free sex, videos sex, porn sex, porn, sex video		
	sexo	videos sexo, videos, videos de sexo, sexo gratis, sexo filmes	0.52	0.85
Russia	sex	sex video, sex free, porno, porno sex, porn sex		
	секс	порно секс (porno sex), порно (porn), видео секс (sex video), фото секс (photo sex), онлайн секс (online sex)	0.61	-0.48
South Africa	sex	porn sex, porn, free sex, sex videos, sex pics		
Spain	sex	sex free, videos, videos sex, sex porn, porn		
	sexo	sexo gratis, videos, videos sexo, videos de sexo, porno	0.46	0.68
Sweden	sex	free sex, porn, sex porn, sex video, sex tube		
Tunisia	sex	sex sex, porno sex, video sex, video, porno		
	جنس	افلام (films), افلام (sex movies), افلام جنس, (arabic sex) **جنس** عربي, سكس (sex), قصص جنس (sex stories)	32.74	0.62
	سكس	افلام سكس (films), افلام (sex arabic), سكس عربي (sex), سكس **سكس** (sex videos), قصص سكس (sex stories)	19.32	0.71
Turkey	sex	porno sex, porno, sex izle, sex hikayeleri, porn		
	seks	porno **seks,** porno, seks hikayeleri (sex stories), seks izle (watch sex), sex	3.29	0.92
UK	sex	free sex, sex porn, porn, sex videos, sex tape		
USA	sex	free sex, sex videos, sex porn, porn, video sex		
Vietnam	sex	phim (movies), phim sex (porn movies), truyen sex (manga sex), truyen (manga), anh sex (he sex)		
Worldwide	sex	sex free, free, sex videos, sex porn, porn		

Table S2. Countries analyzed and categorized according to religion and geographical location (hemisphere) The 1st column shows the international country code, the 2nd columns shows the Country name; the third column (Week) shows the first week for which we could find stable GT™ data. A country

was considered "culturally Christian" when at least half of its population identified as Christian (Catholic, Protestant, Orthodox, or other) according to [13]. A country was considered "culturally Muslim" when at least half of its population identified as Muslim according to [14]. A country was labeled as "Other" when the majority of its population didn't identify as either Christian or Muslim. The 4th column, "Country Set" shows how each country was categorized and the 5th and 6th columns show the percentage of the population that identify as Christian or Muslim, respectively. The 7th and 8th columns show the continent and the Hemisphere to which each country belongs, according to Wikipedia.

Code	Country Name	First Week	Country Set	% Christian	% Muslim	Continent	Hemisphere
AE	United Arab Emirates	04-01-2004	Muslim	2.6 (2.6;)	76	Asia	North
AF	Afghanistan	12-11-2006	Muslim	0.02 (;)	99.8	Asia	North
AL	Albania	06-11-2005	Muslim	17 (7;10)	82.1	Europe	North
AR	Argentina	04-01-2004	Christian	90 (77;13)	2.5	South America	South
AT	Austria	04-01-2004	Christian	68.4 (62.4;6)	5.7	Europe	North
AU	Australia	04-01-2004	Christian	63 (25.8;37)	1.9	Oceania	South
AW	Aruba	04-06-2006	Christian	88 (80.8;7.8)	0	North America	South
BA	Bosnia and Herzegovina	04-01-2004	Christian	52 (15;37)	41.6	Europe	North
BD	Bangladesh	04-01-2004	Muslim	0.3 (0.3;)	90.4	Asia	North
BE	Belgium	04-01-2004	Christian	55.4 (57;7)	6	Europe	North
BG	Bulgaria	04-01-2004	Christian	84 (1;83)	13.4	Europe	North
BH	Bahrain	04-01-2004	Muslim	9 (;9)	81.2	Asia	North
BN	Brunei	08-01-2006	Muslim	11 (;)	51.9	Asia	North
BO	Bolivia	04-01-2004	Christian	89 (76;13)	2.5	South America	South
BR	Brazil	04-01-2004	Christian	90.2 (63;27)	0.1	South America	South
BS	Bahamas	05-06-2005	Christian	81 (13.5;67.6)	0	Central America	North
BY	Belarus	01-01-2006	Christian	55.4 (7.1;48.3)	0.2	Europe	North
CA	Canada	04-01-2004	Christian	67.3 (38.7;29)	2.8	North America	North
CH	Switzerland	04-01-2004	Christian	71 (38;33)	5.7	Europe	North
CL	Chile	04-01-2004	Christian	87.2 (67;20)	0	South America	South
CM	Cameroon	26-08-2007	Christian	65 (38.4;26.3)	18	Africa	North
CN	China	04-01-2004	Other	5 (1;4)	1.8	Asia	North
CO	Colombia	04-01-2004	Christian	90 (75;15)	0	South America	North
CR	Costa Rica	04-01-2004	Christian	83 (69;14)	0	Central America	North
CY	Cyprus	04-01-2004	Christian	79.3 (4.3;75)	22.7	Europe	North
CZ	Czech Republic	04-01-2004	Other	11.2 (10.4;0.8)	0	Europe	North
DE	Germany	04-01-2004	Christian	62 (30;32)	5	Europe	North
DJ	Djibouti	06-01-2008	Muslim	6 (1;5)	97	Africa	North
DK	Denmark	04-01-2004	Christian	81 (1;80)	4.1	Europe	North
Code	**Country Name**	**First Week**	**Country Set**	**% Christian**	**% Muslim**	**Continent**	**Hemisphere**

DO	Dominican Republic	04-01-2004	Christian	95 (95;)		North America	North
DZ	Algeria	04-01-2004	Muslim	2 (1;1)	98.2	Africa	North
EC	Ecuador	04-01-2004	Christian	94 (74;20)	0	South America	South
EE	Estonia	04-01-2004	Other	23.9 (0;23)	0.1	Europe	North
EG	Egypt	04-01-2004	Muslim	18 (0;18)	94.7	Africa	North
ES	Spain	04-01-2004	Christian	73 (71;2)	2.3	Europe	North
ET	Ethiopia	04-01-2004	Christian	63.4 (0;63.4)	33.8	Africa	North
FI	Finland	04-01-2004	Christian	81.6 (0;81)	0.8	Europe	North
FJ	Fiji	03-09-2006	Christian	64.4 (8.9;55.5)	6.3	Oceania	South
FR	France	04-01-2004	Christian	65 (63;2)	7.5	Europe	North
GE	Georgia	01-05-2005	Christian	88.6 (0.9;87.7)	10.5	Europe	North
GH	Ghana	16-10-2005	Christian	68.8 (13.1;55.5)	16.1	Africa	North
GP	Guadalupe	09-03-2008	Christian	96 (95;1)		North America	North
GR	Greece	04-01-2004	Christian	97 (0;97)	4.7	Europe	North
GT	Guatemala	04-01-2004	Christian	87 (47;40)	0	Central America	North
GU	Guam	17-12-2006	Christian	85 (;)	0.1	Oceania	South
HN	Honduras	04-09-2005	Christian	87.6 (47;40)	0.1	Central America	North
HR	Croatia	04-01-2004	Christian	90(70;20)	1.3	Europe	North
HU	Hungary	04-01-2004	Christian	82.7 (70.1;11.6)	0.3	Europe	North
ID	Indonesia	04-01-2004	Muslim	10(3;7)	88.1	Asia	South
IE	Ireland	04-01-2004	Christian	94.1 (82;12)	0.9	Europe	North
IL	Israel	04-01-2004	Other	3.5(;3.5)	17.7	Asia	North
IN	India	04-01-2004	Other	2.6 (1.6;1)	14.6	Asia	North
IQ	Iraq	12-12-2004	Muslim	3(;3)	98.9	Asia	North
IR	Iran	04-01-2004	Muslim	0.4(;)	99.7	Asia	North
IS	Iceland	04-01-2004	Christian	95 (2.5;92.5)	0.1	Europe	North
IT	Italy	04-01-2004	Christian	85.1 (85;0)	2.6	Europe	North
JM	Jamaica	04-01-2004	Christian	65.3 (2;63.3)	0	Central America	North
JO	Jordan	04-01-2004	Muslim	6 (;)	98.8	Asia	North
JP	Japan	04-01-2004	Other	2 (1;1)	0.1	Asia	North
KE	Kenya	04-01-2004	Christian	85.1 (23.4;61.7)	7	Africa	North
KH	Cambodia	05-12-2004	Other	1 (0.15;0.85)	1.6	Asia	North
KR	South Korea	04-01-2004	Other	(;)	0.2	Asia	North
KW	Kuwait	04-01-2004	Muslim	15 (3.2;12.8)	86.4	Asia	North
KZ	Kazakhstan	01-10-2006	Muslim	51 (0.16;50)	56.4	Europe	North
LA	Laos	15-04-2007	Other	2.2 (1;1)	0	Asia	North
Code	**Country Name**	**First Week**	**Country Set**	**% Christian**	**% Muslim**	**Continent**	**Hemisphere**
LB	Lebanon	04-01-2004	Muslim	41 (26;15)	59.7	Asia	North
LK	Sri Lanka	04-01-2004	Other	7.5 (6.1;1.4)	8.5	Asia	North

LT	Lithuania	04-01-2004	Christian	84.9 (77.2;7.6)	0.1	Europe	North
LU	Luxemburg	04-01-2004	Christian	71 (69;2)	2.3	Europe	North
LV	Latvia	04-01-2004	Christian	57 (25;32.2)	0.1	Europe	North
MA	Morocco	04-01-2004	Muslim	2.1 (0.1;2)	99.9	Africa	North
MD	Moldova	02-10-2005	Christian	97.53 (0;93)	0.4	Europe	North
ME	Montenegro	13-11-2005	Christian	78.8 (3.4;72.07)	18.5	Europe	North
MK	Macedonia	04-01-2004	Christian	65.1 (0.3;64.8)	34.9	Europe	North
MM	Myanmar	04-12-2005	Other	7.9 (1;6.9)	3.8	Asia	North
MN	Mongolia	14-08-2005	Other	2.1 (;)	4.4	Asia	North
MT	Malta	04-01-2004	Christian	97 (;)	0.3	Europe	North
MU	Mauritius	10-07-2005	Other	32.2 (-;-)	16.6	Africa	South
MV	Maldives	04-01-2004	Muslim	41 (26;15)	98.4	Asia	North
MX	Mexico	04-01-2004	Christian	92 (;)	0.1	North America	North
MY	Malaysia	04-01-2004	Muslim	12.1 (;)	61.4	Asia	North
MZ	Mozambique	24-02-2008	Christian	56.1 (28.4;27.7)	22.8	Africa	South
NA	Namibia	27-06-2010	Christian	90 (13.7;76.3)	0.4	Africa	South
NG	Nigeria	04-01-2004	Christian	50.01 (14;36)	47.9	Africa	North
NI	Nicaragua	16-08-2009	Christian	89.6 (58.8;30.8)	0	Central America	North
NL	Netherlands	04-01-2004	Other	44 (24;20)	5.5	Europe	North
NO	Norway	04-01-2004	Christian	86.2 (3;83.5)	3	Europe	North
NP	Nepal	04-01-2004	Other	0.9 (0.1;0.8)	4.2	Asia	North
NZ	New Zealand	04-01-2004	Christian	55.6 (28.7;24.9)	0.9	South America	South
OM	Oman	04-01-2004	Muslim	2.5 (2.1;0.4)	87.7	Asia	North
PA	Panama	15-02-2004	Christian	92 (80;12)	0.7	Central America	North
PE	Peru	04-01-2004	Christian	96 (81;15)	0	South America	South
PH	Philippines	04-01-2004	Christian	93 (80;13)	5.1	Asia	North
PK	Pakistan	04-01-2004	Muslim	1.6 (0.8;0.8)	96.4	Asia	North
PL	Poland	04-01-2004	Christian	94.3 (86.3;8)	0.1	Europe	North
PR	Puerto Rico	04-01-2004	Christian	97 (50;47)	0	North America	North
PS	Palestine	04-01-2004	Muslim	(;)	97.5	Asia	North
PT	Portugal	04-01-2004	Christian	95.7 (81;14.7)	0.6	Europe	North
PY	Paraguay	12-02-2006	Christian	96 (88;7.9)	0	South America	South
QA	Qatar	04-01-2004	Muslim	13.8 (;)	77.5	Asia	North
RO	Romania	04-01-2004	Christian	99.5 (5.7;93.8)	0.3	Europe	North
Code	**Country Name**	**First Week**	**Country Set**	**% Christian**	**% Muslim**	**Continent**	**Hemisphere**
RS	Serbia	04-01-2004	Christian	93.5 (4.97;79.4)	3.7	Europe	North
RU	Russia	04-01-2004	Christian	60 (0;60)	11.7	Europe	North
SA	Saudi Arabia	04-01-2004	Muslim	5.5 (3.5;2)	97.1	Asia	North

SD	Sudan	11-01-2004	Muslim	2 (;)	71.4	Africa	North
SE	Sweden	04-01-2004	Christian	67.2 (2;65)	4.9	Europe	North
SG	Singapore	04-01-2004	Other	18 (5.7;12)	14.9	Asia	North
SI	Slovenia	04-01-2004	Christian	79.2 (57;22.2)	2.4	Europe	North
SK	Slovakia	04-01-2004	Christian	86.5 (75.2;11.3)	0.1	Europe	North
SV	El Salvador	04-01-2004	Christian	81.9 (52.6;29.3)	0	Central America	North
SY	Syria	04-01-2004	Muslim	10 (0;10)	92.8	Asia	North
TH	Thailand	04-01-2004	Other	0.7 (0.4;0.3)	5.8	Asia	North
TN	Tunisia	04-01-2004	Muslim	0.2 (;0.2)	99.8	Africa	North
TR	Turkey	04-01-2004	Muslim	0.2 (;)	98.6	Europe	North
TT	Trinidad and Tobago	04-01-2004	Christian	57.6 (21.5;33.4)	5.8	Central America	North
TW	Taiwan	04-01-2004	Other	3.9 (2.6;1.3)	0.1	Asia	North
TZ	Tanzania	04-01-2004	Christian	62 (;)	29.9	Africa	South
UA	Ukraine	04-01-2004	Christian	83.8 (5.9;76.7)	0.9	Europe	North
UG	Uganda	08-01-2006	Christian	88.6 (41.9;46.7)	12	Africa	North
UK	United Kingdom	04-01-2004	Christian	59.3 (8.9;50)	4.6	Europe	North
US	United States of America	04-01-2004	Christian	73 (22;51)	0.8	North America	North
UY	Uruguay	04-01-2004	Christian	58.4 (47;11)	0	South America	South
UZ	Uzbekistan	17-10-2004	Muslim	2.6 (2.6;)	96.5	Asia	North
VE	Venezuela	04-01-2004	Christian	87 (79;8)	9.3	South America	North
VN	Vietnam	04-01-2004	Other	8 (7;1)	0.2	Asia	North
YE	Yemen	04-01-2004	Muslim	0.0013 (0.0013;)	99	Asia	North
ZA	South Africa	04-01-2004	Christian	80 (5;75)	1.5	Africa	South
ZM	Zambia	06-05-2007	Christian	97.6 (25;72)	0.4	Africa	South
ZW	Zimbabwe	05-03-2006	Christian	85 (7;77)	0.9	Africa	South

Table S3. **Correlation Table for the averaged time series of all countries grouped either by hemisphere (Northern or Southern) or by religion (Muslim or Christian).**
Table S3a shows R^2 and Table S3b shows the corresponding p-values.

Table S3a.

	Northern	Southern	Christian	Muslim
Northern	1			
Southern	0.536811	1		
Christian	0.890322	0.627146	1	
Muslim	0.415906	0.309619	0.192213	1

Table S3b.

	Northern	Southern	Christian	Muslim
Northern	1			
Southern	5.89E-90	1		
Christian	1.3E-254	9.1E-115	1	
Muslim	2.07E-63	2.98E-44	3.27E-26	1

Table S4. The three major Muslim holidays, in regard to the Gregorian calendar, for the period under analysis.

Beginning of Ramadan	*Eid-al-Fitr*	*Eid al-Adha*
15 Oct 2004	14 Nov 2004	21 Jan 2005
4 Oct 2005	3 Nov 2005	10 Jan 2006
24 Sep 2006	23 Oct 2006	31 Dec 2006
13 Sep 2007	13 Oct 2007	20 Dec 2007
1 Sep 2008	1 Oct 2008	8 Dec 2008
22 Aug 2009	20 Sep 2009	27 Nov 2009
11 Aug 2010	10 Sep 2010	16 Nov 2010
1 Aug 2011	30 Aug 2011	6 Nov 2011
20 Jul 2012	19 Aug 2012	26 Oct 2012
9 Jul 2013	8 Aug 2013	15 Oct 2013

Table S5 - Starting day of the "Christian Calendar", starting day of the weeks that included December 25th – Christmas (always on week 26), the last week of each centered year and the discarded exception weeks after centering.

1	*26*	*week 52*	*exception week*
-		6/20/2004	-
6/27/2004	12/19/2004	6/19/2005	6/26/2005
7/26/2005	12/25/2005	6/25/2006	
7/2/2006	12/24/2006	6/24/2007	-
7/1/2007	12/23/2007	6/22/2008	-
6/29/2008	12/21/2008	6/21/2009	-
6/28/2009	12/20/2009	6/20/2010	-
6/27/2010	12/19/2010	6/19/2011	26 June 2011
7/3/2011	12/25/2011	6/24/2012	-
7/1/2012	12/23/2012	6/23/2013	-
6/30/2013	12/22/2013	-	-

Table S6. Weeks that included Eid-al-Fitr and the discarded exception weeks after centering.

1	*25*	*week 50*	*exception week*
-		5/23/2004	-
5/30/2004	11/14/2004	5/8/2005	-
5/15/2005	10/30/2005	4/23/2006	4/30/2006
5/7/2006	10/22/2006	4/15/2007	-
4/22/2007	10/7/2007	3/30/2008	4/6/2008
4/13/2008	9/28/2008	3/22/2009	3/29/2009
4/5/2009	9/20/2009	3/14/2010	-
3/21/2010	9/5/2010	2/27/2011	3/6/2011
3/13/2011	8/28/2011	2/19/2012	2/26/2012
3/4/2012	8/19/2012	2/10/2013	-
2/17/2013	8/4/2013	1/26/2014	-
2/2/2014	-	-	-

Table S7. Z-scores on the corresponding centered week for all countries in the dataset, calculated from the each country's average for each week, as detailed in the Methods. When z>1 for both the Christmas and Eid-al-Fitr centered calendars, classification was based on the higher score (bold).

	Country	Country Set	Hemisphere	Christmas	Eid-al-Fitr	June Solstice	Dec Solstice
AE	United Arab Emirates	Muslim	North	1.877	**3.023**	0.179	1.313
AF	Afghanistan	Muslim	North	0.654	0.443	0.587	0.889
AL	Albania	Muslim	North	0.372	1.417	0.399	0.491
AR	Argentina	Christian	South	2.190	-2.066	0.395	1.146
AT	Austria	Christian	North	3.598	-0.089	-0.724	1.879
AU	Australia	Christian	South	3.598	-0.089	-0.724	1.879
AW	Aruba	Christian	South	**1.970**	1.960	-0.570	1.502
BA	Bosnia and Herzegovina	Christian	North	-0.312	0.883	0.658	-0.477
BD	Bangladesh	Muslim	North	1.544	**2.576**	0.701	1.062
BE	Belgium	Christian	North	1.713	0.315	0.770	0.350
BG	Bulgaria	Christian	North	0.843	0.476	1.169	-0.443
BH	Bahrain	Muslim	North	1.151	2.492	1.128	1.879
BN	Brunei	Muslim	North	1.183	2.075	0.912	2.005
BO	Bolivia	Christian	South	3.028	0.831	0.074	1.159
BR	Brazil	Christian	South	3.658	-0.580	0.231	1.921
BS	Bahamas	Christian	North	0.185	-0.069	0.298	0.069
BY	Belarus	Christian	North	0.403	0.217	0.106	-0.534
CA	Canada	Christian	North	2.397	0.327	1.159	0.868
CH	Switzerland	Christian	North	4.012	-0.374	0.553	0.984
CL	Chile	Christian	South	1.966	-2.006	-0.634	1.232
CM	Cameroon	Christian	North	1.410	0.926	1.021	0.650
CN	China	Other	North	-0.650	-0.349	0.300	-1.083
CO	Colombia	Christian	North	2.641	-1.164	0.596	1.995
CR	Costa Rica	Christian	North	3.671	-0.728	-0.038	2.110
CY	Cyprus	Christian	North	2.274	-0.376	0.057	0.390
CZ	Czech Republic	Other	North	2.718	-0.166	0.952	1.020
DE	Germany	Christian	North	3.800	0.043	0.759	0.974
DJ	Djibouti	Muslim	North	-0.506	1.507	0.692	-0.071
DK	Denmark	Christian	North	2.842	-0.558	0.602	0.844
DO	Dominican Republic	Christian	North	2.379	-0.861	0.649	1.240
DZ	Algeria	Muslim	North	0.503	0.872	1.611	0.153
EC	Ecuador	Christian	South	3.203	-0.521	0.513	2.062
EE	Estonia	Other	North	1.302	-0.344	1.598	0.541
EG	Egypt	Muslim	North	1.056	**2.278**	-0.302	0.841
ES	Spain	Christian	North	1.587	-0.063	0.391	0.056
ET	Ethiopia	Christian	North	-0.967	-0.585	-0.164	0.013
FI	Finland	Christian	North	2.260	-0.858	1.690	0.854
	Country	**Country Set**	**Hemisphere**	**Christmas**	**Eid-al-Fitr**	**June Solstice**	**Dec Solstice**

FJ	Fiji	Christian	South	3.087	-0.002	-0.437	1.683
FR	France	Christian	North	2.239	-0.050	0.600	1.242
GE	Georgia	Christian	North	-0.033	0.158	0.674	-0.964
GH	Ghana	Christian	North	3.869	-0.417	0.389	1.850
GP	Guadalupe	Christian	North	1.550	-0.059	1.814	1.751
GR	Greece	Christian	North	1.241	-0.158	-0.156	0.056
GT	Guatemala	Christian	North	3.170	-1.062	0.561	2.496
GU	Guam	Christian	South	0.028	1.379	1.080	-0.229
HN	Honduras	Christian	North	2.903	-0.713	0.279	2.062
HR	Croatia	Christian	North	0.953	0.236	1.712	-0.234
HU	Hungary	Christian	North	1.244	0.588	0.928	0.114
ID	Indonesia	Muslim	South	2.792	**3.584**	-0.415	1.337
IE	Ireland	Christian	North	3.498	0.072	0.477	1.052
IL	Israel	Other	North	-1.235	0.085	1.261	-1.446
IN	India	Other	North	1.850	1.315	-0.363	0.756
IQ	Iraq	Muslim	North	-0.833	0.514	-0.704	-0.066
IR	Iran	Muslim	North	-0.597	0.497	0.714	-1.260
IS	Iceland	Christian	North	1.913	-0.698	0.824	1.064
IT	Italy	Christian	North	1.811	0.107	0.056	0.266
JM	Jamaica	Christian	North	1.255	-0.357	1.799	1.190
JO	Jordan	Muslim	North	-0.169	2.317	1.463	-0.334
JP	Japan	Other	North	1.067	0.257	0.468	-0.734
KE	Kenya	Christian	North	**4.217**	1.686	-0.604	3.297
KH	Cambodia	Other	North	1.064	0.988	-0.475	-0.242
KR	South Korea	Other	North	0.994	-1.400	1.172	-0.305
KW	Kuwait	Muslim	North	1.730	**2.384**	0.145	1.855
KZ	Kazakhstan	Muslim	North	0.151	-0.458	1.537	-0.248
LA	Laos	Other	North	1.559	0.670	0.273	0.290
LB	Lebanon	Muslim	North	1.389	**2.497**	0.843	0.205
LK	Sri Lanka	Other	North	2.505	0.443	-0.390	0.970
LT	Lithuania	Christian	North	0.942	0.594	1.249	-0.277
LU	Luxemburg	Christian	North	4.643	-0.968	0.611	1.418
LV	Latvia	Christian	North	1.087	-0.082	2.154	-0.139
MA	Morocco	Muslim	North	0.148	0.484	1.173	-0.669
MD	Moldova	Christian	North	0.648	-0.115	0.626	-0.154
ME	Montenegro	Christian	North	0.004	-0.514	0.145	0.773
MK	Macedonia	Christian	North	-0.789	-0.233	0.786	-0.920
MM	Myanmar	Other	North	1.753	1.324	-1.771	1.998
MN	Mongolia	Other	North	0.087	-0.694	0.785	-0.143
MT	Malta	Christian	North	1.547	-0.059	1.718	1.145
MU	Mauritius	Other	South	2.627	-0.528	-0.212	1.745
MV	Maldives	Muslim	North	-0.475	0.704	-0.215	0.133
	Country	**Country Set**	**Hemisphere**	**Christmas**	**Eid-al-Fitr**	**June Solstice**	**Dec Solstice**
MX	Mexico	Christian	North	3.092	-1.378	0.739	1.967
MY	Malaysia	Muslim	North	1.838	**3.709**	0.174	0.602

MZ	Mozambique	Christian	South	2.243	-0.531	-0.048	1.702
NA	Namibia	Christian	South	3.757	-1.345	0.064	2.812
NG	Nigeria	Christian	North	**4.650**	1.208	-0.227	3.060
NI	Nicaragua	Christian	North	1.199	-0.917	-0.321	2.106
NL	Netherlands	Other	North	1.692	0.031	0.891	0.197
NO	Norway	Christian	North	3.694	-1.155	0.932	2.015
NP	Nepal	Other	North	1.095	**1.588**	-0.454	0.281
NZ	New Zealand	Christian	South	3.230	-0.254	-0.495	1.660
OM	Oman	Muslim	North	0.873	1.943	0.611	1.054
PA	Panama	Christian	North	1.955	0.914	0.009	1.456
PE	Peru	Christian	South	2.317	-2.338	-0.130	1.514
PH	Philippines	Christian	North	2.444	0.981	-1.614	1.819
PK	Pakistan	Muslim	North	**2.282**	2.126	-0.124	1.787
PL	Poland	Christian	North	1.414	0.083	1.341	0.215
PR	Puerto Rico	Christian	North	2.606	-1.690	1.211	2.274
PS	Palestine	Muslim	North	1.152	**1.609**	0.458	0.215
PT	Portugal	Christian	North	2.226	-0.074	0.699	0.859
PY	Paraguay	Christian	South	1.952	-2.259	-1.242	1.278
QA	Qatar	Muslim	North	1.783	**2.986**	-1.061	0.835
RO	Romania	Christian	North	1.458	0.401	0.960	-0.073
RS	Serbia	Christian	North	-0.163	0.474	1.130	-0.390
RU	Russia	Christian	North	0.042	-0.455	1.443	-0.371
SA	Saudi Arabia	Muslim	North	0.271	2.698	-0.037	0.330
SD	Sudan	Muslim	North	0.460	1.662	0.602	0.682
SE	Sweden	Christian	North	1.764	-0.609	1.547	0.383
SG	Singapore	Other	North	**2.238**	1.525	1.339	1.140
SI	Slovenia	Christian	North	0.742	-0.170	1.275	0.018
SK	Slovakia	Christian	North	2.172	0.125	0.913	0.123
SV	El Salvador	Christian	North	3.076	0.144	-0.263	1.603
SY	Syria	Muslim	North	0.136	2.361	0.845	0.101
TH	Thailand	Other	North	0.658	-0.094	-0.761	-0.361
TN	Tunisia	Muslim	North	0.083	2.042	0.523	1.618
TR	Turkey	Muslim	North	-1.084	2.988	1.447	-1.123
TT	Trinidad Tobago	Christian	North	3.526	1.158	-0.016	1.704
TW	Taiwan	Other	North	1.458	-0.249	0.185	0.382
TZ	Tanzania	Christian	South	2.475	-0.365	1.200	1.710
UA	Ukraine	Christian	North	0.497	0.158	0.270	-0.051
UG	Uganda	Christian	North	3.703	0.921	-1.054	2.327
UK	United Kingdom	Christian	North	3.982	0.208	-0.086	1.559
	Country	**Country Set**	**Hemisphere**	**Christmas**	**Eid-al-Fitr**	**June Solstice**	**Dec Solstice**
US	United States of America	Christian	North	3.100	-0.306	1.009	1.137
UY	Uruguay	Christian	South	2.140	-0.462	-1.259	0.879
UZ	Uzbekistan	Muslim	North	-0.590	2.098	1.472	-0.960
VE	Venezuela	Christian	North	3.768	-0.982	-0.292	2.287

VN	Vietnam	Other	North	-0.033	1.300	0.436	-0.380
YE	Yemen	Muslim	North	-0.367	1.963	0.325	-0.181
ZA	South Africa	Christian	South	3.815	0.048	-0.108	2.375
ZM	Zambia	Christian	South	1.804	0.915	-0.098	2.308
ZW	Zimbabwe	Christian	South	3.783	-0.146	1.001	2.569

Table S8A. Correlation between the Z-scores' time series for all countries in the data set. Calendars were centered around each of the events and the z-cores calculated, as detailed in the Methods. The high correlation between the Z-score variation around Christmas and around the December Solstice is due to the fact that Christmas often falls on the same week or very close to the December Solstice.

	Christmas	*Eid-al-Fitr*	*June Solstice*	*December Solstice*
Christmas	1.00			
Eid-al-Fitr	-0.28	1.00		
June Solstice	-0.29	-0.06	1.00	
December Solstice	0.80	-0.15	-0.36	1.00

Table S8B. Percentage of countries that were originally classified as Christian, Muslim, or as being located in one of the hemispheres (rows) that showed increased sex-searches (z-scores>1) during Christmas, Eid-al-Fitr or the Solstices (columns).

		Increased sex-searches around:			
		Christmas	Eid-al-Fitr	June Sltc	Dec Sltc
Identified as	Christian	80%	6%	25%	56%
	Muslim	40%	77%	23%	30%
	Southern Hemisphere	95%	14%	14%	90%
	Northern Hemisphere	64%	28%	26%	36%

Table S9. Monthly birth data available for countries from Supplementary Table 2. First column, countries that belong to the "Other" country set are marked with a blue background, countries that belong to the "Muslim" country set with a green background, and countries belonging to the "Christian" country set with a white background. At the bottom of the table are the only four countries from the Southern Hemisphere for which we could find birth data, and all four were classified as Christian. Dark shaded area coincides with the period for which we have GT data and these were the years used in all birth plots.

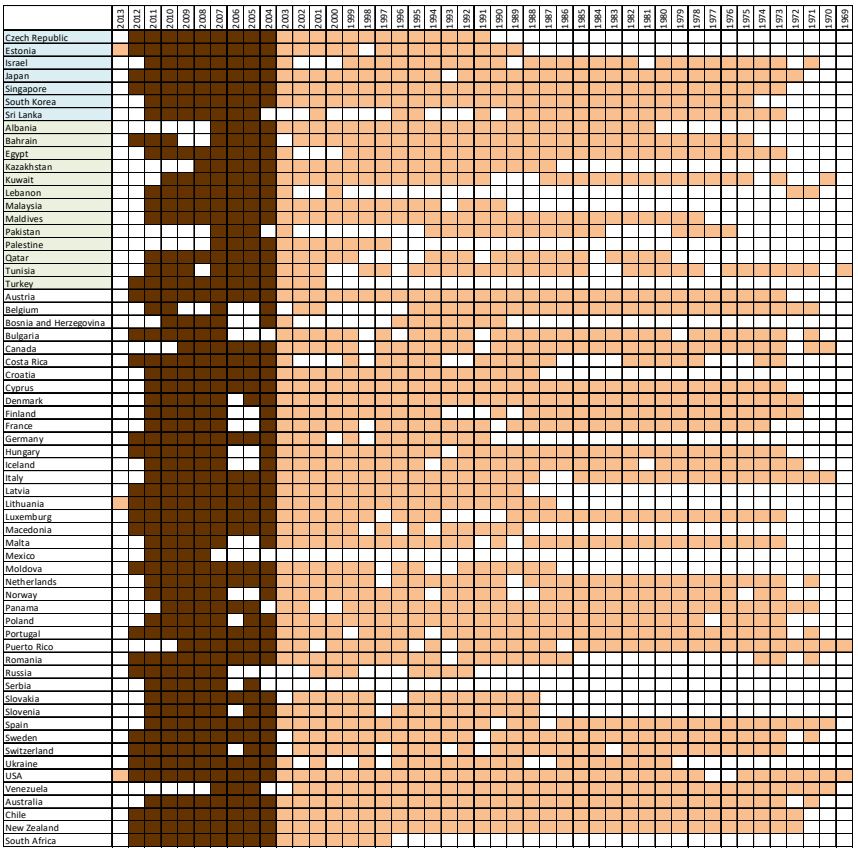


Table S10. Multiple linear regression statistics with all three ANEW dimensions, using weekly ANEW means as independent variables and sex search volume as dependent variable. A) Regression over all years of data. B) Regression over an average year centered on Christmas (USA, Australia, Brazil, Argentina, Chile) and Eid-al-Fitr (Indonesia and Turkey) – Independent variables are: [mean ANEW values averaged across years – the holiday center] (i.e, Christmas is 0,0,0), dependent variable is the number of sex-searches averaged across years of data. R^2 columns indicate the coefficient of determination for the regression, F_p columns indicate the p-value for the F-statistic of the overall model, B columns indicate the coefficients for the independent variables in the regression. t-test p columns indicate the individual t-test p values for the independent variables. Bold values denote significance at α=0.05, italicized values denote Bonferroni corrected significance over countries per variable choice α=0.05/7= 0.00714.

A

Country	R^2	Valence B	Dominance B	Arousal B	F_p	Valence t-test p	Dominance t-test p	Arousal t-test p
USA	0.399	197.69	-379.75	-0.36	***1.18E-20***	***4.76E-18***	***1.06E-12***	0.972
Australia	0.274	55.77	-92.18	-25.05	***2.91E-12***	***1.10E-07***	***4.22E-06***	***9.79E-06***
Brazil	0.401	12.47	37.78	90.74	***1.19E-15***	0.416	0.423	***4.79E-08***
Argentina	0.388	39.22	-36.67	-8.79	***1.40E-14***	***2.59E-09***	***1.91E-03***	0.0786
Chile	0.240	4.93	26.63	-28.93	***1.68E-10***	0.602	0.280	***6.36E-10***
Indonesia	0.187	72.13	-127.96	-12.95	***1.87E-06***	***1.24E-07***	***5.28E-04***	0.366
Turkey	0.135	6.66	-1.72	16.11	***4.22E-04***	0.128	0.893	***1.83E-04***

B

Country	R^2	Valence B	Dominance B	Arousal B	F_p	Valence t-test p	Dominance t-test p	Arousal t-test p
USA	0.426	193.002	-427.958	96.678	***6.20E-06***	***2.94E-07***	***2.77E-05***	0.0632
Australia	0.566	95.519	-128.290	19.318	***8.40E-09***	***4.67E-08***	***1.44E-03***	0.225
Brazil	0.488	90.086	-148.254	35.561	***4.15E-07***	***3.06E-05***	0.0340	0.116
Argentina	0.530	57.468	-65.493	-2.228	***5.51E-08***	***3.61E-07***	0.0145	0.871
Chile	0.697	70.497	-81.955	12.632	***1.73E-12***	***8.21E-08***	0.0123	0.0606
Indonesia	0.267	144.696	-272.516	-53.604	***2.34E-03***	**8.40E-03**	0.0213	0.271
Turkey	0.260	7.835	-81.301	41.880	***2.94E-03***	0.503	0.0103	**0.0220**

Table S11. Linear regression statistics for individual ANEW dimensions, using weekly ANEW means as independent variables and sex search volume as dependent variable. A: Regression over all years of data. B: Regression over an average year centered on Christmas (USA, Australia, Brazil, Argentina, Chile) and Eid-al-Fitr (Indonesia and Turkey) – Independent variables are: [mean ANEW value averaged across years – the holiday center] (i.e, Christmas is 0), dependent variable is the number of sex-searches averaged across years of data. Independent variables from top to bottom: Valence, Dominance, and Arousal. R^2 columns indicate the coefficient of determination for the regression, F_p columns indicate the p-value for the F-statistic of the overall model, B columns indicate the coefficients for the independent variables in the regression. Bold values denote significance at α=0.05, italicized values denote Bonferroni corrected significance over countries per variable choice α=0.05/7 = 0.00714.

A

Country	**Valence R^2**	**Valence F_p**	**Valence B**
USA	0.057	***8.99E-04***	54.80
Australia	0.065	***5.34E-04***	-10.74
Brazil	0.004	0.434	6.57
Argentina	0.255	***1.78E-10***	13.25
Chile	0.019	0.0680	8.52
Indonesia	0.091	***2.29E-04***	27.35
Turkey	0.008	0.3.07	3.16

Country	**Dominance R^2**	**Dominance F_p**	**Dominance B**
USA	0.052	***1.46E-03***	-101.86
Australia	0.120	***1.84E-06***	-30.11
Brazil	0.141	***3.21E-06***	115.42
Argentina	0.143	***3.67E-06***	15.86
Chile	0.001	0.711	4.19
Indonesia	0.007	0.302	21.45
Turkey	0.031	**4.84E-02**	18.09

Country	**Arousal R^2**	**Arousal F_p**	**Arousal B**
USA	0.102	***6.20E-06***	-41.47
Australia	0.148	***9.78E-08***	-15.07
Brazil	0.347	***6.28E-15***	90.83
Argentina	0.026	0.0573	6.66
Chile	0.186	***1.48E-09***	-23.98
Indonesia	0.007	0.323	11.16
Turkey	0.105	***1.95E-04***	13.97

B

Country	**Valence R^2**	**Valence F_p**	**Valence B**
USA	0.166	***2.76E-03***	80.924
Australia	0.459	***3.39E-08***	56.364
Brazil	0.437	***9.41E-08***	51.052
Argentina	0.418	***2.25E-07***	31.673
Chile	0.652	***4.82E-13***	43.522
Indonesia	0.008	0.541	19.959
Turkey	0.043	0.150	8.871

Country	**Dominance R^2**	**Dominance F_p**	**Dominance B**
USA	0.002	0.778	-20.481
Australia	0.167	***2.66E-03***	74.668
Brazil	0.214	***5.48E-04***	121.891
Argentina	0.138	**6.68E-03**	39.978
Chile	0.426	***1.55E-07***	97.791
Indonesia	0.049	0.123	-94.162
Turkey	0.005	0.612	-11.387

Country	**Arousal R^2**	**Arousal F_p**	**Arousal B**
USA	0.000	0.945	-3.900
Australia	0.165	***2.85E-03***	42.919
Brazil	0.010	0.490	-14.894
Argentina	0.006	0.598	7.270
Chile	0.000	0.948	-0.640
Indonesia	0.146	***6.16E-03***	-112.497
Turkey	0.125	0.0119	28.478

Table S12 – Ordinary least squares linear regression statistics for sex-searches v.s proximity in eigenmood to Christmas. The components selected were the two components (eigenbins) that most distinguish the holiday week from other weeks (see Methods S11). In the Components column, v stands for valence, d for dominance, and a for arousal. R^2 is the coefficient of determination, F_P is the p-value of the overall F-test for the regression, and the Slope is the slope of regressions. ρ is the Pearson's correlation coefficient between proximity and sex searches, ρ_D is the Brownian distance correlation coefficient, and $DCov_p$ is the p-value for the Brownian distance covariance calculated from a permutation test of the data. Bold denotes significance at α=0.05, italicized values denote Bonferroni corrected significance over countries per variable choice α=0.05/7 = 0.00714, underlined denote Bonferroni corrected significance over all table possibilities α=0.05/21 = 0.00238.

Christmas

Country	Components	R^2	Fp	Slope	ρ	ρD	$DCov_p$
USA	v4, v5	0.38	***5.08E-06***	6.50E+04	0.616	0.559	0.001
Australia	d5, d8	0.392	***2.52E-06***	2.44E+04	0.626	0.576	0.001
Brazil	a3, v2	0.504	***3.35E-08***	9.47E+03	0.71	0.624	0.001
Argentina	v5, d3	0.577	***6.11E-10***	5.35E+03	0.759	0.712	0.001
Chile	v3, d8	0.419	***1.16E-06***	7.96E+03	0.647	0.646	0.001
Indonesia	a3, v3	0.448	***2.66E-07***	9.95E+03	0.67	0.657	0.001
Turkey	a3, d3	0.373	***6.46E-06***	-1.42E+03	-0.611	0.618	0.001

Eid-al-Fitr without Ramadan

Country	Components	R^2	Fp	Slope	ρ	ρD	$DCov_p$
USA	a6, v3	0.065	0.107	1.57E+05	0.256	0.328	0.118
Australia	v3, v4	0.02	0.381	-1.62E+03	-0.141	0.317	0.154
Brazil	a3, d8	0.147	**0.0147**	-4.07E+04	-0.383	0.539	0.001
Argentina	v9, d3	0.598	***3.08E-09***	-2.32E+04	-0.773	0.735	0.001
Chile	a6, d2	0.189	***5.00E-03***	-1.15E+04	-0.435	0.461	0.005
Indonesia	v3, d3	0.637	***6.87E-10***	8.70E+03	0.798	0.712	0.001
Turkey	a3, d3	0.737	***6.94E-13***	4.81E+02	0.859	0.858	0.001

Eid-al-Fitr

Country	Components	R^2	Fp	Slope	ρ	ρD	$DCov_p$
USA	a6, v3	0.077	0.0645	1.75E+05	0.278	0.343	0.061
Australia	v3, v4	0.038	0.198	-2.30E+03	-0.196	0.333	0.085
Brazil	a3, d8	0.124	**0.0204**	-3.54E+04	-0.353	0.516	0.001
Argentina	v9, d3	0.593	***6.23E-10***	-2.31E+04	-0.77	0.73	0.001
Chile	a6, d2	0.191	***3.03E-03***	-1.04E+04	-0.437	0.489	0.001
Indonesia	v3, d3	0.407	***3.19E-06***	9.85E+03	0.638	0.621	0.001
Turkey	a3, d3	0.339	***3.42E-05***	3.32E+02	0.582	0.634	0.001

Table S13– List of words and expressions removed from the Twitter/ANEW analysis.

"merry christmas"
"merry xmas"
"happy christmas"
"happy xmas"
"happy new year"
"happy newyear"
"happy thanksgiving"
"happy ramadan"
"happy easter"
"happy holidays"
"happy hanukkah"
"happy hanukah"
"happy ramadan"
"happy eid"
"happy halloween"
"happy valentines day"
"happy valentine's day"
"feliz natal"
"feliz ano"
"feliz pascoa"
"pascoa feliz"
"feliz thanksgiving"
"feliz navidad"
"feliz ano nuevo"
"feliz ano novo"
"feliz ramadan"
"feliz año"
"feliz páscoa"
"páscoa feliz "
"feliz año nuevo"
"happy anzac day"
"feliz anzac day"
"happy adelaide cup"
"feliz adelaide cup"
"happy all saints day"
"feliz all saints day"
"happy all souls day"
"feliz all souls day"
"happy andalucia day"
"feliz andalucia day"
"happy arafat day"
"feliz arafat day"
"happy armistice day"
"feliz armistice day"
"happy army day"
"feliz army day"
"happy asahna bucha day"
"feliz asahna bucha day"
"happy ascension day"
"feliz ascension day"
"happy ash monday"
"feliz ash monday"
"happy ash wednesday"
"feliz ash wednesday"
"happy ashura"
"feliz ashura"
"happy assumption day"
"feliz assumption day"
"happy asturias"
"feliz asturias"
"happy auckland province"
"feliz auckland province"
"happy august bank holiday"
"feliz august bank holiday"
"happy august holiday"
"feliz august holiday"
"happy australia day"
"feliz australia day"
"happy australia day holiday"
"feliz australia day holiday"
"happy autumnal equinox day"
"feliz autumnal equinox day"
"happy awal muharram"
"feliz awal muharram"
"happy balearic islands"
"feliz balearic islands"
"happy bank holiday"
"feliz bank holiday"
"happy bastille day"
"feliz bastille day"
"happy battle of the boyne"
"feliz battle of the boyne"
"happy benito juarezs birthday"
"feliz benito juarezs birthday"
"happy berchtolds day"
"feliz berchtolds day"
"happy bettagsmontag"
"feliz bettagsmontag"
"happy bhogi"
"feliz bhogi"
"happy bicentennial of the constituent assembly of 1813"
"feliz bicentennial of the constituent assembly of 1813"
"happy birthday of muhammad iqbal"
"feliz birthday of muhammad iqbal"
"happy birthday of prophet muhammad"
"feliz birthday of prophet muhammad"
"happy birthday of quaid-e-azam muhammad ali jinnah"
"feliz birthday of quaid-e-azam muhammad ali jinnah"
"happy birthday of spb yang di pertuan agong"
"feliz birthday of spb yang di pertuan agong"
"happy birthday of the sultan of selangor"
"feliz birthday of the sultan of selangor"
"happy boxing day"
"feliz boxing day"

"happy bridge public"
"feliz bridge public"
"happy buddha purnima"
"feliz buddha purnima"
"happy buddhas birthday"
"feliz buddhas birthday"
"happy canada day"
"feliz canada day"
"happy canary islands"
"feliz canary islands"
"happy canberra day"
"feliz canberra day"
"happy canterbury"
"feliz canterbury"
"happy carnival"
"feliz carnival"
"happy castile-la mancha"
"feliz castile-la mancha"
"happy catalonia"
"feliz catalonia"
"happy celebration of the golden spurs"
"feliz celebration of the golden spurs"
"happy ceuta"
"feliz ceuta"
"happy chanukah"
"feliz chanukah"
"happy chatham islands"
"feliz chatham islands"
"happy childrens day"
"feliz childrens day"
"happy chinese new year"
"feliz chinese new year"
"happy chinese new year eve"
"feliz chinese new year eve"
"happy ching ming"
"feliz ching ming"
"happy christmas day"
"feliz christmas day"
"happy christmas eve"
"feliz christmas eve"
"happy christmas eve day"
"feliz christmas eve day"
"happy christmas"
"feliz christmas"
"happy chulalongkorn day"
"feliz chulalongkorn day"
"happy chung yeung festival"
"feliz chung yeung festival"
"happy cinco de mayo"
"feliz cinco de mayo"
"happy civic day"
"feliz civic day"
"happy columbus day"
"feliz columbus day"
"happy coming of age day"
"feliz coming of age day"
"happy community day"
"feliz community day"
"happy community festival of madrid"
"feliz community festival of madrid"
"happy constitution day"
"feliz constitution day"
"happy constitution memorial day"
"feliz constitution memorial day"
"happy corpus christi"
"feliz corpus christi"
"happy culture day"
"feliz culture day"
"happy day after christmas"
"feliz day after christmas"
"happy day after new years day"
"feliz day after new years day"
"happy day of atonement"
"feliz day of atonement"
"happy day of good will"
"feliz day of good will"
"happy day of national sovereignty"
"feliz day of national sovereignty"
"happy day of reconciliation"
"feliz day of reconciliation"
"happy day of reformation"
"feliz day of reformation"
"happy day of unity"
"feliz day of unity"
"happy day of respect for cultural diversity"
"feliz day of respect for cultural diversity"
"happy day of the battle of salta"
"feliz day of the battle of salta"
"happy day of the constitution of the slovak republic"
"feliz day of the constitution of the slovak republic"
"happy day of the dead"
"feliz day of the dead"
"happy day of the establishment of the slovak republic"
"feliz day of the establishment of the slovak republic"
"happy day of the german-speaking community of belgium"
"feliz day of the german-speaking community of belgium"
"happy day of the virgin of guadalupe"
"feliz day of the virgin of guadalupe"
"happy day of victory over fascism"
"feliz day of victory over fascism"
"happy declaration of independence"
"feliz declaration of independence"
"happy deepavali"
"feliz deepavali"

"happy deewali"
"feliz deewali"
"happy defence of the motherland"
"feliz defence of the motherland"
"happy discovery day"
"feliz discovery day"
"happy double ninth day"
"feliz double ninth day"
"happy dragon boat festival"
"feliz dragon boat festival"
"happy dussehra"
"feliz dussehra"
"happy early may bank holiday"
"feliz early may bank holiday"
"happy easter"
"feliz easter"
"happy easter monday"
"feliz easter monday"
"happy easter sunday"
"feliz easter sunday"
"happy eid al adha"
"feliz eid al adha"
"happy eid al fitr"
"feliz eid al fitr"
"happy eid milad un-nabi"
"feliz eid milad un-nabi"
"happy eid ul-azha day 1"
"feliz eid ul-azha day 1"
"happy eid ul-azha day 2"
"feliz eid ul-azha day 2"
"happy eid-ul-fitr"
"feliz eid-ul-fitr"
"happy emancipation day"
"feliz emancipation day"
"happy epiphany"
"feliz epiphany"
"happy extremadura"
"feliz extremadura"
"happy family & community day"
"feliz family & community day"
"happy family day"
"feliz family day"
"happy fathers day"
"feliz fathers day"
"happy feast of st ambrose"
"feliz feast of st ambrose"
"happy feast of st anthony"
"feliz feast of st anthony"
"happy feast of st john the baptist"
"feliz feast of st john the baptist"
"happy federal territory day"
"feliz federal territory day"
"happy fiesta de san isidro"
"feliz fiesta de san isidro"
"happy foundation day"
"feliz foundation day"
"happy foundation of the independent czechoslovak state"
"feliz foundation of the independent czechoslovak state"
"happy freedom day"
"feliz freedom day"
"happy french community"
"feliz french community"
"happy ganesh chaturthi"
"feliz ganesh chaturthi"
"happy general prayer day"
"feliz general prayer day"
"happy german unity day"
"feliz german unity day"
"happy good friday"
"feliz good friday"
"happy greenery day"
"feliz greenery day"
"happy groundhog day"
"feliz groundhog day"
"happy guru nanak birthday"
"feliz guru nanak birthday"
"happy guy fawkes night"
"feliz guy fawkes night"
"happy h.m. kings birthday"
"feliz h.m. kings birthday"
"happy h.m. queens birthday"
"feliz h.m. queens birthday"
"happy hangeul day"
"feliz hangeul day"
"happy hari hol almarhum sultan iskandar"
"feliz hari hol almarhum sultan iskandar"
"happy hari raya haji"
"feliz hari raya haji"
"happy hari raya nyepi"
"feliz hari raya nyepi"
"happy hari raya puasa"
"feliz hari raya puasa"
"happy harvest festival"
"feliz harvest festival"
"happy hawkes bay"
"feliz hawkes bay"
"happy health-sports day"
"feliz health-sports day"
"happy heritage day"
"feliz heritage day"
"happy hijri new years day"
"feliz hijri new years day"
"happy hispanic day"
"feliz hispanic day"
"happy holi"
"feliz holi"
"happy holy spirit monday"
"feliz holy spirit monday"

"happy human rights day"
"feliz human rights day"
"happy idul adha"
"feliz idul adha"
"happy idul fitr"
"feliz idul fitr"
"happy idul juha"
"feliz idul juha"
"happy immaculate conception day"
"feliz immaculate conception day"
"happy independence day"
"feliz independence day"
"happy independence day of chile"
"feliz independence day of chile"
"happy independence day"
"feliz independence day"
"happy independence of cartagena"
"feliz independence of cartagena"
"happy isra miraj"
"feliz isra miraj"
"happy israa & miaraj night"
"feliz israa & miaraj night"
"happy jan hus day"
"feliz jan hus day"
"happy janmashtami"
"feliz janmashtami"
"happy june holiday"
"feliz june holiday"
"happy kannada rajyothsava"
"feliz kannada rajyothsava"
"happy kashmir day"
"feliz kashmir day"
"happy kings feast"
"feliz kings feast"
"happy knabenschiessen"
"feliz knabenschiessen"
"happy korean new year"
"feliz korean new year"
"happy la rioja"
"feliz la rioja"
"happy labor day"
"feliz labor day"
"happy labour day"
"feliz labour day"
"happy labour thanksgiving day"
"feliz labour thanksgiving day"
"happy labour day"
"feliz labour day"
"happy lady of aparecida"
"feliz lady of aparecida"
"happy lantern festival"
"feliz lantern festival"
"happy late mid autumn festival"
"feliz late mid autumn festival"
"happy liberation day"
"feliz liberation day"
"happy liberation day czech republic"
"feliz liberation day czech republic"
"happy maha shivratri"
"feliz maha shivratri"
"happy maharashtra day"
"feliz maharashtra day"
"happy mahatma gandhi birthday"
"feliz mahatma gandhi birthday"
"happy mahavir jayanti"
"feliz mahavir jayanti"
"happy makha bucha day"
"feliz makha bucha day"
"happy malaysia day"
"feliz malaysia day"
"happy malvinas day"
"feliz malvinas day"
"happy march 1st movement"
"feliz march 1st movement"
"happy marine day"
"feliz marine day"
"happy marlborough"
"feliz marlborough"
"happy martin luther king day"
"feliz martin luther king day"
"happy maulidur rasul"
"feliz maulidur rasul"
"happy maundy thursday"
"feliz maundy thursday"
"happy may bank holiday"
"feliz may bank holiday"
"happy may day"
"feliz may day"
"happy may day revolution"
"feliz may day revolution"
"happy melbourne cup day"
"feliz melbourne cup day"
"happy memorial day"
"feliz memorial day"
"happy mid autumn festival"
"feliz mid autumn festival"
"happy midsummer day"
"feliz midsummer day"
"happy milad-un-nabi"
"feliz milad-un-nabi"
"happy mothering sunday"
"feliz mothering sunday"
"happy mothers day"
"feliz mothers day"
"happy muharram"
"feliz muharram"
"happy murcia"
"feliz murcia"
"happy national day"
"feliz national day"

"happy national flag day"
"feliz national flag day"
"happy national foundation day"
"feliz national foundation day"
"happy national remembrance day"
"feliz national remembrance day"
"happy national sovereignty and children's day"
"feliz national sovereignty and children's day"
"happy national womens day"
"feliz national womens day"
"happy national holiday"
"feliz national holiday"
"happy navy day"
"feliz navy day"
"happy nelson"
"feliz nelson"
"happy new year"
"feliz new year"
"happy new years day"
"feliz new years day"
"happy new years eve"
"feliz new years eve"
"happy new years"
"feliz new years"
"happy orthodox christmas day"
"feliz orthodox christmas day"
"happy orthodox easter monday"
"feliz orthodox easter monday"
"happy orthodox good friday"
"feliz orthodox good friday"
"happy otago province"
"feliz otago province"
"happy our lady of mount carmel"
"feliz our lady of mount carmel"
"happy our lady of the almudena"
"feliz our lady of the almudena"
"happy pakistan day"
"feliz pakistan day"
"happy pancake tuesday"
"feliz pancake tuesday"
"happy parsi new year"
"feliz parsi new year"
"happy passover"
"feliz passover"
"happy peace memorial day"
"feliz peace memorial day"
"happy pentecost"
"feliz pentecost"
"happy picnic day"
"feliz picnic day"
"happy pongal"
"feliz pongal"
"happy portugal day"
"feliz portugal day"
"happy presidential elections"
"feliz presidential elections"
"happy presidents day"
"feliz presidents day"
"happy public holiday"
"feliz public holiday"
"happy purim"
"feliz purim"
"happy queens birthday"
"feliz queens birthday"
"happy race day"
"feliz race day"
"happy ram navami"
"feliz ram navami"
"happy ramazan feast"
"feliz ramazan feast"
"happy reformation day"
"feliz reformation day"
"happy remembrance day"
"feliz remembrance day"
"happy repentance day"
"feliz repentance day"
"happy republic day"
"feliz republic day"
"happy respect for the aged day"
"feliz respect for the aged day"
"happy restoration day"
"feliz restoration day"
"happy restoration day of the independent czech state"
"feliz restoration day of the independent czech state"
"happy restoration of independence"
"feliz restoration of independence"
"happy revolution day"
"feliz revolution day"
"happy sacred heart"
"feliz sacred heart"
"happy sacrifice feast"
"feliz sacrifice feast"
"happy saint leopold"
"feliz saint leopold"
"happy saint nicholas"
"feliz saint nicholas"
"happy saint peter and saint paul"
"feliz saint peter and saint paul"
"happy saint stephens day"
"feliz saint stephens day"
"happy sechselauten"
"feliz sechselauten"
"happy second day of christmas"
"feliz second day of christmas"
"happy showa day"
"feliz showa day"
"happy simchat torah"
"feliz simchat torah"

"happy slovak national uprising anniversary"
"feliz slovak national uprising anniversary"
"happy songkran festival"
"feliz songkran festival"
"happy south canterbury"
"feliz south canterbury"
"happy southland"
"feliz southland"
"happy special administration region (sar) day"
"feliz special administration region (sar) day"
"happy st andrews day"
"feliz st andrews day"
"happy st cyril and methodius day"
"feliz st cyril and methodius day"
"happy st davids day"
"feliz st davids day"
"happy st georges day"
"feliz st georges day"
"happy st james day"
"feliz st james day"
"happy st josephs day"
"feliz st josephs day"
"happy st martins day"
"feliz st martins day"
"happy st patricks day"
"feliz st patricks day"
"happy st stephens day"
"feliz st stephens day"
"happy st wenceslas day"
"feliz st wenceslas day"
"happy struggle for freedom and democracy day"
"feliz struggle for freedom and democracy day"
"happy sukkot"
"feliz sukkot"
"happy swiss federal fast"
"feliz swiss federal fast"
"happy taranaki"
"feliz taranaki"
"happy thaipusam"
"feliz thaipusam"
"happy thanksgiving"
"feliz thanksgiving"
"happy buddhas birthday"
"feliz buddhas birthday"
"happy emperors birthday"
"feliz emperors birthday"
"happy national holiday of quebec"
"feliz national holiday of quebec"
"happy ochi day"
"feliz ochi day"
"happy patron saint of turin"
"feliz patron saint of turin"
"happy thiruvalluvar day"
"feliz thiruvalluvar day"
"happy tiradentes day"
"feliz tiradentes day"
"happy tomb sweeping festival"
"feliz tomb sweeping festival"
"happy tomb sweeping holiday"
"feliz tomb sweeping holiday"
"happy truth and justice memorial day"
"feliz truth and justice memorial day"
"happy uae national day"
"feliz uae national day"
"happy ugadi"
"feliz ugadi"
"happy urs mubarak of hazrat data gunj bakhsh"
"feliz urs mubarak of hazrat data gunj bakhsh"
"happy v-e day"
"feliz v-e day"
"happy valencia"
"feliz valencia"
"happy vernal equinox day"
"feliz vernal equinox day"
"happy vesak day"
"feliz vesak day"
"happy veterans day"
"feliz veterans day"
"happy victoria day"
"feliz victoria day"
"happy victory day"
"feliz victory day"
"happy visakha bucha day"
"feliz visakha bucha day"
"happy waisak day"
"feliz waisak day"
"happy waitangi day"
"feliz waitangi day"
"happy wellington province"
"feliz wellington province"
"happy wesak day"
"feliz wesak day"
"happy westland"
"feliz westland"
"happy whitmonday"
"feliz whitmonday"
"happy womens day"
"feliz womens day"
"happy youth day"
"feliz youth day"
"happy zumbi dos palmares"
"feliz zumbi dos palmares"
"christmas"
"navidad"
"natal"
"valentine"
"san valentín"
"valentín"
"san valentin"

“valentin”
“valentim

	1	2	3	4	5	6
variable g-cause sex-search	0.350	0.0103*	0.0512	0.172	0.000 383***	0.0261*
sex-search g-cause variable	0.0643	0.726	0.473	0.319	0.418	0.808

TABLE B.1: Arousal Eigenbin Granger-Causality p-values. * indicates $p < 0.05$, ** indicates table Bonferroni corrected $p < 0.00417$, *** indicates all tests Bonferroni corrected $p < 0.00179$

	1	2	3	4	5	6
variable g-cause sex-search	0.689	0.0143*	0.006 46*	0.798	0.448	0.0262
sex-search g-cause variable	0.646	0.316	0.376	0.133	0.006 57*	0.000 143***

TABLE B.2: Dominance Eigenbin Granger-Causality p-values. * indicates $p < 0.05$, ** indicates table Bonferroni corrected $p < 0.00417$, *** indicates all tests Bonferroni corrected $p < 0.00179$

	1	2	3	4	5	6
variable g-cause sex-search	0.477	0.0586	0.120	0.0279*	0.0541	0.750
sex-search g-cause variable	0.802	0.937	0.0367*	0.218	0.985	0.378

TABLE B.3: Valence Eigenbin Granger-Causality p-values. * indicates $p < 0.05$, ** indicates table Bonferroni corrected $p < 0.00417$, *** indicates all tests Bonferroni corrected $p < 0.00179$

B.1 Additional Granger Causality Analysis for United States

In Table B.1 we see that of the first 6 eigenbins for arousal, three eigenbins have a signficant Granger-causal relationship with sex searches, while sex-searches do not Granger-cause these arousal. The fifth component is significant even after a Bonferroni correction against all Granger tests in this chapter.

In Table B.2 we see that two of the first six dominance eigenbins Granger-cause sex searches, but not at a level significant after a Bonferroni correction. However, sex-searches Granger-cause the fifth and the sixth dominance eigenbin, the sixth at a level that is signficant even after a Bonferroni correction for all Granger tests in this chapter.

In Table B.3 We see that only the fourth valence eigenbin is seen to Granger-cause

	arousal	dominance	valence
arousal		0.469	0.684
dominance	0.358		0.278
valence	0.128	0.0518	

TABLE B.4: Mean Values Granger-Causality p-values. The values in the table are p-values for whether the variable specified by the Row Granger-causes the variable specified by the Column. * indicates $p < 0.05$, ** indicates table Bonferroni corrected $p < 0.00833$, *** indicates all tests Bonferroni corrected $p < 0.00179$

sex-searches, but not after Bonferroni correction. The third component is Granger-caused by sex searches. We should note here that the selected Christmas eigenmood for the US, selected by how much they distinguish Christmas, is the set of valence components 4 and 5. This includes the most significant Granger-causal variable in this table, and a component (the fifth) that is marginally Granger-causal at $p < 0.1$.

We can also examine whether mean sentiments Granger-cause each other. As you can see in Table B.4 there is no significant Granger-causality between the mean sentiment time series, but valence g-causing dominance is marginally significant ($p < 0.1$)

Overall we see that the mean values of sentiment detect some of the mood that drives human reproductive cycles and is in turn driven by it, but selecting particular eigenmoods reveals much deeper drivers – a similarity to holiday mood and a component of arousal that significantly drive sex searches, and a component of dominance that is significantly driven by sex searches.

Appendix C

Chapter 4 Appendix

C.1 Modeled significance over time

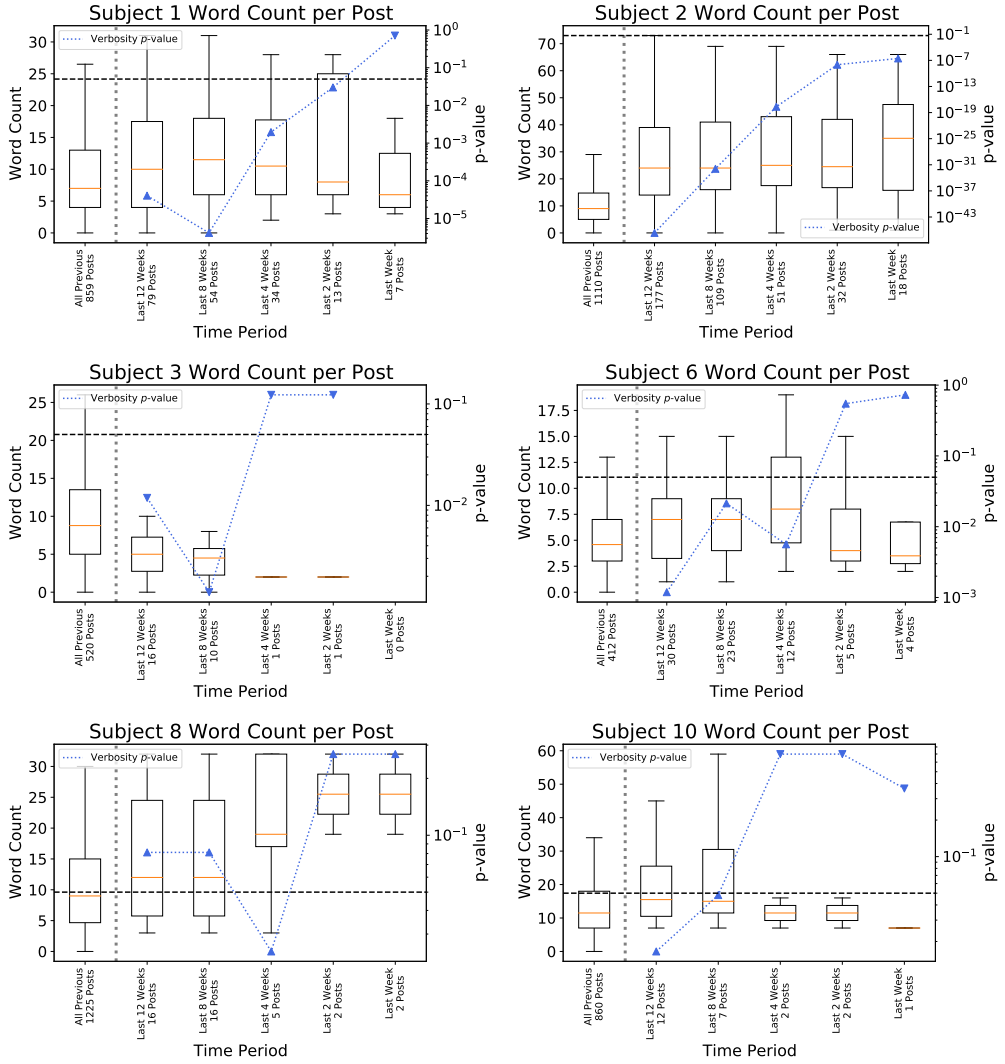


FIGURE C.1: **Subject verbosity per post over different epochs**. Difference between word count per post in the period immediately preceding SUDEP compared to word count per post during earlier posting periods. Different selections of the time window for the last posting period are displayed on the x-axis. The box plot on the far left represents all posts before the 12 weeks preceding SUDEP. The blue line represents the p-value of the time coefficient for the negative binomial regression. The direction of the arrow represents the sign of the coefficient, up indicates an increase in wordcount during the period preceding SUDEP and down indicates a decrease. The horizontal black line represents p=0.05

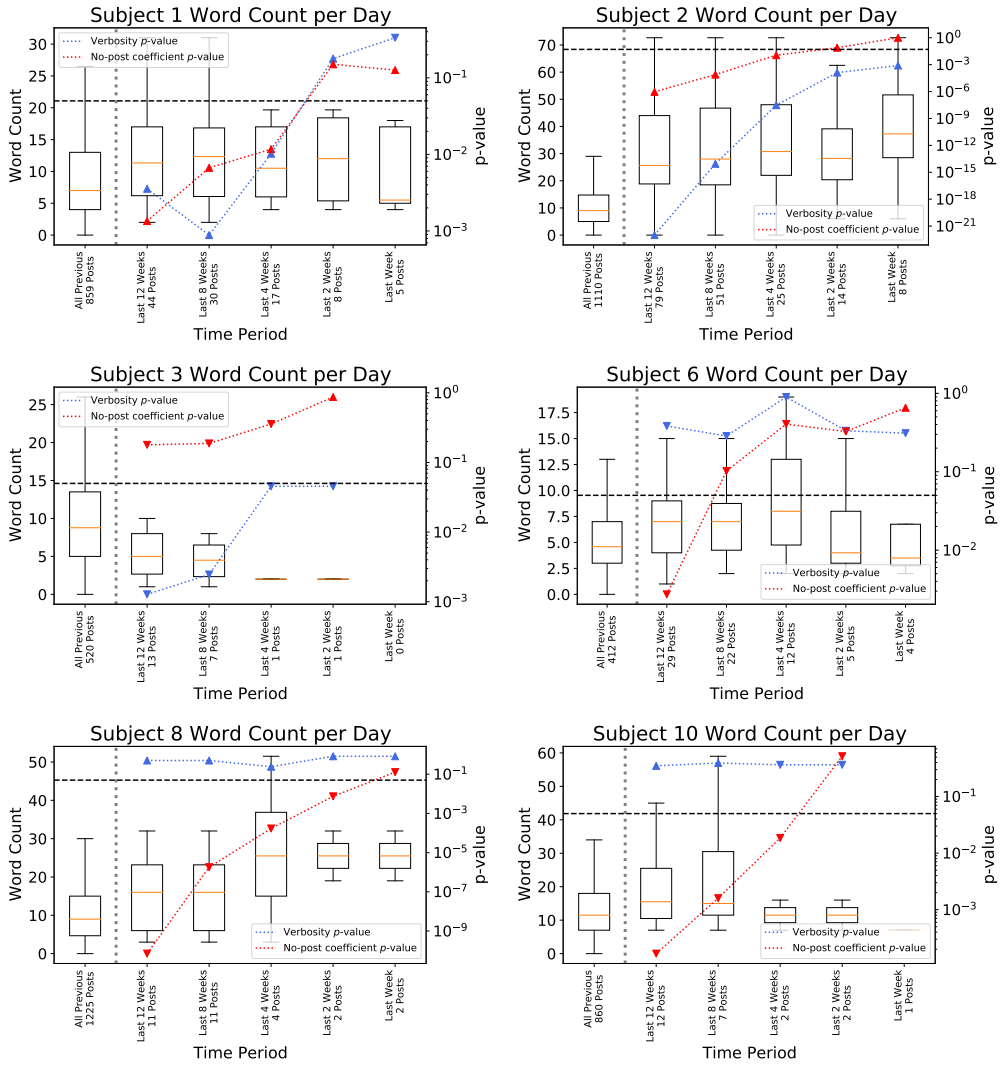


FIGURE C.2: **Subject verbosity per day over different epochs**. Difference between word count per day in the period immediately preceding SUDEP compared to word count per day during earlier posting periods. Different selections of the time window for the last posting period are displayed on the x-axis. The box plot on the far left represents all posts before the 12 weeks preceding SUDEP. The blue line represents the p-value of the word count time coefficient for the zero-inflated negative binomial regression. The direction of the blue triangle represents the sign of the coefficient, up indicates an increase in wordcount during the period preceding SUDEP and down indicates a decrease. The red line represents the p-value of the zero post time coefficient of the regression, with red triangles representing whether there is an increase in the likelihood of any post on a day (up) or a decrease (down). The horizontal black line represents $p = 0.05$.

C.2 Regression statistics

Subject	μ_1	n_1	μ_2	n_2	*intercept*	$time_{coef}$	$time_{se}$	$time_p$	θ	θ_{se}
2	12.431	2162	34.413	109	2.520	1.018	0.086	**1.197e-32**	1.373	0.043
1	9.592	1547	17.889	54	2.261	0.623	0.135	**4.146e-06**	1.113	0.043
8	12.070	2185	18.375	16	2.491	0.420	0.241	0.081	1.153	0.036
6	5.252	717	7.304	23	1.659	0.330	0.143	**0.021**	3.136	0.269
10	13.983	1147	23.571	7	2.638	0.522	0.264	**0.048**	2.254	0.105
3	11.125	834	4.100	10	2.409	-0.998	0.312	**0.001**	1.385	0.072

TABLE C.1: Statistics from a Negative Binomial Regression on Word Count per Post. μ_1 and n_1 correspond to the mean word count and number of posts before the last two months, while μ_2 and n_2 correspond to the mean word count and number of posts during the last two months before SUDEP. Also included are the *intercept* of the regression, the coefficient on the last month indicator variable $time_{coef}$, its standard error $time_{se}$, the p-value of the coefficient $time_p$, and the dispersion parameter θ with its standard error θ_{se}.

A negative binomial model is often used to model over-dispersed count data, i.e. when the variance is considerably larger than the mean [166]. Here a negative binomial model is estimated through a generalized linear regression with log link function on word count per post over a dummy variable representing whether the post's word count occurs during the last month. The significance of the time-indicator dummy variable estimates the significance of the change in the last month over all other posts. As shown in Table C.1 we see significant increases in the word count per post for four subjects at $p < 0.05$. The table is ordered according to the rank product of the number of posts before and during the last two months preceding SUDEP, and the two with the greatest number of posts in both periods by rank product are also the two with the greatest increase in word count, subjects 2 and 1, with two additional subjects showing significant increases, subject 6 and 10. There are five subjects with decreases in word count per post, with subject 11 and subject 3 with significant decreases.

An alternative formulation is to examine word count per day rather than per post. Perhaps some subjects additionally start posting short posts with increased frequency during periods of stress. However, many days contain zero posts, thus zero words, for

Subject	$intercept$	$time_{coef}$	$time_{se}$	$time_p$	$0_{intercept}$	$0_{time_{coef}}$	0_{time_p}
2	3.114	1.185	0.153	**8.802e-15**	-0.275	-1.854	**7.779e-05**
1	2.821	0.621	0.187	**8.828e-04**	0.586	-0.755	**0.007**
8	3.028	0.213	0.318	0.503	-0.261	1.637	**1.802e-06**
6	2.170	-0.213	0.199	0.285	-0.183	0.490	0.102
10	2.914	0.240	0.281	0.393	0.513	1.295	**0.002**
3	2.829	-1.234	0.408	**0.002**	0.991	0.571	0.187

TABLE C.2: Statistics of a Zero-Inflated Negative Binomial Regression on word count per day. This is similar to Table C.1, but models the word count per day rather than per post, with the addition of a logistic regression model representing the likelihood of no post at all. Included are the *intercept* of the regression, the coefficient on the last month indicator variable $time_{coef}$, its standard error $time_{se}$, the p-value of the coefficient $time_p$. Additionally, parameters of the logistic regression on no-post probabilities are shown: the intercept $0_{intercept}$, the coefficient on the time indicator $0_{time_{coef}}$ and the significance of this coefficient 0_{time_p}.

most subjects. We can model this with a zero-inflated negative binomial model that also estimates a probability that no words will be posted [166, 167]. As shown in Table C.2 we see that subject 2 and 1 still have significant increases in word count per day (columns $time_{coef}$ and $time_p$) and both are significantly more likely to post during the last 2 months (columns $0_{time_{coef}}$ and 0_{time_p}, note the negative coefficient corresponds to a lower probability of having no posts on a given day). Subjects 8 and 10 are significantly less likely to post during the last two months. Subject 11, however, is significantly more likely to post during the last two months, although with significantly fewer words per day. Subject 3 is seen to have a significant drop in word count per day. Additionally, subjects 7 and 5 are significantly more likely to post in the last two months, but with non-significant changes in word count. This view of the posting behavior also reveals interesting patterns but is not particularly more informative than the negative binomial model per post.

Ian Wood

https://www.linkedin.com/in/iwood2/
ibwood@indiana.edu

Education

Indiana University August, 2011 – Present
PhD Candidate, Informatics - Complex Systems Track *Bloomington, Indiana*

Clemson University August, 2007 – May, 2011
BS Computer Science *Clemson, SC*

Work Experience

Senior Software Engineer / Software Engineer September, 2018 – Present
LinkedIn *San Francisco, CA*

- Designed and developed new out-of-network trending content recommendation systems launched to millions of members
- Supervised engineers developing embedding models for cohort-based content recommendations
- Mentored new engineers and AI training for established engineers
- Developed training data pipelines for Event recommendations and added recommendations to the Feed leading to 1.5% increase in event attendees
- Conducted a counterfactual research project to identify the impact of Follow Edges as 2.22% of Daily Unique Contributors and 1.2% of sponsored revenue
- Trained models and fixed systems for Group recommendations leading to 20% increases in unique Group joins for established members, 120% increases for new members, and 3.5% increases to group contributions
- Trained new Feed recommendation model resulting in a 0.12% increase in engaged Feed users

Research Assistant August, 2016 – May, 2018; August, 2012 – May, 2015
Indiana University *Bloomington, IN*

- Conducted research into sentiment analysis of large social media corpora, text classification of research corpora, and robotic input modeling

Software Engineering Intern May, 2017 – August, 2017; May, 2016 – August, 2016
LinkedIn *Sunnyvale, CA*

- Trained Feed models on actor affinity features
- Developed particle-swarm hyperparameter exploration for nearline model training

Assistant Instructor August, 2015 – May, 2016; August 2011 – May, 2012
Indiana University *Bloomington, IN*

- Gave lectures, graded assignments, led lab sessions, and held office hours for classes I400 - Large-Scale Social, Phenomena and I501 - Introduction to Informatics, and I211- Information Infrastructure II

Research Assistant May, 2008 – May, 2011
Clemson University *Clemson, SC*

- Built web visualizations for trends in Engineering Education, educational environments for middle school students in 3rd Rock Grid, and immersive VR prototypes

Patents

Data flow based feature vector clustering. YY Ahn, AN Chekuvar, **IB Wood**, J Park, Y Jing, MD Conover. US Patent 10,592,535

Online hyperparameter tuning in distributed machine learning. **IB Wood**, X Miao, CM Tsai, JD Young. US Patent Pending US20220027359A1

Publications

Wood, Ian B., Rion Brattig Correia, Wendy R. Miller, and Luis M. Rocha. "Small cohort of patients with epilepsy showed increased activity on Facebook before sudden unexpected death." *Epilepsy & Behavior* 128 (2022): 108580. https://doi.org/10.1016/j.yebeh.2022.108580

Correia, Rion Brattig, **Ian B. Wood**, Johan Bollen, and Luis M. Rocha. "Mining social media data for biomedical signals and health-related behavior." *Annual review of biomedical data science* 3 (2020): 433-458. https://doi.org/10.1146/annurev-biodatasci-030320-040844

Park, Jaehyuk, **Ian B. Wood**, Elise Jing, Azadeh Nematzadeh, Souvik Ghosh, Michael D. Conover, and Yong-Yeol Ahn. "Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters." *Nature communications* 10, no. 1 (2019): 3449. https://doi.org/10.1038/s41467-019-11380-w

Gates, Alexander J., **Ian B. Wood**, William P. Hetrick, and Yong-Yeol Ahn. "Element-centric clustering comparison unifies overlaps and hierarchy." *Scientific reports* 9, no. 1 (2019): 8574. https://doi.org/10.1038/s41598-019-44892-y

Wood, Ian B., Pedro L. Varela, Johan Bollen, Luis M. Rocha, and Joana Gonçalves-Sá. "Human sexual cycles are driven by culture and match collective moods." *Scientific reports* 7, no. 1 (2017): 1-11. https://doi.org/10.1038/s41598-017-18262-5

Francisco, Matthew R., **Ian Wood**, Selma Šabanović, and Luis Rocha. "Designing a Minimalist Socially Aware Robotic Agent for the Home." *Proceedings of the ALIFE 14: The Fourteenth International Conference on the Synthesis and Simulation of Living Systems* (2014): 876-883. https://dx.doi.org/10.7551/978-0-262-32621-6-ch144

Ten Thij, Marijn, **Ian B. Wood**, Johan Bollen, Luis M. Rocha. "Decomposition of online sentiment reveals societal eigenmoods" *In Preparation*

Wang, X., **Ian B. Wood**, Heng-Yi Wu, Shijun Zhang, Hagit Shatkay-Reshef, Lang Li, Luis M. Rocha. "Systematic prediction of drug-drug-interaction study types and discovery of evidence gaps in the literature" *In Preparation*

Contributed Talks

"Community Detection with Selective Zooming." September 19, 2017. *Conference on Complex Systems*. Cancun, Quintana Roo, Mexico. **Ian Wood**, Xiaoran Yan, Xiaozhong Liu and Yong-Yeol Ahn

"Community Detection with Selective Zooming." June, 2017. *International School and Conference on Network Science*. Indianapolis, Indiana. **Ian Wood**, Xiaoran Yan, Xiaozhong Liu and Yong-Yeol Ahn

"Eigenmood Twitter Analysis: measuring collective mood variation." September, 2015. *Conference on Complex Systems*. Tempeh, Arizona. **Ian B. Wood**, Joana Gonçalves-Sá, Johan Bollen and Luis M. Rocha

"Eigenday Twitter analysis." June 11, 2015. *International Conference on Computational Social Science*. Finlandia Hall, Helsinki, Finland. **Ian Wood**, Johan Bollen and Luis Rocha.