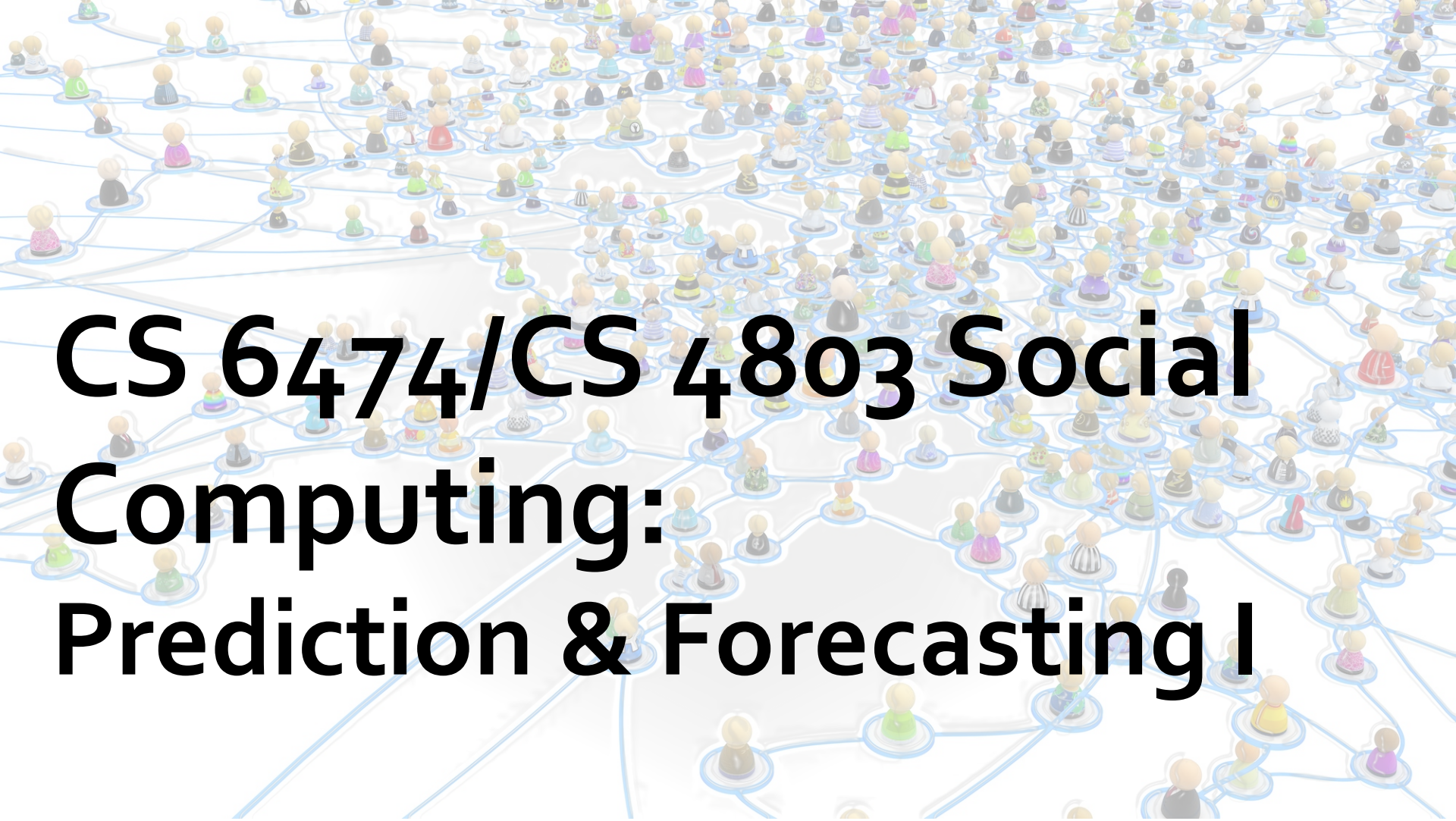


CS 6474/CS 4803 Social Computing: Prediction & Forecasting I

Munmun De Choudhury

munmund@gatech.edu

Week 12 | November 6, 2019

A little background…

- Human beings are fascinated with what will happen in the future and, indeed, we even associate intelligence with an ability to predict future events (Hawkins 2004).
- In ancient times, several techniques were invented including inspecting bird flights, haruspicy, and astrology. ‘
- Later, predictions were done mostly through experts who had developed their own intuitions and methods of prediction.


Machine Learning is Everywhere

Discovering of Health Risks and Case-Based Forecasting of Epidemics in a Health Surveillance System *

M. Bull G. Kundt L. Gierl

University of Rostock, Department for Medical Informatics and Biometry
Rembrandtstr. 16/17, D-18055 Rostock, Germany
{mathias.bull|guenther.kundt|lothar.gierl}@medizin.uni-rostock.de

Abstract. In this paper we present the methodology and the architecture of an early warning system which fulfills the following tasks. (1) discovering of health risks, (2) forecasting of the temporal and spatial spread of epidemics and (3) estimating of the consequences of an epidemic w.r.t. the personnel load and costs of the public health service. To cope this three task methods from knowledge discovery and data mining, case-based reasoning, and statistics are applied.

Keywords: knowledge discovery and data mining, case-based reasoning and forecasting,

Applying Text Mining to Protest Stories as Voice against Media Censorship

Tahsin Mayeesha
North South University
Dhaka, Bangladesh

Zareen Tasnim
North South University
Dhaka, Bangladesh

Jasmine Jones
University of Minnesota
author3@anotherco.com
jazzi@umn.edu

Nova Ahmed
North South University
Dhaka, Bangladesh
nova@northsouth.edu

Abstract
Data driven activism attempts to collect, analyze and visualize data to foster social change. However, during media censorship it is often impossible to collect such data. Here we demonstrate that data from personal stories can also help us to gain insights about protests and activism which can work as a voice for the activists.

Author Keywords
Protest; data mining; social justice; text analysis;media restriction.

ACM Classification Keywords
H.5.m. Information interfaces and presentation: Miscellaneous

Introduction
Many social movements like "Occupy Wall Street" [1] or "Arab Spring"[2] has been quantified and modeled extensively using data from social media like Twitter. Despite lacking in social media data due to censorship, here we demonstrate that analyzing social movements by text analysis of personal stories can also help us to learn the emotional effects and entities involved in a social movement and can act as a voice for the activists. We use the data from a recent student driven protest in Bangladesh for road safety for this purpose.

Artificial Intelligence — The Revolution Hasn't Happened Yet

Michael Jordan Follow
Apr 19, 2018 · 16 min read

Artificial Intelligence (AI) is the mantra of the current era. The phrase is intoned by technologists, academicians, journalists and venture capitalists alike. As with many phrases that cross over from technical academic fields into general circulation, there is significant misunderstanding accompanying the use of the phrase. But this is not the classical case of the public not understanding the scientists—here the scientists are often as befuddled as the public. The idea that our era is somehow seeing the emergence of an intelligence in silicon that rivals our own entertains all of us—enthralling us and frightening us in equal measure. And, unfortunately, it distracts us.

LETTER

https://doi.org/10.1038/s41586-018-0438-y

Deep learning of aftershock patterns following large earthquakes

Phoebe M. R. DeVries[1,2]*, Fernanda Viégas[3], Martin Wattenberg[3] & Brendan J. Meade[1]

Aftershocks are a response to changes in stress generated by large earthquakes and represent the most common observations of the triggering of earthquakes. The maximum magnitude of aftershocks and their temporal decay are well described by empirical laws (such as Bath's law[1] and Omori's law[2]), but explaining and forecasting the spatial distribution of aftershocks is more difficult. Coulomb failure stress change[3] is perhaps the most widely used criterion to explain the spatial distributions of aftershocks[4–8], but its applicability has been disputed[9–11]. Here we use a deep-learning approach to identify neuron may be interpreted as the predicted probability that a grid cell generates one or more aftershocks.

The stress changes and aftershock locations associated with about 75% of randomly selected distinct mainshocks were used as training data; the remaining 25% were reserved to test the trained neural networks. The training and testing datasets both consist of the elements of the stress-change tensor as features and the corresponding labels of either 0, for grid cells without aftershocks, or 1, for grid cells with aftershocks.

HEALTH CARE

AI WILL MAKE QUICKER DIAGNOSES, CREATE BETTER TREATMENT PLANS, AND ENABLE NEW APPROACHES TO INSURANCE

Health care is a promising market for AI. There is enormous potential in its ability to draw inferences and recognize patterns in large volumes of patient histories, medical images, epidemiological statistics, and other data. AI has the potential to help doctors improve their diagnoses, forecast the spread of diseases, and customize treatments. Artificial intelligence combined with health care digitization can allow providers to monitor or diagnose patients remotely as well as transform the way we treat the chronic diseases that account for a large share of health-care budgets.

NITI Aayog

Introduction

#AIforAll: Technology Leadership for Inclusive Growth

Artificial Intelligence (AI) is poised to disrupt our world. With intelligent machines enabling high-level cognitive processes like thinking, perceiving, learning, problem solving and decision making, coupled with advances in data collection and aggregation, analytics and computer processing power, AI presents opportunities to complement and supplement human intelligence and enrich the way people live and work.

Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

***Abstract*—In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.**

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter [1], a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of users who actively participate in the creation and propagation of content.

We have focused on movies in this study for two main reasons.

A Long List of Predictions….

- Predicting the H1N1 flu outbreak using Twitter data
- Predicting the outcome of the 2016 Presidential elections in the US using social media data
- Predicting people's home location from their geo-located social media posts
- Predicting traffic conditions every day during rush hour based on geo-located social media posts
- Predicting people's diurnal moods based on social media activity
- Predicting an individual's mental health state from the language of social media posts

Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”

Summary

- The paper predicts stock market indicators such as Dow Jones, NASDAQ and S&P 500 by analyzing Twitter posts
- Main measure of interest – emotionality of tweets (positive and negative words like hope, fear and so on)
  - Goal to use the emotionality measures at day *t* to predict stock market indices at day *t*+1
- Emotionality is negatively correlated with Dow Jones, NASDAQ and S&P 500 but positively with VIX

- One of the earliest works on financial prediction using social media

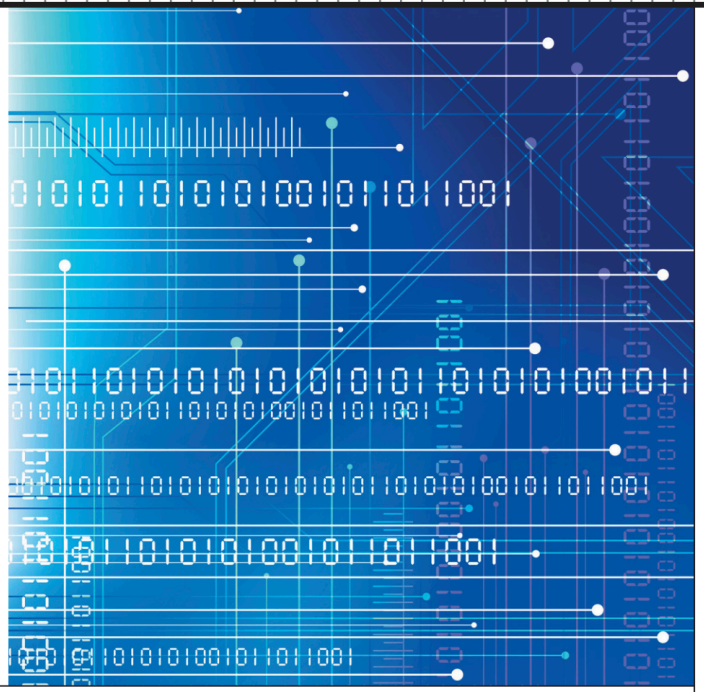
	Dow	NASDAQ	S&P 500	VIX
Hope %	– 0.381**	– 0.407**	– 0.373**	0.337**
Happy %	– 0.107	– 0.105	– 0.103	0.114
Fear %	– 0.208*	– 0.238*	– 0.200	0.235*
Worry %	– 0.300**	– 0.305**	– 0.295**	0.305**
Nervous %	– 0.023	– 0.054	– 0.021	0.015
Anxious %	– 0.261*	– 0.295**	– 0.262*	0.320**
Upset %	– 0.185	– 0.188	– 0.184	0.126
Positive %	– 0.192	– 0.197	– 0.187	0.188
Negative %	– 0.294**	– 0.323**	– 0.288**	0.301**

Table 2. Correlation Coefficient of emotional tweets percentage and stock market indicators (N=93) with total number of tweets per day as a baseline

	Dow	NASDAQ	S&P 500	VIX
Hope%	– 0.381**	– 0.407**	– 0.373**	0.337*
Hope%-2 mean	– 0.618**	– 0.631**	– 0.607**	0.518**
Hope%-3-mean	– 0.737**	– 0.738**	– 0.724**	0.621**
Fear%	– 0.208 *	– 0.238 *	– 0.2	0.235*
Fear%-2-mean	– 0.259*	– 0.285**	– 0.253*	0.312**
Fear%-3-mean	– 0.346**	– 0.368**	– 0.342**	0.403**
Worry%	– 0.3**	– 0.305**	– 0.295**	0.305*
Worry%-2-mean	– 0.421**	– 0.415**	– 0.414**	0.410**
Worry%-3-mean	– 0.472**	– 0.460**	– 0.467**	0.459**
Hope+Fear+Worry%	– 0.379**	– 0.405**	– 0.37**	0.347*
Hope+Fear+Worry%-2-mean	– 0.612**	– 0.625**	– 0.6**	0.532**
Hope+Fear+Worry%-3-mean	– 0.726**	– 0.728**	– 0.713**	0.633**

Table 6. Correlation Coefficient of average emotional tweets percentage and stock market indicators (N=93)

Twitter Mood as a Stock Market Predictor



Johan Bollen and Huina Mao
Indiana University Bloomington

Behavioral finance researchers can apply computational methods to large-scale social media data to better understand and predict markets.

It has often been said that stock markets are driven by "fear and greed"—that is, by psychological as well as financial factors. The tremendous volatility of stock markets across the globe in recent years underscores the need to better understand the role that emotions play in shaping stock prices and other economic indices.

A stock market is a large-scale, complex information processing of rational considerations and that their behavior is subject to particular psychological biases and emotions. Consequently, predicting market behavior requires understanding the factors that shape investors' individual as well as collective behavior.

PREDICTING MARKET BEHAVIOR

Behavioral finance and investor sentiment theory have firmly sentiment affect stock prices, as it was a few decades ago, but rather how we can best measure and model their effects.

Historically, surveys have been the most direct way to measure social mood and investor sentiment. For example, the Conference Board's Consumer Confidence Index, the University of Michigan's Consumer Sentiment Index, and Gallup's Economic Confidence Index measure

Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

Widespread Worry and the Stock Market

Eric Gilbert and Karrie Karahalios

Department of Computer Science
University of Illinois at Urbana Champaign
[egilber2, kkarahal]@cs.uiuc.edu

Abstract

Our emotional state influences our choices. Research on how it happens usually comes from the lab. We know rela tively little about how real world emotions affect real world settings, like financial markets. Here, we demonstrate that estimating emotions from weblogs provides novel informa tion about future stock market prices. That is, it provides information not already apparent from market data. Specifi cally, we estimate anxiety, worry and fear from a dataset of over 20 million posts made on the site LiveJournal. Using a Granger causal framework, we find that increases in expres sions of anxiety, evidenced by computationally identified linguistic features, predict downward pressure on the S&P 500 index. We also present a confirmation of this result via Monte Carlo simulation. The findings show how the mood of millions in a large online community, even one that pri marily discusses daily life, can anticipate changes in a seem ingly unrelated system. Beyond this, the results suggest new ways to gauge public opinion and predict its impact.

risk-averse. Still, this thread of research comes from the lab. How do real world emotions affect real world markets, like the stock market?

In this paper, we take a step toward answering this question. From a dataset of over 20 million LiveJournal posts, we construct a metric of anxiety, worry and fear called the Anxiety Index. The Anxiety Index is built on the judgements of two linguistic classifiers trained on a LiveJournal mood corpus from 2004. The major finding of this paper is that the Anxiety Index has information about future stock market prices not already apparent from market data. We demonstrate this result using an econometric technique called Granger causality. In particular, we show that the Anxiety Index has novel information about the S&P 500 index over 174 trading days in 2008, roughly 70% of the trading year. We estimate that a one standard deviation rise in the Anxiety Index corresponds to S&P 500 returns 0.4% lower than otherwise expected.

Journal of Computational Science 2 (2011) 1–8

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Twitter mood predicts the stock market

Johan Bollen[a,*,1], Huina Mao[a,1], Xiaojun Zeng[b]

[a] School of Informatics and Computing, Indiana University, 919 E. 10th Street, Bloomington, IN 47408, United States
[b] School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, United Kingdom

ARTICLE INFO

Article history:
Received 15 October 2010
Received in revised form 2 December 2010
Accepted 5 December 2010
Available online 2 February 2011

Keywords:
Social networks
Sentiment tracking
Stock market
Collective mood

ABSTRACT

Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 86.7% in predicting the daily up and down changes

Expert Systems With Applications 42 (2015) 9603–9611

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Sentiment analysis on social media for stock movement prediction

Thien Hai Nguyen [a,*], Kiyoaki Shirai [a], Julien Velcin [b]

[a] School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[b] University of Lyon (ERIC, Lyon 2), 5 Avenue Pierre Mendes-France, 69676 Bron Cedex, France

ARTICLE INFO

Keywords:
Sentiment analysis
Opinion mining
Classification
Prediction
Stock
Social media
Message board

ABSTRACT

The goal of this research is to build a model to predict stock price movement using the sentiment from social media. Unlike previous approaches where the overall moods or sentiments are considered, the sentiments of the specific topics of the company are incorporated into the stock prediction model. Topics and related sentiments are automatically extracted from the texts in a message board by using our proposed method as well as existing topic models. In addition, this paper shows an evaluation of the effectiveness of the sentiment analysis in the stock prediction task via a large scale experiment. Comparing the accuracy average over 18 stocks in one year transaction, our method achieved 2.07% better performance than the model using historical prices only. Furthermore, when comparing the methods only for the stocks that are difficult to predict, our method achieved 9.83% better accuracy than historical price method, and 3.03% better than human sentiment method.

International Review of Financial Analysis 48 (2016) 272–281

Contents lists available at ScienceDirect

International Review of Financial Analysis

Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction ☆

Andrew Sun [a], Michael Lachanski [b], Frank J. Fabozzi [c,*]

[a] Consultant, United States
[b] University of Tokyo, Graduate School of Economics, Japan
[c] EDHEC Business School, United States

ARTICLE INFO

Article history:
Received 6 August 2016
Accepted 17 October 2016
Available online 24 October 2016

Keywords:
Tweets
Social media text mining
Sparse matrix factorization

ABSTRACT

We investigate the potential use of textual information from user-generated microblogs to predict the stock market. Utilizing the latent space model proposed by Wong et al. (2014), we correlate the movements of both stock prices and social media content. This study differs from models in prior studies in two significant ways: (1) it leverages market information contained in high-volume social media data rather than news articles and (2) it does not evaluate sentiment. We test this model on data spanning from 2011 to 2015 on a majority of stocks listed in the S&P 500 Index and find that our model outperforms a baseline regression. We conclude by providing a trading strategy that produces an attractive annual return and Sharpe ratio.

International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016

Sentiment Analysis of Twitter Data for Predicting Stock Market Movements

Venkata Sasank Pagolu
School of Electrical Sciences
Computer Science and Engineering
Indian Institute of Technology,
Bhubaneswar, India 751013
Email: vp12@iitbbs.ac.in

Kamal Nayan Reddy Challa
School of Electrical Sciences
Computer Science and Engineering
Indian Institute of Technology,
Bhubaneswar, India 751013
Email: kc11@iitbbs.ac.in

Ganapati Panda
School of Electrical Sciences
Indian Institute of Technology
Bhubaneswar, India 751013
Email: gpanda@iitbbs.ac.in

Babita Majhi
Department of Computer Science and IT
G.G Vishwavidyalaya, Central University
Bilaspur, India 495009
Email: babita.majhi@gmail.com

***Abstract*—Predicting stock market movements is a well-known problem of interest. Now-a-days social media is perfectly representing the public sentiment and opinion about current events. Especially, twitter has attracted a lot of attention from researchers for studying the public sentiments. Stock market prediction on the basis of public sentiments expressed on twitter has been an intriguing field of research. Previous studies have concluded that the aggregate public mood collected from twitter may well be correlated with Dow Jones Industrial Average Index (DJIA). The thesis of this work is to observe how well the changes in stock prices of a company, the rises and falls, are correlated with the public opinions being expressed in tweets about that company. Understanding author's opinion from a piece of text is the objective of sentiment analysis. The present paper have employed two different textual representations, Word2vec and N-gram, for analyzing the public sentiments in tweets. In this paper, we have applied sentiment analysis and supervised machine learning principles to the tweets extracted from twitter and**

random walk pattern and cannot be predicted with more than 50% accuracy [1].

With the advent of social media, the information about public feelings has become abundant. Social media is transforming like a perfect platform to share public emotions about any topic and has a significant impact on overall public opinion. Twitter, a social media platform, has received a lot of attention from researchers in the recent times. Twitter is a micro-blogging application that allows users to follow and comment other users thoughts or share their opinions in real time [3]. More than million users post over 140 million tweets every day. This situation makes Twitter like a corpus with valuable data for researchers [4].Each tweet is of 140 characters long and speaks public opinion on a topic concisely. The information exploited from tweets are very useful for making predictions [5].

Class Discussion

Discuss how can social media based stock market predictors could be helpful. Who will it help?

The paper does not answer the "why"? Why was emotionality negatively correlated with Dow and S&P, but positively with VIX?

Class Discussion

If social media is such a great predictor of stock market indices, why are we not preventing bad financial outcomes? And why is anybody ever losing money on the market?

Computer Science > Computers and Society

"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" Balanced Survey on Election Prediction using Twitter Data

Daniel Gayo-Avello

(Submitted on 28 Apr 2012)

Predicting X from Twitter is a popular fad within the Twitter research subculture. It seems both appealing and relatively easy. Among s electoral prediction is maybe the most attractive, and at this moment there is a growing body of literature on such a topic. This is not research problem but, above all, it is extremely difficult. However, most of the authors seem to be more interested in claiming positiv providing sound and reproducible methods. It is also especially worrisome that many recent papers seem to only acknowledge those s the idea of Twitter predicting elections, instead of conducting a balanced literature review showing both sides of the matter. After rea papers I have decided to write such a survey myself. Hence, in this paper, every study relevant to the matter of electoral prediction usi commented. From this review it can be concluded that the predictive power of Twitter regarding elections has been greatly exaggerate research problems still lie ahead.

Comments: 13 pages, no figures. Annotated bibliography of 25 papers regarding electoral prediction from Twitter data
Subjects: **Computers and Society (cs.CY)**; Computation and Language (cs.CL); Social and Information Networks (cs.SI); Physics and Society (physics.
Cite as: arXiv:1204.6441 [cs.CY]
(or arXiv:1204.6441v1 [cs.CY] for this version)

Submission history

Why Watching Movie Tweets Won't Tell the Whole Story?

Felix Ming Fai Wong
EE, Princeton University
mwthree@princeton.edu

Soumya Sen
EE, Princeton University
soumyas@princeton.edu

Mung Chiang
EE, Princeton University
chiangm@princeton.edu

ABSTRACT

Data from Online Social Networks (OSNs) are providing analysts with an unprecedented access to public opinion on elections, news, movies etc. However, caution must be taken to determine whether and how much of the opinion extracted from OSN user data is indeed reflective of the opinion of the larger online population. In this work we study this issue in the context of movie reviews on Twitter and compare the opinion of Twitter users with that of the online population of IMDb and Rotten Tomatoes. We introduce new metrics to show that the Twitter users can be characteristically different from general users, both in their rating and their relative preference for Oscar-nominated and non-nominated movies. Additionally, we investigate whether such data can truly predict a movie's box-office success.

Categories and Subject Descriptors

this study because marketers consider brand interaction and information dissemination as a major aspect of Twitter. The focus on movies in this paper is also driven by two key factors:

(a) *Right in the Level of Interest:* Movies tend to generate a high interest among Twitter users as well as in other online user population (e.g., IMDb).

(b) *Right in Timing:* We collected Twitter data during Academy Award season (Oscars) to obtain a unique dataset to analyze characteristic differences between Twitter and IMDb or Rotten Tomatoes users in their reviews of Oscar-nominated versus non-nominated movies.

We collected data from Twitter between February-March 2012 and manually labeled 10K tweets as training data for a set of classifiers based on SVM. We focus on the following questions to investigate whether Twitter data is sufficiently representative and indicative of future outcomes:

Limitations of stock market prediction with social media?

Social media predictions and traditional forecasting

- Data is collected through past logs of experiences.
- Data could also be collected on demand.
- A traditional and direct way to do that is by using polling, asking the public directly for their opinion or behavioral intentions, as is done with survey models.
  - Statistical models are often employed to make sense of them.
- Social media data is observational.
  - Prediction market model – use the wisdom of the crowds.

- But a sound market architecture guaranteeing the heterogeneity of participants is needed (Surowiecki 2004).

Treading with caution

Attention to noise, bias, and "provenance" — broadly, where did data arise, what inferences were drawn from the data, and how relevant are those inferences to the present situation?

Listen to this story
--:-- 22:31

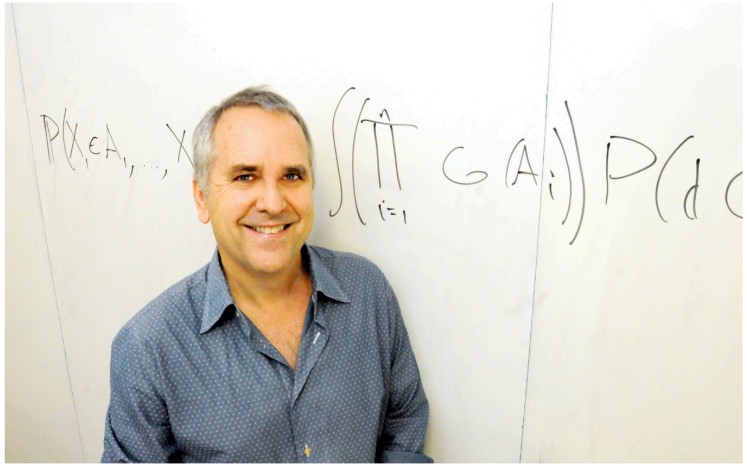


Photo credit: Peg Skorpinski

Artificial Intelligence — The Revolution Hasn't Happened Yet

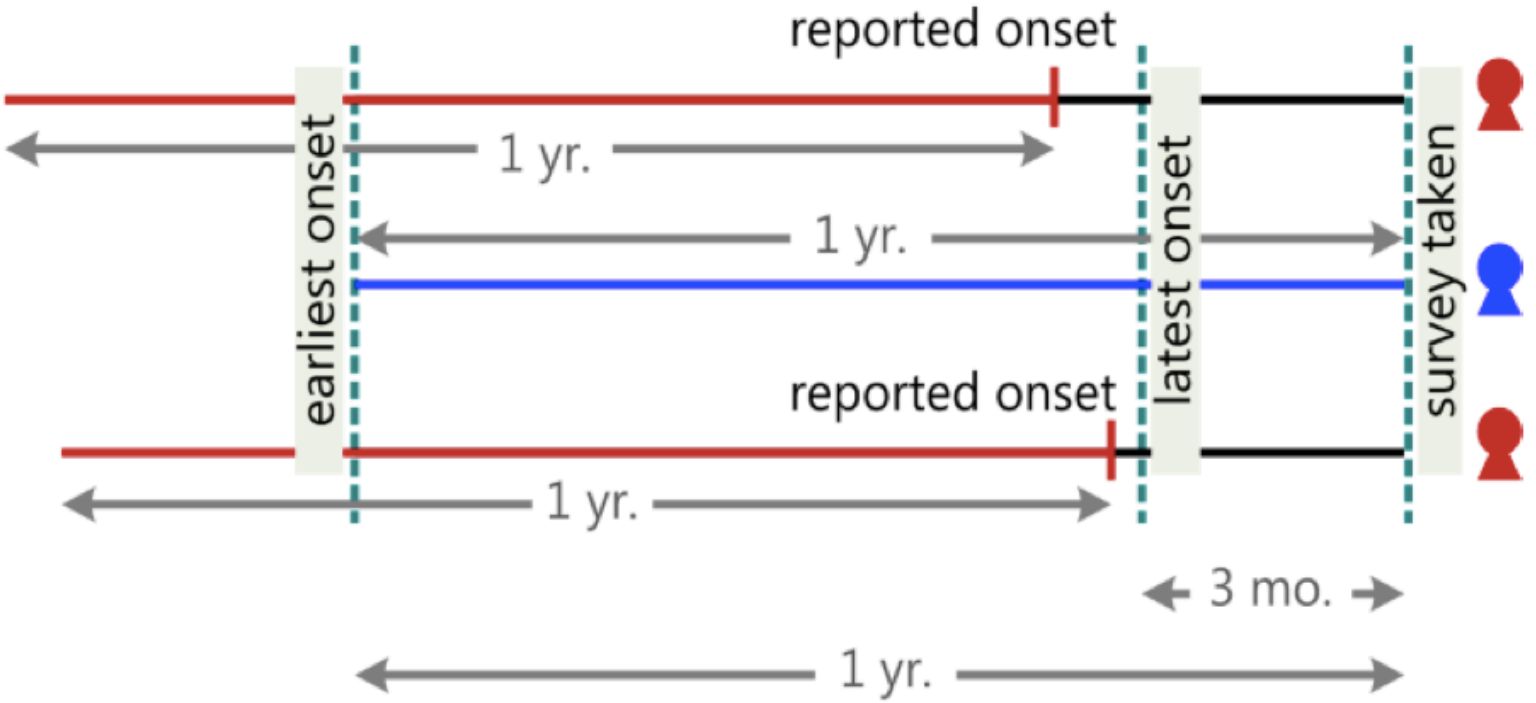
Michael Jordan Follow
Apr 19, 2018 · 16 min read

Artificial Intelligence (AI) is the mantra of the current era. The phrase is intoned by technologists, academicians, journalists and venture capitalists alike. As with many phrases that cross over from technical academic fields into general circulation, there is significant misunderstanding

Predicting Depression via Social Media

Summary

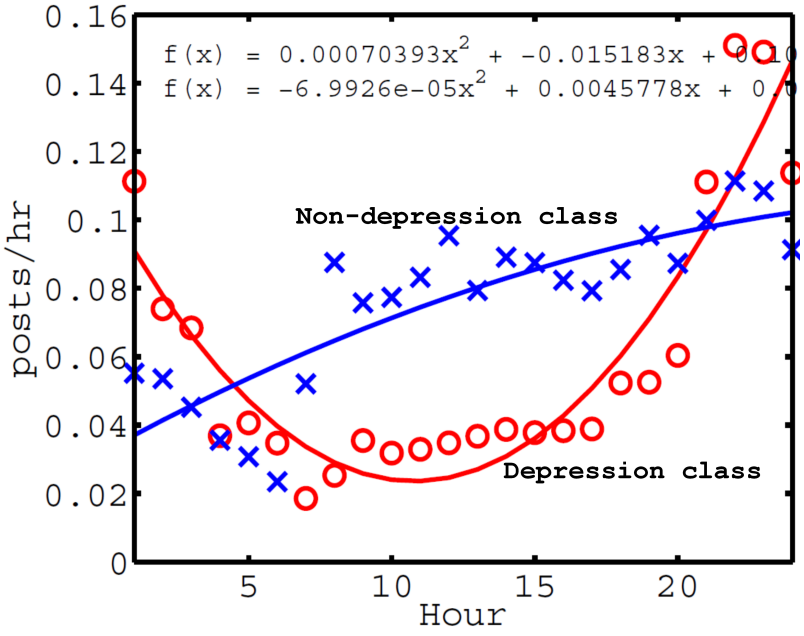
- Can social media activities and connectedness predict risk to major depressive disorder?
- Recruitment of a sample of Twitter users through a survey methodology over Amazon's Mechanical Turk
  - ~40% provided access to Twitter data

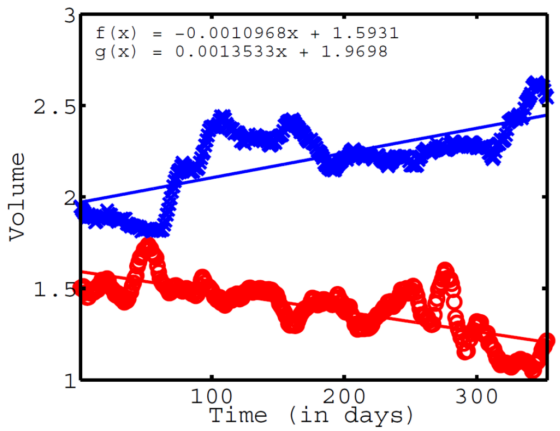


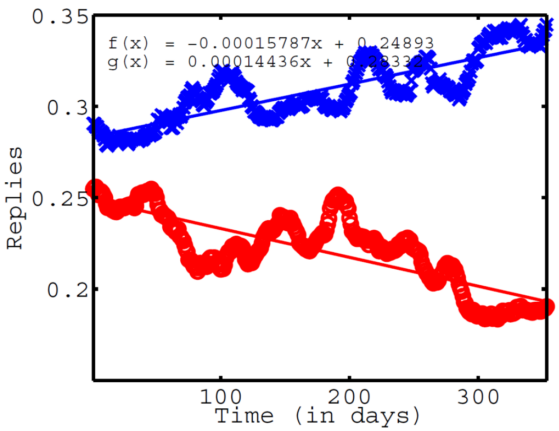
Summary

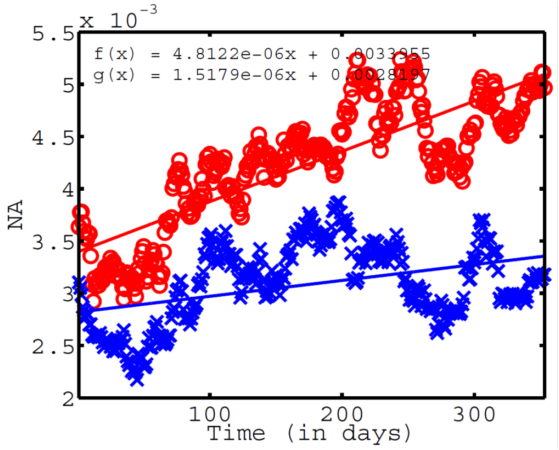
- Social engagement
- "Insomnia index" – mean *z*-score of an individual's volume of Twitter activity per hour
- Ego-centric social graph – nodal properties (*inlinks, outlinks*); dyadic properties (*reciprocity, interpersonal exchange*); neighborhood properties (*density, clustering coefficient, two-hop neighborhood, embeddedness, number of ego components*)
- Language
  - Depression lexicon – top uni- and bigrams compiled from Yahoo! Answers category on mental health
  - Linguistic style

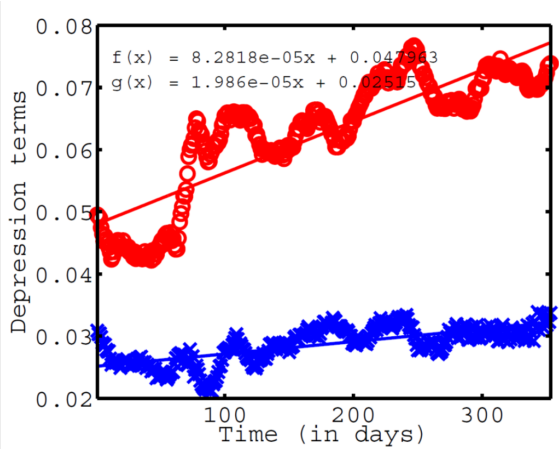
Summary


f(x) = 0.00070393x^2 + -0.015183x + 0.10
f(x) = -6.9926e-05x^2 + 0.0045778x + 0.0
Non-depression class
Depression class
posts/hr
Hour


f(x) = -0.0010968x + 1.5931
g(x) = 0.0013533x + 1.9698
Volume
Time (in days)


f(x) = -0.00015787x + 0.24893
Replies
Time (in days)


x 10^-3
f(x) = 4.8122e-06x + 0.0033855
g(x) = 1.5179e-06x + 0.0028197
NA
Time (in days)


f(x) = 8.2818e-05x + 0.047963
g(x) = 1.986e-05x + 0.02515
Depression terms
Time (in days)

Summary

Egonetwork measures	Depres. class	Non-depres. class
#followers/inlinks	26.9 (σ=78.3)	45.32 (σ=90.74)
#followees/outlinks	19.2 (σ=52.4)	40.06 (σ=63.25)
Reciprocity	0.77 (σ=0.09)	1.364 (σ=0.186)
Prestige ratio	0.98 (σ=0.13)	0.613 (σ=0.277)
Graph density	0.01 (σ=0.03)	0.019 (σ=0.051)
Clustering coefficient	0.02 (σ=0.05)	0.011 (σ=0.072)
2-hop neighborhood	104 (σ=82.42)	198.4 (σ=110.3)
Embeddedness	0.38 (σ=0.14)	0.226 (σ=0.192)
#ego components	15.3 (σ=3.25)	7.851 (σ=6.294)

Class Exercise I

In the depression prediction paper, the ground truth was obtained from Amazon mechanical turk workers. Anything unique about this population that may have affected the findings? What would be alternative ways of recruiting people or gathering high quality ground truth?

Class Exercise II

Depression is not an online condition, but one that spans both the online and the offline life. The paper does not take offline attributes into their models.

Is there a way to that into account? What would be the most significant offline attributes to consider?

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance

Mauricio Santillana[1,2,3]*, André T. Nguyen[1], Mark Dredze[4], Michael J. Paul[5], Elaine O. Nsoesie[6,7], John S. Brownstein[2,3]

1 Harvard School of Engineering and Applied Sciences, Cambridge, Massachusetts, United States of America, **2** Boston Children's Hospital Informatics Program, Boston, Massachusetts, United States of America, **3** Harvard Medical School, Boston, Massachusetts, United States of America, **4** Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, United States of America, **5** Department of Information Science, University of Colorado, Boulder, Colorado, United States of America, **6** Department of Global Health, University of Washington, Seattle, Washington, United States of America, **7** Institute for Health Metrics and Evaluation, Seattle, Washington, United States of America

* msantill@fas.harvard.edu

Citation: Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS (2015) Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol 11(10): e1004513. doi:10.1371/journal.pcbi.1004513

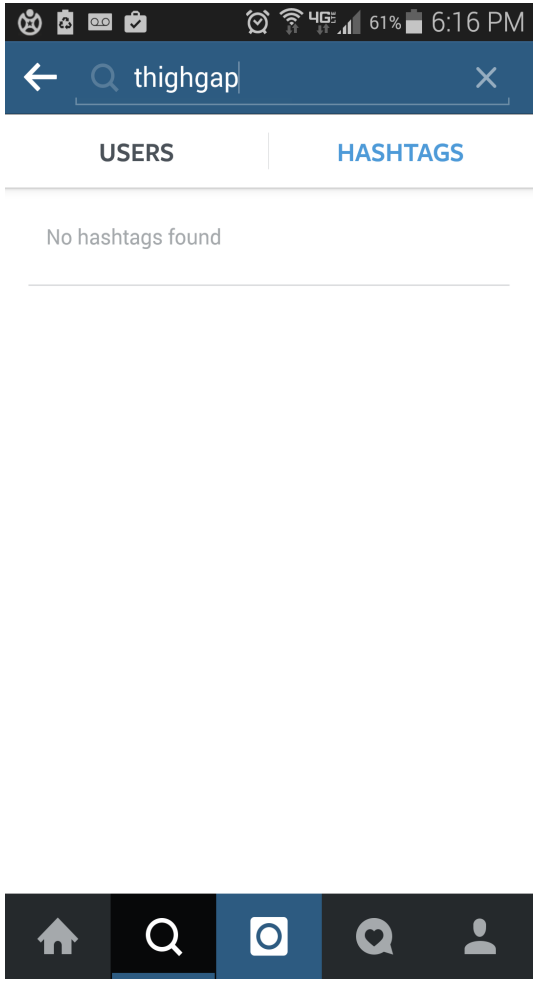
Abstract

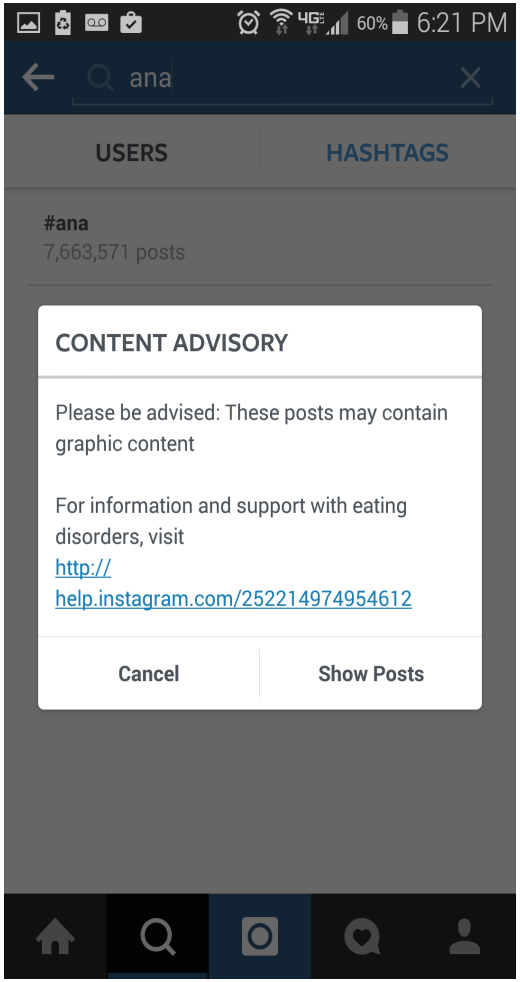
We present a machine learning-based methodology capable of providing real-time ("now-cast") and forecast estimates of influenza activity in the US by leveraging data from multiple data sources including: Google searches, Twitter microblogs, nearly real-time hospital visit records, and data from a participatory surveillance system. Our main contribution consists of combining multiple influenza-like illnesses (ILI) activity estimates, generated indepen-

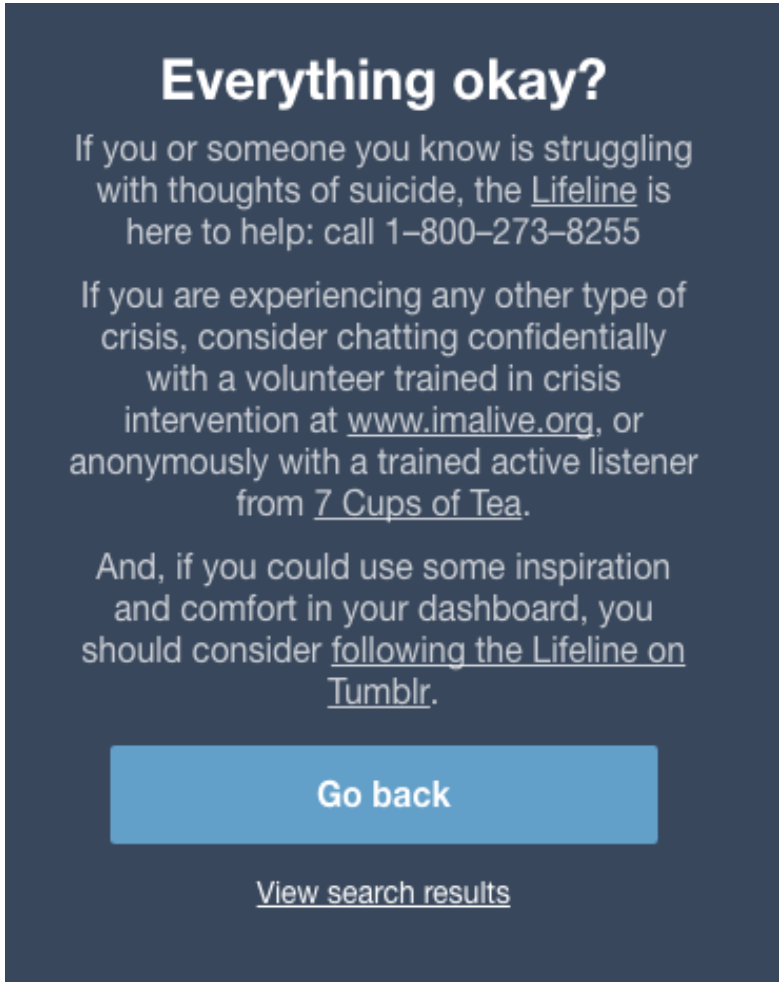
Class Exercise III

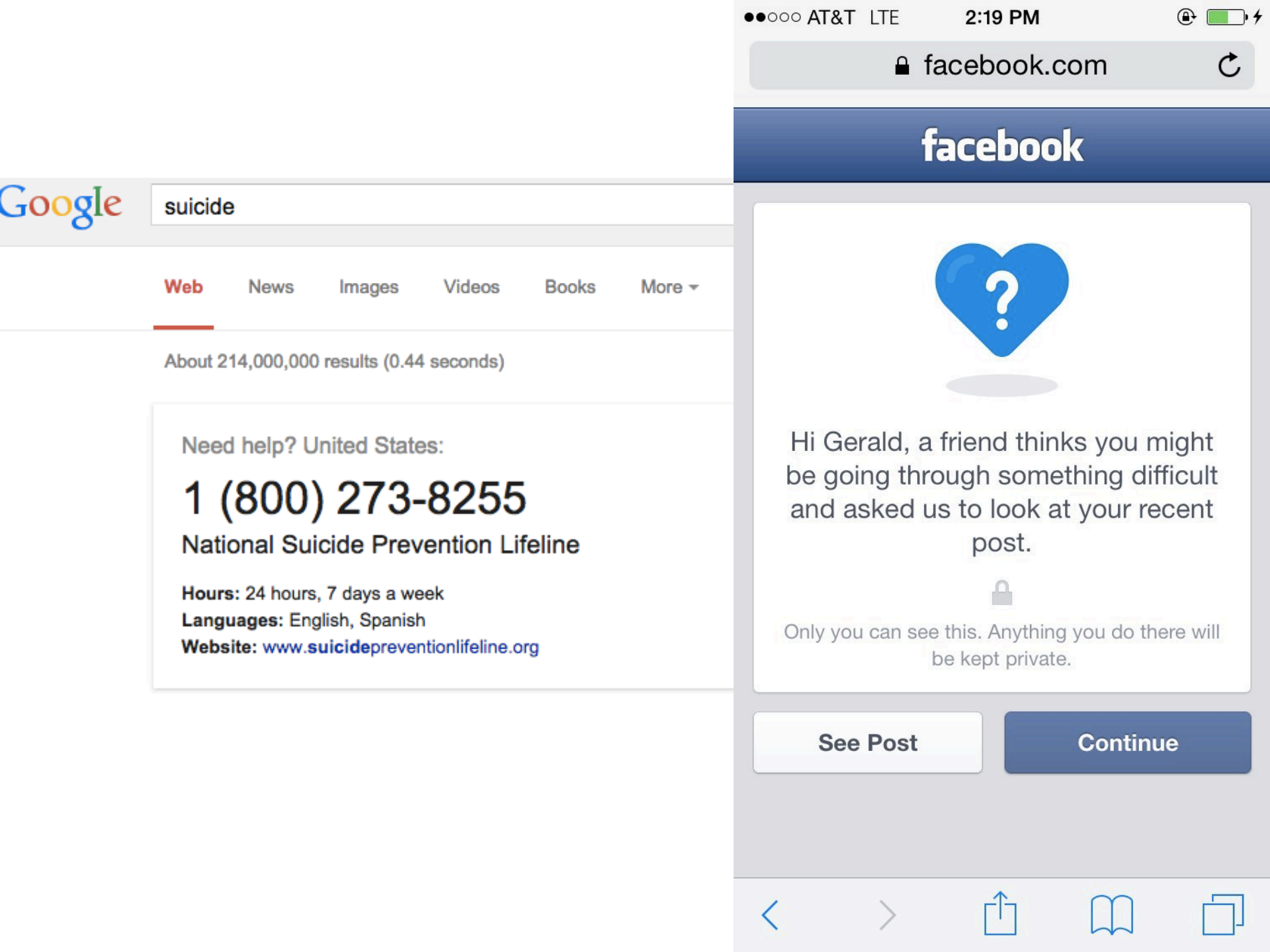
Describe a design idea where we can use social media based depression (or other mental health condition) predictors to help people.

Improving “Blanket” Interventions

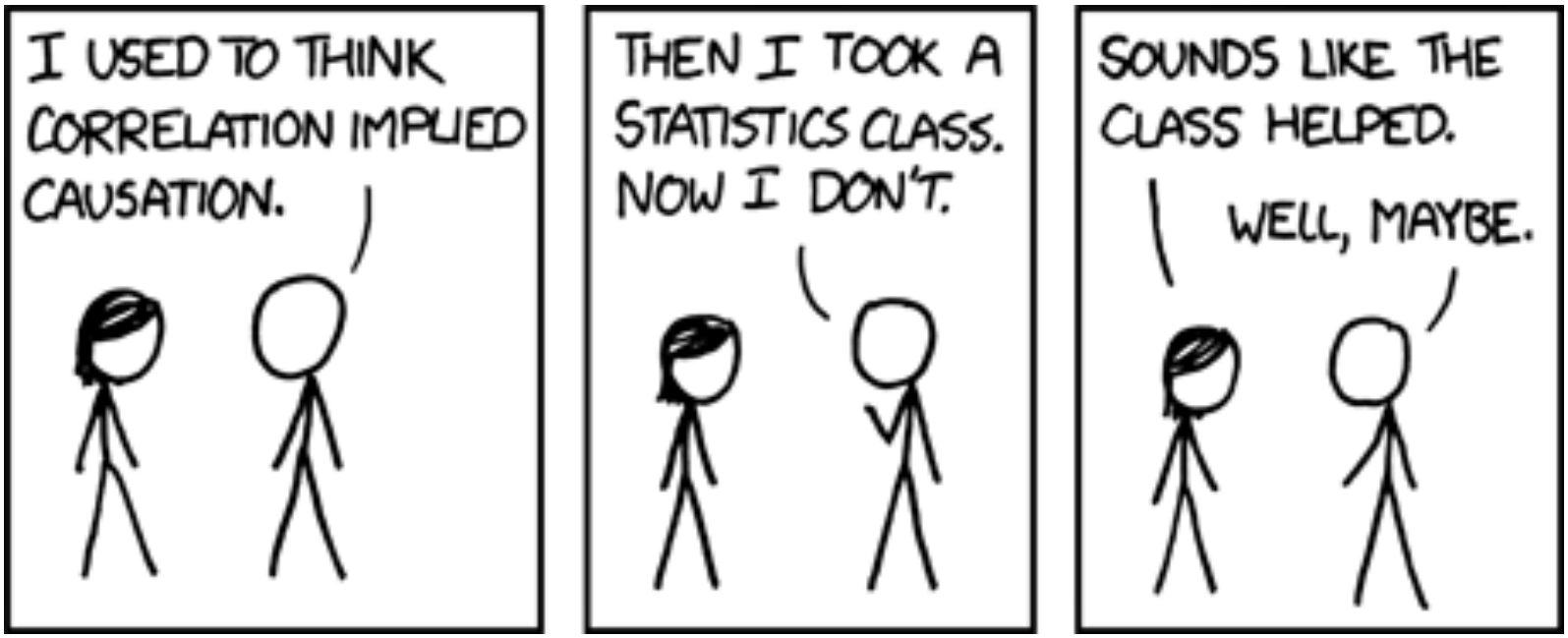




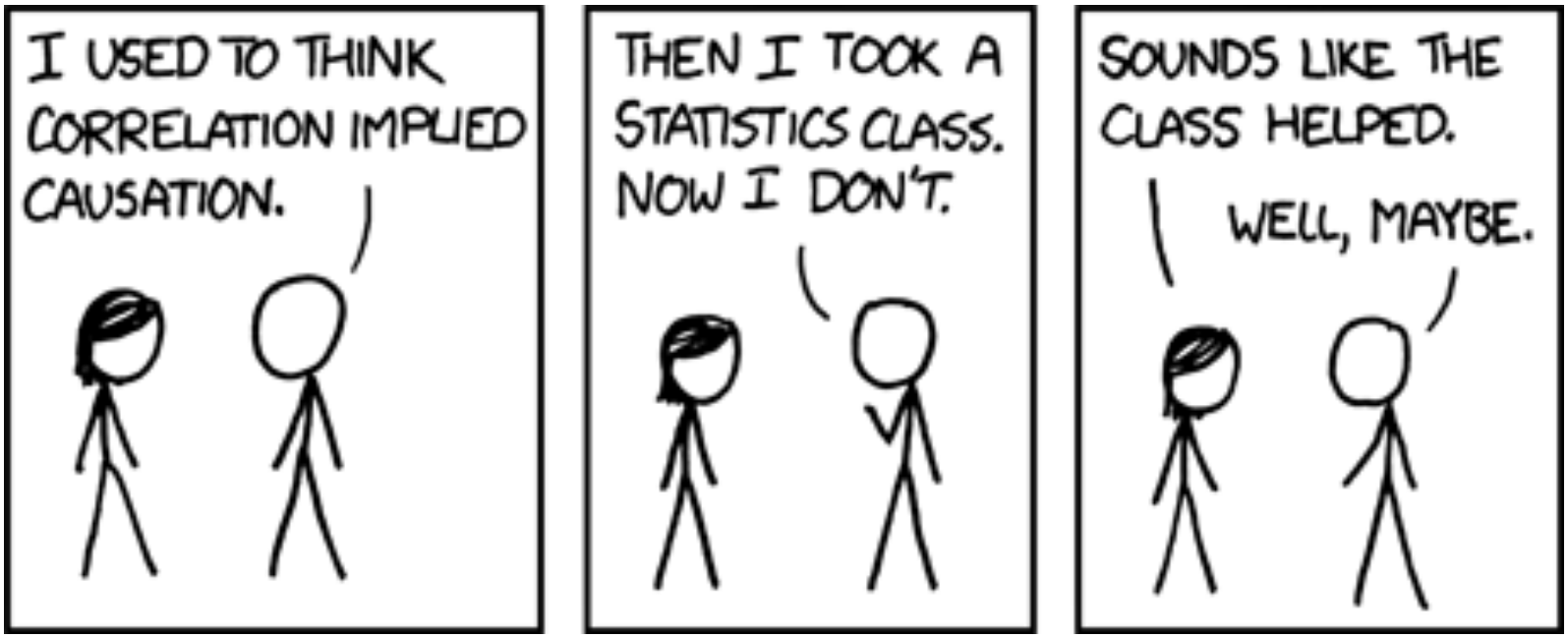



Google
suicide
Web
News
Images
Videos
Books
More
About 214,000,000 results (0.44 seconds)
Need help? United States:
1 (800) 273-8255
National Suicide Prevention Lifeline
Hours: 24 hours, 7 days a week
Languages: English, Spanish
Website: www.suicidepreventionlifeline.org
AT&T LTE
2:19 PM
facebook.com
facebook
Hi Gerald, a friend thinks you might be going through something difficult and asked us to look at your recent post.
Only you can see this. Anything you do there will be kept private.
See Post
Continue

But are models trained on aggregated group-level differences useful at the individual level?



Correlation and causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Social media data-based models cannot predict the future perfectly, because real-world outcomes can change in ways that are not anticipated by these data-based models.