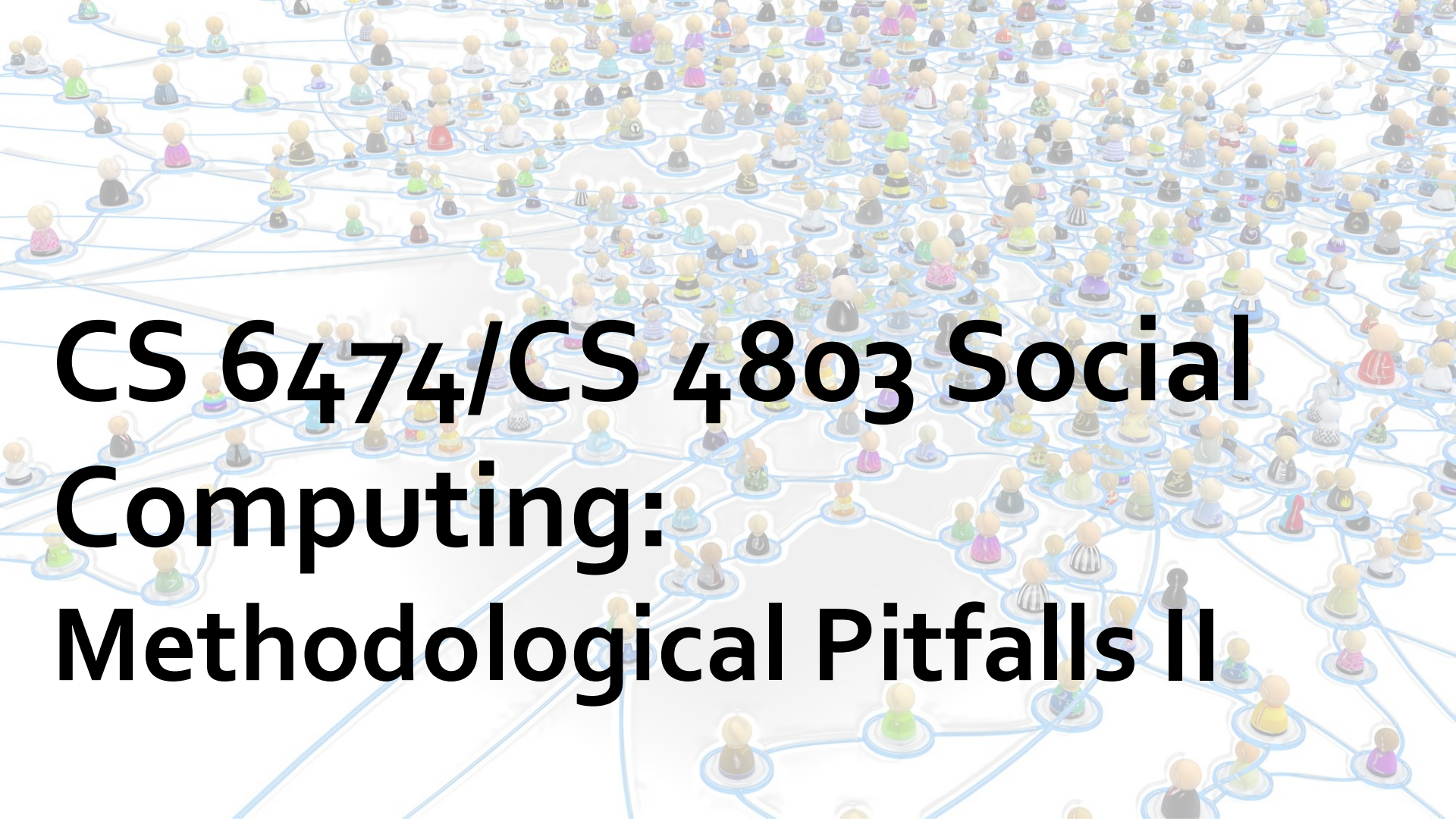


CS 6474/CS 4803 Social Computing: Methodological Pitfalls II

Munmun De Choudhury

munmund@gatech.edu

Week 13 | April 8, 2026

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

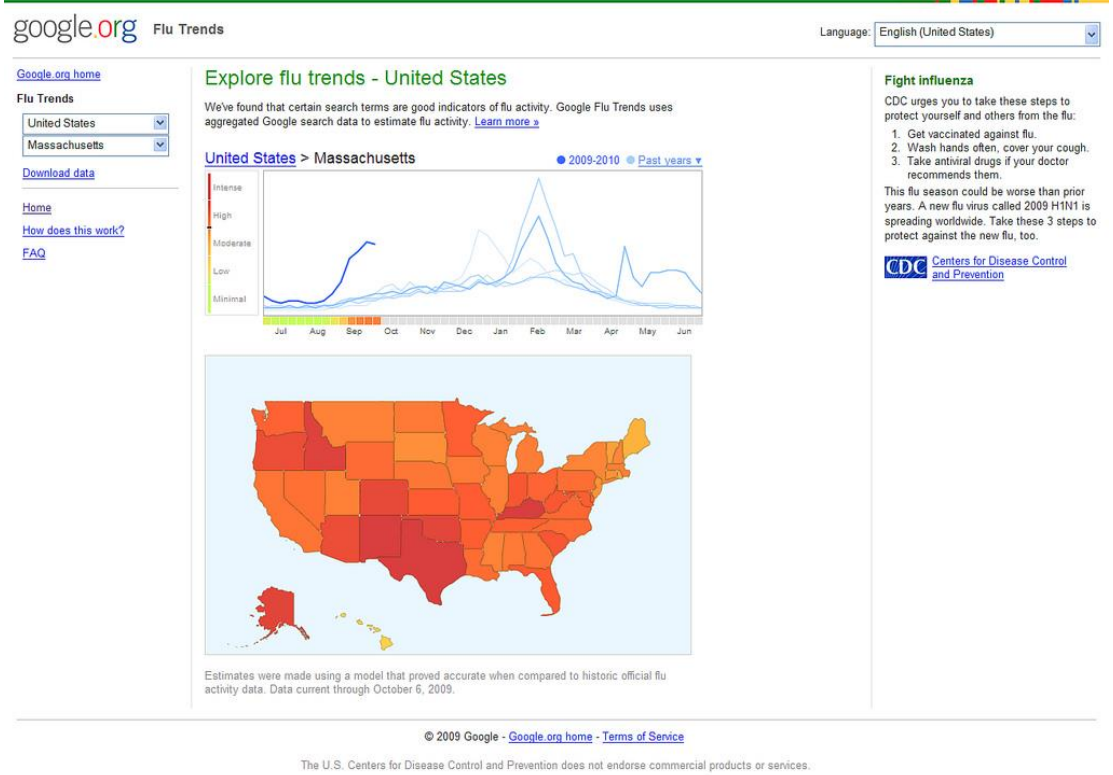
Daniel Gayo-Avello

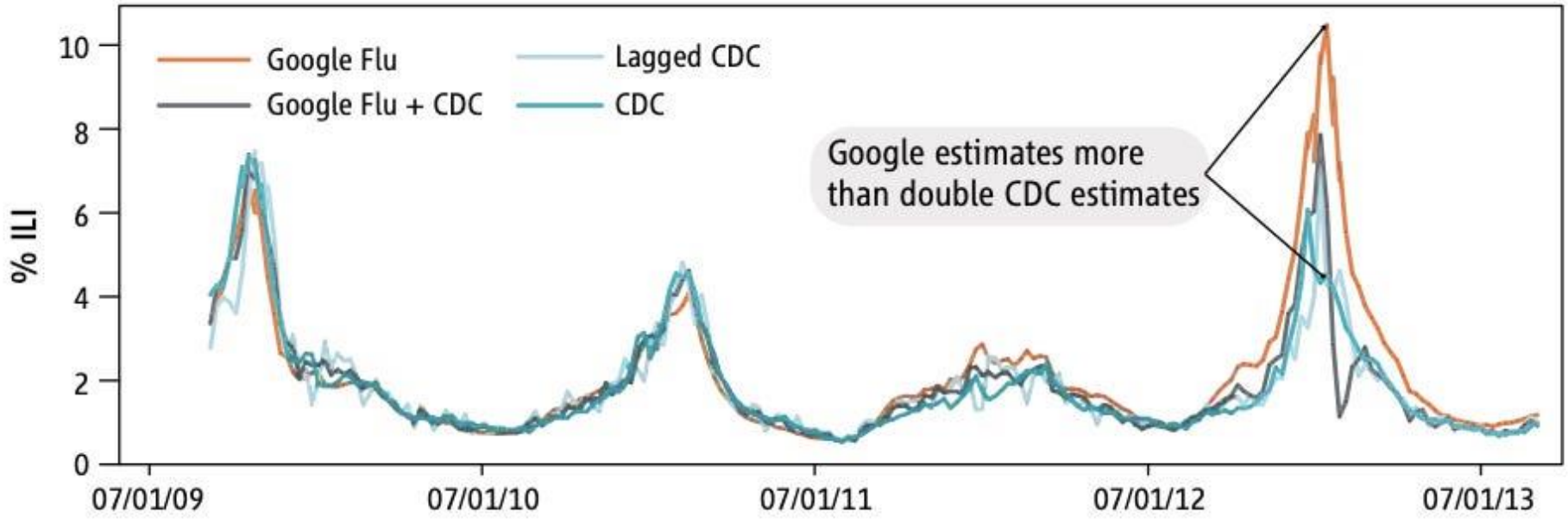
(Submitted on 28 Apr 2012)

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


google.org Flu Trends
Language: English (United States)
Google.org home
Flu Trends
United States
Massachusetts
Download data
Home
How does this work?
FAQ
Explore flu trends - United States
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »
United States > Massachusetts
2009-2010
Past years
Intense
High
Moderate
Low
Minimal
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through October 6, 2009.
Fight influenza
CDC urges you to take these steps to protect yourself and others from the flu:
1. Get vaccinated against flu.
2. Wash hands often, cover your cough.
3. Take antiviral drugs if your doctor recommends them.
This flu season could be worse than prior years. A new flu virus called 2009 H1N1 is spreading worldwide. Take these 3 steps to protect against the new flu, too.
CDC Centers for Disease Control and Prevention
© 2009 Google - Google.org home - Terms of Service
The U.S. Centers for Disease Control and Prevention does not endorse commercial products or services.


Google Flu
Google Flu + CDC
Lagged CDC
CDC
Google estimates more than double CDC estimates
% ILI
10
8
6
4
2
0
07/01/09
07/01/10
07/01/11
07/01/12
07/01/13

A little background…


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
jazzij@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.

Major trends in social media prediction

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

	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)

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.

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

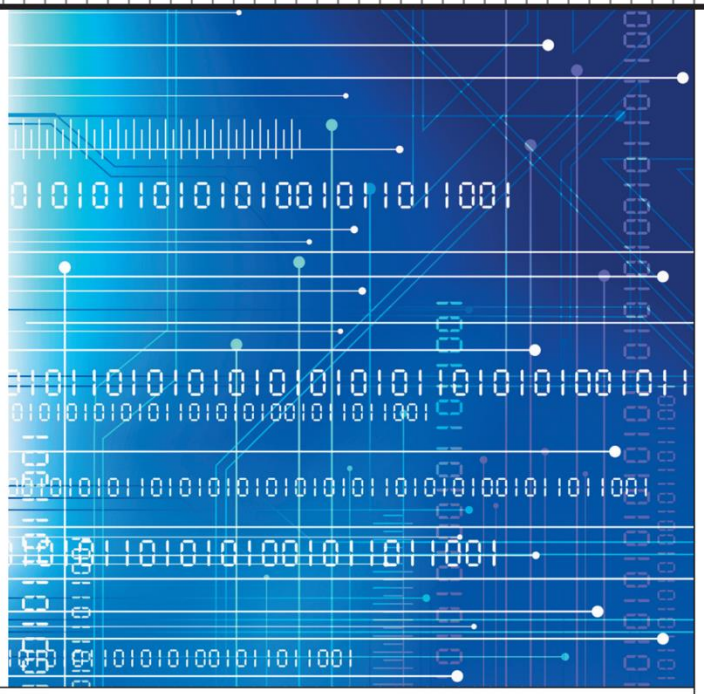
Keywords:
Tweets
Social media text mining
Sparse matrix factorization

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.

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

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?

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:

Published: 14 December 2019

Social media prediction: a literature review

Dimitrios Rousidis, Paraskevas Koukaras & Christos Tjortjis

Multimedia Tools and Applications **79**, 6279–6311(2020) | Cite this article

Abstract

Social Media Prediction (SMP) is an emerging powerful tool attracting the attention of researchers and practitioners alike. Despite its many merits, SMP has also several weaknesses, as it is limited by data issues, like bias and noise, and the lack of confident predictions and generalizable results. The goal of this paper is to survey popular and trending fields of SMP from 2015 and onwards and discuss the predictive models used. We elaborate on results found in the literature, while categorizing the forecasting attempts, based on specific values (source of data, algorithm used, outcome of prediction etc.). Finally, we present our findings and conduct statistical analysis on our dataset and critique the outcome of the attempted prediction reported by the reviewed papers. Our research indicates that results are ambiguous, accuracy, and prediction seems dependable

Social media predictions and traditional forecasting

Moreover…

- So many things can be predicted with social media, but when and what should be predicted and with what goal in mind?
- What are the implications of these predictions in the world and on people?
- When not to predict.

Private Traits and Attributes are Predictable from Digital Records of Human Behavior

Your Friends` Personalities

Most Like Me

Your Personality Soulmate

Sofie Jansson

Similarity Score: **85.77%**
(How was this calculated?)

Trait	0	50	100	% (diff.)
O				50% (-)
C				56% (-25%)
E				88% (+13%)
A				56% (-7%)
N				69% (+13%)

Least Like Me

Maybe Opposites Attract?

Damon Alexander Young

Similarity Score: **75.87%**
(How was this calculated?)

Trait	0	50	100	% (diff.)
O				94% (+44%)
C				56% (-25%)
E				88% (+13%)
A				69% (+6%)
N				44% (-12%)

Friend`s Name	Personality					Similarity Score	View Comparison Graph	View Full Personality Profile	Friend Rating Status
	O	C	E	A	N				
You	50%	81%	75%	63%	56%				
Sofie Jansson	50%	56%	88%	56%	69%	86%			
Sara Lee	88%	63%	63%	69%	56%	80%			
Damon Alexander Young	94%	56%	88%	69%	44%	76%			

The Plus-es First…

- Largest FB study outside of FB
- Provocative findings
  - It is indeed amazing so many (sensitive) things can be predicted
  - Validity of Facebook data as reflective of people's underlying traits and behaviors
  - Possibility of educating people how simple metrics like "likes" can be profusely revealing of deeply personal information
- Data collected with consent (?)

- But at what cost?

Who benefits from this research?

11 March 2013 Last updated at 15:00 ET

Facebook 'likes' predict personality

Sexuality, political leanings and even intelligence can be gleaned from the things you choose to "like" on Facebook, a study suggests.

What do your Facebook likes say about you?

Researchers at Cambridge University used algorithms to predict religion, politics, race and sexual orientation.

The research, **published in the journal PNAS**, forms surprisingly accurate personal portraits, researchers said.

The findings should "ring alarm bells" for users, privacy campaigners said.

The study used 58,000 volunteers who alongside their Facebook "likes" and demographic information also provided psychometric testing results - designed to highlight personality traits.

The Facebook likes were fed into algorithms and matched with the

Related Stories

Facebook changes privacy settings

Privacy groups lobby Facebook

Top stories

Soldier killed in Canada shootings

WHO crisis team holds Ebola talks

Sweden could use force in sub hunt

U-boat found with ship it sank

Total plane crash driver 'got lost'

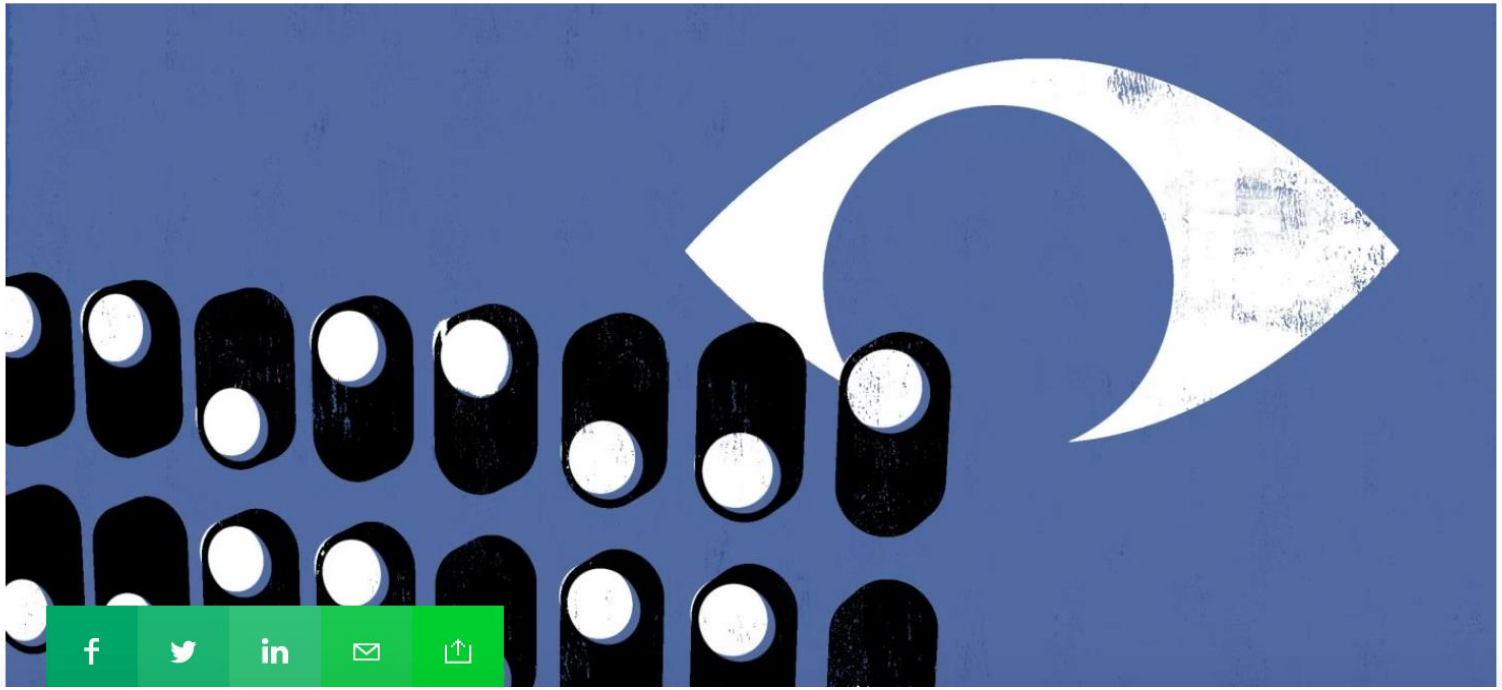
"This research should ring alarm bells for anyone who thinks that privacy settings are the solution to protecting information online. We need to fundamentally re-think how much data we are voluntarily sharing," said Nick Pickles, director of privacy campaign group Big Brother Watch.

"Yet again, it is clear the lack of transparency about how users' data is being used will lead to entirely justified fears about our data being exploited for commercial gain."

Facebook bans first app since Cambridge Analytica, myPersonality, and suspends hundreds more

Devin Coldewey @techcrunch / 9:14 pm EDT • August 22, 2018

Comment



sociological science

Gaydar and the Fallacy of Decontextualized Measurement

Andrew Gelman,[a] Greggor Mattson,[b] Daniel Simpson[c]

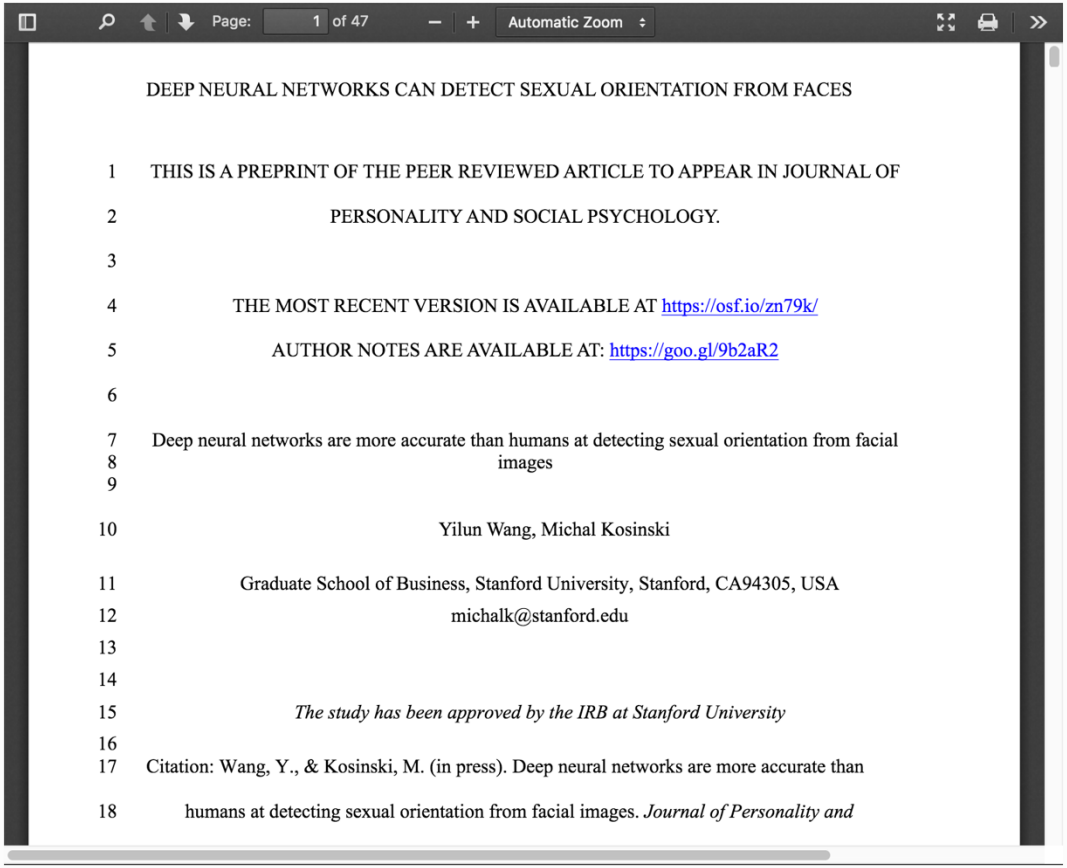
a) Columbia University; b) Oberlin College; c) University of Toronto

Abstract: Recent media coverage of studies about "gaydar," the supposed ability to detect another's sexual orientation through visual cues, reveal problems in which the ideals of scientific precision strip the context from intrinsically social phenomena. This fallacy of objective measurement, as w term it, leads to nonsensical claims based on the predictive accuracy of statistical significance. We interrogate these gaydar studies' assumption that there is some sort of pure biological measu of perception of sexual orientation. Instead, we argue that the concept of gaydar inherently exist within a social context and that this should be recognized when studying it. We use this case as an example of a more general concern about illusory precision in the measurement of social phenomena

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.

Yilun Wang, Michal Kosinski

Created on: September 07, 2017 | Last edited: October 16, 2017



DEEP NEURAL NETWORKS CAN DETECT SEXUAL ORIENTATION FROM FACES

THIS IS A PREPRINT OF THE PEER REVIEWED ARTICLE TO APPEAR IN JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY.

THE MOST RECENT VERSION IS AVAILABLE AT https://osf.io/zn79k/

AUTHOR NOTES ARE AVAILABLE AT: https://goo.gl/9b2aR2

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images

Yilun Wang, Michal Kosinski

Graduate School of Business, Stanford University, Stanford, CA94305, USA
michalk@stanford.edu

The study has been approved by the IRB at Stanford University

Citation: Wang, Y., & Kosinski, M. (in press). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and*

wang_kosinski.pdf

Download preprint Version: 10

Download preprint

Downloads: 24829

Abstract

We show that faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain. We used deep neural networks to extract features from 35,326 facial images. These features were entered into a logistic regression aimed at classifying sexual orientation. Given a single facial image, a classifier ...

See more

Preprint DOI

10.17605/OSF.IO/HV28A

License

Disciplines

Social and Behavioral Sciences | Psychology

Tags

Artificial Intelligence | Big Data | Computational Social Science | Computer Vision | Facial recognition | Prenatal Hormone Theory | Privacy | Sexual orientation

Citations

APA

Wang, Y., & Kosinski, M. (2017, October 16). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.. Retrieved from psyarxiv.com/hv28a

Automatic Crime Prediction using Events Extracted from Twitter Posts

Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown

Department of Systems and Information Engineering, University of Virginia
{xw4u,msg8u,brown}@virginia.edu

Abstract. Prior work on criminal incident prediction has relied primarily on the historical crime record and various geospatial and demographic information sources. Although promising, these models do not take into account the rich and rapidly expanding social media context that surrounds incidents of interest. This paper presents a preliminary investigation of Twitter-based criminal incident prediction. Our approach is based on the automatic semantic analysis and understanding of natural language Twitter posts, combined with dimensionality reduction via latent Dirichlet allocation and prediction via linear modeling. We tested our model on the task of predicting future hit-and-run crimes. Evaluation results indicate that the model comfortably outperforms a baseline model that predicts hit-and-run incidents uniformly across all days.

1 Introduction

Traditional crime prediction systems (e.g., the one described by Wang and Brown [14]) make extensive use of historical incident patterns as well as layers of in-

Class Discussion

How do we decide what should and shouldn't be predicted?

Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development

Why Study Dataset Politics?

- Datasets shape how machine learning systems see and act on the world.
- Critical past work has exposed biases—racial, gendered, geographic—in datasets.
- Prior critiques often focused on outcomes (e.g., bias in facial recognition).
- This paper shifts focus: **What values guide dataset creation itself?**

Method

- **Corpus**: 113 datasets sampled from ~500 collected; spanning face, body, object recognition, etc.
- **Data Sources**: Papers, websites, documentation, code repositories.
- **Analysis Methods**:
  - Structured content analysis: What is documented?
  - Thematic analysis: How is it documented?
- **Goal**: Reveal explicit and *silenced* values in dataset practices.

Key Disciplinary Values in Dataset Development

Four dominant values in how datasets are described:

1. **Efficiency over Care**
   - Speed, cost, and scale prioritized over thoughtful, ethical collection.
2. **Universality over Contextuality**
   - Data often stripped of cultural, geographic, or social specificity.
3. **Impartiality over Positionality**
   - Researchers avoid acknowledging their own perspectives or biases.
4. **Model Work over Data Work**
   - Papers focus more on algorithms than on data design processes.

What's Missing in Dataset Culture?

- **Human annotators**: Rarely acknowledged or compensated transparently.
- **Data subjects**: Often dehumanized or treated as technical inputs.
- **Documentation**: Few standards for describing provenance, consent, or fairness.
- **Data stability**: Most datasets hosted informally (lab websites, dead links).

"Datasets are treated as natural reflections of the world, not constructed artifacts."

Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media

STEVIE CHANCELLOR, Georgia Tech, USA
ERIC P.S BAUMER, Lehigh University, USA
MUNMUN DE CHOUDHURY, Georgia Tech, USA

"Human-centered machine learning" (HCML) combines human insights and domain expertise with data-driven predictions to answer societal questions. This area's inherent interdisciplinarity causes tensions in the obligations researchers have to the humans whose data they use. This paper studies how scientific papers represent human research subjects in HCML. Using mental health status prediction on social media as a case study, we conduct thematic discourse analysis on 55 papers to examine these representations. We identify five discourses that weave a complex narrative of who the human subject is in this research: Disorder/Patient, Social Media, Scientific, Data/Machine Learning, and Person. We show how these five discourses create paradoxical subject and object representations of the human, which may inadvertently risk dehumanization. We also discuss the tensions and impacts of interdisciplinary research; the risks of this work to scientific rigor, online communities, and mental health; and guidelines for stronger HCML research in this nascent area.

Additional Key Words and Phrases: human-centered machine learning; machine learning; social media; research ethics; mental health

ACM Reference Format:
Stevie Chancellor, Eric P.S Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 147 (November 2019), 32 pages. https://doi.org/10.1145/3359249

1 INTRODUCTION

> The two hardest things in Computer Science are: People, and convincing others that "People" is the hardest thing in Computer Science. – attributed to Brad Grzesiak[1]

"Human-centered machine learning" (HCML)[2] is a rising subfield of computer science (CS) that combines the expertise of data-driven predictions and outside domain knowledge to make headway on questions of societal importance. These approaches have become popular in predicting elections [171], understanding criminal justice [175], and detecting fake news [163]; in HCI and CSCW, HCML has examined questions such as abusive content detection [40] and crisis [166].

HCML is focused on impacts to individuals, communities, and society, made explicit by its contributions to human-centered domains and challenges and self-stated goals within papers [25].

[1] https://twitter.com/listrophy/status/876129823130869760
[2] This emergent field has many names, including but not limited to, human-centered machine learning, human-centered AI, and data science for social good. For simplicity's sake, we use HCML as the umbrella term throughout this paper.

Authors' addresses: Stevie Chancellor, Georgia Tech, Atlanta, GA, USA, schancellor3@gatech.edu; Eric P.S Baumer, Lehigh University, Bethlehem, PA, USA, ericpsb@lehigh.edu; Munmun De Choudhury, Georgia Tech, Atlanta, GA, USA, munmund@gatech.edu.

2573-0142/2019/11-ART147 $15.00
https://doi.org/10.1145/3359249

Discourse	*Terms (number of documents/papers)*
Disorder/Patient	patient (17), depression (10), depressed (9), sufferer (9), behavior (7), condition (4), distressed (4), PTSD (4), neurotypicals (3), non-depressed (3), suicide (3), normal (3), victim (3), clinical (2), anxiety (2), bipolar (2), mentally ill (2), non-stressed (2), pro-anorexic (2), stressed (2), suicidal ideation (2), score (2), standard (2), state (2)
Social Media	user (55), post (25), tweets (16), content (15), account (14), author (14), community (10), microblog (7), text (7), document (6), member (6), activity (4), followers (4), message (3), poster (3), tweeter (3), corpus (3), blog (2), item (2), networks (2), publisher (2), profiles (2), lexicon (2)
Scientific	population (29), control (21), participant (16), subject (10), cohort (8), candidate (6), respondents (6), observation (2), pool (2)
Data/Machine Learning	data (31), sample (25), dataset (18), class (16), example (8), subset (8), test set (5), category (4), positive/negative (3), task (3), data point (2), model (2), prediction (2)
Person	people/person (47), individual (40), she/he (11), woman (7), one's (5), man (5), youth (5), student (5), mother (4), worker (4), crowdworker (4), female (3), someone (3), peers (3), friends (2), others (2), they (2), adolescents (2)

High-level discursive categories from the analysis ordered in decreasing order by appearance in unique documents. Words used in only one paper are excluded.

Humans as a disorder/patient

"We perform an empirical study...of potentially depressed users against a differential control group of normal users."

Humans as data

"distress is an important risk factor in suicide, one that is observable from microblog text"

Humans as a machine learning object

"Positive: the tweet content indicates the presence of one of the studied diseases/states in the person who has written the tweet."

Towards Ethical, Reflexive Dataset Practices

- Embrace reflexivity: Acknowledge the social and political contexts of data.
- Value data work: Elevate the labor of curation, annotation, and documentation.
- Design for care: Prioritize consent, well-being, and contextual accuracy.
- Support infrastructure: Create stable, accountable repositories and standards.

The remainder of the course…

- How are these aspects of data collection or model training getting replicated, mitigated, or exacerbated by genAI?
- How can institutions support better ways forward?
- What responsibilities do researchers have to data subjects?