Introduction to Social Media Analytics (Lec 1)

Hao PENG
Department of Data Science
City University of Hong Kong
https://haoopeng.github.io/

Hao PENG (彭浩)

- Assistant Professor
  - Department of Data Science
  - College of Computing, CityU
- **Postdoc**: Kellogg School of Management, Northwestern U
- **PhD**: Information Science, U of Michigan
- **BS**: Information Management, Sun Yat-sen U
- **Research interests**:

Computational Social Science, Social Networks, Science of Innovation, Social Media, Management Science, Complex Systems, Network Science, LLM, NLP, AI, Gender, Diversity

About this course

- You will learn full pipeline of conducting social media analytics (**SMA**) in business, economics, politics, public health, etc.
  - Data, methods, visualization
  - Domain applications, ethics, etc.
- The main coding environment is Python & Jupyter
- Instructor: PENG Hao
  - 16F-274, AC3
  - Office hour: Email or by appointment
- TA: SHI Menghan
  - Email: menghashi4-c@my.cityu.edu.hk

Key topics

- Weeks 1-4: methodological foundations of SMA
  - Data foundations, machine learning (supervised, unsupervised)
- Weeks 5-6: text analytics
  - Sentiment analysis, topic modeling, misinformation, etc.
- Week 7: mid-term
- Weeks 8-9: modeling user behavior
  - Demographics, psychological traits, health, engagement, etc.
- Weeks 10-11: social network analysis
  - Strong & weak tie, small world, homophily, influence
  - Community detection, info diffusion, viral marketing
- Weeks 12-13: visualization, ethics
  - NetworkX, D3, Matplotlib
  - Limitations of SMA

Grading

- Take-home assignments: 30%
  - 3 homework (individual work) in total
  - Due two weeks after release
- Paper-presentation “assignment”: 10%
  - Each team (4-5 students) presents one paper
  - Paper (week) selection is first come first serve
- Midterm test: 20%
  - In class, week 7
  - Close-book written test
  - CheatSheet allowed (3 A4 size papers)
- Final exam: 40%
  - Close-book written test
  - Coding task is minimal
  - Bring CheatSheet (5 A4 papers)

Paper presentation (teamwork)

- Form/join your team on Canvas – People – Week:
  - Pick only one paper from your selected week's papers
  - Email me your choice one week before your presentation
  - Submit presentation slides under Canvas - "Assignments"
- The pres. (~30 mins) should contain five sections:
  - study context & motivation
  - research questions
  - data & methods
  - key findings & takeaways
  - implications & limitations

Prepare at least 3 interesting questions to spark discussion among your audiences.
All students need to read the paper before class & actively participate in discussion.

Late Policy

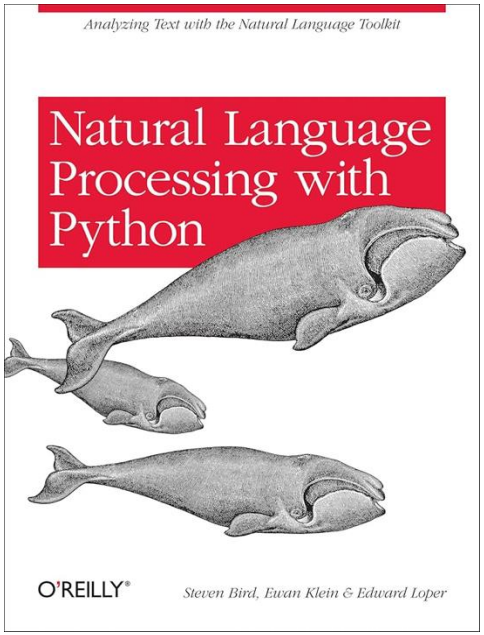
- Late submission is not graded
  - **1 min late is still late**
  - You can still submit after due date (but no grading)
- 1 no-penalty late submission allowed
  - No excuse needed
  - Acceptable up to 3 days after due date
  - Can only use it for 1 assignment
  - Use it wisely (for sick, family issues, etc.)

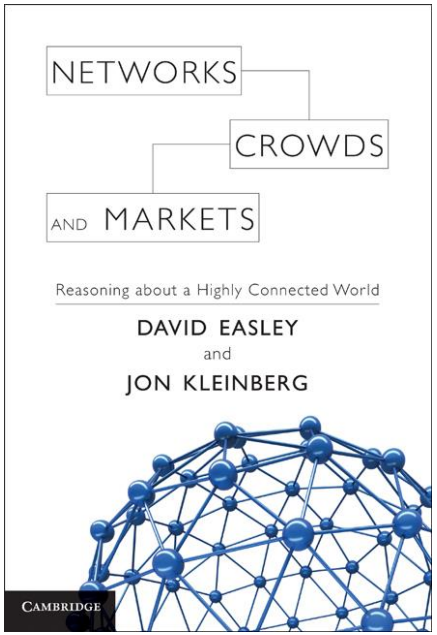
Original work policy

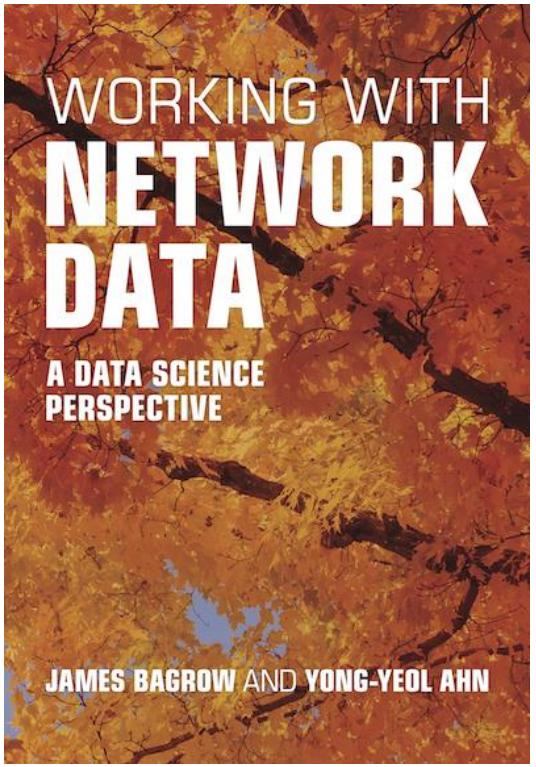
- Unless otherwise specified in an assignment, all submitted work must be your own, original work.
- You may discuss approaches with others on individual assignments, but you may not copy code or other work wholesale.
- Any excerpt or reference from the work of others must be clearly identified as a quotation, and a proper citation provided.
- Any violation of the Department's policy on Academic Integrity will result in severe penalties, which might range from:
  - failing an assignment
  - failing the course
  - being expelled from the program

Suggested books

- [Natural Language Processing with Python]
- [Networks, Crowds, and Markets]
- [Working with Network Data]







Questions about class logistics?

Agenda for today

- What, Why, Where, and How of SMA
- Overview of social media platforms
- Applications of SMA in various domains
- Challenges and opportunities of social media
- Emerging technologies in SMA
- Setup Python coding environment

What is Social Media Analytics (SMA)?

- Definition: SMA is the practice of collecting, analyzing, and interpreting data from social media platforms to extract insights on user behavior, trends, sentiment, engagement, and performance, often to support the decision makings in business, marketing, policy, and research.
- Importance: SMA is crucial for businesses, marketers, and researchers as it transforms raw social data into actionable insights, driving strategic decisions and growth.

Primer

- Social media has fundamentally transformed our daily lives and the very fabric of modern society, revolutionizing how information is created, shared, and consumed.
- Today, billions of individuals gravitate towards platforms like X, Instagram, TikTok, YouTube, engaging in a digital footprint of scrolling, posting, sharing, and liking.
- This digital revolution stands as one of the most significant technological advancements in the modern history, weaving itself so deeply into our human existence.

The evolution of the Internet

Year	Event
1950s	After WWII, the Cold War spawns the need for new ways to communicate more quickly—hence, Arpanet originated for military uses.
1960s	Arpanet became available to researchers around the world.
1970s	Apple and IBM developed personal computers in the late 1970's, which appeal to the masses who can run simple word processing and spreadsheet applications on them.
1980s	TCP/IP was developed as a language for computers to communicate over the Internet. Web 1.0, or the initial version of the World Wide Web, developed and was entirely made up of static, non-interactive Web pages connected by hyperlinks.
1990s	The creation of Web 1.0 led to new Internet-based business models.

What are the differences between Internet and WWW?

The evolution of WWW

Feature	Web 1.0	Web 2.0	Web 3.0	Web 4.0
Definition	The "read-only" web.	The "participatory" or "social" web.	The "semantic" / "data" web.	The "symbiotic" web.
Content Creation	Static content created by webmasters.	Dynamic content and user-generated content.	Content linked across platforms, understandable by machines.	AI-generated content, highly personalized and anticipatory.
User Interaction	Limited to reading.	Interactive, with users contributing content and engaging with each other.	Users and machines interact seamlessly, with intelligent search and data integration.	Deeply immersive experiences with AR, VR, and AI enabling natural interactions.
Technologies	HTML, CSS.	AJAX, RSS, Web APIs.	RDF, OWL, SPARQL, Semantic Web technologies.	AI, IoT, blockchain, advanced encryption.
Connectivity	Low. Web pages are mostly isolated.	High. Social media and wikis link people and content.	Very high. Data from multiple sources is connected.	Ubiquitous. Everything is connected, often in real-time.
Data Flow	One-way.	Two-way, with feedback loops.	Multi-directional, context-aware.	Autonomous, predictive.
Privacy and Security	Basic.	Increased focus due to social nature.	Significant concern, with emphasis on standards and protocols.	Critical, with advanced solutions required for safety and ethics.

Understanding social media

- Social media is built on the *Web 2.0 philosophy*, that is, to give more control to the user over the content. It can be defined as an easy-to-use Internet-based platform that provides users with opportunities to create and exchange content in a one-to-one, one-to-many, and many-to-many context.
- Key features of Web 2.0
  - Interactivity and Participation
  - Everyone can be a producer
  - Collaboration & Crowdsourcing
  - Rich User Experiences

History of social media

Year(s)	Description
1990s:	**The Internet Goes Public**
1991	The World Wide Web goes live to the public, transforming the internet from a scientific and academic tool into a comprehensive network accessible to the masses.
1994	GeoCities, a web hosting service, allows users to create their own websites, grouping them into "neighborhoods" based on content themes, resembling early social networking.
1997	SixDegrees.com launches as one of the first recognizable social networking sites, enabling users to create profiles and friend other users.
2000s:	**Social Networking Expansion**
2002	Friendster is launched, aimed at creating a safer, more realistic online community environment, influencing future social networking sites.
2003	LinkedIn goes live, offering a networking solution for business professionals, while MySpace quickly becomes the most popular social network in the U.S.
2004	Facebook launches, initially restricted to Harvard students before expanding to other universities and eventually the public in 2006.
2005	YouTube is created, revolutionizing online video sharing and consumption.
2006	X launches, introducing a microblogging platform that limits posts to 140 characters, later expanded to 280.
2007	The introduction of iPhone marks a significant shift toward mobile internet use, greatly influencing social media accessibility and design.
2010s:	**The Mobile and Multimedia Era**
2010	Instagram launches, focusing on mobile photo and later video sharing, quickly becoming one of the most popular social platforms.
2011	Snapchat is released, introducing ephemeral content and multimedia messaging, popularizing the concept of stories.
2012	Facebook acquires Instagram, further consolidating its presence in the social media sphere.
2013-2014	The rise of messaging apps like WhatsApp and Telegram, emphasizing privacy and encryption, diversifies the social media landscape.
2016	TikTok launches internationally, bringing short-form video content to the forefront of social media trends.
2020s:	**Evolution and New Horizons**
2020-2021	The COVID-19 pandemic leads to unprecedented growth in social media use, as individuals and businesses rely more heavily on digital platforms for communication, entertainment, and commerce.
2022-2023	Augmented reality (AR) and virtual reality (VR) integrations in social media platforms enhance user engagement, paving the way for more immersive social experiences.

Characteristics of social media

- User-Generated Content (UGC)
- Connectivity and Interactivity
- Personalization & Privacy
- Real-Time Communication
- Multimedia Content
- Accessibility

Major social media platforms

- Top platforms by user base in 2025
  - Facebook, YouTube, Instagram, WeChat, TikTok, LinkedIn, X, Reddit
- Each has > 500 million monthly active users (MAU)
- Owned mostly by Tech companies
  - Meta, Google, ByteDance, Microsoft, Tencent

Types of social media platforms

- Social networking sites
  - Facebook, Twitter (X), Weibo, WeChat
- Media sharing (blogging)
  - YouTube, Instagram, TikTok, RedNote
- Content communities
  - Discussion forum (Reddit, Douban, Flickr)
  - Customer review (Yelp, DianPing, TripAdvisor)
  - Community Q&A (Quora, Stack Overflow, Zhihu)
- Online collaborative platforms
  - Wikipedia, Github, Overleaf
- Purpose-build platforms
  - Dating apps, LinkedIn, Ebay, etc.

Social networking sites

- Network is the key
- Node: user
  - Multiple types of users, such org. vs individual, verified vs non-verified
  - Attributes of users, such as gender, location, age, occupation, interests, joined date, some statistics/metrics, profile photo
  - Any others?

Social networking sites

- Edge: relations between users
  - Follower
  - Mention
  - Reply
  - Repost
  - Like
  - Comment
- Attributes of edges?
  - Frequency, Time, Strength, Direction
- Publicly hosted Twitter datasets:
  - https://github.com/shaypal5/awesome-twitter-data

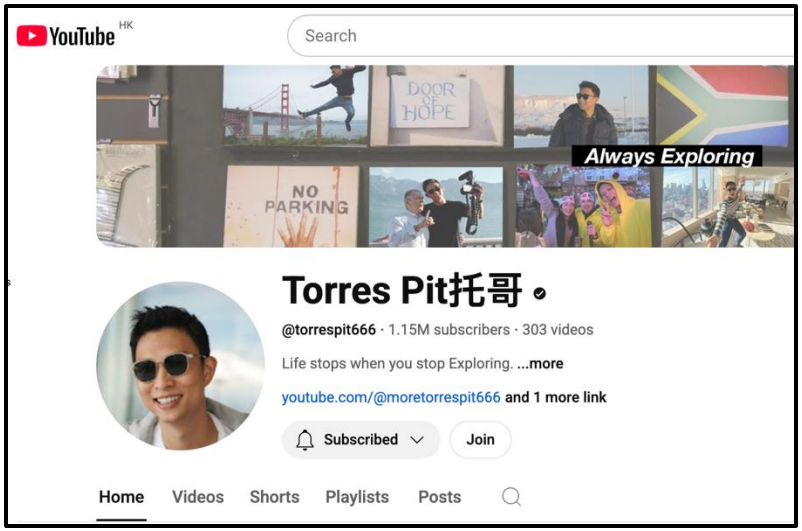
Social networking sites

- What's unique about it?
  - Full of UGC, real-time information flow
  - Algorithmic feeds, instant messaging, and multimedia sharing
  - Complex user interactions and dynamics
- Platform usage trends 2025
  - Short-form video dominates (e.g., RedNote)
  - E-commerce integration growing (e.g., FB marketplace)
- User demographics
  - Millennials: Facebook, LinkedIn
  - Gen Z: TikTok, Instagram

Media sharing sites

- Media is more important than social interactions
- Mostly for business and marketing purposes
- Most users are un-verified / anonymized

TikTok



Paper of next week

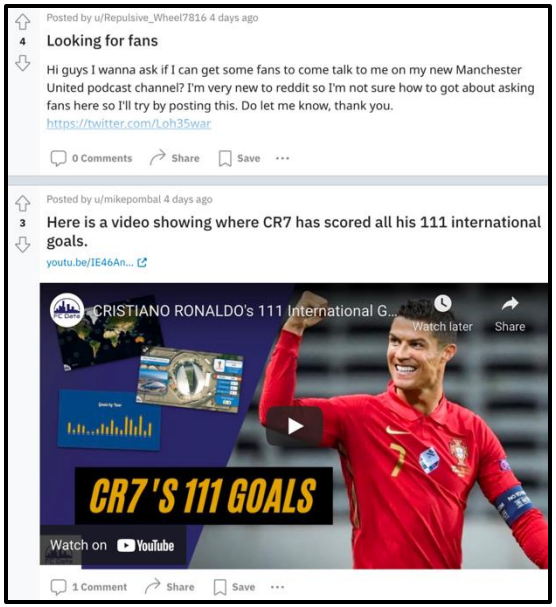
WWW 2010 • Full Paper April 26-30 • Raleigh • NC • USA

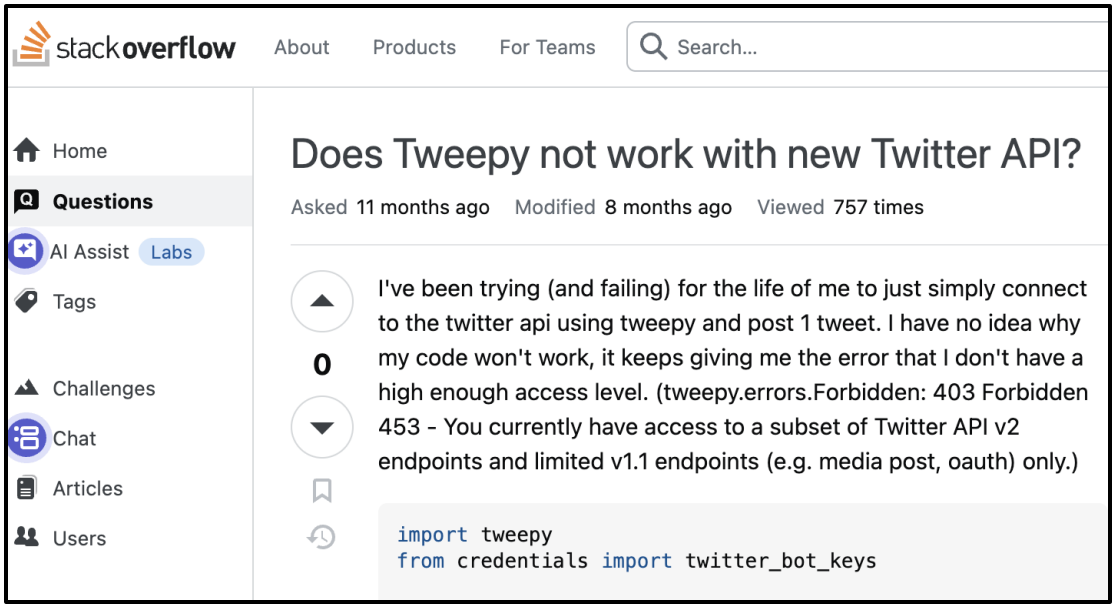
What is Twitter, a Social Network or a News Media?

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon
Department of Computer Science, KAIST
335 Gwahangno, Yuseong-gu, Daejeon, Korea
{haewoon, chlee, hosung}@an.kaist.ac.kr, sbmoon@kaist.edu

Content communities

- They are communities organized based on common interest
  - Discussion forum (organized by topics)
  - Online review sites (restaurants, hotels, travel, product)
  - Q&A sites (up vote answers, share knowledge/skills)
- Far fewer interactions, but very useful resource-wise

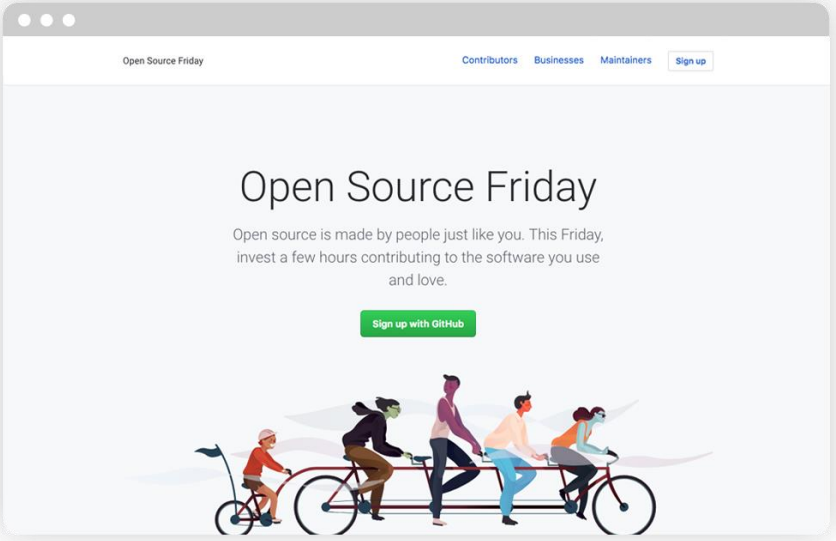




Online collaborative platforms

- Crowdsourcing, crowd wisdom, team collaboration
- Many are used for non-commercial purposes

WIKIPEDIA
The Free Encyclopedia



Why should we study SMA?

- Business insights: Companies leverage social media analytics to understand consumer preferences, optimize marketing campaigns, and enhance brand engagement.
- Advancing research: In computational social science, social media data enables researchers to study topics like polarization, cultural trends, or network dynamics at scale, often in real time, complementing traditional methods like surveys.
- Informing public policy: Governments and non-profit organizations use social media data and analytics to gauge public sentiment, monitor misinformation, predict social unrest, and mobilize political protests.
- Personal development: learn technical skills (e.g., NLP tools, network analysis, machine learning) and critical thinking, which are transferable across careers in data science including academia and industry positions.

Where can we apply SMA to inform decision-making and drive economic value?

SMA in business and economics

- Consumer behavior and spending prediction
  - Box office revenue, retail sales, market share, etc.
- Stock market prediction
  - Price movement, trading volumes, market index, etc.
- Macroeconomic indicators
  - Unemployment rates, GDP growth, consumer confidence, etc.
- Socioeconomic status and income prediction
  - Education, age, gender, race, marital status, income level, etc.
- More…

Housing market analysis

What Makes a Good Image? Airbnb Demand Analytics Leveraging Interpretable Image Features

Shunyuan Zhang,[a] Dokyun Lee,[b] Param Vir Singh,[c] Kannan Srinivasan[c]

[a] Harvard Business School, Harvard University, Cambridge, Massachusetts 02163; [b] Questrom School of Business, Boston University, Boston, Massachusetts 02215; [c] Tepper School University, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Contact: szhang@hbs.edu, https://orcid.org/0000-0002-9044-1621 (SZ); dokyun@bu.edu, https://orcid.org/0000-0002-3186-3349 (DL); psidhu@cmu.edu, https://orcid.org/0000-0002-0211-7849 (PVS); kannans@cmu.edu, https://orcid.org/0000-0001-6449-9750 (KS)

How can we leverage better images to increase Airbnb bookings?

- Dataset: images of 7K properties listed in 16 months
- Key findings:
  - verified images boost occupancy by 9% (images taken by host)
  - identified 12 attributes that are common signature of verified images

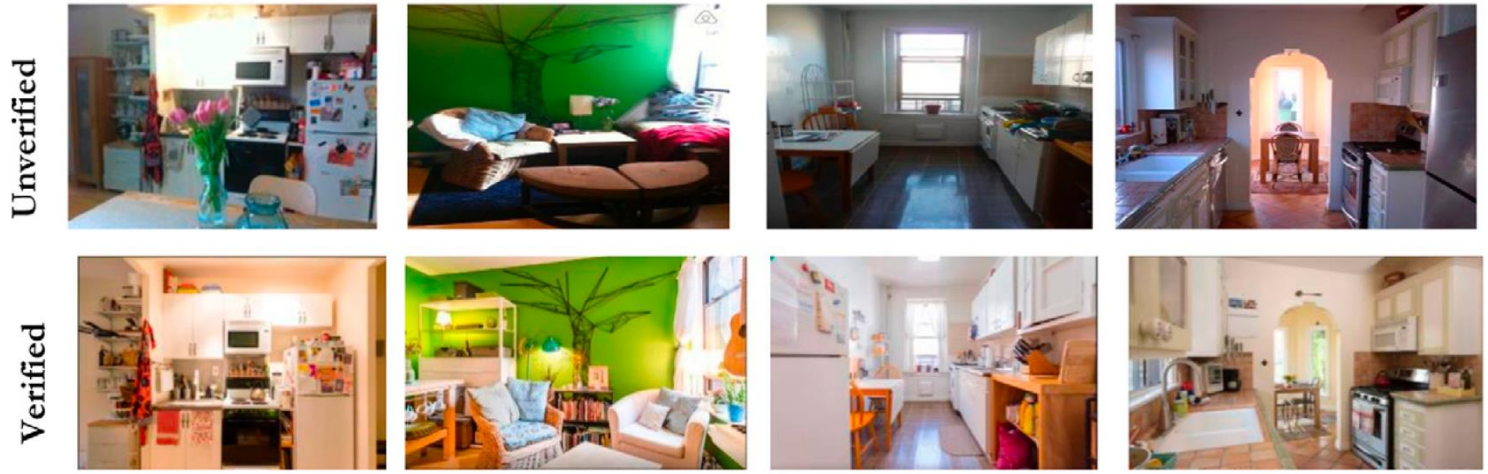
Housing market analysis

Table 6. The 12 Image Attributes and Their Descriptions

Component		Attribute	Description
Composition	1	Diagonal dominance	Alignment of the key objects with the diagonals
	2	Visual balance-intensity	Symmetry of key objects around the imaginary vertical central line
	3	Visual balance-color	Symmetry of colors around the imaginary vertical central line
	4	Rule of thirds	Alignment of the key objects with the intersections of two imaginary horizontal and two vertical lines
Color	5	Warm hue	Dominance of warm colors (e.g., red and yellow) over cold colors (e.g., blue and purple)
	6	Saturation	The richness/vividness of image colors
	7	Brightness	Overall level of illumination
	8	Contrast of brightness	The variance in the illumination flow across the image
	9	Image clarity	Whether image colors have sufficient intensity
Figure-ground relationship	10	Size difference	Difference in area between the figure and ground
	11	Color difference	Difference in color between the figure and ground
	12	Texture difference	Difference in texture between the figure and ground

Housing market analysis

Figure 1. (Color online) Comparison of Unverified and Verified Photos



Key difference between verified and unverified images:

- Composition features
  - diagonal dominance, rule of thirds, visual balance
- Color features
  - warm hue, high saturation, high contrast, high clarity
- Subject-background relationship
  - more area size, color, and texture difference increase figure salience

Other business applications

Temporal Orientation of Tweets for Predicting Income of Users

Mohammed Hasanuzzaman[1], Sabyasachi Kamila[2], Mandeep Kaur[2], Sriparna Saha[2] and Asif Ekbal[2]

The Effects of Twitter Sentiment on Stock Price Returns

Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, Miha Grčar, Igor Mozetič

Published: September 21, 2015 • https://doi.org/10.1371/journal.pone.0138441

- Twitter users' "future" words usage predicts income level
- Sentiment expressed in tweets correlates with abnormal stock returns during the peaks of Twitter volume.

SMA in politics / policy

- Political campaign
  - Voter sentiment analysis
  - Predict offline activities (protest, shooting, etc.)
- US presidential election
  - Predict supporting rate via social data
  - Estimate temporal and geographical trends
- Observatory of online social movement (OSM)
  - #MeToo
  - #BLM
  - #HongKongProtest
  - #AbortionRight

BLM OSM predicts offline events

Social Media Participation in an Activist Movement for Racial Equality

Munmun De Choudhury,[†] Shagun Jhaver,[†] Benjamin Sugar,[†] Ingmar Weber,[§]
[†] Georgia Institute of Technology, [§] Qatar Computing Research Institute, HBKU
{munmund, jhaver.shagun, bsugar}@gatech.edu, iweber@qf.org.qa

Event-Driven Analysis of Crowd Dynamics in the *Black Lives Matter* Online Social Movement

Hao Peng
University of Michigan
haopeng@umich.edu

Ceren Budak
University of Michigan
cbudak@umich.edu

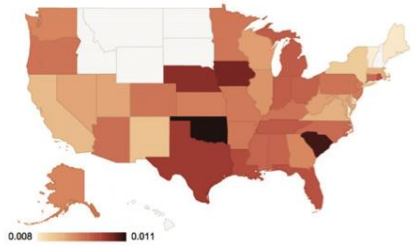
Daniel M. Romero
University of Michigan
drom@umich.edu

BLM OSM predicts offline events

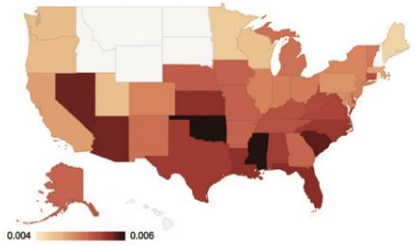
- BLM: a social movement of police violence against Blacks
- Research Qs:
  - Are there geographical variations in social media language used?
  - Can engagement and linguistic attributes predict offline protests?
- Online and offline datasets:
  - Collected ~30M tweets by 6M users with #blm related hashtags
  - Police shooting data: https://fatalencounters.org/our-visualizations/
  - Protest data: https://elephrame.com/charts/BLM
- Methodology:
  - Use NLP tools (LIWC) to calculate linguistic features of online Tweets
  - Leverage regression models to estimate correlations with offline events

BLM OSM analysis

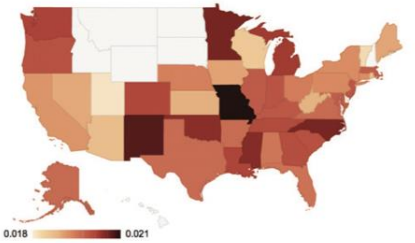
- High PK states are associated with:
  - High negative reactions
  - Low social orientation
  - High psychological distance



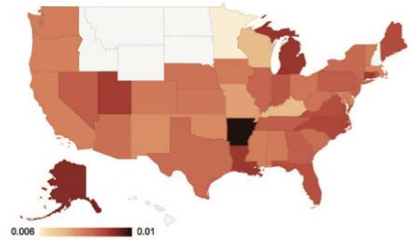
(a) "death"



(b) 1st person singular



(c) NA



(d) PA

	β	[95% conf. interval]		p
Activity and engagement				
# @-replies	1.6386	0.939	4.216	*
Affective attributes				
PA	-20.278	-37.63	-11.07	**
NA	28.54	16.306	33.39	***
anger	9.757	7.5289	12.043	***
anxiety	14.071	10.369	21.512	***
sadness	13.268	7.84	25.376	***
swear	12.101	9.4505	14.752	***
Cognitive attributes				
cognitive mech	-1.1461	-5.981	-0.688	*
negation	-7.1799	-15.847	-3.4876	**
Perception attributes				
see	10.489	3.722	22.699	***
hear	10.27	7.52	18.06	***
feel	4.2771	2.7379	14.292	**
death	40.036	22.271	52.34	***
Social orientation				
social	-7.4167	-11.643	-5.81	**
family	-5.5384	-16.059	-1.982	**
friends	-4.8271	-10.897	-1.2425	***
Interpersonal awareness				
1st p. singular	39.213	18.93	58.505	***
1st p. plural	-12.3689	-16.225	-5.4877	***
2nd p.	4.6663	1.227	14.559	**
Psychological distancing				
Temporal references				
past tense	4.4593	1.24	9.16	**
present tense	3.8796	1.368	14.127	*
Function words				
article	-2.7275	-9.641	-0.1859	**
adverbs	-4.2942	-14.168	-1.5798	**
conjunction	-3.1092	-7.832	-0.6138	**
pseudo R^2 = .618, LL = 141.7; LR χ^2= 22.86, $p < -10^4$				

Table 4: Summary of a Poisson regression model with PK in states as dependent variable. Significance is estimated following Bonferroni correction: $^*\alpha = .05/34; ^{**}\alpha = .01/34; ^{***}\alpha = .001/34$.

BLM OSM analysis

- High PV states are associated with:
  - High level of engagement
  - High collective identities
  - More futuristic inclination

	RMSE	MAPE	SMAPE (%)	Correct @ $\leq$ 20% (%)
Constant model	9496.5	84.63	34.59	42
Next day MA	8379.2	70.68	26.35	48
Activity, Engagement	7155.4	57.26	20.40	59
Affective Attributes	5775.6	36.05	8.34	74
Cognition, Perception	6100.4	38.04	9.64	71
Social Orientation	6532.5	54.15	15.39	62
Interpers. Awareness	6311.3	42.57	10.43	69
Psychological Dist.	6519.9	49.74	12.56	65
All	5528.1	32.62	6.37	81

Table 6: Performance metrics of predicting daily PV. Here (1) RMSE is root mean squared error; (2) MAPEis median absolute percentage error; (3) SMAPE is symmetric mean absolute percentage error; and (4) Correct @ $\leq$ 20% is the percent of PV estimates within 20% of the true values.

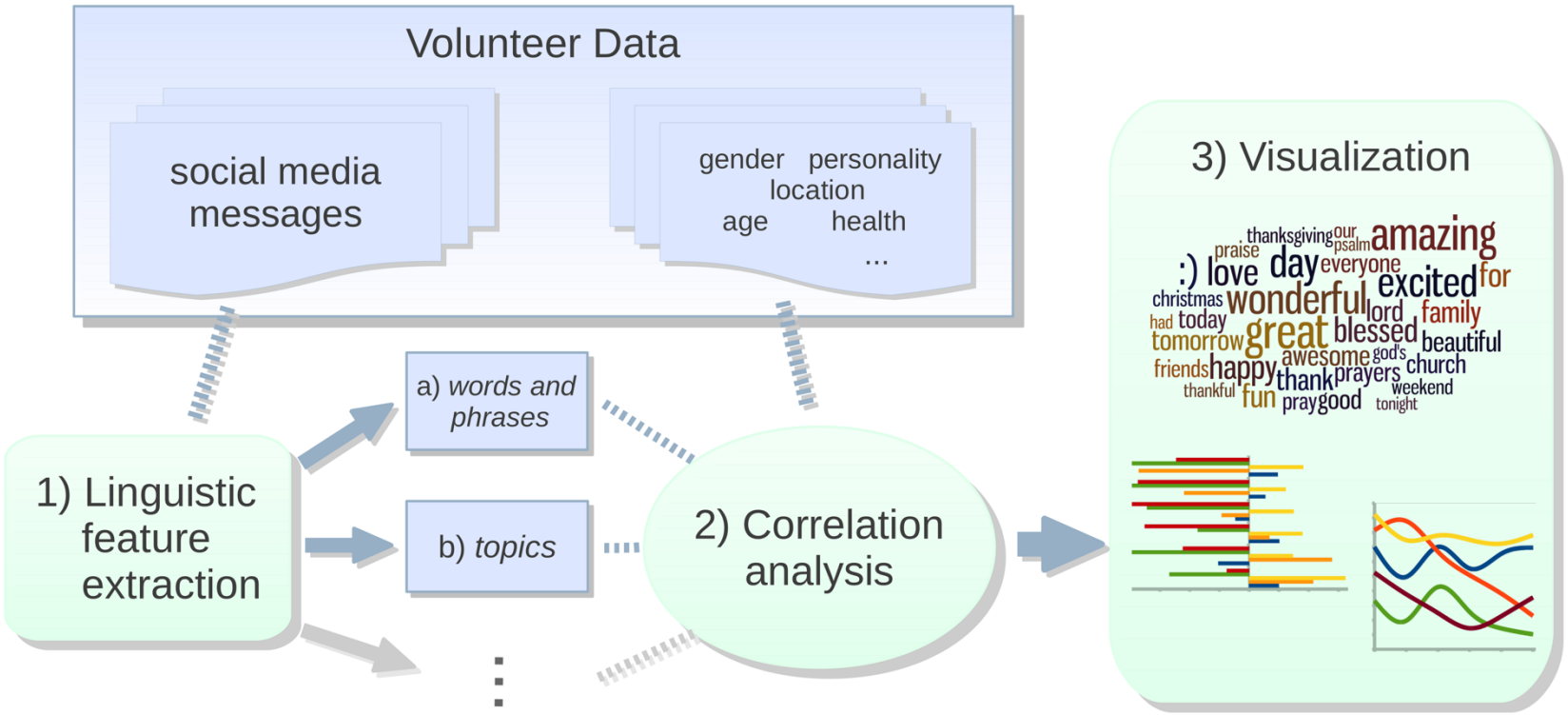
	β	[95% conf. interval]		p
[intercept]	1.453	0.795	2.110	***
Activity and engagement				
# posts	7.010	1.244	15.264	***
# @-replies	6.635	4.258	10.528	***
# retweets	2.727	0.4844	3.138	***
# posts w/ link	1.831	0.2574	2.236	**
Affective attributes				
PA	-0.2985	-1.104	-0.1072	*
NA	21.74	12.624	36.103	***
anger	-0.7518	-0.957	-0.4533	**
anxiety	-3.9124	-6.3191	-1.5057	***
sadness	1.7109	0.8238	3.4021	**
swear	-2.0366	-4.5444	-1.4712	***
Cognitive attributes				
discrepancies	1.0604	0.4118	2.3327	***
negation	2.8294	0.97	4.3113	***
Perception attributes				
hear	0.9437	0.675	2.587	**
feel	1.0153	0.5229	2.0923	**
death	-15.144	-24.726	-5.437	**
Social orientation				
social	0.2811	0.0922	1.485	*
family	1.8024	1.1568	4.961	***
friends	1.9103	1.0357	5.2563	***
Interpersonal awareness				
1st p. singular	-4.3174	-14.178	-1.5428	***
1st p. plural	2.0803	1.7301	5.8907	***
2nd p.	0.8822	0.3311	2.7957	**
Psychological distancing				
Temporal references				
past tense	-1.1051	-2.3229	-0.1127	**
present tense	0.3832	0.1706	1.9382	*
future tense	0.2803	0.0301	2.8627	*
Function words				
article	-0.3403	-1.6725	-0.1924	*
verbs	-0.3344	-1.54	-0.1709	*
pseudo R^2 = .422, LL = 826.1; LR χ^2= 81.24, $p < -10^6$				

Table 5: Summary of negative binomial regression with daily protest volume (PV) as dependent variable. Significance is estimated following Bonferroni correction: $^{*}\alpha = .05/34$; $^{**}\alpha = .01/34$; $^{***}\alpha = .001/34$.

SMA in public health

- Health surveillance and outbreak detection
  - Tracking disease outbreaks via posts (e.g., COVID-19 monitoring)
- Mental health monitoring and intervention
  - Identify depression, suicide, loneliness, etc.
- Health behavior and lifestyle
  - Public opinion of “mask wearing”
  - Track diet, exercise, fitness, smoking, substance use, etc.
- Disaster management
  - Real-time event tracking
  - Hashtag analysis for aid coordination during natural disasters

Predict demographic identities

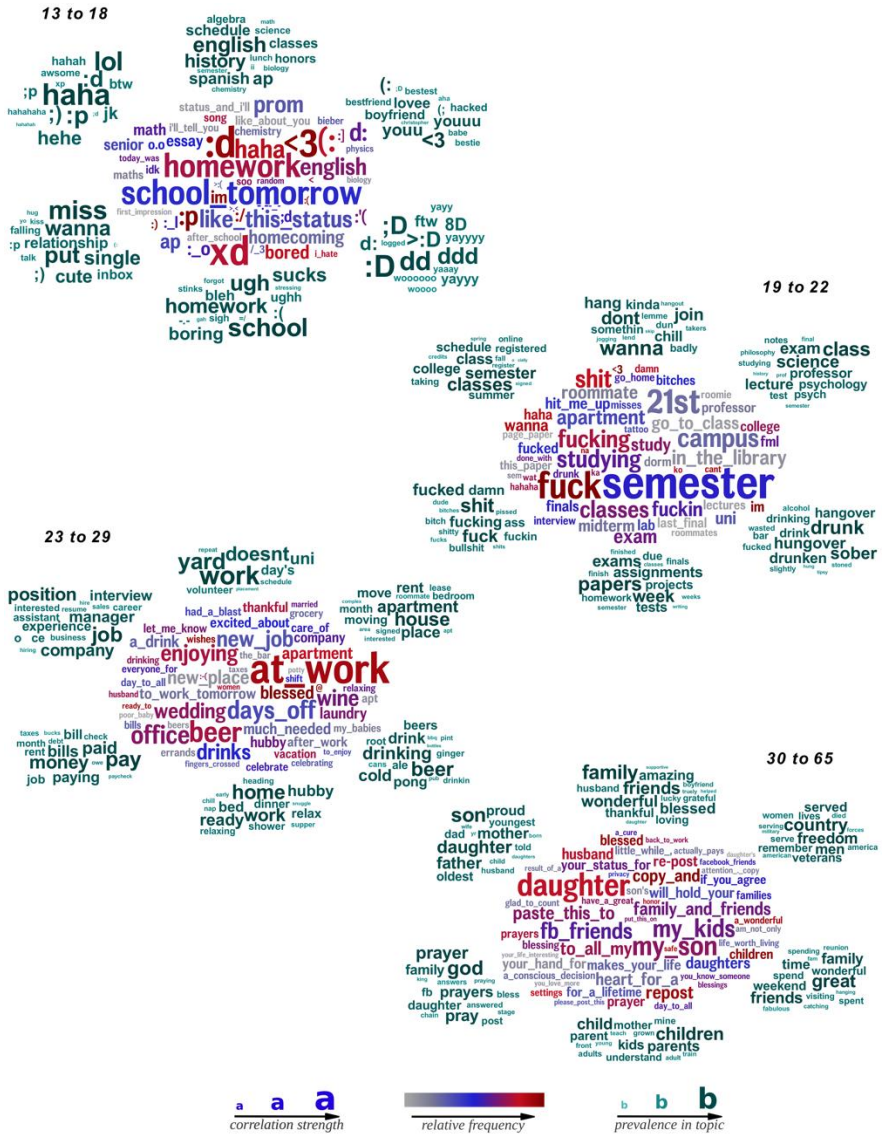


Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*(9), e73791.

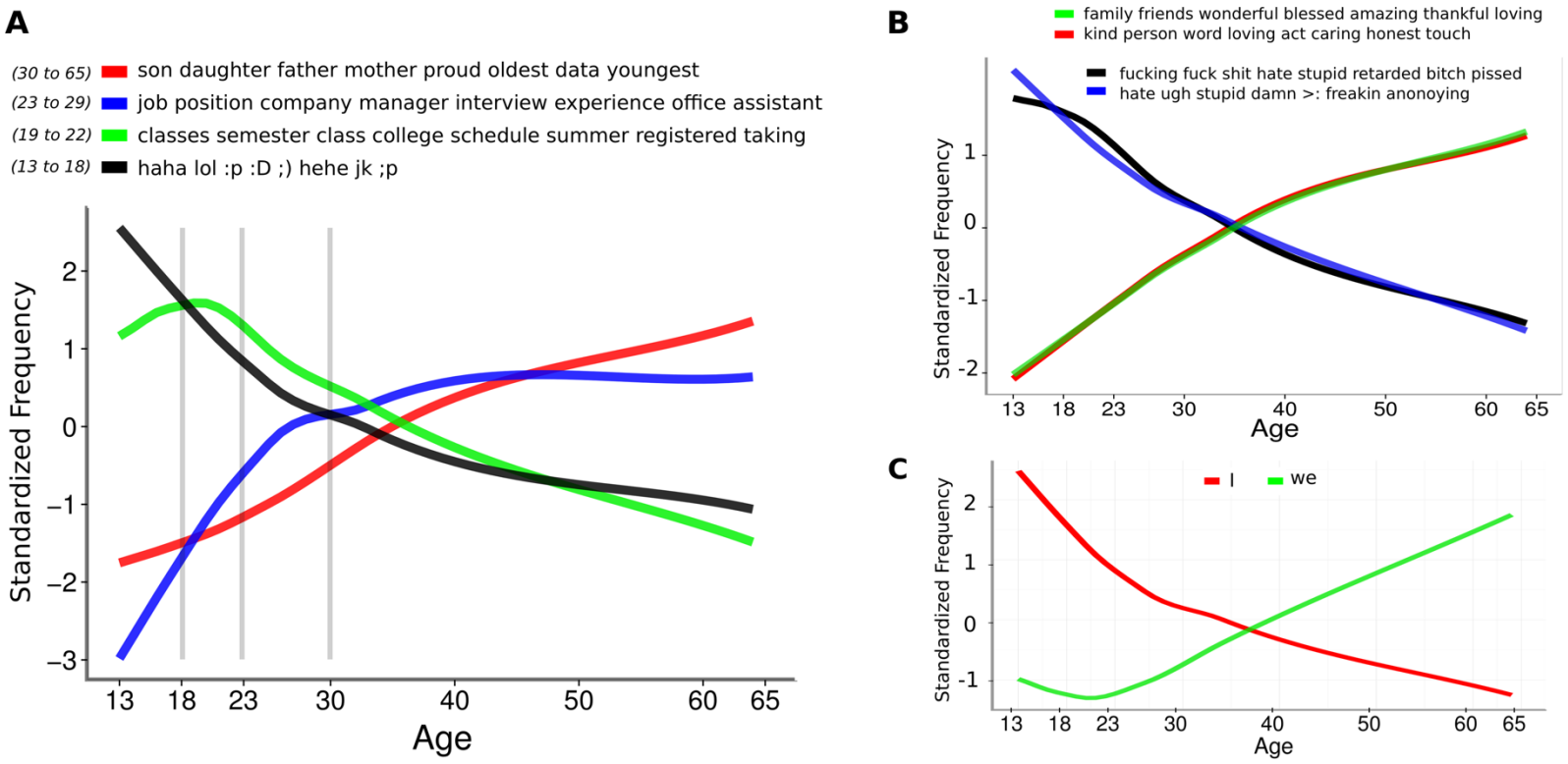
Our key topics change as we grow up

We will learn the LDA topic modeling later in the class.

No worry for now!



We become more positive as we age



Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*(9), e73791.

The power of language in predicting demographic and psychological traits

features	Gender *accuracy*	Age *R*	Extraversion *R*	Agreeableness *R*	Conscientious. *R*	Neuroticism *R*	Openness *R*
LIWC	78.4%	.65	.27	.25	.29	.21	.29
Topics	**87.5%**	**.80**	**.32**	**.29**	**.33**	**.28**	**.38**
WordPhrases	**91.4%**	**.83**	**.37**	**.29**	**.34**	**.29**	**.41**
WordPhrases + Topics	**91.9%**	**.84**	**.38**	**.31**	**.35**	**.31**	**.42**
Topics + LIWC	**89.2%**	**.80**	**.33**	**.29**	**.33**	**.28**	**.38**
WordPhrases + LIWC	**91.6%**	**.83**	**.38**	**.30**	**.34**	**.30**	**.41**
WordPhrases + Topics + LIWC	**91.9%**	**.84**	**.38**	**.31**	**.35**	**.31**	**.42**

accuracy: percent predicted correctly (for discrete binary outcomes). *R*: Square-root of the coefficient of determination (for sequential/continuous outcomes). *LIWC*: *A priori* word-categories from Linguistic Inquiry and Word Count. *Topics*: Automatically created *LDA* topic clusters. *WordPhrases*: words and phrases (n-grams of size 1 to 3 passing a collocation filter). Bold indicates significant ($p<.01$) improvement over the baseline set of features (use of *LIWC* alone).
doi:10.1371/journal.pone.0073791.t002

Especially accurate for predicting gender and age, less so for personality.

Predict mental health & psychology

Facebook language predicts depression in medical records

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, +4, and H. Andrew Schwartz Authors Info & Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved September 11, 2018 (received for review February 26, 2018)

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Computer-based personality judgments are more accurate than those made by humans

Wu Youyou, Michal Kosinski, and David Stillwell Authors Info & Affiliations

Detect earthquake using Twitter

Earthquake Shakes Twitter Users:
Real-time Event Detection by Social Sensors

Takeshi Sakaki
The University of Tokyo
Yayoi 2-11-16, Bunkyo-ku
Tokyo, Japan
sakaki@biz-model.t.u-tokyo.ac.jp

Makoto Okazaki
The University of Tokyo
Yayoi 2-11-16, Bunkyo-ku
Tokyo, Japan
m_okazaki@biz-model.t.u-tokyo.ac.jp

Yutaka Matsuo
The University of Tokyo
Yayoi 2-11-16, Bunkyo-ku
Tokyo, Japan
matsuo@biz-model.t.u-tokyo.ac.jp

Detecting earthquake

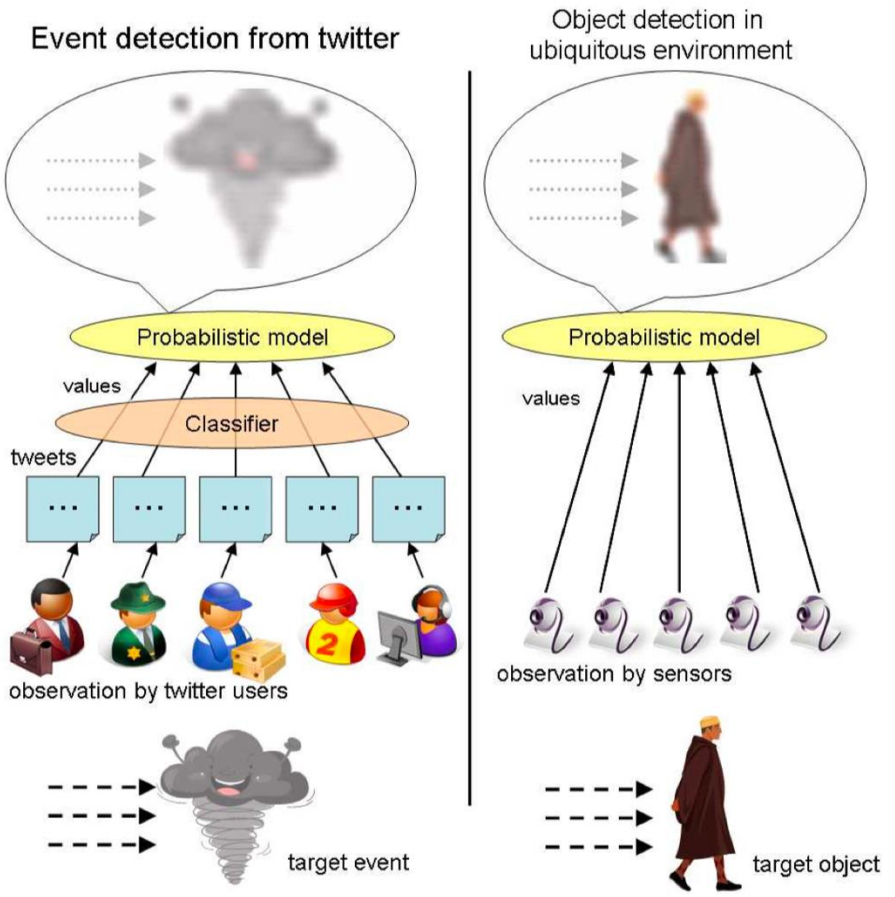


Figure 3: Correspondence between event detection from Twitter and object detection in a ubiquitous environment.

Algorithm 2 Event detection and location estimation algorithm.

1. Given a set of queries Q for a target event.
2. Put a query Q using search API every s seconds and obtain tweets T.
3. For each tweet $t \in T$, obtain features A, B, and C. Apply the classification to obtain value $v_t = \{0, 1\}$.
4. Calculate event occurrence probability p_{occur} using $v_t, t \in T$; if it is above the threshold p_{occur}^{thre}, then proceed to step 5.
5. For each tweet $t \in T$, we obtain the latitude and the longitude l_t by i) utilizing the associated GPS location, ii) making a query to Google Map the registered location for user u_t. Set l_t = null if both do not work.
6. Calculate the estimated location of the event from $l_t, t \in T$ using Kalman filtering or particle filtering.
7. (optionally) Send alert e-mails to registered users.

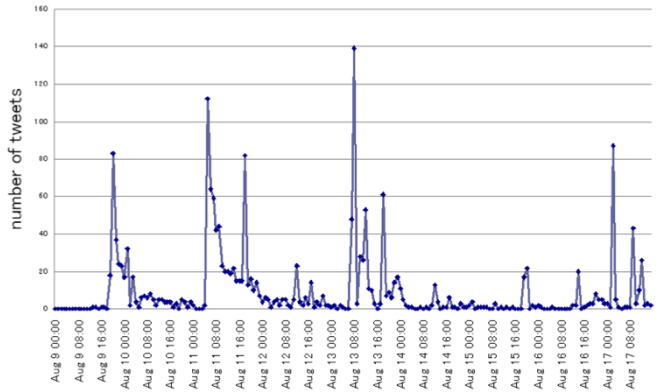


Figure 4: Number of tweets related to earthquakes.

Detecting earthquake

Classify earthquake-related tweets using:

- Semantic features (length, etc.)
- Keywords features
- Word context features

Table 1: Performance of classification.

(i) *earthquake* query:

Features	Recall	Precision	F-value
A	87.50%	63.64%	73.69%
B	87.50%	38.89%	53.85%
C	50.00%	66.67%	57.14%
All	87.50 %	63.64%	73.69%

(ii) *shaking* query:

Features	Recall	Precision	F-value
A	66.67%	68.57%	67.61%
B	86.11%	57.41%	68.89%
C	52.78%	86.36%	68.20%
All	80.56 %	65.91%	72.50%

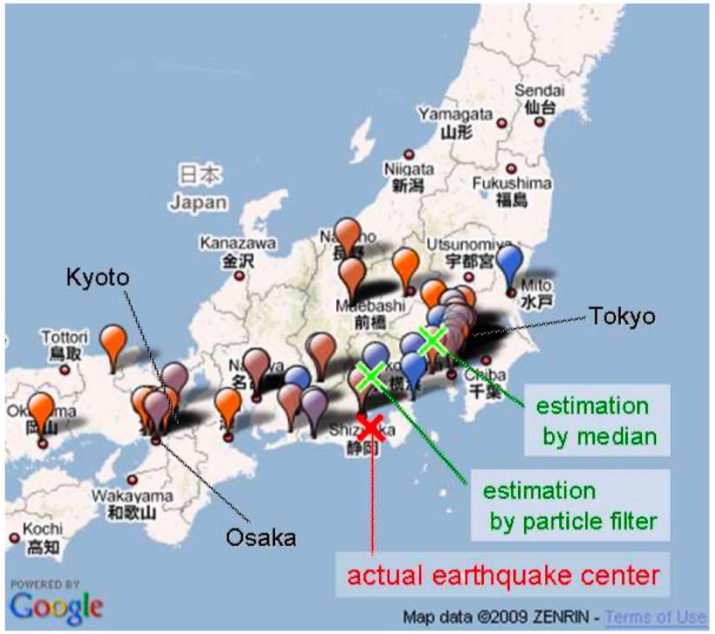


Figure 9: Earthquake location estimation based on tweets. Balloons show the tweets on the earthquake. The cross shows the earthquake center. Red represents early tweets; blue shows later tweets.

Hard case: "Is this an *earthquake* or a truck passing?"

Detecting earthquake

Real-world performance is very good:

- Can detect 96% earthquake with JMA seismic intensity scale +3
- Recency close to 1 minute; better than JMA!

Table 5: Earthquake detection performance for two months from August 2009. 'Promptly detected' denotes detection within a minutes.

JMA intensity scale	2 or more	3 or more	4 or more
Num. of earthquakes	78	25	3
Detected	70(89.7%)	24 (96.0%)	3 (100.0%)
Promptly detected	53 (67.9%)	20 (80.0%)	3 (100.0%)

Table 4: Facts about earthquake detection.

Date	Magnitude	Location	Time	E-mail sent time	#tweets within 10 min	Announce of JMA
Aug. 18	4.5	Tochigi	6:58:55	7:00:30	35	07:08
Aug. 18	3.1	Suruga-wan	19:22:48	19:23:14	17	19:28
Aug. 21	4.1	Chiba	8:51:16	8:51:35	52	8:56
Aug. 25	4.3	Uraga-oki	2:22:49	2:23:21	23	02:27
Aug. 25	3.5	Fukushima	22:21:16	22:22:29	13	22:26
Aug. 27	3.9	Wakayama	17:47:30	17:48:11	16	17:53
Aug. 27	2.8	Suruga-wan	20:26:23	20:26:45	14	20:31
Aug. 31	4.5	Fukushima	00:45:54	00:46:24	32	00:51
Sep. 2	3.3	Suruga-wan	13:04:45	13:05:04	18	13:10
Sep. 2	3.6	Bungo-suido	17:37:53	17:38:27	3	17:43

SMA for social media industry

Social media companies have the most interest in SMA:

- User engagement & retention & adoption
- Reduce misinformation / polarization / segregation
- Target marketing
- Personalization

Key metrics in SMA

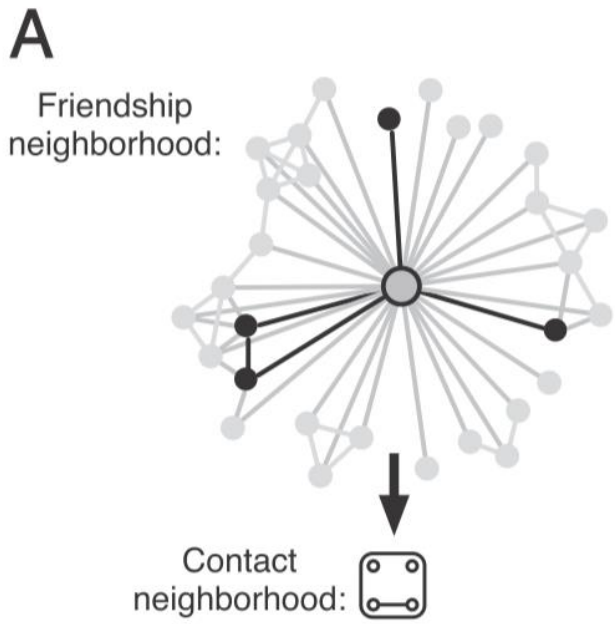
- **Reach**: number of unique users who see the content
  - E.g., A post seen by 10,000 unique users
- **Impressions**: total number of times a post was displayed
  - Same post can by viewed multiple times by the same person
- **Engagement**: number of interactions
  - E.g., Likes, comments, shares
- Engagement rate: (Engagements / Reach) * 100
- Click-Through Rate: Clicks / Impressions

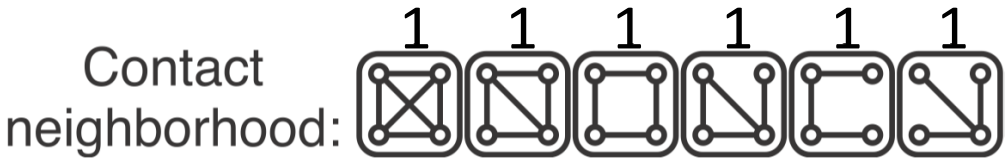
Discussion

- We would like to know who are the top pop stars followed by the younger generations in Hong Kong.
- Which is the most important social media platform to study?
- What data / information should we collect?
  - Users
  - Relations
  - Attributes
  - Content/Posts
- How about comparing the pattern with Mainland China?

How to increase platform adoption?

(1) Num. of friends (size) vs.
(2) Num. of their connected components (structure)
Which matters more? What if (1) > (2) or the other way around?





How to increase platform adoption?

Structural diversity in social contagion

Johan Ugander[a], Lars Backstrom[b], Cameron Marlow[b], and Jon Kleinberg[c,1]

[a]Center for Applied Mathematics and [c]Department of Computer Science, Cornell University, Ithaca, NY 14853; and [b]Facebook, Menlo Park, CA 94025

Edited by Ronald L. Graham, University of California at San Diego, La Jolla, CA, and approved February 21, 2012 (received for review October 6, 2011)

Network structure is a stronger predictor of adoption than size.

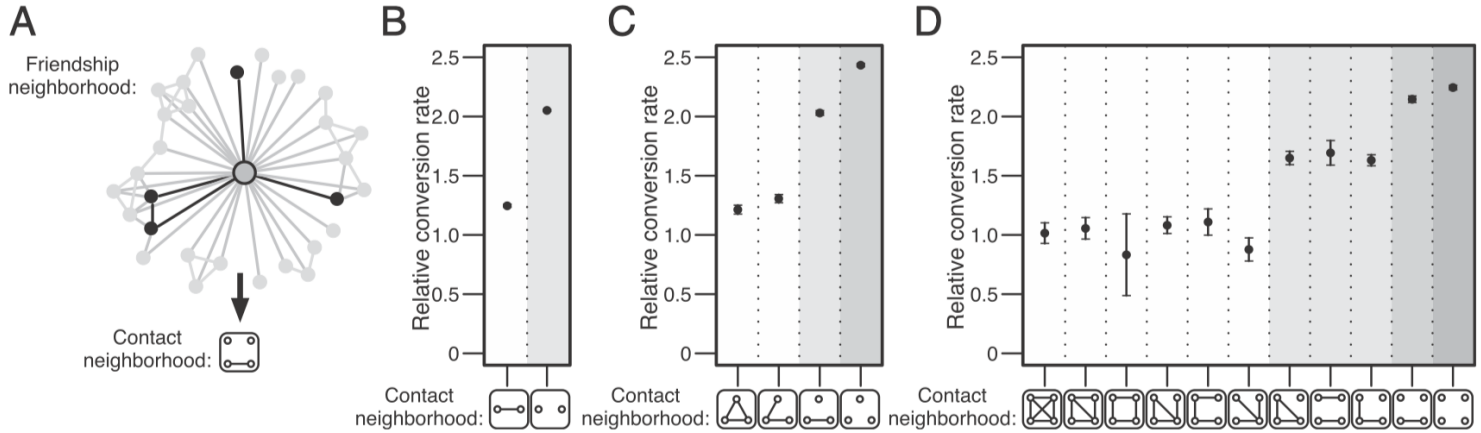


Fig. 1. Contact neighborhoods during recruitment. (*A*) An illustration of a small friendship neighborhood and a highlighted contact neighborhood consisting of four nodes and three components. (*B–D*) The relative conversion rates for two-node, three-node, and four-node contact neighborhood graphs. Shading indicates differences in component count. For five-node neighborhoods, see Fig. S1. Invitation conversion rates are reported on a relative scale, where 1.0 signifies the conversion rate of one-node neighborhoods. Error bars represent 95% confidence intervals and implicitly reveal the relative frequency of the different topologies.

Bot detection

Online Human-Bot Interactions:
Detection, Estimation, and Characterization

Onur Varol,[1*] Emilio Ferrara,[2] Clayton A. Davis,[1] Filippo Menczer,[1] Alessandro Flammini[1]
[1]Center for Complex Networks and Systems Research, Indiana University, Bloomington, US
[2]Information Sciences Institute, University of Southern California, Marina del Rey, CA, US

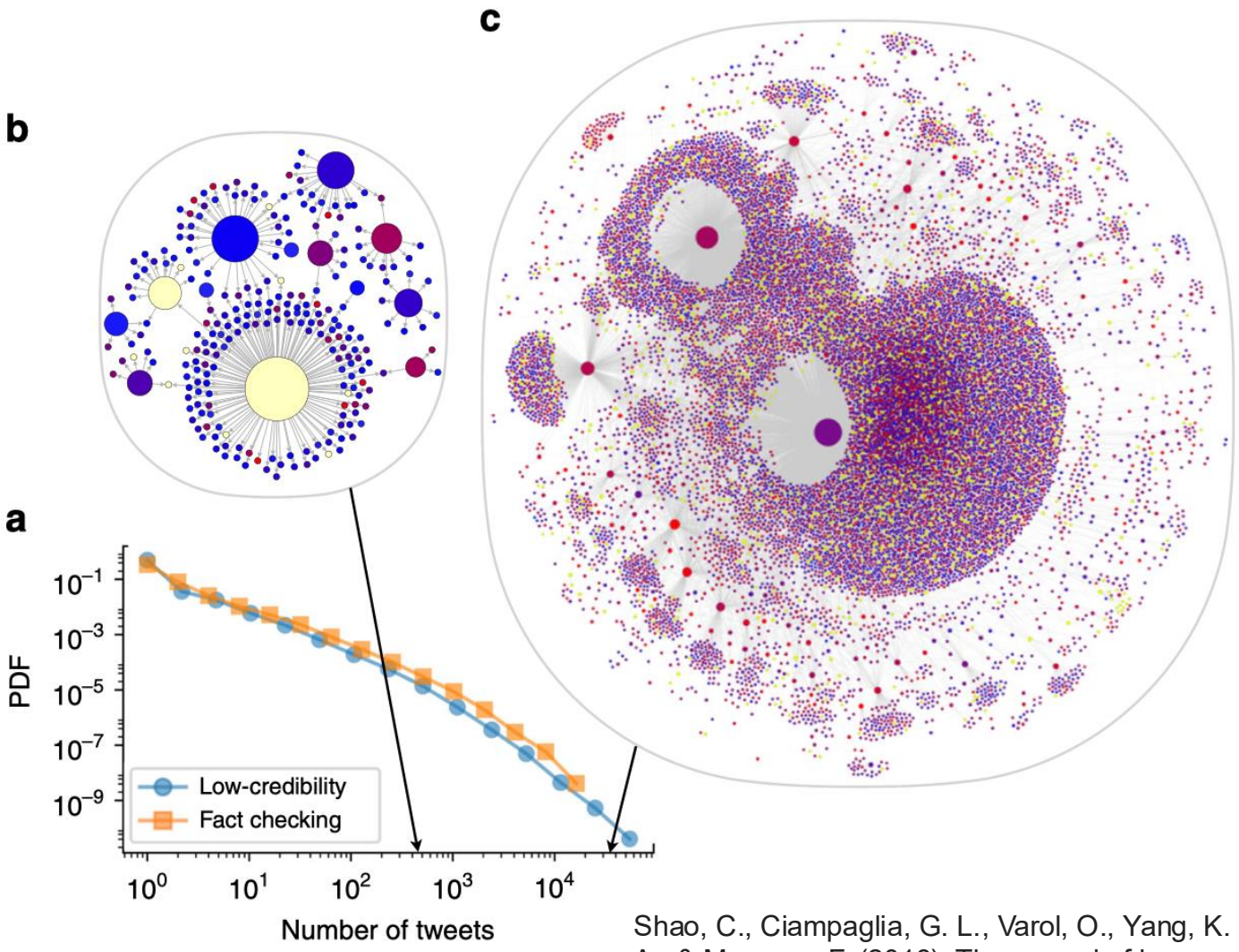
Social bots: SM accounts controlled by software

- Amplifying misinformation
- Swaying public attention
- Committing financial fraud
- Suppressing or disrupting speech
- Spreading malware or spam
- Trolling/attacking victims

Botometer X
An OSoMe project (bot•o•meter)

Demo: https://botometer.osome.iu.edu/

Spread of misinformation



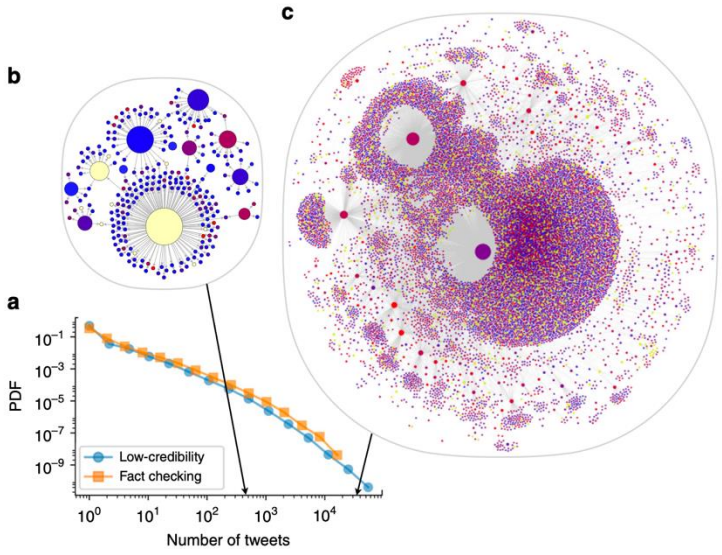
Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787.

Spread of misinformation

Fake articles vs. true ones:

- Similar popularity distribution (hard to tell apart just based on popularity)
- False news is slightly even more viral
- Popularity is very heterogeneous across articles in both groups
- Some extremely viral fake news is largely driven by bots (red nodes)

Can we utilize the network structure of spreading dynamics to detect fake news?



Challenges in SMA

- Privacy concerns (consent, anonymity)
- Data bias / unrepresentative
- Misinformation / bots
- Proprietary algorithms
- Decreasing well-being

Emerging technologies

- Intelligent LLM agents
  - Replaces human moderators
- Immersive technologies (VR/AR/MR)
  - Enables new forms of engagement, such as virtual events
- Blockchain for data security
  - Enhancing data privacy and security in SMA
  - Ensuring ethical data collection and compliance with regulations
  - Addressing growing user concerns about data protection
- Video analytics
  - Provides deeper insights into performance metrics, including engagement, watch time, and audience retention
- Real-time analytics
  - Responds swiftly to trends or crises; enables proactive strategies

Getting prepared

- Setup Python programming environment
  - Install Python
  - Install Jupyter Notebook for interactive coding
  - Install libraries: numpy, pandas, matplotlib, etc.
- Practice time
  - Follow steps to setup computing tools
  - Consider using virtual environments
  - Ask questions to resolve issues now