Introduction to Social Media Analytics (Lec 7)

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

Agenda for this week

- Audience analytics
- User demographics
- Psychological traits
- Political affiliations
- Health informatics

Audience analytics

- Sentiment and emotion are “text-level” characteristics
- Lots of user-level social / psychological attributes
  - Demographics (age, race, gender, income, etc)
  - Psychological traits (personality, sexual orientation, prosociality)
  - Political affiliation, user interests, communities, incivility, etc.
  - Mental health/state (empathy, depression, stress, etc)
  - Healthy lifestyle (weight lose, eating disorder, early riser)
  - …

Audience analytics

- Audience analytics aims to understand the demographic & psychographic characteristics of social media users such as audience size, location, gender, age, and interests, etc.

- By leveraging audience analytics, businesses & org. can gain invaluable insights into *who their followers are, where they come from, and what they care about*, enabling the creation of more targeted and effective marketing strategies.

Examples of audience analytics

- **Product Development Insights**—A gaming company discovers through audience analytics that a substantial number of its X followers are female gamers interested in role-playing games (RPGs). This insight leads to the development of a new RPG with strong female lead characters, directly catering to this audience segment.

- **Geographical Targeting**—A coffee chain uses audience analytics to find out that a large number of its Facebook followers are located in northeastern of United States. This insight guides the chain's decision to run a special promotion in those regions, capitalizing on the high concentration of followers there.

Demographic analysis

Demographics refer to the statistical characteristics of human populations and particular groups within it. This includes:

- **Age:** Knowing the age distribution of audience can help tailor content to be more relatable and engaging.
- **Gender:** Understanding the gender breakdown can inform product offerings and effective marketing messages.
- **Race/Ethnicity**: One of the most important social identities in racially diverse countries such as the U.S.
- **Location:** Location data helps in understanding geographical representation and can help design marketing campaigns.

Gender

- Biological/sex-based definition
  - Refers to biological characteristics (female vs. male)
  - Often used in medical treatments
- Social/cultural definition
  - Refers to social roles, behaviors, and expectations societies assign to individuals based on their perceived sex.
  - They can vary across cultures and history.
- Gender Identities
  - Binary (women vs. men)
  - Non-binary, mixed, unisex
- Gender prediction: https://genderize.io/

Race, nationality, ethnicity

- **Race**:
  - Based on appearance (Asian, Black, Hispanic, White, etc.)
  - Could be broken down into finer categories (East vs. South Asian)
- **Nationality**:
  - Objective, but hard to know without self-identities
  - Asian --> Chinese, Japanese, Korean, etc.
  - White --> England, French, Germany, Italian, etc.
- **Ethnicity**:
  - Nationality is about country's flag; ethnicity is about cultural roots
  - Can differ from one's nationality due to migration and immigration
- **True vs. perceived identity**
  - Perceived attributes can often be inferred from name
  - In many cases, they are not the same (e.g., Michael Jackson)
  - Which one should you use? In what contexts?

Socioeconomic status (SES)

- Education
- Occupation
- Income level
- **SES**: a concept used by economists & sociologists. The measure combines a person‘s work experience and their or their family’s access to economic resources & social position relative to others.
- **Social class**: refers to a person’s relatively stable cultural background, whereas *SES refers to one’s current social & economic situation which is more changeable over time*.

https://en.wikipedia.org/wiki/Socioeconomic_status

Case study on user demographics

How to understand the demographic composition of a brand's social media followers using techniques in SMA?

- **Tool Setup:** Use APIs to access audience data from a brand’s Facebook or Instagram account.
- **Data Collection:** Extract data on age, gender, and location.
- **Data Analysis:** Create visualizations (pie charts for gender distribution, age histograms, and heat maps for location) to display the demographic data.
- **Reporting:** Write a report on how the brand‘s content might be tailored to better suit the diverse demographics of its users.

Psychographic analysis

Psychographics delve into the psychological attributes of an audience, including personality traits, values, interests, and lifestyles. This can significantly enhance user understanding:

- **Interests:** By analyzing what your audience is interested in, you can tailor content to their preferences. A travel agency, for instance, may discover through audience analytics that their followers are keen on sustainable travel, prompting them to share more eco-friendly travel options.
- **Lifestyle:** Understanding the lifestyle of your audience can help in creating relatable content. If a luxury watch brand finds that its Instagram followers are interested in luxury lifestyle beyond just watches, including fine dining and high-end cars, it might expand its content to cover these areas as well.

Big Five personality model

The Big Five model is widely used in psychology for its reliability and ability to predict user behavior across various contexts.

- **Openness**: Creativity and curiosity; high openness means embracing new ideas, low means preferring routine.
- **Conscientiousness**: Organization and responsibility; high is diligent and goal-focused, low is impulsive.
- **Extraversion**: Sociability and energy; extraverts are outgoing, introverts are reserved.
- **Agreeableness**: Compassion and cooperation; high is kind and trusting, low is competitive.
- **Neuroticism**: Emotional stability; high is anxious and moody, low is calm and resilient.

Predicting personality traits

Supervised learning (e.g., logistic regression, random forests) can be used to predict Big Five traits from social media activity:

- **Self-assessment tools**: The Big Five is typically measured using standardized questionnaires like the Big Five Inventory.
- **Data collection**: Use APIs to collect user data with permission.
- **Feature engineering**: Extract features like linguistic patterns (e.g., LIWC, positive/negative sentiment for Agreeableness), posting frequency (Extraversion), topic diversity (Openness), and profile metadata (e.g., bio descriptions and avatars).
- **Accuracy**: Studies (e.g., Kosinski et al., 2013) show that models trained on social media data can predict personality traits with correlations of ~0.4-0.6 to self-reported scores.

Applications of Big Five

- **User segmentation**: Cluster users by personality traits to tailor content or ads. For example, high Openness users may respond to creative, innovative campaigns, while high Conscientiousness users prefer structured, goal-oriented content.
- **Engagement prediction**: Forecast user interactions (e.g., likes) by including features of personality traits. Extraverts are more likely to share content, while conscientious users may engage more often with professional or educational posts.
- **Example**: Build a ML model to predict which users will retweet based on their Extraversion and Openness scores.

Political affiliation

Political affiliation refers to an individual's self-alignment with a political party, ideology, or movement.

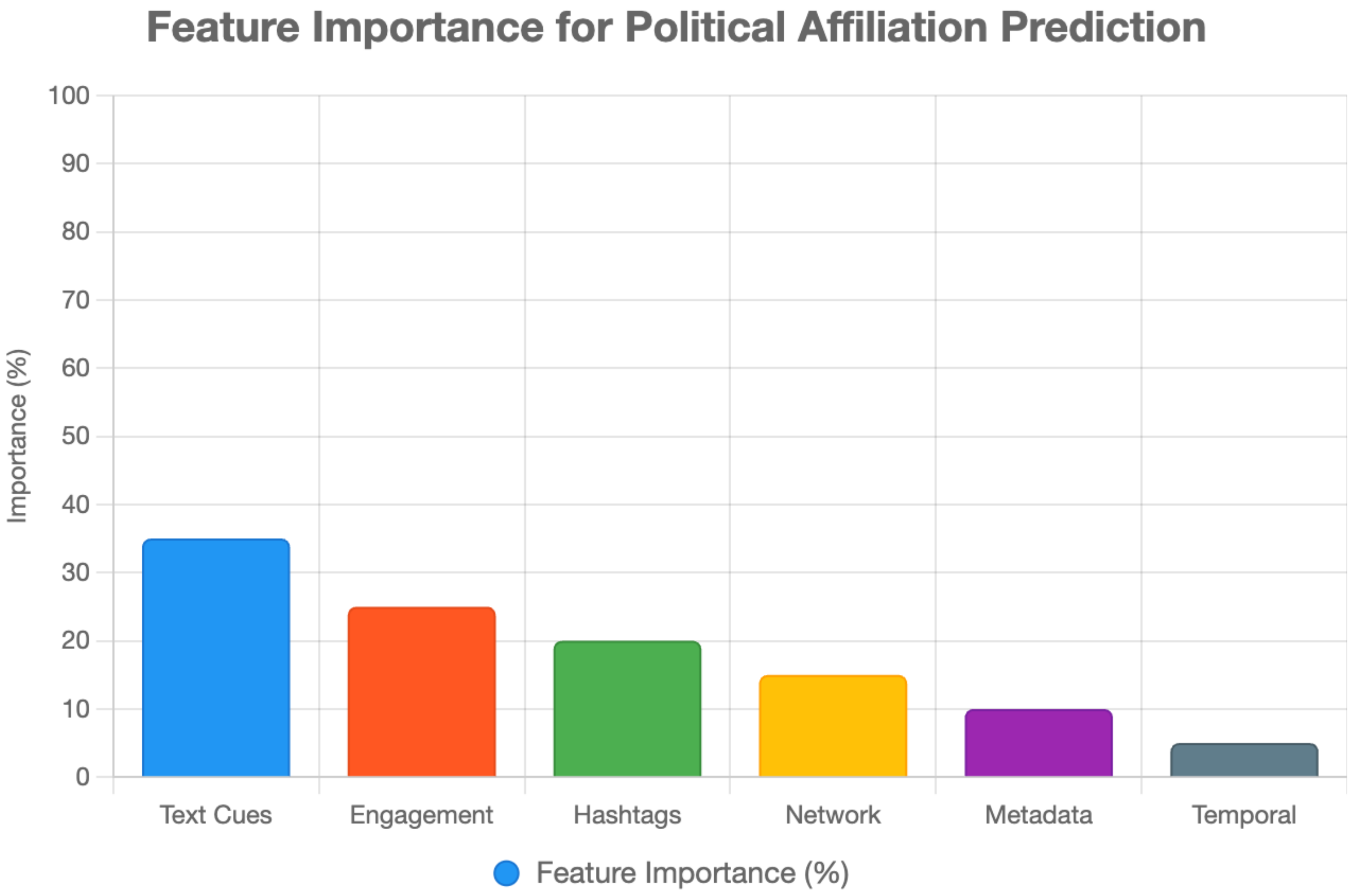
- **Liberal**: Advocates for systemic change, social equality, and policy intervention to promote welfare and rights. On social media, these users often engage with progressive hashtags.
- **Conservative**: Emphasizes tradition, limited government, and preserving established norms. Often prioritizes social stability, economic freedom, and cultural norms.
- **Independent**: Resists rigid ideological alignment, favoring issue-specific positions over ideological loyalty.
- **Populist**: Focuses on representing ordinary people against perceived elites, often emphasizing anti-establishment.

Predicting political affiliations

Useful features for predicting political affiliations:

- **Linguistic cues**: Words, phrases, and sentiment in user posts. Liberals often use inclusive terms (e.g., “equity,” “freedom”); Conservatives favor patriotic language (e.g., “heritage”).
- **Hashtag and topics**: Frequency of interactions with political content. Left-leaning users may engage with social justice hashtags. Topics can signal affiliation (e.g., #BLM for Liberals).
- **Network features**: Following or interacting with accounts of known affiliations (e.g., following progressive politicians for Liberals, authoritative leaders for Conservatives).
- **Profile Metadate**: Bio description & profile image. Terms like “patriot” (Conservative) & “activist” (Liberal) are strong signal.

Predicting political affiliations

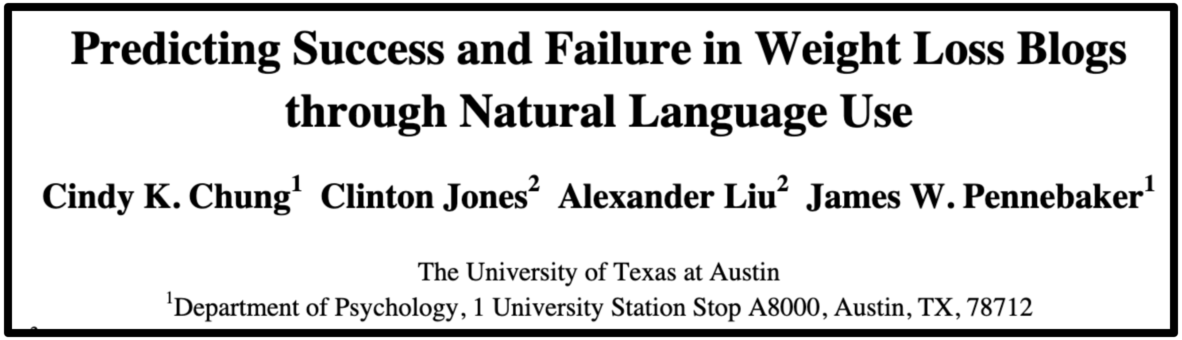


Health informatics

Research in this area focuses on using social media data to predict physical & mental health issues:

- Depression, anxiety, suicide risk, etc.
- Empathy, sexual abuse, anonymity
- Social support, help seeking on Reddit, etc.
- Healthy lifestyle (weight losing, pro-eating disorder)

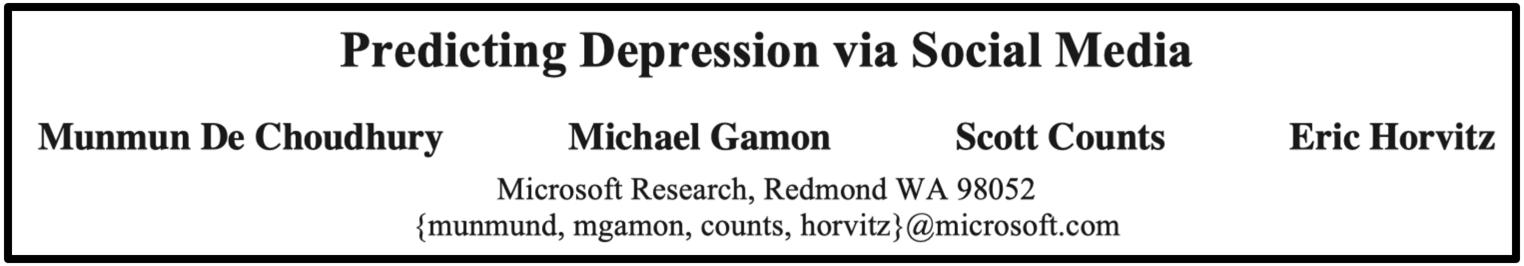
Empathy, depression, weight loss



Predicting Success and Failure in Weight Loss Blogs through Natural Language Use

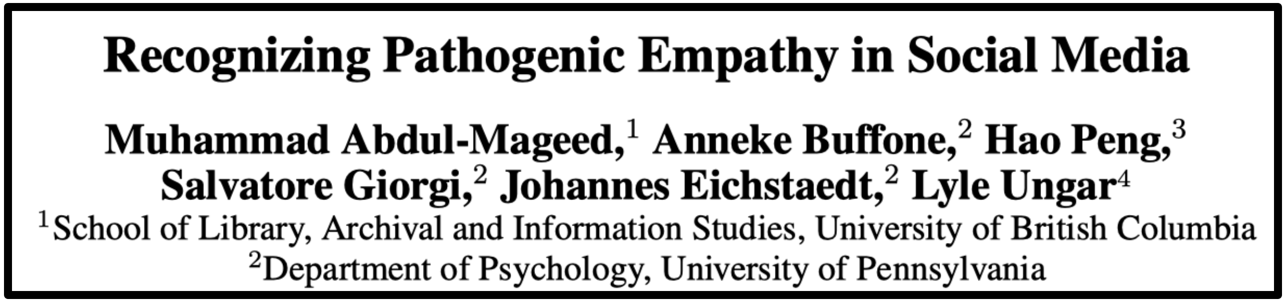
Cindy K. Chung[1] Clinton Jones[2] Alexander Liu[2] James W. Pennebaker[1]

The University of Texas at Austin
[1]Department of Psychology, 1 University Station Stop A8000, Austin, TX, 78712



Predicting Depression via Social Media

Munmun De Choudhury Michael Gamon Scott Counts Eric Horvitz
Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com



Recognizing Pathogenic Empathy in Social Media

Muhammad Abdul-Mageed,[1] Anneke Buffone,[2] Hao Peng,[3]
Salvatore Giorgi,[2] Johannes Eichstaedt,[2] Lyle Ungar[4]
[1]School of Library, Archival and Information Studies, University of British Columbia
[2]Department of Psychology, University of Pennsylvania

Predicting mental health outcomes

Leverage machine learning & NLP tools to predict mental health using social media data (e.g., posts, comments, engagement):

- **Self-assessment tools**: use crowdsourcing to compile a set of social media users who report being diagnosed as depressed; measure their DV based on standard psychometric instrument.
- **Data collection**: Use APIs to collect user data with permission.
- **Feature engineering**: Extract and measure their behavioral attributes such as *social engagement, emotion, language use, linguistic styles, ego network, mentions of using medications*.
- **Correlational analysis**: Most studies found that social media contains useful signals for predicting onset of mental issues.

Accuracy is not good enough! Can be used as monitoring tools.

Course notes

- Midterm questions review now!
  - Raise grading issues in class and via email
  - Return exam paper after class (or 0 score)
  - Don’t spread the question paper online!
- User modeling V2 next week (W9)
- HW2 will be released next week
  - Due in two weeks (W11)
- Social network analysis V1 in Week 10
  - Will use Team-Based Learning (TBL) setting
  - New location: CIC G-001 (map)