Lecture 7:
Data Analysis 2: Data Exploration

IS 618: Social Media Data Analysis
Indira Sen, FSS 2026

Recap: Wikipedia and Social Media Data Cleaning

What makes Wikipedia different from other Social Media Platforms?

- Open access
  - Commitment to Open Access Policy https://w.wiki/7APQ
- Wikimedia Research Team aims to make any work produced with our support openly available to the public and reusable on Wikimedia projects.
  - Open licenses
  - Content can be freely re-used https://w.wiki/8Bur
  - e.g. CC-BY-SA 4.0
- Transparency
  - Data pipelines are public
  - Example: clickstream dataset https://w.wiki/ACVn
- Privacy
  - No 3rd-party collection or sharing of user data https://w.wiki/6awx
  - Private user data removed after 90 days

Source: https://meta.wikimedia.org/wiki/Wikimedia_Data_Tutorial_ICWSM_2024

What is data cleaning and why do we do it?

- **Data cleaning** is the process of detecting and correcting or removing corrupt, inaccurate, or irrelevant records from a dataset to improve its quality before analysis.
- also called data processing, data preprocessing, data wrangling or data munging
- Deduplication: Near-duplicates (e.g., retweets with slight edits) require fuzzy matching.

Figure 1. Standard social media data cleaning pipeline.

https://htmlpreview.github.io/?https://github.com/dess-mannheim/SMDA/blob/main/Handouts/social_media_data_cleaning.html

Text Data Cleaning

- tokenization
- remove stopwords
- Stemming / Lemmatization
- remove numbers
- remove headers and footers
- remove rare words
- Beyond Text:
    - dropping or imputing missing values
    - dropping columns with missing values

Faces Engage Us (Bakshi et al.,)

- **Construct**: how do the presence of faces, their gender and age impact engagement?
- **Data collection**: start with 2,000 popular posts and then snowball. Automatically infer faces, gender, and age using Face++ API
- **Data cleaning**: ??
- **Data analysis**: ne binomial regression wh engagement content (face faces)
- **Result**: posts engagement, though age and gender do not have much effect

Is this data collection, processing, or analysis?

Session: Personal Values and Preferences CHI 2014, One of a CHInd, Toronto, ON, Canada

Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram

Saeideh Bakhshi
Georgia Tech
sbakhshi@cc.gatech.edu

David A. Shamma
Yahoo Labs
aymans@acm.org

Eric Gilbert
Georgia Tech
gilbert@cc.gatech.edu

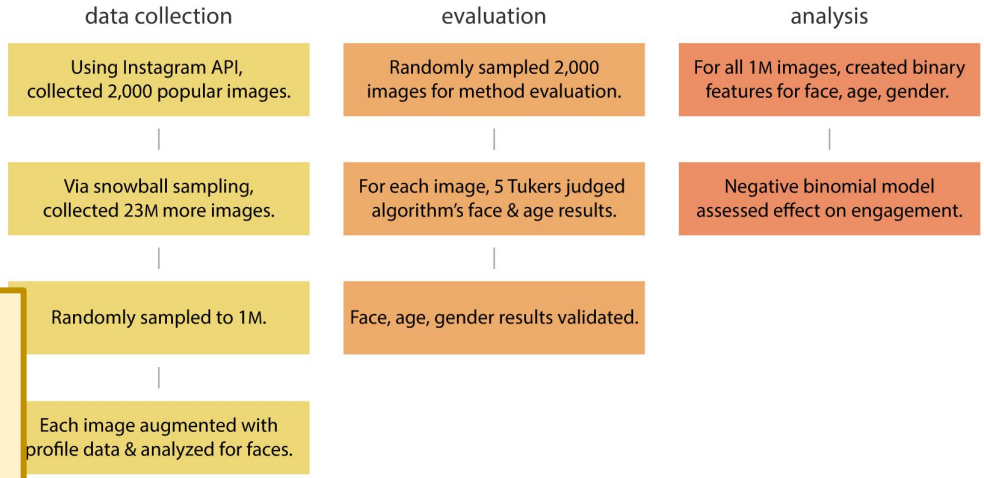


Figure 2. An overview of the steps taken in this paper for collection, evaluation and analysis of the data.

Bakhshi, Saeideh, David A. Shamma, and Eric Gilbert. "Faces engage us: Photos with faces attract more likes and comments on instagram." Proceedings of the SIGCHI conference on human factors in computing systems. 2014.

Data Analysis vs. Collection/Processing

- sometimes, there isn't a clear separation between the steps; collection, processing, and analysis may go hand in hand
- Data augmentation can sometimes be a part of data collection or data cleaning — more of a stylistic choice
- steps may range from text cleaning (parsing the data, removing stopwords) to complex analysis (create text networks or get stance of the post)

Data Cleaning has Consequences

Element	Typical treatment	Caution
URLs	Replace with `[URL]` token or remove	URL domain can be a signal (news vs. spam)
Mentions (`@user`)	Replace with `[USER]` or remove	Preserve for network analysis
Hashtags (`#topic`)	Segment and keep text; remove `#`	Often contain key semantic content
Emojis	Convert to text description or keep as-is	Emojis are critical for sentiment tasks
HTML entities	Decode (`&` → `&`)	Always decode before any analysis
Lowercasing	Convert to lowercase	May destroy named entity casing signals
Punctuation	Task-dependent; often keep for sentiment	"!!!" carries intensity information

Therefore, we should always

- Validate data cleaning steps
- Document data cleaning steps

Discuss: Data Collection and Cleaning Decisions

- Choose one of the following use cases:
  - Studying the factuality of nutrition-related posts on Instagram
  - Studying the factuality of movie-related posts on Wikipedia

- How would you collect data for the chosen use case?
  - Could there be any data quality problems with your approach?

- How would you clean the data?
  - Could there be any data quality problems with your approach?

Today:

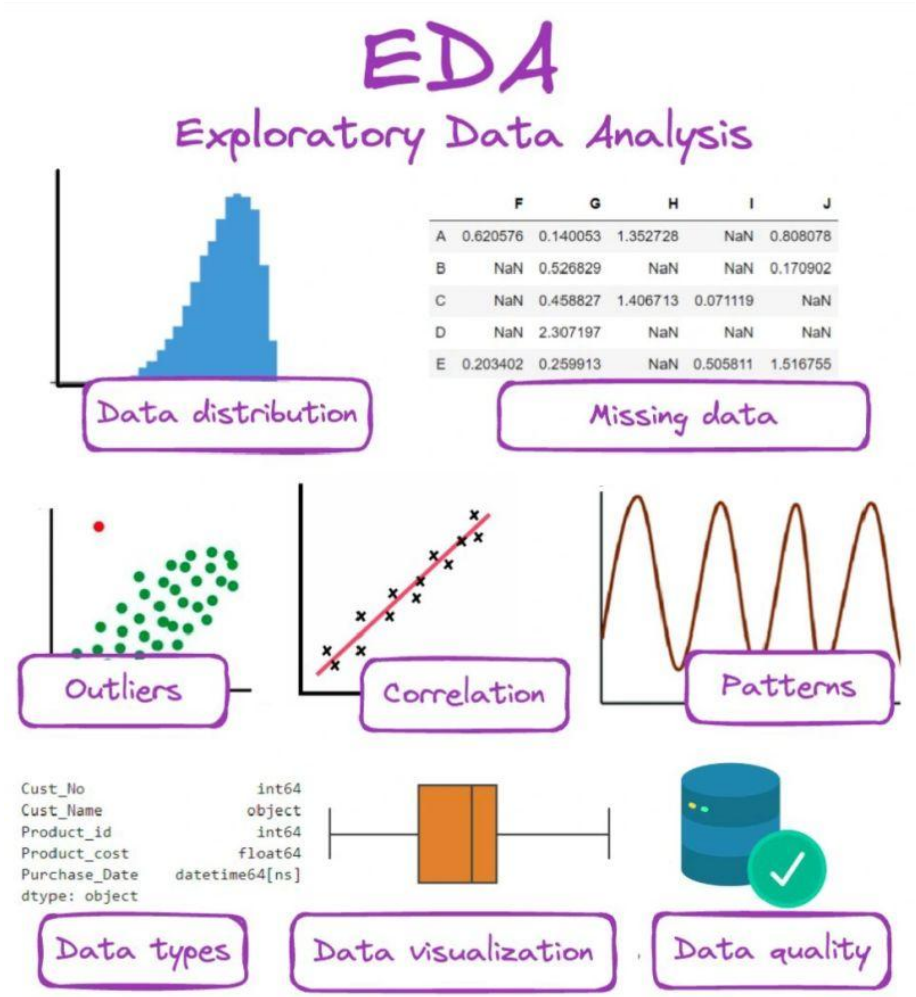
1. Data Exploration, or Exploratory Data Analysis (EDA)
2. Data Visualization

Today:

1. **Data Exploration, or Exploratory Data Analysis (EDA)**
2. Data Visualization

Exploratory Data Analysis

- Summary and descriptive statistics (mean, median, mode, quantiles, min, max)
- Computing distributions of various variables
- Creating new variables from existing ones
    - Length of content
    - # entities mentioned in the content (using Named Entity Recognition)
    - # links in the content (using regex)
    - (and their domains...)
- Bivariate interactions, e.g.,:
    - Length of content across different user groups
    - Difference in content length by user groups (verified vs. unverified)
- ....



	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

https://www.markovml.com/blog/exploratory-data-analysis

EDA for different types of data: TF-IDF

- Tokenize the text into words or n-grams (sequences of words).
- Term Frequency Analysis: Calculate the frequency of each **term** (word or multiple words) in the corpus.
- Document: each text in your dataset, e.g., a single tweet

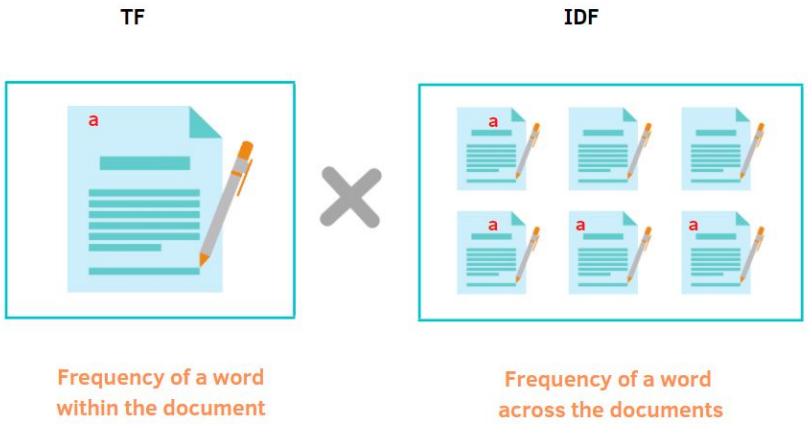
EDA for different types of data: TF-IDF

- Tokenize the text into words or n-grams (sequences of words).
- Term Frequency Analysis: Calculate the frequency of each **term** (word or multiple words) in the corpus.
- Document: each text in your dataset, e.g., a single tweet
- Term frequency (TF) of a term *t* in a document *d*: number of times a term appears in a document

https://betterprogramming.pub/a-friendly-guide-to-nlp-tf-idf-with-python-example-5fcb26286a33

EDA for different types of data: TF-IDF

- Tokenize the text into words or n-grams (sequences of words).
- Term Frequency Analysis: Calculate the frequency of each **term** (word or multiple words) in the corpus.
- Document: each text in your dataset, e.g., a single tweet
- Term frequency (TF) of a term *t* in a document *d*: number of times a term appears in a document
- Inverse document frequency (IDF) of a term *t*: weightage to each term based on its frequency in the entire dataset (*D*)



https://betterprogramming.pub/a-friendly-guide-to-nlp-tf-idf-with-python-example-5fcb26286a33

EDA for different types of data: TF-IDF

- Tokenize the text into words or n-grams (sequences of words).
- Term Frequency Analysis: Calculate the frequency of each **term** (word or multiple words) in the corpus.
- Document: each text in your dataset, e.g., a single tweet
- Term frequency (TF) of a term *t* in a document *d*: number of times a term appears in a document
- Inverse document frequency (IDF) of a term *t*: weightage to each term based on its frequency in the entire dataset (*D*)

https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/

Frequency of a word within the document

$$TF(t, d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

$$IDF(t, D) = \log \frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

Frequency of a word across the documents

We use TF * IDF to upweigh terms which are specific to certain documents rather than simply relying on frequency. Pure frequency would upweigh frequent stopwords like ‘is’, ‘the’

TF-IDF Example

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

$$TF(t, d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

Words	TF (for A)	TF (for B)
Jupiter	1/5	0
Is	1/5	1/8
The	1/5	2/8
largest	1/5	0
Planet	1/5	1/8
Mars	0	1/8
Fourth	0	1/8
From	0	1/8
Sun	0	1/8

Depend on a specific document, so needs to calculated for every term and every document

https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d/

TF-IDF Example

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

Words	TF (for A)	TF (for B)	IDF
Jupiter	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
largest	1/5	0	ln(2/1) = 0.69
Planet	1/5	1/8	ln(2/2) = 0
Mars	0	1/8	ln(2/1) = 0.69
Fourth	0	1/8	ln(2/1) = 0.69
From	0	1/8	ln(2/1) = 0.69
Sun	0	1/8	ln(2/1) = 0.69

$$\text{TF}(t, d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

$$\text{IDF}(t, D) = \log \frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d/

TF-IDF Example

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

$$TF(t, d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

$$IDF(t, D) = \log \frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

Words	TF (for A)	TF (for B)	IDF
Jupiter	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
largest	1/5	0	ln(2/1) = 0.69
Planet	1/5	1/8	ln(2/2) = 0
Mars	0	1/8	ln(2/1) = 0.69
Fourth	0	1/8	ln(2/1) = 0.69
From	0	1/8	ln(2/1) = 0.69
Sun	0	1/8	ln(2/1) = 0.69

Does not depend on a specific document, so only calculated once for every term

https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d/

TF-IDF Example

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

$$TF(t, d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

$$IDF(t, D) = \log \frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

Words	TF (for A)	TF (for B)	IDF	TFIDF (A)	TFIDF (B)
Jupiter	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
largest	1/5	0	ln(2/1) = 0.69	0.138	0
Planet	1/5	1/8	ln(2/2) = 0	0.138	0
Mars	0	1/8	ln(2/1) = 0.69	0	0.086
Fourth	0	1/8	ln(2/1) = 0.69	0	0.086
From	0	1/8	ln(2/1) = 0.69	0	0.086
Sun	0	1/8	ln(2/1) = 0.69	0	0.086

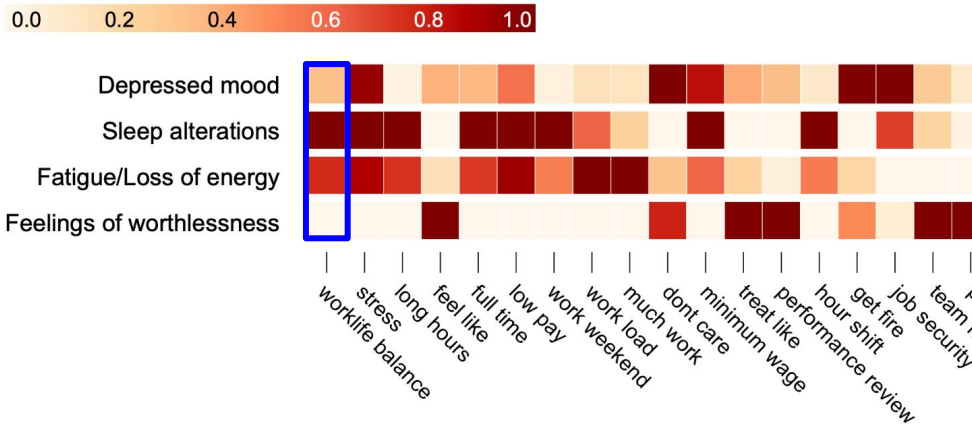
https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d/

What TF-IDF tell us about our data

- After calculating the TF-IDF of terms
  - rank them to find important terms
  - across different types of documents
  - one set of tweets could be considered a single document, e.g., tweets from verified users vs. 'regular' users

- Can extend to multi-words: **n-grams (e.g., bigrams, trigrams)** to capture phrases or collocations.

- TF-IDF can also be used as features for document classification (more in lecture 8)



Sen, Indira, et al. "Depression at work: exploring depression in major US companies from online reviews." *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (2022): 1-21.

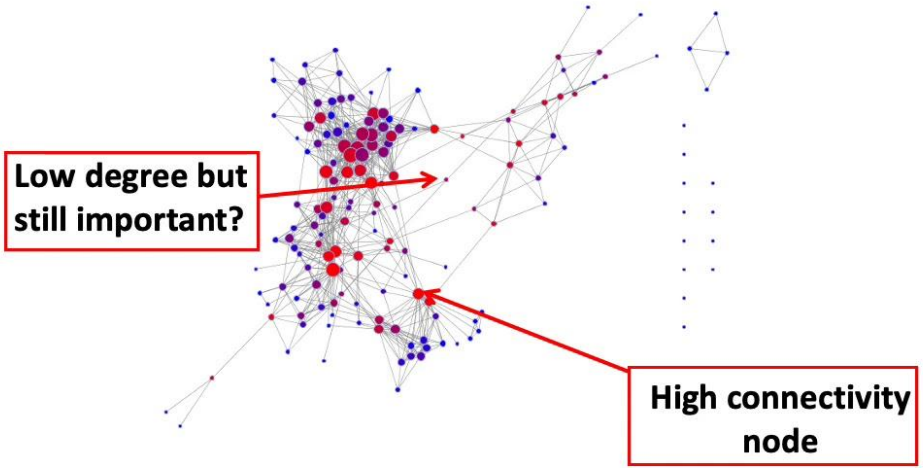
Text Data Exploration

- Language Identification (for multilingual studies)
- Only look at certain types of tokens: e.g., named entities, nouns, adjectives, verbs (i.e., do parts-of-speech tagging or named entity recognition on the raw data)
- Type-token ratio (TTR): # unique words over all words —> richness of vocabulary
- Keyword extraction
- Topic modeling (more in lecture 9)

EDA for different types of data: Network

- **Degree Distribution:**
  - Plot the degree distribution to understand how node degrees are distributed in the network. This can help identify important nodes or hubs.

- **Centrality Measures:**
  - degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality
  - Use centrality scores on the network to identify key nodes.

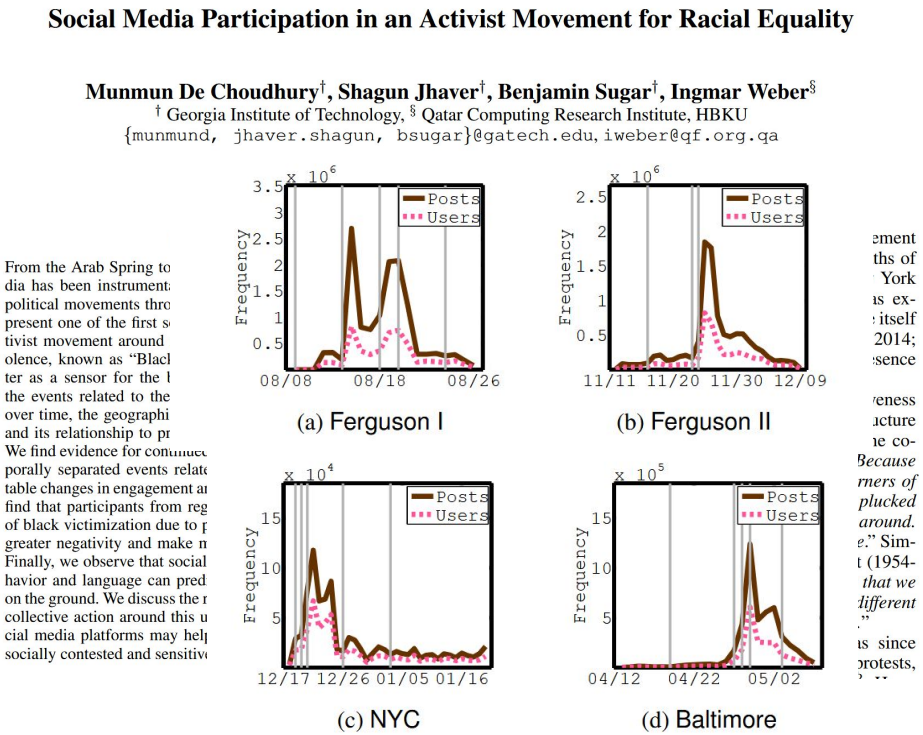


EDA for different types of data: Images and Videos

- Low-Level Visual Feature Exploration
  - Color distribution — histograms of RGB/HSV channels; dominant color palettes (K-means clustering on pixel colors)
  - Brightness and contrast
  - Texture
- Aspect ratio & resolution: platform norms and content types (portrait vs. landscape vs. square)
- Object detection
- Scene classification: indoor/outdoor, urban/nature, etc
- Face detection
- Text in images: Optical Character Recognition (OCR)
- Keyframe analysis

General EDA for Social Media Data

- **Temporal patterns:** post frequency, time-of-day distributions, viral spread timelines
- **Engagement metrics:** likes, shares, comments, views as a function of media type/content
- **Geolocation:** EXIF data or tagged locations for spatial clustering
- **Platform-specific signals:** hashtags, captions, alt-text, account follower counts



Tweet and user volume over time and geography about "Black Lives Matter"

De Choudhury, Munmun, et al. "Social media participation in an activist movement for racial equality." *Proceedings of the international aaai conference on web and social media.* Vol. 10. No. 1. 2016.

Examples of Exploratory Data Analysis in Social Media Research: Mental Health on Reddit

- **Construct and Platform:** Mental health discussions on Reddit

- **Data collection:**
  - use official Reddit API to collect posts, comments, and metadata from several mental health related subreddits
  - Subreddit identified using keywords ('mental health'), snowballing, and manual checks
  - 20,411 posts, 97,661 comments, 27,102 users

De Choudhury, Munmun, and Sushovan De. "Mental health discourse on reddit: Self-disclosure, social support, and anonymity." *Proceedings of the international AAAI conference on web and social media*. Vol. 8. No. 1. 2014.

Mental Health Discourse on reddit:
Self-Disclosure, Social Support, and Anonymity

Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology Atlanta GA
munmund@gatech.edu

Sushovan De
Department of Computer Science
Arizona State University Tempe AZ
sushovan@asu.edu

Abstract

Social media is continually emerging as a platform of information exchange around health challenges. We study mental health discourse on the popular social media: reddit. Building on findings about health information seeking and sharing practices in online forums, and social media like Twitter, we address three research challenges. First, we present a characterization of self-disclosure in mental illness communities on reddit. We observe individuals discussing a variety of concerns ranging from the daily grind to specific

online fora, these social systems are more holistic, in the sense that millions of people use them to post about the mundane goings on of their lives. Additionally, unlike many online communities, most social media sites have a personal permanent identity associated with user profiles. Consequently, these platforms provide a rich ecosystem to study the variety of self-disclosure, social support, and disinhibition that engender health related discourse.

Specifically, we examine a highly popular social news and entertainment media: reddit (*http://www.reddit.com/*).

n interesting online social system that has the
f a forum: it allows sharing blurbs of text and
osts that invite votes and commentary. At the
it is often used as a social feed of information
d from people's contacts and audiences. Another
aspect of reddit is its "throwaway" accounts.
emporary identities often used as an "anonymity
iscuss uninhibited feelings, sensitive information,
unacceptable thoughts momentarily; information
onsidered unsuitable for the mainstream.
e unique characteristics of reddit, in this paper
rested in studying the nature of discourse around
ant health challenge of mental illness. Mental

Posts per day	247.1 (±41.3)
Average posts per user	0.75 (±2.21)
Mean, Median length of posts (words)	96.4, 64
Average upvotes per post	8.37 (±23.2)
Average downvotes per post	1.26 (±3.71)
Average comments per post	5.03 (±7.52)
Mean, Median comment length (words)	27.03, 16
Average comments per user	2.69 (±8.49)
Average upvotes per comment	2.39 (±4.16)
Average downvotes per comment	2.06 (±3.58)
Total number of posts	72,236

Table 1. Basic descriptive statistics of the reddit dataset.

Examples of Exploratory Data Analysis in Social Media Research: Mental Health on Reddit

- **Tip for project presentations and reports:** Show **examples** of the content you've collected
- Helps readers understand more clearly what type of data you're working with
- And how you will analyze it
- Examples provide a complementary and qualitative perspective

Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity

Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology Atlanta GA
munmund@gatech.edu

Sushovan De
Department of Computer Science
Arizona State University Tempe AZ
sushovan@asu.edu

Scammed out of $2600, so depressed
Just sharing... and maybe some advice will come.
How did you feel when your antidepressant started working?
Inability to concentrate destroying my grades
Has anyone else gained weight due to depression?
I can't do this anymore negative thoughts have taken over my mind these past few months
Am I depressed, I don't think I am but I keep sabotaging everything I do.
i'm afraid i might have depression that won't go away by itself and that it needs to be treated professionally...
loneliness, therapy, and just a sharing of weird feelings

Table 2. Titles of a few mental health oriented reddit posts.

Average downvotes per comment	2.06 (±3.58)
Total number of posts	72,236

Table 1. Basic descriptive statistics of the reddit dataset.

holistic, in the
post about the
lly, unlike many
have a personal
user profiles.
ch ecosystem to
al support, and
scourse.
lar social news
w.reddit.com/).
em that has the
urbs of text and
mentary. At the
of information
iences. Another
way" accounts.
an "anonymity
ive information,
ily; information
nstream.
it, in this paper
iscourse around
illness. Mental

De Choudhury, Munmun, and Sushovan De. "Mental health discourse on reddit: Self-disclosure, social support, and anonymity." *Proceedings of the international AAAI conference on web and social media*. Vol. 8. No. 1. 2014.

Data Visualization

General Data Visualization

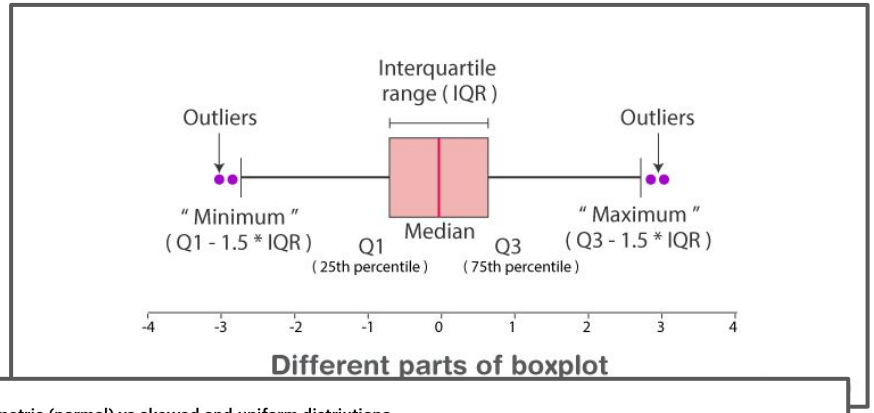
Histograms

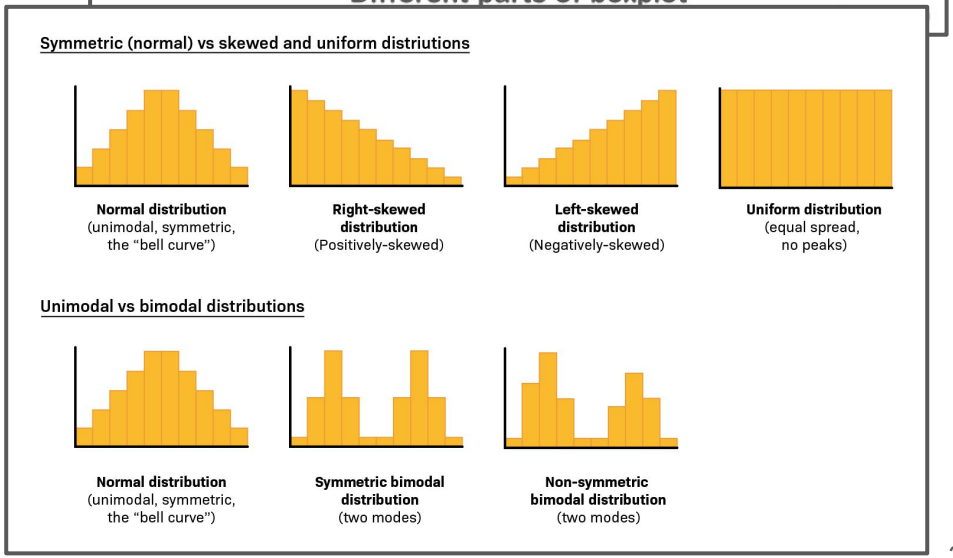
Bar Charts

Box Plots

Pie Charts

...





Examples of Data Visualization in Social Media Research: Mental Health on Reddit

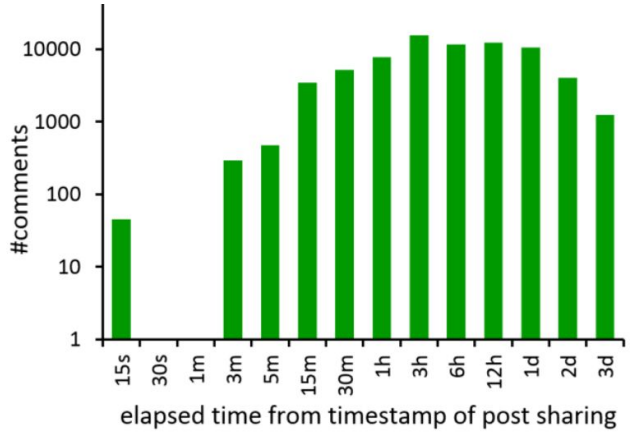


Figure 2. Comment distribution over time. Each bin shows the number of comments shared within a certain time interval from the timestamp of post.

De Choudhury, Munmun, and Sushovan De. "Mental health discourse on reddit: Self-disclosure, social support, and anonymity." *Proceedings of the international AAAI conference on web and social media*. Vol. 8. No. 1. 2014.

Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity

Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology Atlanta GA
munmund@gatech.edu

Sushovan De
Department of Computer Science
Arizona State University Tempe AZ
sushovan@asu.edu

Abstract

Social media is continually emerging as a platform of information exchange around health challenges. We study mental health discourse on the popular social media: reddit. Building on findings about health information seeking and sharing practices in online forums, and social media like Twitter, we address three research challenges. First, we present a characterization of self-disclosure in mental illness communities on reddit. We observe individuals discussing a variety of concerns ranging from the daily grind to specific

online fora, these social systems are more holistic, in the sense that millions of people use them to post about the mundane goings on of their lives. Additionally, unlike many online communities, most social media sites have a personal permanent identity associated with user profiles. Consequently, these platforms provide a rich ecosystem to study the variety of self-disclosure, social support, and disinhibition that engender health related discourse.

Specifically, we examine a highly popular social news and entertainment media: reddit (*http://www.reddit.com/*).

n interesting online social system that has the
f a forum: it allows sharing blurbs of text and
osts that invite votes and commentary. At the
it is often used as a social feed of information
d from people's contacts and audiences. Another
aspect of reddit is its "throwaway" accounts.
emporary identities often used as an "anonymity
iscuss uninhibited feelings, sensitive information,
unacceptable thoughts momentarily; information
onsidered unsuitable for the mainstream.
e unique characteristics of reddit, in this paper
rested in studying the nature of discourse around
ant health challenge of mental illness. Mental

Posts per day	247.1 (±41.3)
Average posts per user	0.75 (±2.21)
Mean, Median length of posts (words)	96.4, 64
Average upvotes per post	8.37 (±23.2)
Average downvotes per post	1.26 (±3.71)
Average comments per post	5.03 (±7.52)
Mean, Median comment length (words)	27.03, 16
Average comments per user	2.69 (±8.49)
Average upvotes per comment	2.39 (±4.16)
Average downvotes per comment	2.06 (±3.58)
Total number of posts	72,236

Table 1. Basic descriptive statistics of the reddit dataset.

Data Visualization for Text

- Word clouds: easy to make, but often too vague.

Data Visualization for Text

- Word clouds: easy to make, but often too vague.

- Variants of word clouds
  - Accentuated word clouds

English

Arabic

Data Visualization for Text

- Word clouds: easy to make, but often too vague.

- Variants of word clouds
  - Accentuated word clouds
  - Semantically grouped word clouds

English

abuse
toxicity
directed
derogatory
comment
trolling
targeted
tweet
form
domains
bias
stereotypes
users
sexual
toxic
haras
used
protected
cause
emotion
words
communit
keywords
basis
across
tweets
harm
category
categories
expresses
clearly
express
using
media
task
hostile
antisemitism
incidents
comments
offensive
types
detection
data
resulting
image
causing
hatespeech
abusive
system
may
work
orientation
information
conversation
problem
warning
define
carried
etc
direct
harful
technologies
explicitly
communication
can
not
detect
annotated
automatically
automatic
classifier
classifiers
classification
classifications
classified
classifications
classifying
classifying
classifying

Data Visualization for Text

- Word clouds: easy to make, but often too vague.

- Variants of word clouds
  - Accentuated word clouds
  - Semantically grouped word clouds

- Word Shift Graphs

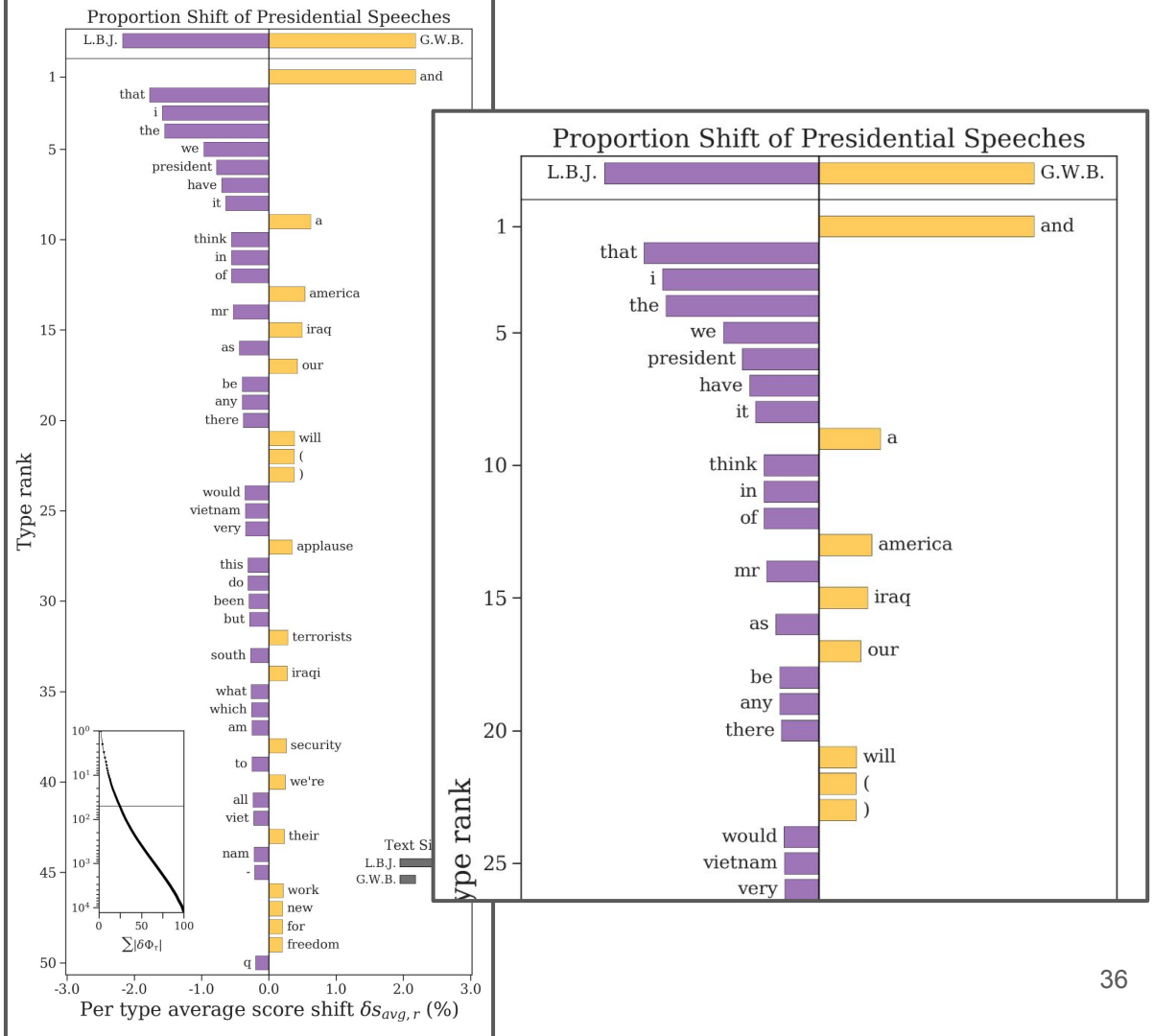
English

Arabic

The Origin of Species, by Charles Darwin

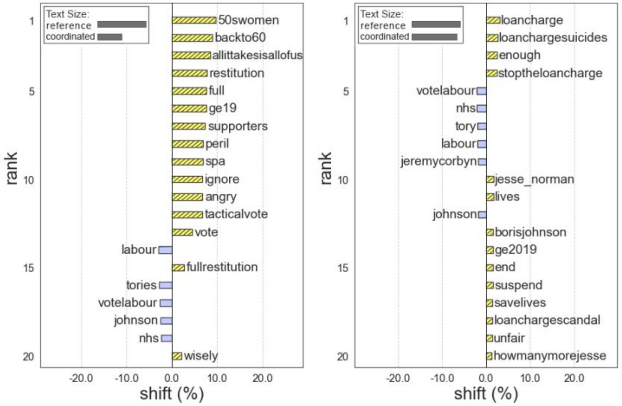
Word Shift Graphs

- Used for visualizing **pairwise comparisons** between texts through word shifts
- Word shift graphs extract which words contribute to a difference between two texts



https://shifterator.readthedocs.io/en/latest/

Using Word Shift Graphs to Study Information Manipulation



(a) B60. (b) LCH.

Figure 8: Application of word shift graphs for highlighting narratives that characterize coordinated communities. Words are ranked based on their contribution towards differentiating coordinated and non-coordinated users. Yellow-colored words (right-hand side of each word shift graph) are in-formative for coordinated users, while blue-colored words(left-hand side) are informative for non-coordinated users.

Coordinated Behavior on Social Media in 2019 UK General Election

Leonardo Nizzoli,[12*] Serena Tardelli,[12*†] Marco Avvenuti,[2] Stefano Cresci,[1] Maurizio Tesconi[1]
[1]Institute of Informatics and Telematics, National Research Council, Pisa, Italy.
[2]Department of Information Engineering, University of Pisa, Italy
{leonardo.nizzoli, serena.tardelli, stefano.cresci, maurizio.tesconi}@iit.cnr.it, marco.avvenuti@unipi.it

Abstract

Coordinated online behaviors are an essential part of information and influence operations, as they allow a more effective disinformation's spread. Most studies on coordinated behaviors involved manual investigations, and the few existing computational approaches make bold assumptions or oversimplify the problem to make it tractable.

Here, we propose a new network-based framework for uncovering and studying coordinated behaviors on social media. Our research extends existing systems and goes beyond limiting binary classifications of coordinated and uncoordinated behaviors. It allows to expose different coordination patterns and to estimate the degree of coordination that characterizes diverse communities. We apply our framework to a dataset collected during the 2019 UK General Election, detecting and characterizing coordinated communities that participated in the electoral debate. Our work conveys both theoretical and practical implications and provides more nuanced and fine-grained results for studying online information manipulation.

efforts, researchers debate the efficacy of the proposed solutions, and IOs still appear to pose a severe threat to our democracies and societies (Barrett 2019).

Meanwhile, groundbreaking advances in specific areas of computing are causing profound changes to the online information landscape. Advances in artificial intelligence brought to the rise of deepfakes – synthetic media where the source has been modified via deep learning techniques. Deepfakes allow crafting arbitrary texts that resemble a target person's writing style or audio and video samples where a target person's face and voice could be made to do or say anything. Unsurprisingly, these powerful techniques have already been used to create fake news (Zellers et al. 2019) and fake profile pictures for fraudulent accounts[1]. However, each IO must spread to "infect" many users in order to be successful, independently of its aims and tools used to deceive. This goal often mandates extensive and coordinated social media efforts for the campaign to obtain a significant outreach, exert influence, and thus have an impact. In light of this consideration, since 2018, all leading platforms showed great interest in studying *coordinated inauthentic behavior* (CIB)[2]. Despite often appearing together, coordination and inauthenticity are two distinct concepts. For example, activists and other grassroots initiatives typically feature coordinated but authentic behaviors. Conversely, a single fake account manipulating the intent to mislead might exhibit inauthentic but uncoordinated behavior. The majority of existing efforts for studying CIB involved plenty of manual investigations, while the computational approaches are still few and far between. Among the challenges, there are the ambiguity and fuzziness of CIB itself: What exactly is a coordinated behavior? What is inauthentic behavior? How many organized

Alassad, Spann, amos 2017). IOs can take different shapes, target various individuals, online crowds, or communities, and have diverse goals (Starbird, Arif, and Wilson 2019). Among the strategic tools used by perpetrators are fake news, propaganda, hateful speech, as

Nizzoli, Leonardo, et al. "Coordinated behavior on social media in 2019 UK general election." Proceedings of the international AAAI conference on web and social media. Vol. 15. 2021.

Related Approaches

- TF-IDF is not the only approach for highlighting significant words
- Word Shift graphs use different options for computing text comparison, e.g., Shannon entropy, Kullback-Leibler divergence
- Check Monroe et al., 2008 for a nice overview
- Subtle differences in the results of different approaches

Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict

Burt L. Monroe
Department of Political Science, Quantitative Social Science Initiative, The Pennsylvania State University, e-mail: burtmonroe@psu.edu (corresponding author)

Michael P. Colaresi
Department of Political Science, Michigan State University, e-mail: colaresi@msu.edu

Kevin M. Quinn
Department of Government and Institute for Quantitative Social Science, Harvard University, e-mail: kevin_quinn@harvard.edu

Entries in the burgeoning "text-as-data" movement are often accompanied by lists or visualizations of how word (or other lexical feature) usage differs across some pair or set of documents. These are intended either to establish some target semantic concept (like the content of partisan frames) to estimate word-specific measures that feed forward into another analysis (like locating parties in ideological space) or both. We discuss a variety of techniques for selecting words that capture partisan, or other, differences in political speech and for evalu
size
several new
the
relative utility
speech
in the U.S. S

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16.4 (2008): 372-403.

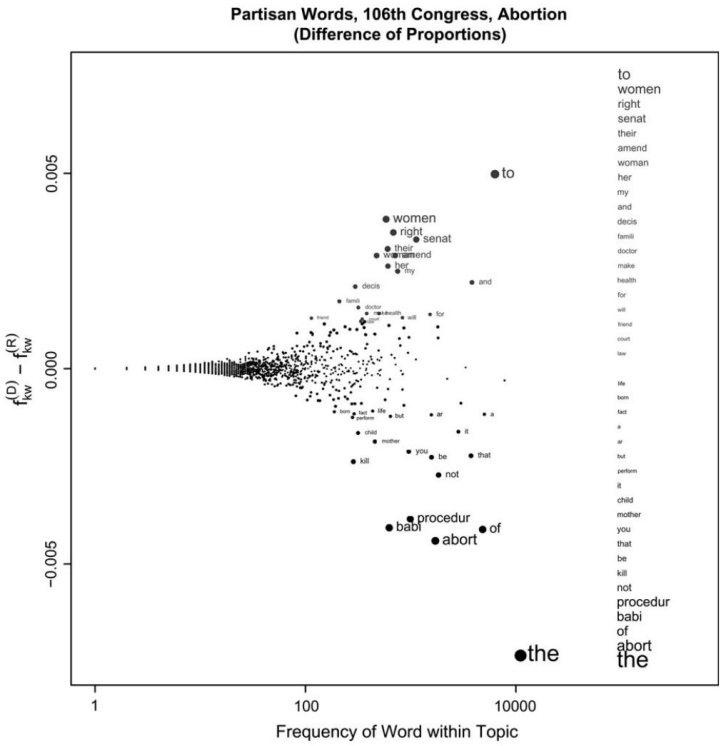


Fig. 1 Feature evaluation and selection using $f_{kw}^{(D)} - f_{kw}^{(R)}$. Plot size is proportional to evaluation weight, $|f_{kw}^{(D)} - f_{kw}^{(R)}|$. The top 20 Democratic and Republican words are labeled and listed in rank order to the right. The results are almost identical for two other measures discussed in the text: unlogged *tf.idf* and frequency-weighted WordScores.

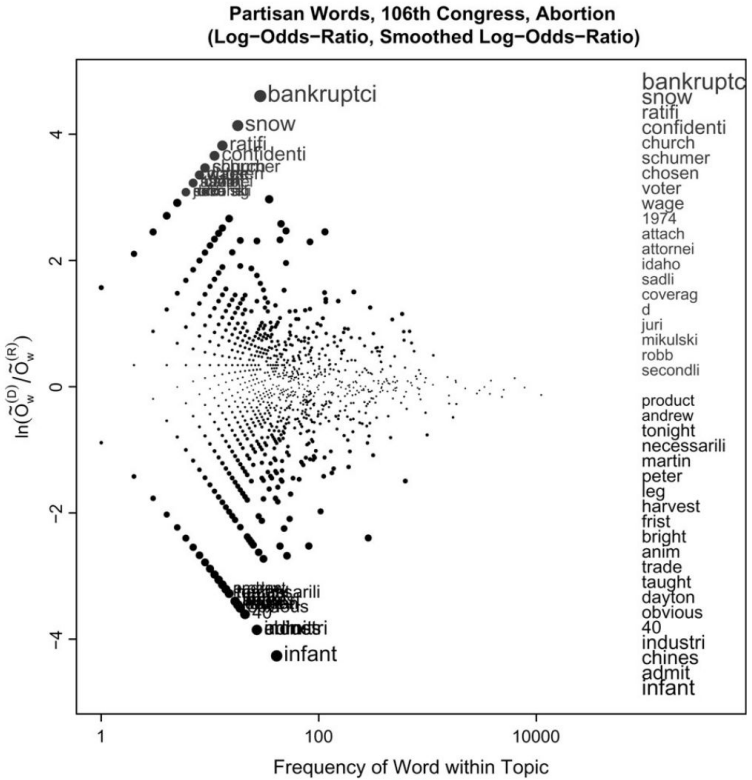


Fig. 2 Feature evaluation and selection using $\hat{\delta}_{kw}^{(D-R)}$. Plot size is proportional to evaluation weight, $|\hat{\delta}_{kw}^{(D-R)}|$. Top 20 Democratic and Republican words are labeled and listed in rank order. The results are identical to another measure discussed in the text: the log-odds-ratio with uninformative Dirichlet prior.

Data Visualization for Text

- Word clouds: easy to make, but often too vague.
- Variants of word clouds
  - Accentuated word clouds
  - Semantically grouped word clouds
- Word Shift Graphs
- Which ones should you use and which approach to find distinctive words?
  - Ideally multiple as robustness checks
- More complex:
  - Word trees[1]
  - Bubble charts[2]

General tips, especially for your reports and presentations

Bakhshi, Saeideh, David A. Shamma, and Eric Gilbert. "Faces engage us: Photos with faces attract more likes and comments on instagram." Proceedings of the SIGCHI conference on human factors in computing systems. 2014.

- Descriptive statistics:
  - Tables or figures showing summary of your data
  - especially the variables you are mainly studying (distributions, min, max, averages)

Type	Variable	Description	Distribution
Engagement	likes*	Number of likes on each photo.	
	comments*	Number of comments on each photo.	
Audience & Activity	followers*	Number of users who follow the photo's owner.	
	photos*	Number of photos shared by photo's owner.	
Faces	has face	1 if the photo contains a face, 0 otherwise.	
	has face $<$ 18 years old	1 if there is at least one face younger than 18, 0 otherwise.	
	has face $\in$ [18, 35] years old	1 if there is at least one face with age between 18 and 35, 0 otherwise.	
	has face $>$ 35 years old	1 if there is at least one face older than 35, 0 otherwise.	
	has female face	1 if there is at least one female face in the photo, 0 otherwise.	
	has male face	1 if there is at least one male face in the photo, 0 otherwise.	

Table 1. Distributions of quantitative and binary variables used in this paper. Variables marked with '*' are log-transformed. The red and blue lines identify mean and median of the distribution, respectively. Orange refers to 1's in the bar graphs. The engagement variables are our dependent measures. Audience and activity variables are used as controls, and faces variables are the focal point of this study.

General tips, especially for your reports and presentations

- Data visualization:
  - All axes should have labels
  - Captions should be self-contained
  - The text in the figure should be as big as the text in your paper
  - Using colors to convey information is nice, but also good to use markers or textures for this
  - Also keep in mind accessibility:
  - https://towardsdatascience.com/two-simple-steps-to-create-colorblind-friendly-data-visualizations-2ed781a167ec

Summary and Next Week

- This session we learned:
  - Exploratory Data Analysis
  - Deeper dive into Term Frequency-Inverse Document Frequency (TF-IDF)
  - Word Clouds, Word Shift Graphs, and other related visualizations

- Next time:
  - Text Analysis 1: supervised analysis
  - Readings:
    - Kevin Markham's Introduction to Linear Regression https://www.dataschool.io/linear-regression-in-python/
    - Mitra, Tanushree, and Eric Gilbert. "The language that gets people to give: Phrases that predict success on kickstarter." Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. 2014.

- Please register your teams: Project Registration IS 618 Social Media Data Analysis
  - Deadline: 16.04
  - Project pitches during the lecture slot on 21.04

Questions?