

# New Directions in Analyzing Text as Data 2023

University of Massachusetts Amherst

November 9–10, 2023

<https://tada2023.org>

**Thursday, November 9, 2023**

---

8:00am	Continental Breakfast Commonwealth Honors College Events Hall
9:00am	Welcome Remarks
9:15am	<b><i>Social Media in China &amp; Taiwan</i></b>  China's Global Voice: A Multimodal Analysis of Image and Text Propaganda Strategies. <i>Author:</i> Hannah Bailey (Oxford).  Narratives of Foreign Media Ecosystems in Chinese Social Media Discussions of the Russo-Ukrainian War. <i>Authors:</i> Hans Hanley (Stanford), Yingdan Lu (Northwestern), and Jennifer Pan (Stanford).  Characterizing the 2024 Taiwanese Presidential Elections on the News and Social Media. <i>Authors:</i> Sunny Fang (Barnard) and Herbert Chang (Dartmouth).  <i>Discussant:</i> Brandon Stewart (Princeton)
10:30am	Break
10:45am	<b><i>The Lifecycles of Text Data: A discussion on access, circulation, &amp; control</i></b>  Ethan Zuckerman. <i>Panelist.</i> Associate Professor of Public Policy, Communication, and Information (UMass Amherst); Founder, Institute for Digital Public Infrastructure.  Gregory Crane. <i>Panelist.</i> Professor and Winnick Family Chair in Technology & Entrepreneurship (Tufts University), Editor in Chief, Perseus Digital Library.  Laure Thompson, <i>Moderator.</i> Assistant Professor, UMass Amherst.
12:00pm	Lunch, including doctoral consortium mentoring.

1:00pm	<p><b><i>Cultural Analytics</i></b></p> <p>Topic, Genre, Race: Computational Approaches to the Lyrical History of Black-face Minstrelsy. <i>Authors:</i> Samuel Backer and Tom Lippincott (Johns Hopkins).</p> <p>Using Austen’s Characters to Evaluate Computational Pipelines for Literature. <i>Authors:</i> Funing Yang (Microsoft) and Carolyn Jane Anderson (Wellesley).</p> <p>Mapping Inventions in the Space of Ideas, 1836–2022: Representation and Measurement. <i>Authors:</i> Ina Ganguli (UMass), Jeff Lin (Federal Reserve Bank of Philadelphia), Vitaly Meursault (Federal Reserve Bank of Philadelphia), Nicholas Reynolds (Essex).</p> <p><i>Discussant:</i> Maria Antoniak (Allen Institute for AI)</p>
2:15pm	Break
2:30pm	<p><b><i>LLMs &amp; Annotation</i></b></p> <p>Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. <i>Authors:</i> Naoki Egami (Columbia), Musashi Hinck (Princeton), Brandon Stewart (Princeton), and Hanying Wei (Columbia).</p> <p>Mapping Subjectivity in Press Discourse: Human vs. Transformer Comprehension of Belgian French Press Articles. <i>Authors:</i> Louis Escoufflaire, Jérémie Bogaert, Antonin Descampe, and Cédric Fairon (UCLouvain).</p> <p>Automated Annotation with Generative AI Requires Validation. <i>Authors:</i> Nicholas Pangakis, Sam Wolken, and Neil Fasching (Penn).</p> <p><i>Discussant:</i> Katherine Keith (Williams College)</p>
3:45pm	Break & Move to Campus Center Hotel
4:15pm	<p><b><i>Poster Session I</i></b></p> <p>Amherst Room, Campus Center Hotel (Cash Bar)</p>
5:15pm	End of day, dinner on your own

## Friday, November 10, 2023

---

8:30am	Continental Breakfast Amherst Room, Campus Center Hotel
8:45am	<b><i>Poster Session II</i></b> Amherst Room, Campus Center Hotel
9:45am	Break & Walk to Commonwealth Honors College Events Hall
10:15am	<b><i>LLMs</i></b>  Large Language Models Can Argue in Convincing and Novel Ways About Politics: Evidence from Experiments and Human Judgement. <i>Authors:</i> Alexis Palmer (NYU) and Arthur Spirling (Princeton).  Applying Linguistic Theory to Large Language Models: Using Speech Act Theory to Understand How Language Causes Representational Harms. <i>Authors:</i> Emily Corvi, Stefanie Reed, Hannah Washington, Chad Atalla, Emily Sheng, Dan Vann, and Hanna Wallach (Microsoft).  How well do pretrained LLMs handle variant literary orthography? <i>Authors:</i> Craig Messner, Tom Lippincott (Johns Hopkins).  <i>Discussant:</i> Phil Resnik (Maryland)
11:30am	Break
11:45am	<b><i>Text Measurement and Application</i></b>  Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. <i>Authors:</i> Li Lucy (UC Berkeley), Jesse Dodge (Allen Institute), David Bamman (UC Berkeley), and Katherine Keith (Williams).  Understanding Personal Drug Experiences Shared on Reddit: A Multilabel Classification System for Addiction Recovery Research. <i>Authors:</i> Layla Bouzoubaa and Shadi Rezapour (Drexel).  From text to measure: Creating trustworthy measures of latent concepts from text with supervised machine learning. <i>Authors:</i> Ju Yeon Park (Ohio State) and Jacob Montgomery (Washington University in St. Louis).  <i>Discussant:</i> Kelsey Shoub (UMass Amherst)
1:00pm	Closing Remarks & Lunch A boxed lunch will be available. For those flying out earlier, it's approximately 55 minutes to Bradley Airport (Hartford).

**Poster Session I. Thursday, 4:15pm.**

*Farewell Victoria: Quantifying the Value of Foreign Names.* Zhaowen Guo (Washington); Chenyue Cao (Chicago)

*Are Women's Works and Claims Received with More Uncertainty?* Jina Lee (Arizona)

*You Are What You Annotate: Towards Better Models through Annotator Representations.* Naihao Deng, Siyang Liu, Xinliang Frederick Zhang, Winston Wu, Lu Wang, Rada Mihalcea (Michigan)

*Topics, Authors, and Networks in Large Language Model Research: Trends from 17K arXiv Papers.* Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, Emma Pierson (Cornell)

*Populism and Polarization in Comparative Perspective.* Yujin Julia Jung (Missouri)

*Decreasing the human coding burden in randomized trials with text-based outcomes via model-assisted impact analysis.* Reagan Mozer (Bentley), Luke Miratrix (Harvard)

*Evaluating the impact of lexical gender bias on gender bias measurement with word embeddings.* Shintaro Sakai (Nagoya University)

*Combining GPT Annotations and Vector Search Methods to Measure Obstructive Speech in Legislatures.* Mitchell Bosley (Michigan)

*Gendering the Subject in TED Talks.* K.M. Kinnaird (Smith), Allison J.B. Chaney (Researcher), John Laudun (Louisiana)

*Towards Personalistic Rule: Evidence from Textbooks in China.* Rosemary Pang (UMass Amherst), Yuehong Cassandra Tai (Penn State), Erico Yu (UMass Amherst)

*Safety Regulation in the U.S. Poultry Industry: For Whom?* Bridget Diana (UMass Amherst)

*A Monte Carlo Language Model Pipeline for Zero-Shot Sociopolitical Event Extraction.* Erica Cai (UMass Amherst), Brendan O'Connor (UMass Amherst)

*On nonprofits and neoliberalism: computational approaches to the study of ideology.* Isaac Dalke (UC Berkeley)

*Creating Custom Event Data Without Dictionaries: A Bag-of-Tricks.* Andrew Halterman (Michigan State), Philip A. Schrodtt (Parus Analytics), Andreas Beger (Basil Analytics), Benjamin E. Bagozzi (Delaware), Grace I. Scarborough (Leidos)

*Improving Mental Health Classifier Generalization with Pre-Diagnosis Data.* Yujian Liu (UC Santa Barbara), Laura Biester (Middlebury), Rada Mihalcea (Michigan)

*Inclusion-Moderation Hypothesis Reconsidered: Evidence from Erdogan's Political Speech.*

Humeyra Biricik (Oxford)

*Independencies in Causal Graphs with Text Variables: Embeddings vs. Frequencies.* Angus Li (Williams), Katie Keith (Williams)

*Manifesting a Coup: Cross-Website Extremism and Violence on the Dot Win Network Prior to January 6, 2021.* Alex Newhouse (Colorado), Meghan Conroy (Digital Forensics Research Lab)

*Cheep Talk: Investigating Corporate Social Responsibility Speech on Twitter.* Nile Phillips (Harvey Mudd), Michelle Lum (Harvey Mudd), Manisha Goel (Pomona), Michelle Zemel (Pomona), Alexandra Schofield (Harvey Mudd)

*Discovering Works in Historic and Hybrid Languages with Language Models.* Hale Sirin (Johns Hopkins), Ali Bolcakan (Michigan), Tom Lippincott (Johns Hopkins)

*Exploring shifts in women-directed stereotypes across Wikipedia time periods.* Cheryl Wang (Amazon), Carolyn Jane Anderson (Wellesley)

*The traces of repression. Recovery and processing of military archives produced during the Uruguayan dictatorship (1973-1985).* Elina Gomez (Universidad de la Republica de Uruguay)

*A Narrative Graph Approach for Analyzing Framing in News Articles.* Aditya Chandra (Colorado), Rohan Das (Colorado), Chih-Hao Fang (Purdue), Ananth Grama (Purdue), I-Ta Lee (META), Maria Leonor Pacheco (Colorado)

*LLMs for Few-shot Information Extraction with Climate Action Plans.* Nupoor Gandhi (Carnegie Mellon), Tom Corringham (UC San Diego), Emma Strubell (Carnegie Mellon University)

*TopicGPT: A Prompt-Based Framework for Topic Modeling.* Chau Minh Pham (UMass Amherst), Alexander Hoyle (Maryland), Simeng Sun (UMass Amherst), Mohit Iyyer (UMass Amherst)

*Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language.* Jimin Mun (Carnegie Mellon), Emily Allaway (Columbia), Akhila Yerukola (Carnegie Mellon), Laura Vianna (Washington), Sarah-Jane Leslie (Princeton), Maarten Sap (Carnegie Mellon)

*RIVETER: Measuring Power and Social Dynamics Between Entities.* Maria Antoniak (AI2), Anjalie Field (Johns Hopkins), Jimin Mun (Carnegie Mellon), Melanie Walsh (Washington), Lauren F. Klein (Emory), Maarten Sap (Carnegie Mellon)

*Lost in Translation: Using Language Models to Investigate Cultural Adaptation of Chinese Immigrants from Parallel Chinese Restaurant Names in the United States.* Hang Jiang (MIT), Nanxi Liu (Wellesley), Jad Kabbara (MIT), Deb Roy (MIT)

*The Language of US Partisan Newspapers from 1869 to 1925.* Si Wu (Northeastern), David Smith (Northeastern)

*The Operational Code of World Leaders: A New Comparison Sample Using the United Nations General Debate Corpus 1970-2021.* Sercan Canbolat (UConn), Baris Kesgin (Elon), Leah Windsor (Memphis)

*“One-size-fits-all”? Observations and Expectations of NLG Systems across Identity-related Language Features.* Li Lucy (UC Berkeley), Su Lin Blodgett (Microsoft), Milad Shokouhi (Microsoft), Hanna Wallach (Microsoft), Alexandra Olteanu (Microsoft)

*Decoding Disparities: A Textual Analysis of Clinical Notes Across Identity Attributes.* Afia Khan (Chicago)

*XiaoshuoNLP: An Computational Pipeline for Processing Chinese Literary Texts.* Kiara Meng Hui Liu (Cornell), Xiaomeng Zhu (Yale), Carolyn Jane Anderson (Wellesley)

*Exploring Ideal Point Estimation through Twitter Social Networks and Text Data.* Yicheng Shen, Alexander Volfovsky (Duke)

**Poster Session II.** Friday, 8:45am.

*GPT Detects Dissent Between Hawks and Doves in the Federal Open Market Committee.* Denis Peskoff (Princeton), Adam Visokay (Washington), Sander Schulhoff (Maryland), Benjamin Wachspress (Princeton), Alan Blinder (Princeton), Brandon Stewart (Princeton)

*Comparing morphosyntactic indicators of power in Classical Latin texts and their English translations.* Marisa Hudspeth, Brendan O'Connor, Laure Thompson (UMass Amherst)

*The Generalized Topic Model.* Elliott Ash, Germain Gauthier, Philine Widmer (PostDoc, ETH Zurich)

*T5 meets Tybalt: author attribution in Early Modern English drama using large language models.* Rebecca Hicke, David Mimno (Cornell)

*Using Cross-Encoder to Construct Sequences of Socio-Political Events.* Han Zhang, Zhiyu Li (The Hong Kong University of Science and Technology)

*Diachronic Syntactic and Semantic Shifts in Latin: Corpus, Method, and Relationships to Auerbach's Literary Scholarship.* Hale Sirin, Tom Lippincott (Johns Hopkins)

*A Computational Model of Selection and Framing in News about the US Economy.* Alexandria Leto (Colorado), Elliot Pickens (Wisconsin), Coen Needell (Penn), David Rothschild (Microsoft), Maria Leonor Pacheco (Colorado)

*On the Limits of Linguistic Answers to Social Problems.* Justine Zhang (Michigan)

*Excavating the Evolving Self in Autobiographies.* Avra Janz (New York University)

*Supply and demand: How social norms are constructed in Chinese-language job advertisements.* Anna Zhang (Northeastern)

*Using Large Language Models to detect the transmission of Darwinian Ideas.* Lucian Li (Illinois)

*Leveraging LLMs for cleaning text data with privacy constraints.* Andrea W. Wang (Cornell), Allison Koenecke (Cornell), David Mimno (Cornell)

*Russian War Propaganda – Evidence from Rossiyskaya Gazeta.* Yufan Yang (Augusta)

*A Computational Linguistic Measure of Anthropomorphism.* Myra Cheng, Kristina Gligoric, Tiziano Piccardi, Dan Jurafsky (Stanford)

*Tracing Accounts of Racial Terror in Historical Newspapers.* Shijia Liu, David A. Smith (Northeastern)

*An embarrassingly simple method for attributed network embedding.* Jacob A. Matthews, Imane Terhmina, Laurent Dubreuil, Marten van Schijndel (Cornell)

*International News Synchrony During the Start of the COVID-19 Pandemic.* Xi Chen (UMass Amherst), Scott Hale (Oxford), David Jurgen (Michigan), Mattia Samory (Sapienza University of Rome), Ethan Zuckerman (UMass Amherst), Przemyslaw Grabowicz (UMass Amherst)

*Inside the Echo Chamber: Leveraging Language Models to Measure Community Consensus.* Jeremiah Milbauer, Emma Strubell (Carnegie Mellon)

*What Do Readers Value? Manually Tagging and Quantitatively Measuring Aspects in Goodreads Reviews.* Maria Antoniak (AI2), Yujia Gao (Coinbase), Melanie Walsh (Washington), David Mimno (Cornell)

*Framing Social Movements on Social Media: Unpacking Diagnostic, Prognostic, and Motivational Strategies.* Julia Mendelsohn, Maya Vijan, Dallas Card, Ceren Budak (Michigan)

*Understanding social identity bias in large language models.* Tiancheng Hu, Yaroslava Kyrychenko, Sander van der Linden, Nigel Collier, Jon Roozenbeek (Cambridge)

*The Online #StopAsianHate Movement: More Global and BTS-Driven Than You'd Think.* Tessa Masis, Zhangqi Duan, Weiai Xu, Jonathan Corpus Ong, Ethan Zuckerman, Jane Pyo, Brendan O'Connor (UMass Amherst)

*Exploring GPT-4's Capabilities in Discourse Analysis for Aphasia: Advancing Language Treatment Outcomes through AI.* Vishnupriya Varadharaju, Jacquie Kurland, Anna Liu, Brendan O'Connor (UMass Amherst)

*Large Language Models Can Be Used to Estimate the Latent Positions of Politicians.* Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, Solomon Messing (New York University)

*Eliciting Ideological Scales from Large Language Models.* Sean O'Hagan, Aaron Schein (Chicago)

*Natural Language Decompositions of Implicit Content Enable Better Text Representations.* Alexander Hoyle (Maryland), Rupak Sarkar (Maryland), Pranav Goel (Northeastern), Philip Resnik (Maryland)

*Large Language Models in the Limelight: Testing their Effectiveness through Disinformation Detection.* Matyas Bohacek (Stanford), Michal Bravansky (University College London)

*Selection, appraisal, and provenance of pretraining datasets for pretrained language models.* Meera Desai, Dallas Card, Abigail Jacobs (Michigan)

*Measuring Diversity in Online News.* Samar Haider (Penn), David Rothschild (Microsoft), Chris Callison-Burch (Penn), Duncan Watts (Penn)

*Modeling Legal Reasoning: Annotation at the Edge of Human Agreement.* Rosamond Thalken, Edward H. Stiglitz, David Mimno, Matthew Wilkens (Cornell)