

TrustNLP 2023

**The Third Workshop on Trustworthy Natural Language  
Processing**

**Proceedings of the Workshop (TrustNLP 2023)**

July 14, 2023

The TrustNLP organizers gratefully acknowledge the support from the following sponsors.

**Gold**



©2023 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
209 N. Eighth Street  
Stroudsburg, PA 18360  
USA  
Tel: +1-570-476-8006  
Fax: +1-570-476-0860  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 978-1-959429-86-9

## Introduction

We welcome all participants of TrustNLP 2023, the third Workshop on Trustworthy Natural Language Processing. This year, we are embracing a hybrid format for the workshop, scheduled for July 14, 2023, and is co-located with ACL 2023.

Recent advances in Natural Language Processing, and the emergence of pretrained Large Language Models (LLM) specifically, have made NLP systems omnipresent in various aspects of our everyday life. In addition to traditional examples such as personal voice assistants, recommender systems, etc, more recent developments include content-generation models such as ChatGPT, text-to-image models (Dall-E), and so on. While these emergent technologies have an unquestionable potential to power various innovative NLP and AI applications, they also pose a number of challenges in terms of their safe and ethical use.

In response to these challenges, NLP researchers have formulated various objectives, e.g., intended to make models more fair, safe, and privacy-preserving. However, these objectives are often considered separately, which is a major limitation since it is often important to understand the interplay and/or tension between them. For instance, meeting a fairness objective might require access to users' demographic information, which creates tension with privacy objectives. The goal of this workshop is to move toward a more comprehensive notion of Trustworthy NLP, by bringing together researchers working on those distinct yet related topics, as well as their intersection.

Our agenda features four keynote speeches, a panel session, a presentation session, and a poster session. This year, we were delighted to receive 57 submissions, out of which 41 papers were accepted. Among these, 28 have been included in our proceedings. These papers span a wide array of topics including fairness, robustness, factuality, privacy, explainability, and model analysis in NLP.

We would like to express our gratitude to all the authors, committee members, keynote speakers, panelists, and participants. We also gratefully acknowledge the generous sponsorship provided by Amazon.

## Program Committee

### Organizers

Ninareh Mehrabi, Amazon Alexa AI  
Anaelia Ovalle, University of California Los Angeles  
Trista Cao, University of Maryland  
Jwala Dhamala, Amazon Alexa AI  
Apurv Verma, Amazon Alexa AI  
Anoop Kumar, Amazon Alexa AI  
Yada Pruksachatkun, Infinitus Systems  
Aram Galystan, University of Southern California, Amazon Alexa AI  
Rahul Gupta, Amazon Alexa AI  
Kai-Wei Chang, University of California Los Angeles, Amazon Alexa AI

### Program Committee

Griffin Adams, Columbia University  
Stefan Arnold, FAU Erlangen-Nürnberg  
Connor Baumler, University of Maryland  
Keith Burghardt, USC Information Sciences Institute  
Yang Trista Cao, University of Maryland  
Jwala Dhamala, Amazon Alexa AI-NLU  
Jacob Eisenstein, Google  
Katja Filippova, Google  
Aram Galstyan, USC Information Sciences Institute  
Umang Gupta, University of Southern California  
Devamanyu Hazarika, Amazon  
Zihao He, University of Southern California  
William Held, Georgia Tech  
Qian Hu, Amazon.com  
Fatimah Husain, Kuwait University  
Anoop Kumar, Amazon  
Sasha Luccioni, Hugging Face  
Pranav Narayanan Venkit, Pennsylvania State University  
Isar Nejadgholi, National Research Council Canada  
Aishwarya Padmakumar, Amazon  
Ashwinee Panda, Princeton University  
Anirudh Raju, Amazon, Alexa  
Anthony Rios, University of Texas at San Antonio  
Robik Shrestha, RIT  
Anna Sotnikova, University of Maryland  
Arjun Subramonian, University of California, Los Angeles  
Jialu Wang, University of California, Santa Cruz  
Chhavi Yadav, UCSD  
Kiyoon Yoo, Seoul National University

## Table of Contents

<i>Towards Faithful Explanations for Text Classification with Robustness Improvement and Explanation Guided Training</i>	
Dongfang Li, Baotian Hu, Qingcai Chen and Shan He . . . . .	1
<i>Driving Context into Text-to-Text Privatization</i>	
Stefan Arnold, Dilara Yesilbas and Sven Weinzierl . . . . .	15
<i>Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models</i>	
Pranav Narayanan Venkit, Mukund Srinath and Shomir Wilson . . . . .	26
<i>Pay Attention to the Robustness of Chinese Minority Language Models! Syllable-level Textual Adversarial Attack on Tibetan Script</i>	
Xi Cao, Dolma Dawa, Nuo Qun and Trashi Nyima . . . . .	35
<i>Can we trust the evaluation on ChatGPT?</i>	
Rachith Aiyappa, Jisun An, Haewoon Kwak and Yong-yeol Ahn . . . . .	47
<i>Improving Factuality of Abstractive Summarization via Contrastive Reward Learning</i>	
I-chun Chern, Zhiruo Wang, Sanjan Das, Bhavuk Sharma, Pengfei Liu and Graham Neubig . . . . .	55
<i>Examining the Causal Impact of First Names on Language Models: The Case of Social Commonsense Reasoning</i>	
Sullam Jeoung, Jana Diesner and Halil Kilicoglu . . . . .	61
<i>Reliability Check: An Analysis of GPT-3's Response to Sensitive Topics and Prompt Wording</i>	
Aisha Khatun and Daniel Brown . . . . .	73
<i>Sample Attackability in Natural Language Adversarial Attacks</i>	
Vyas Raina and Mark Gales . . . . .	96
<i>A Keyword Based Approach to Understanding the Overpenalization of Marginalized Groups by English Marginal Abuse Models on Twitter</i>	
Kyra Yee, Alice Schoenauer Sebag, Olivia Redfield, Matthias Eck, Emily Sheng and Luca Belli . . . . .	108
<i>An Empirical Study of Metrics to Measure Representational Harms in Pre-Trained Language Models</i>	
Saghar Hosseini, Hamid Palangi and Ahmed Hassan Awadallah . . . . .	121
<i>Linguistic Properties of Truthful Response</i>	
Bruce W. Lee, Benedict Florance Arockiaraj and Helen Jin . . . . .	135
<i>Debunking Biases in Attention</i>	
Shijing Chen, Usman Naseem and Imran Razzak . . . . .	141
<i>Guiding Text-to-Text Privatization by Syntax</i>	
Stefan Arnold, Dilara Yesilbas and Sven Weinzierl . . . . .	151
<i>Are fairness metric scores enough to assess discrimination biases in machine learning?</i>	
Fanny Jourdan, Laurent Risser, Jean-michel Loubes and Nicholas Asher . . . . .	163
<i>DEPTH+: An Enhanced Depth Metric for Wikipedia Corpora Quality</i>	
Saied Alshahrani, Norah Alshahrani and Jeanna Matthews . . . . .	175

<i>Distinguishing Fact from Fiction: A Benchmark Dataset for Identifying Machine-Generated Scientific Papers in the LLM Era.</i>	
Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci and Georg Groh .....	190
<i>Detecting Personal Information in Training Corpora: an Analysis</i>	
Nishant Subramani, Sasha Luccioni, Jesse Dodge and Margaret Mitchell .....	208
<i>Enhancing textual counterfactual explanation intelligibility through Counterfactual Feature Importance</i>	
Milan Bhan, Jean-noel Vittaut, Nicolas Chesneau and Marie-jeanne Lesot .....	221
<i>Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization</i>	
Oleksandr Yermilov, Vipul Raheja and Artem Chernodub .....	232
<i>GPTs Don't Keep Secrets: Searching for Backdoor Watermark Triggers in Autoregressive Language Models</i>	
Evan Lucas and Timothy Havens .....	242
<i>Make Text Unlearnable: Exploiting Effective Patterns to Protect Personal Data</i>	
Xinzhe Li and Ming Liu .....	249
<i>Training Data Extraction From Pre-trained Language Models: A Survey</i>	
Shotaro Ishihara .....	260
<i>Expanding Scope: Adapting English Adversarial Attacks to Chinese</i>	
Hanyu Liu, Chengyuan Cai and Yanjun Qi .....	276
<i>IMBERT: Making BERT Immune to Insertion-based Backdoor Attacks</i>	
Xuanli He, Jun Wang, Benjamin Rubinstein and Trevor Cohn .....	287
<i>On The Real-world Performance of Machine Translation: Exploring Social Media Post-authors' Perspectives</i>	
Ananya Gupta, Jae Takeuchi and Bart Knijnenburg .....	302
<i>Enabling Classifiers to Make Judgements Explicitly Aligned with Human Values</i>	
Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab and Pascale Fung ...	311
<i>Strength in Numbers: Estimating Confidence of Large Language Models by Prompt Agreement</i>	
Gwenyth Portillo Wightman, Alexandra Delucia and Mark Dredze .....	326

# Program

## Friday, July 14, 2023

- 09:00 - 09:10     *Opening Remarks*
- 09:10 - 09:50     *Keynote 1 Hal Daume III*
- 09:50 - 10:30     *Keynote 2 Ramprasaath R. Selvaraju*
- 10:30 - 11:00     *Break*
- 11:00 - 11:40     *Keynote 3 Rachel Rudinger*
- 11:40 - 12:30     *Poster Session*
- 12:30 - 14:00     *Lunch*
- 14:00 - 14:40     *Keynote 4 Sunipa Dev*
- 14:40 - 15:30     *Panel Discussion*
- 15:30 - 16:00     *Break*
- 16:00 - 17:10     *Oral Presentation*
- 17:10 - 17:30     *Closing Session/Best Paper Announcement*