Third Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)

Toronto, Canada
14 July 2023

ISBN: 978-1-7138-8236-7

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2023) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Towards Faithful Explanations for Text Classification with Robustness Improvement and Explanation Guided Training
Dongfang Li, Baotian Hu, Qingcai Chen and Shan He 1

Driving Context into Text-to-Text Privatization
Stefan Arnold, Dilara Yesilbas and Sven Weinzierl 15

Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models
Pranav Narayanan Venkit, Mukund Srinath and Shomir Wilson 26

Pay Attention to the Robustness of Chinese Minority Language Models! Syllable-level Textual Adversarial Attack on Tibetan Script
Xi Cao, Dolma Dawa, Nuo Qun and Trashi Nyima 35

Can we trust the evaluation on ChatGPT?
Rachith Aiyappa, Jisun An, Haewoon Kwak and Yong-yeol Ahn 47

Improving Factuality of Abstractive Summarization via Contrastive Reward Learning
I-chun Chern, Zhiruo Wang, Sanjan Das, Bhavuk Sharma, Pengfei Liu and Graham Neubig . . 55

Examining the Causal Impact of First Names on Language Models: The Case of Social Commonsense Reasoning
Sullam Jeoung, Jana Diesner and Halil Kilicoglu 61

Reliability Check: An Analysis of GPT-3's Response to Sensitive Topics and Prompt Wording
Aisha Khatun and Daniel Brown 73

Sample Attackability in Natural Language Adversarial Attacks
Vyas Raina and Mark Gales 96

A Keyword Based Approach to Understanding the Overpenalization of Marginalized Groups by English Marginal Abuse Models on Twitter
Kyra Yee, Alice Schoenauer Sebag, Olivia Redfield, Matthias Eck, Emily Sheng and Luca Belli 108

An Empirical Study of Metrics to Measure Representational Harms in Pre-Trained Language Models
Saghar Hosseini, Hamid Palangi and Ahmed Hassan Awadallah 121

Linguistic Properties of Truthful Response
Bruce W. Lee, Benedict Florance Arockiaraj and Helen Jin 135

Debunking Biases in Attention
Shijing Chen, Usman Naseem and Imran Razzak 141

Guiding Text-to-Text Privatization by Syntax
Stefan Arnold, Dilara Yesilbas and Sven Weinzierl 151

Are fairness metric scores enough to assess discrimination biases in machine learning?
Fanny Jourdan, Laurent Risser, Jean-michel Loubes and Nicholas Asher 163

DEPTH+: An Enhanced Depth Metric for Wikipedia Corpora Quality
Saied Alshahrani, Norah Alshahrani and Jeanna Matthews 175

Distinguishing Fact from Fiction: A Benchmark Dataset for Identifying Machine-Generated Scientific Papers in the LLM Era.
Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci and Georg Groh . . . 190

Detecting Personal Information in Training Corpora: an Analysis
Nishant Subramani, Sasha Luccioni, Jesse Dodge and Margaret Mitchell . . . 208

Enhancing textual counterfactual explanation intelligibility through Counterfactual Feature Importance
Milan Bhan, Jean-noel Vittaut, Nicolas Chesneau and Marie-jeanne Lesot . . . 221

Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization
Oleksandr Yermilov, Vipul Raheja and Artem Chernodub . . . 232

GPTs Don't Keep Secrets: Searching for Backdoor Watermark Triggers in Autoregressive Language Models
Evan Lucas and Timothy Havens . . . 242

Make Text Unlearnable: Exploiting Effective Patterns to Protect Personal Data
Xinzhe Li and Ming Liu . . . 249

Training Data Extraction From Pre-trained Language Models: A Survey
Shotaro Ishihara . . . 260

Expanding Scope: Adapting English Adversarial Attacks to Chinese
Hanyu Liu, Chengyuan Cai and Yanjun Qi . . . 276

IMBERT: Making BERT Immune to Insertion-based Backdoor Attacks
Xuanli He, Jun Wang, Benjamin Rubinstein and Trevor Cohn . . . 287

On The Real-world Performance of Machine Translation: Exploring Social Media Post-authors' Perspectives
Ananya Gupta, Jae Takeuchi and Bart Knijnenburg . . . 302

Enabling Classifiers to Make Judgements Explicitly Aligned with Human Values
Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab and Pascale Fung . . . 311

Strength in Numbers: Estimating Confidence of Large Language Models by Prompt Agreement
Gwenyth Portillo Wightman, Alexandra Delucia and Mark Dredze . . . 326