JIAZHAO LI

(+1)734-604-1596, jiazhaol@umich.edu
Linkedin, Google Scholar, Personal Website

EDUCATION

University of Michigan, Ann Arbor U.S.
Ph.D. in Informatics (Natural Language Processing) *2020 – 2024*
M.S. in Electrical Computer Engineering (Computer Vision) *2017 – 2019*
Nankai University China
B.S. in Electrical Engineering *2013 – 2017*

RESEARCH INTEREST

Natural Language Processing & CyberSecurity & Health Informatics
Robustness of NLP models (Attack and Defense Strategy), Large Language Model.

WORK EXPERIENCE

Research Scientist Intern (**Yahoo Research**) May.2023 - Aug.2023
Research Associate (**Michigan Medicine**) Aug.2019 - Aug.2020

CONFERENCE PAPER

Li, J, Yang Y., Wu, Z., Vydiswaran, V. G., Xiao, C. *ChatGPT as an Attack Tool: Stealthy Textual Backdoor Attack via Blackbox Generative Model Trigger* ***(Under Review)*** [pdf]
Li, J, Wu, Z., Ping, W., Xiao, C., Vydiswaran, V. G. *Defending against Insertion-based Textual Backdoor Attacks via Attribution* ***(Findings of ACL'23)*** [pdf]
Li, J, Lester C., Zhao X., Ding Y., Jiang Y., and Vydiswaran, V. G. *PharmMT: A Neural Machine Translation Approach to Simplify Prescription Directions.* ***(Findings of EMNLP'20)*** [pdf]
Li, J, Murali A., Mei Q., Vydiswaran, V. G. *Re-ranking biomedical literature for precision medicine with pre-trained neural models. (**ICHI'20**)*[pdf]

JOURNAL PAPER

Li, J., Vydiswaran VGV. et al. Accelerating Theme Analysis on clinical tele-visit narrows via Active Learning *(Under Review).*
Lester, C.A., **Li, J.**, Ding, Y. et al. Performance evaluation of a prescription medication image classification model: an observational cohort. *Nature Partner Journals Digit. Med. 4, 118 (2021).*[pdf]
Lester CA, Ding Y, **Li J**, Jiang Y, Rowell B, Vydiswaran VGV, Comparing Human versus Machine Translation of Electronic Prescription Directions *Journal of the American Pharmacists Association (2021)* [pdf]
Chang T, DeJonckheere M, Vydiswaran VGV, **Li J**, Buis L, Guetterman T. Accelerating Mixed Methods Research with Natural Language Processing of Big Text Data. *Journal of Mixed Methods Research (2021).*[pdf]

RESEARCH EXPERIENCE

Robustness of LLM under backdoor attack.
In Process with Yahoo *Sep 2023 - Now*

- Applied backdoor attack during instruction fine-tuning and task-specific fine-tuning against LLaMA-2 7B via Parameter-Efficient Fine-Tuning, LoRA.
- Check Attack Transferability of LLM under same label space.
- Check Attack Transferability of backdoor attack under different label spaces among different tasks.

Defending against Insertion-based Textual Backdoor Attacks via Attribution
Findings of ACL'23 *Feb 2022 - Sep 2022*

- Build a defense framework against backdoor attacks on text classifier (pre-training and post-training)
- Apply a poisoned sample detector ELECTRA to identify poisoned samples.
- Identify triggers by calculating the attribution score of tokens using Partial LRP (trigger word contributes most to mislabeling)
- Achieve SOTA performance, an average accuracy of 79.97% (56.59%↑) and 48.34% (3.99%↑) on 4 benchmarks against pre-training attack and post-training attack respectively.

ChatGPT as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger
Under Review *Nov 2022 - Jan 2023*

- Propose a stealthy, input-dependent backdoor attack method to mislead textual classifiers utilizing paraphrasing models (ChatGPT, mBART, BART) as LM-based triggers, making the generated backdoor examples less noticeable for human cognition.
- ChatGPTAttack is easily accessible and avoids detected by both GPT-detection and defense methods.

PharmMT: A Neural Machine Translation Approach to Simplify Prescription Directions.
Findings of EMNLP'20 *Sept 2019 - Feb 2020*

- Built Seq-to-Seq Text Simplification Model between parallel prescription and pharmacy directions corpus using OpenNMT framework.
- Archive 60.27 BLEU score against pharmacists' reference and 94.3% of the simplified directions could be used as-is or with minimal changes evaluated by pharmacists.

Open-domain Aspects Exploration for Qualitative Analysis via Active Learning
Under Review *Feb 2020 - Sep 2022*

- Build a framework to explore diverse aspects of selected theme (open-domain classification task)
- Use keyword-based filtering and binary text-classifier to collect the relevant sentence-level corpus.
- Select 'difficulty' samples (on classifier decision boundary) to the label instead of random sampling to accelerate diverse aspect exploration.

Re-ranking biomedical literature for precision medicine with pre-trained neural models.
ICHI'20 *Jan 2019 - May 2019*

- TREC precision medicine information retrieval challenge on ontology topics. Combining two relevant score using Rank Fusion.
- 6.2% improvement on inferred NDCG and 6.8% improvement on R-precision against SOTA models .

PageRank++: European Soccer Team Ranking Prediction [Report] Jan. 2018 - Apr. 2018
Best poster of Graph Data Mining Course Project *University of Michigan*

· Designing PageRank++ algorithm to make a prediction on team ranking in the league before the season begin through re-defined directed graph of team.

- Applied K-mean to classify soccer player into 4 groups based on scores in 33 features in 5 fields of performance.
- Used PageRank iteration to predict rank where nodes representing feature vector of scores of players component for each team and edges of graph representing probability of win (directed).

PUBLIC SERVICE

- PC / Reviewer: ACL '23, EMNLP '23'22'21, NAACL '21, EACL '22
- External Journal Reviewer: Frontiers in Big Data, section Cybersecurity and Privacy.