Limitations and Vulnerabilities of LLMs

Penn State University, University of Mississippi

Thai Le, Ph.D.

IARPA BENGAL / October 24, 2023

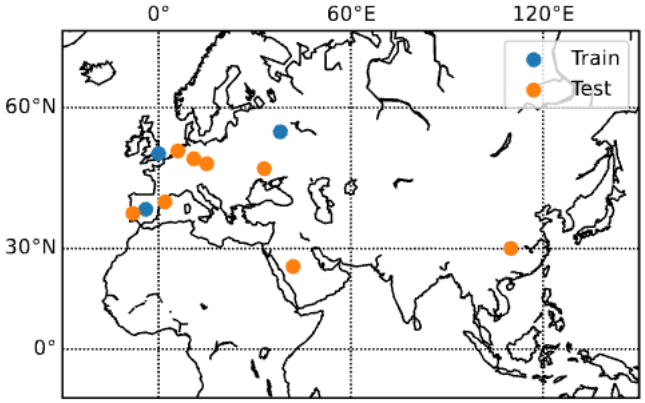
Penn State + U. Mississippi

- Looking for a BEGAL team to collaborate with!
- PIs with 16 Ph.D. students (11@PSU, 5@UM) and National Research Lab
  - https://pike.psu.edu/, https://lethaiq.github.io/,
- Active research in Data Science and AI areas
- # pubs in top CS venues for last 3 years
  - AI: AAAI (6), AAMAS (2)
  - NLP: ACL (3), EMNLP (6), NAACL (1)
  - DS: KDD (6), ICDM (4), ICDE (1)
  - Web/IR: WWW (4), CIKM (4), SIGIR (1)
  - HCI: CHI (2), CSCW (1), ICWSM (4)

Team's Expertise

- **TuringBench** (EMNLP'20, 21), **MULTITuDE** (EMNLP'23), **Hansen** (EMNLP'23): benchmark environments to study neural authorship attribution problem in both written and spoken texts and *multilingual* neural text detection

AA Model	P	R	F1	Accuracy
Random Forest	0.5893	0.6053	0.5847	0.6147
SVM (3-grams)	0.7124	0.7223	0.7149	0.7299
WriteprintsRFC	0.4578	0.4851	0.4651	0.4943
OpenAI detector	0.7810	0.7812	0.7741	0.7873
Syntax-CNN	0.6520	0.6544	0.6480	0.6613
N-gram CNN	0.6909	0.6832	0.6665	0.6914
N-gram LSTM-LSTM	0.6694	0.6824	0.6646	0.6898
BertAA	0.7796	0.7750	0.7758	0.7812
BERT-Multinomial	0.8031	0.8021	0.7996	0.8078
RoBERTa-Multinomial	**0.8214**	**0.8126**	**0.8107**	**0.8173**



Train	Test: **en**	Test: **non-en**	Difference
en	0.9292	0.6903	↓ 25.7%

Team's Expertise

- **Human Detection of LLM texts (HCOMP'23):** human evaluation of neural text detection
- **Fighting Fire with Fire (EMNLP'23):** crafting and detecting misinformation with LLMs
- **Do Language Models Plagiarize? (WWW'23):** investigating plagiarism behaviors of LLMs
- **UPTON (EMNLP'23):** privacy leakage of human identities from LLMs trained on social media
- **SHIELD, ANTHRO, CrypText, NoisyHate (ACL'22, ICDE'23)**: attacking and defending LLMs against machine-generated and human-written perturbations in the wild

Contact Info

- Dongwon Lee (PSU)
  - https://pike.psu.edu/
  - dongwon@psu.edu

- Thai Le (UMississippi)
  - https://lethaiq.github.io/tql3/
  - thaile@olemiss.edu