Neural Authorship Attribution and Obfuscation

The Pennsylvania State University

Dongwon Lee, Ph.D.

dongwon@psu.edu

HIATUS / Jan 19, 2022

Penn State PIKE Team

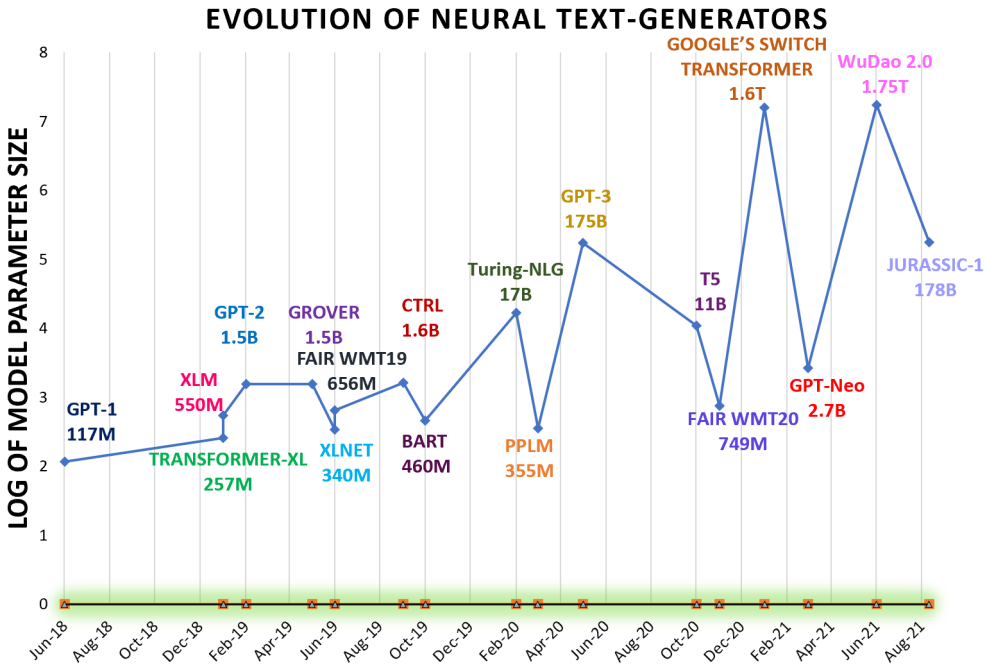
- Looking for a HIATUS team to join !

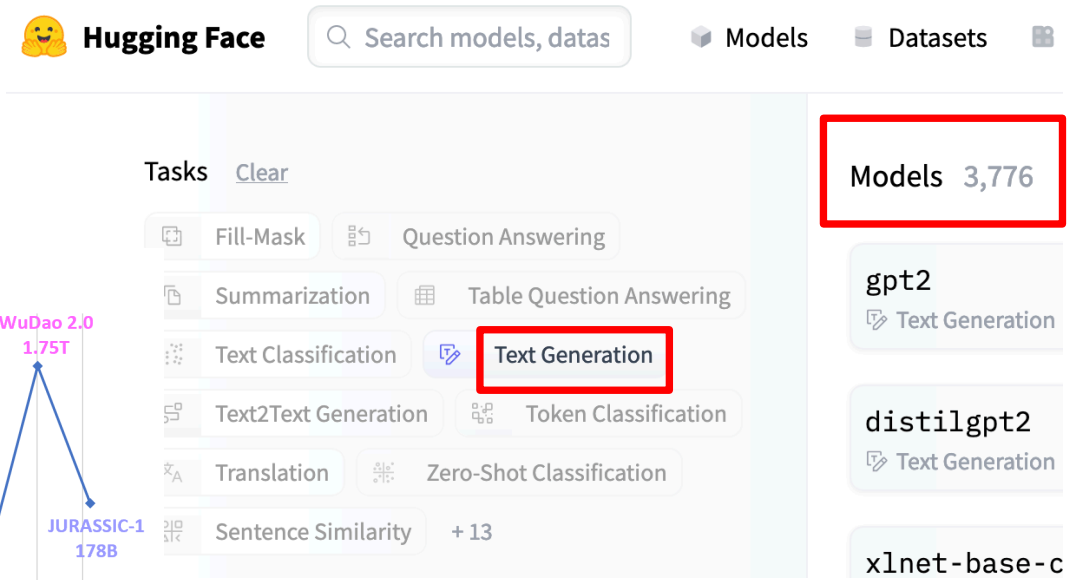
- PI with 11 Ph.D. students
  - https://pike.psu.edu/
- Active research in Data Science and AI areas
- # pubs in top CS venues for last 3 years
  - *Data Science*: KDD (6), ICDM (4), WWW (4), CIKM (4), SIGIR (1), ICDE (1)
  - *AI*: AAAI (4), AAMAS (1)
  - *NLP*: ACL (1), EMNLP (2), NAACL (1)
  - *HCI*: CHI (2), CSCW (1)

Old Problem, New Spin !

Neural Authorship

~~Authorship~~ Attribution and Obfuscation





Team's Expertise

HIATUS TA1 & TA2

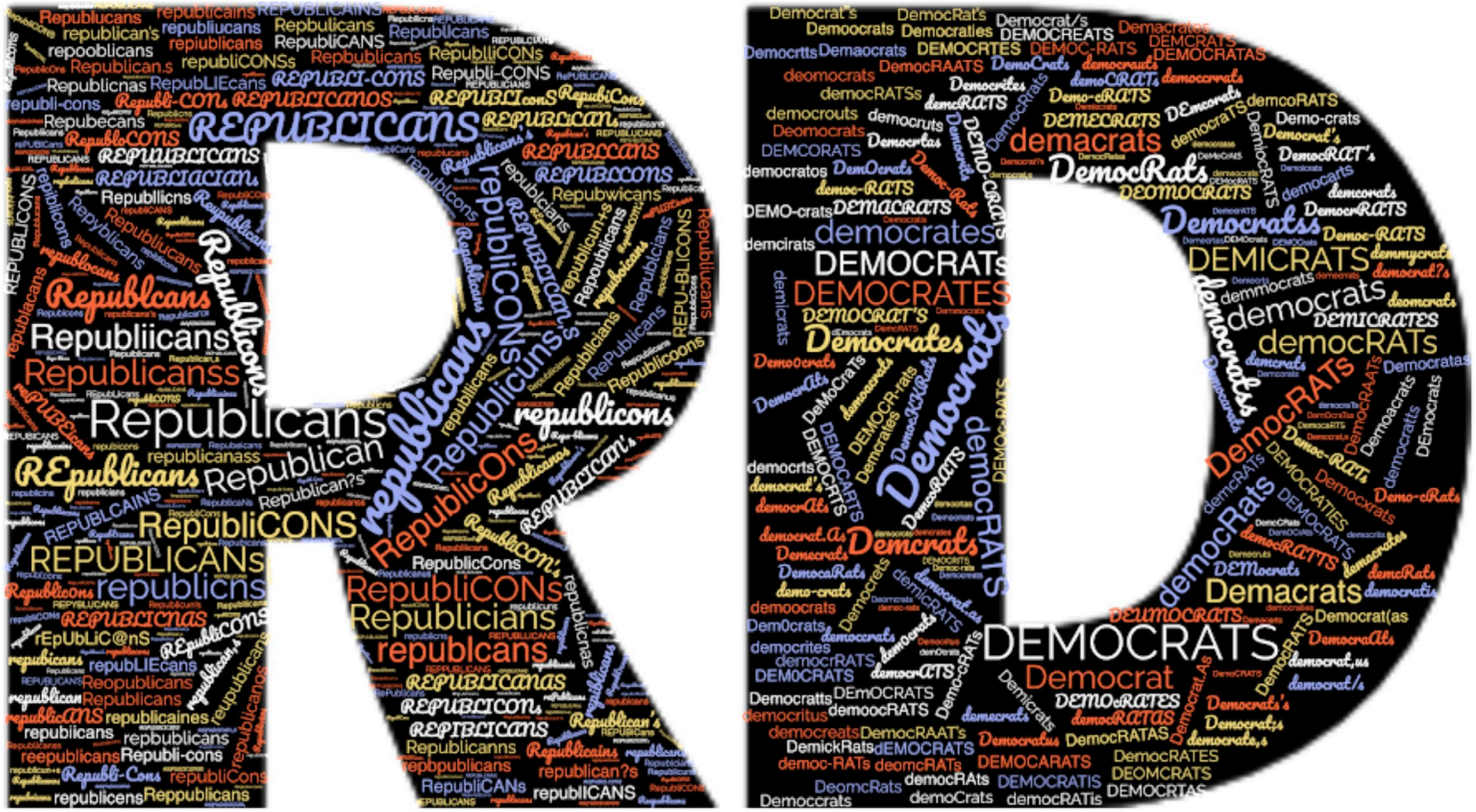
- EMNLP 2020 & 2021
  - **TuringBench**: public benchmark environment to study neural Authorship Attribution (AA) problem
    - Created 20 corpus on news genre (200K)—ie, 19 generated by language models and 1 human-written
  - Compared 10 AA detection models

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**

Team's Expertise

HIATUS TA3

- **Authorship Obfuscation (AO) in the Wild**
  - What do real human-written AO look like?
  - What are AO less likely to be detected?



Contact Information

Dongwon Lee, Ph.D.

dongwon@psu.edu

https://pike.psu.edu/