STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Interpretable and Explainable AI for

Prof. K.P (Suba) Subbalakshmi, FNAI

Dept. of E.C.E, Stevens Institute of Technology

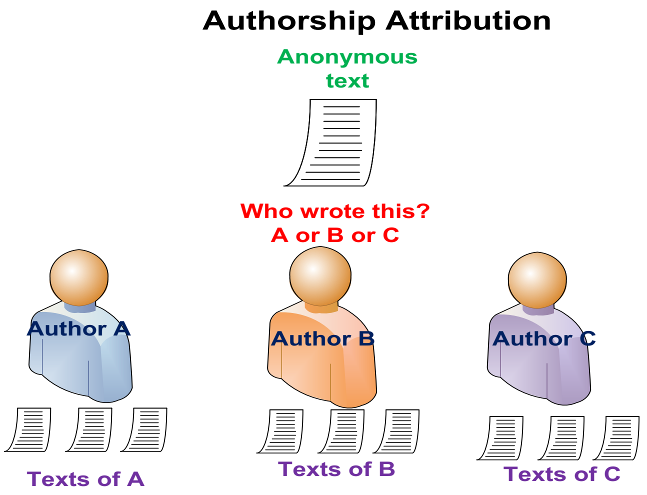
Jefferson Science Fellow

ksubbala@stevens.edu

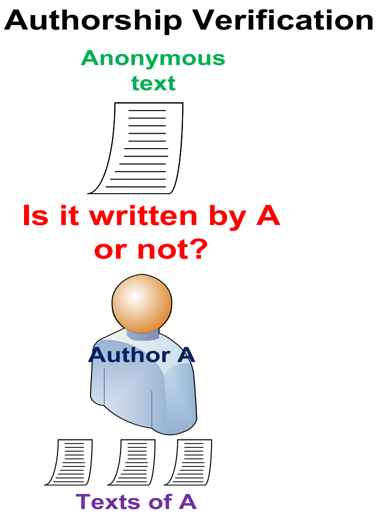
Https://www.kpsuba.com

About the PI

- Prof. with over 21 years of experience
- Fellow, National Academy of Inventors, 2018
- NJ Inventors Hall of Fame Award for work on deception detection from text
- NASEM's Science and Technology Experts Group for ODNI

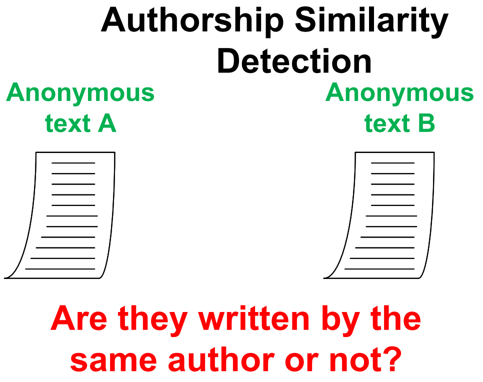


Authorship Attribution



Authorship Verification

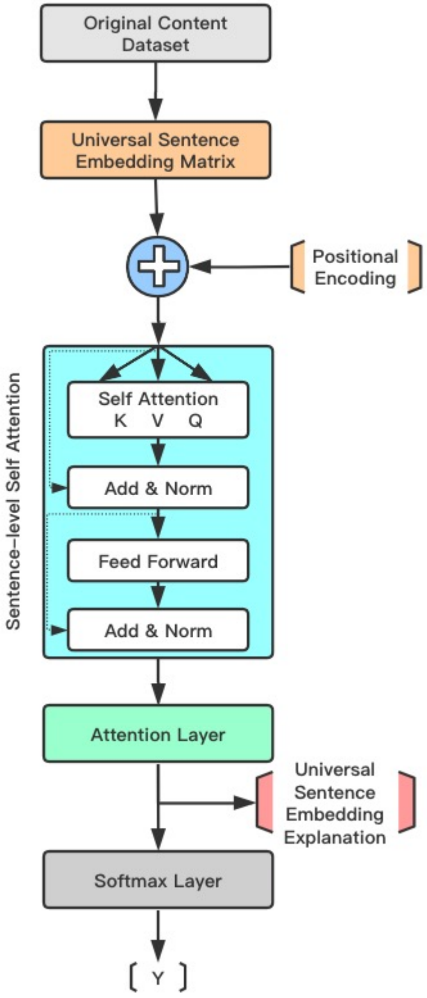
- Interpretable and Explainable NLP
- Fake News Detection
- Gender identification



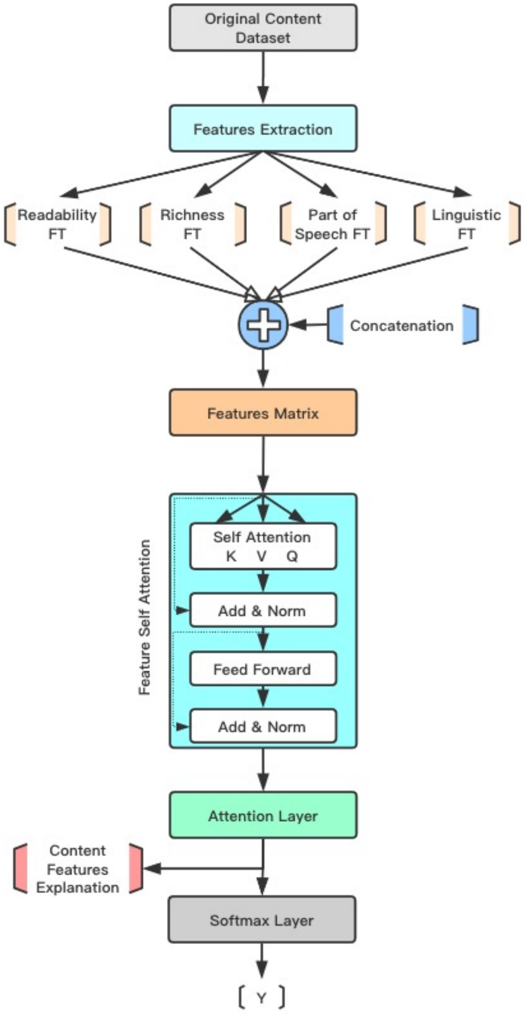
Authorship Similarity Detection

Example Interpretable Architectures

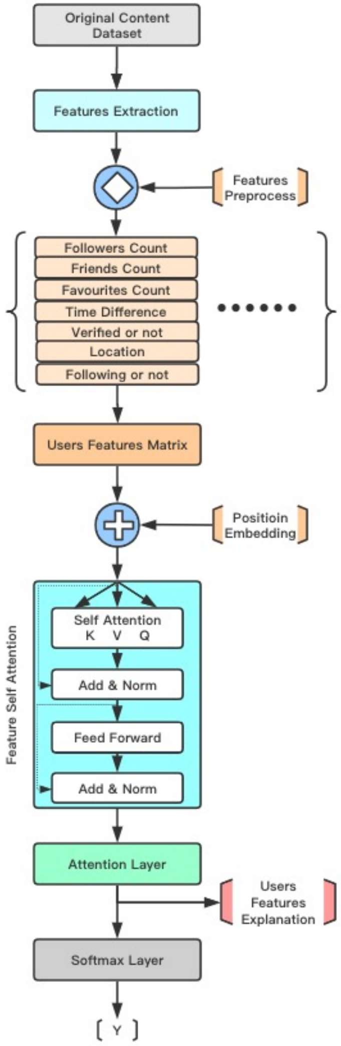
Universal Sentence Embedding Attention Network



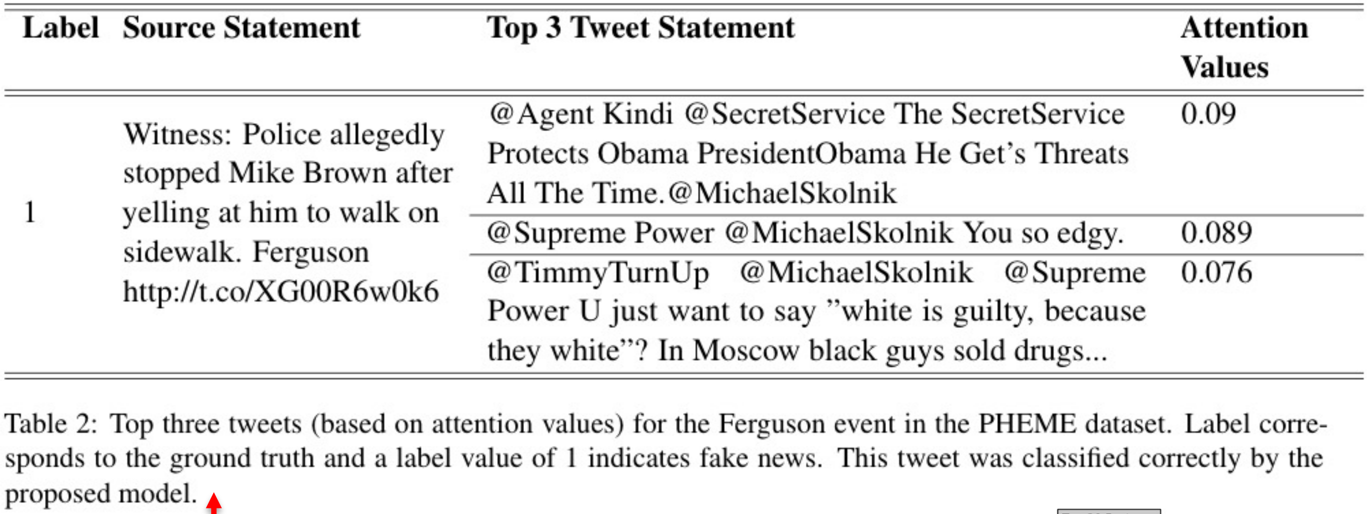
Content Features Attention Network



Users Features Attention Network



Example Explanations

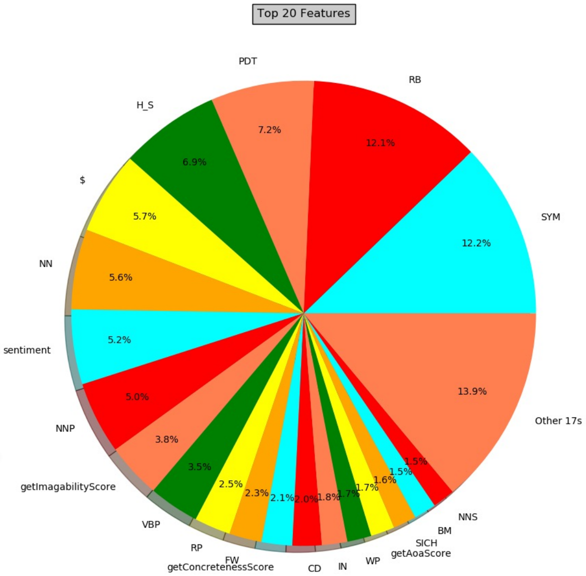


Label	Source Statement	Top 3 Tweet Statement	Attention Values
1	Witness: Police allegedly stopped Mike Brown after yelling at him to walk on sidewalk. Ferguson http://t.co/XG00R6w0k6	@Agent Kindi @SecretService The SecretService Protects Obama PresidentObama He Get's Threats All The Time.@MichaelSkolnik	0.09
		@Supreme Power @MichaelSkolnik You so edgy.	0.089
		@TimmyTurnUp @MichaelSkolnik @Supreme Power U just want to say ”white is guilty, because they white”? In Moscow black guys sold drugs...	0.076

Table 2: Top three tweets (based on attention values) for the Ferguson event in the PHEME dataset. Label corresponds to the ground truth and a label value of 1 indicates fake news. This tweet was classified correctly by the proposed model.

Tweet level explanations

Content feature level explanations



Example Results

Goal	Performance
Authorship identification	Over 90% accuracy
Fake News Detection	Beats SOTA performance while also providing layers of explanation
Gender identification	Accuracy between 75% and 85%

Sample Relevant Publications and Patents

- Mingxuan Chen, Ning Wang, K. P. Subbalakshmi, "Explainable Rumor Detection using Inter and Intra-feature Attention Networks", TrueFact KDD Workshop, 2020 [

- Ning Wang, Mingxuan Chen, K. P. Subbalakshmi, “Explainable CNN-attention Networks (C-Attention Network) for Automated Detection of Alzheimer's Disease”, BioKDD, 2020.

- Mingxuan Chen, Xinqiao Chu and K.P. Subbalakshmi, "MMCoVaR: Multimodal COVID-19 Vaccine Focused Data Repository for Fake News Detection and a Baseline Architecture for Classification", ASONAM 2021

- Constanine Boyadjiev, R. Chandramouli, K.P. Subbalakshmi, Zongru Shao, "Machine learning for authenticating voice", US Patent 10,593,336, 2020.

- Constantine Boyadijev, R. Chandramouli, K.P. Subbalakshmi and Zongru (Doris) Shao, "Natural Language Processing Artificial Intelligence Network and Data Security System", US Patent No: , December 24, 2019.

- R. Chandramouli, Xiaoling Chen, K.P. Subbalakshmi and R. Perera, “Systems and methods for automatically detecting deception in human communications expressed in digital form", US Patent 9292493, PCT/US11/033936, Awarded March 22, 2016

- R. Chandramouli, X. Chen and K.P. Subbalakshmi, “Psycho-linguistic statistical deception detection from text content”, PCT/US11/020390, US Patent Number 9116877, Issue Date: August 25, 2015.

- More at https://www.kpsuba.com

1870

STEVENS

INSTITUTE of TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu