STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Using Explainable Architectures for Reason Extraction from Documents

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

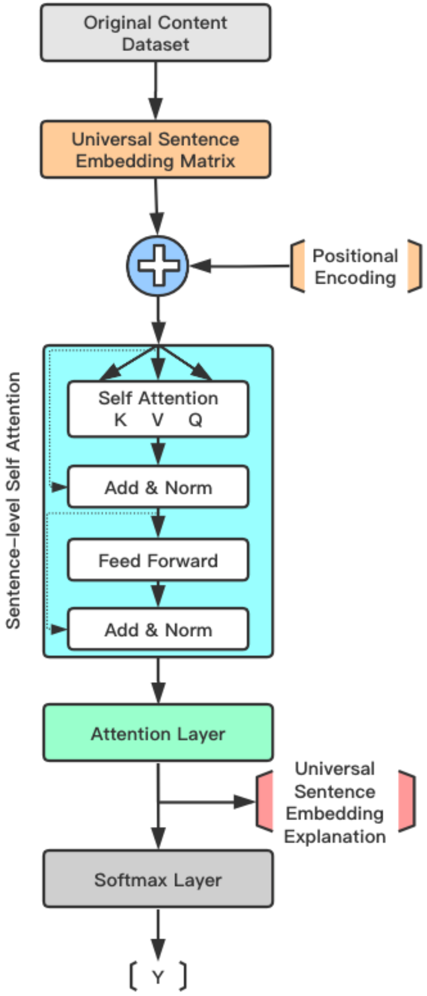
- Prof. with over 22 years of experience
- Fellow, National Academy of Inventors, 2018
- NASEM’s Science and Technology Experts Group for ODNI
- Technical expertise:
  - Explainable AI
    - Bingyang Wen, K.P. Subbalakshmi and Fan Yang, "Revisiting Attention Weights as Explanations from an Information Theoretic Perspective”, **NeurIPS** 2022
    - Bingyang Wen, K.P. Subbalakshmi and R. Chandramouli, “Revealing the Roles of Part-of-Speech Taggers in Alzheimer’s Disease Detection: A Scientific Discovery Using One-intervention Causal Explanation”, **JMIR**, Frontier Research, accepted, 2022
    - 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.
  - Causality and generative models
    - Bingyang Wen, Yupeng Cao, Fan Yang, K.P. Subbalakshmi, R. Chandramouli, "Causal-TGAN: Modeling Tabular Data Using Causally-Aware GAN ", **ICLR 2022** Workshop on Deep Generative Models for Highly Structured Data.

Technical Expertise (Contd.)

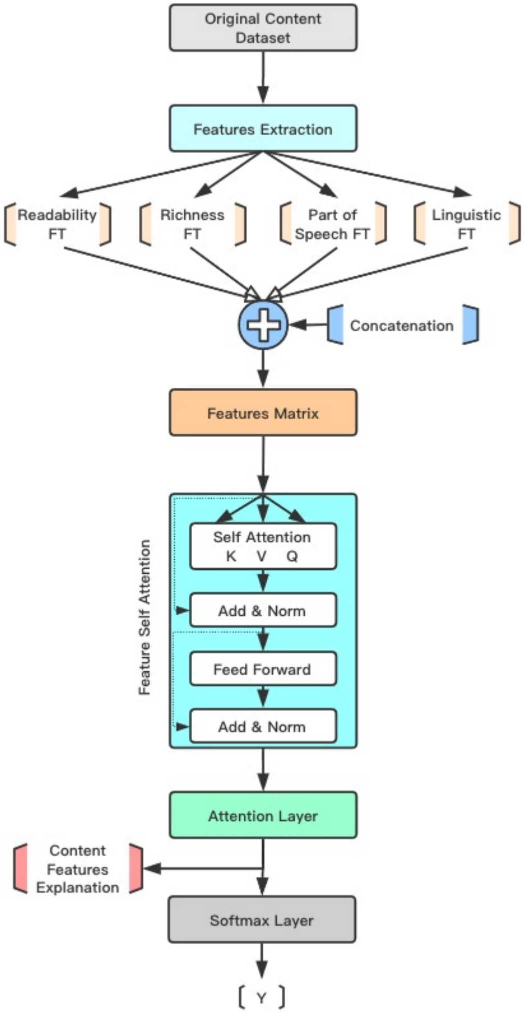
- Technical expertise:
  - Explainable Fake News Detection
    - Mingxuan Chen, Xinqiao Chu and K.P. Subbalakshmi, "MMCoVaR: Multimodal COVID-19 Vaccine Focused Data Repository for Fake News Detection and a Baseline Architecture for Classification", ASONAM 2021
Causality and generative models
    - Harish Sista and K.P. Subbalakshmi, ”Fake News Identification by Extracting Relevant Information from Verified Publications”, under preparation
    - Mingxuan Chen, Yupeng Cao and K.P. Subbalakshmi, “REIHAN: Relevant Information Enhanced Hierarchical Attention Network for Automated Claim Verification” under preparation

Example Explainable Architectures

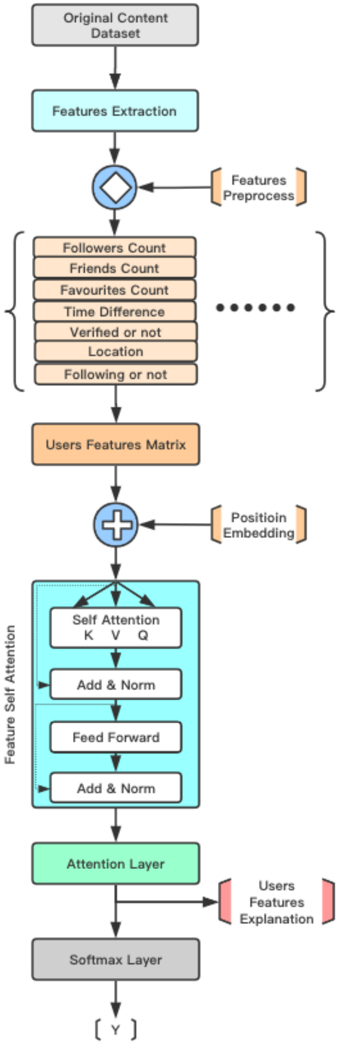
Universal Sentence Embedding Attention Network



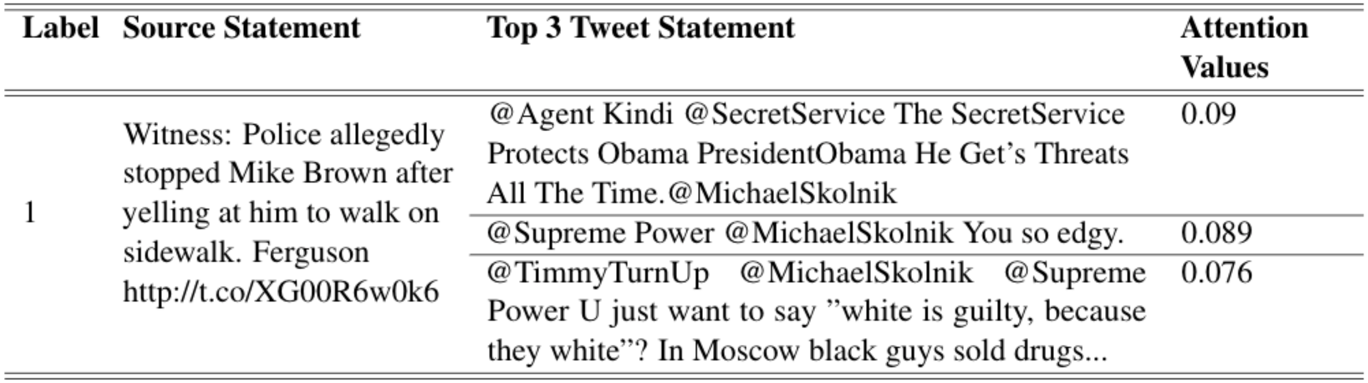
Content Features Attention Network



Users Features Attention Network



Example Explanations

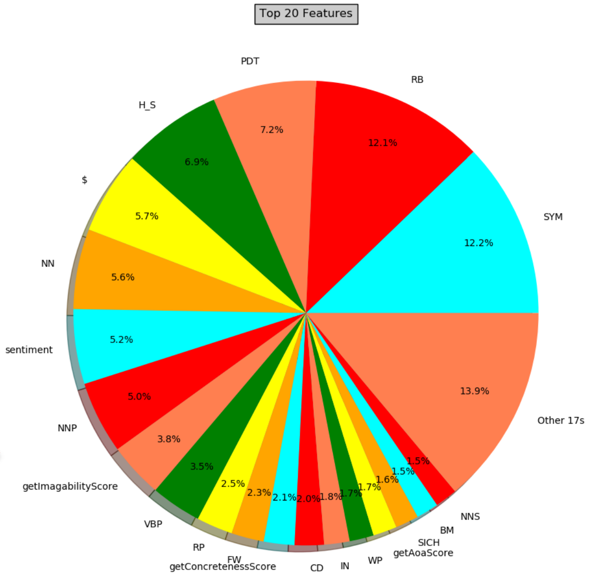


Label	Source Statement	Top 3 Tweet Statement	Attention Values
1	Witness: Police allegedly stopped Mike Brown after yelling at him to walk on sidewalk. Ferguson http://t.co/XG00R6w0k6	@Agent Kindi @SecretService The SecretService Protects Obama PresidentObama He Get's Threats All The Time.@MichaelSkolnik	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



1870

STEVENS

INSTITUTE *of* TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu