Transparent Authorship Verification

Stephanie Wan
Mentored by Gil Alterovitz and Ning Xie

MIT PRIMES October Conference 10/12/24

(Same) Authorship Verification

- Determine if two texts have the same author

- Counter
  - Misinformation
  - Plagiarism
  - Harassment
  - Impersonation
  - Criminal activities
  - Falsified text
- Privacy & Anonymity

For Martin Luther King Jr. weekend, the Knudsons and the Myers are going up north to a cross-country ski resort called Maplelag. This is a yearly tradition and I won't have any new posts to my blog until at least Sunday night (I'll probably have it updated on Monday maybe). Well I hope all of you have a great three-day weekend and have tons of fun!

one of my exes got a tattoo while i was dating him, it was probably the stupidest one i've ever seen. he wanted superman (big deal, right?).. only he didn't want superman the way everyone else wanted superman. he wanted the whole superman. he got it. it looked.. silly. to be honest, it was one of the poorest attempts at being different that i've ever seen. 'It's like.. instead of saying, 'i like superman', you're saying, 'wow, i really like that super man'.'

For Martin Luther King Jr. weekend, the Knudsons and the Myers are going up north to a cross-country ski resort called Maplelag. This is a yearly tradition and won't have any new posts to my blog until at least Sunday night (I'll probably have it updated on Monday maybe). Well I hope all of you have a great three-day weekend and have tons of fun!

Different

one of my exes got a tattoo while i was dating him, it was probably the stupidest one i've ever seen. he wanted superman (big deal, right?).. only he didn't want superman the way everyone else wanted superman. he wanted the more superman. he got it. it looked.. silly. to be honest, it was one of the poorest attempts at being different that i've ever seen. 'It's like.. instead of saying, 'i like superman', you're saying, 'wow, i really like that super man'.'

Authorship Verification

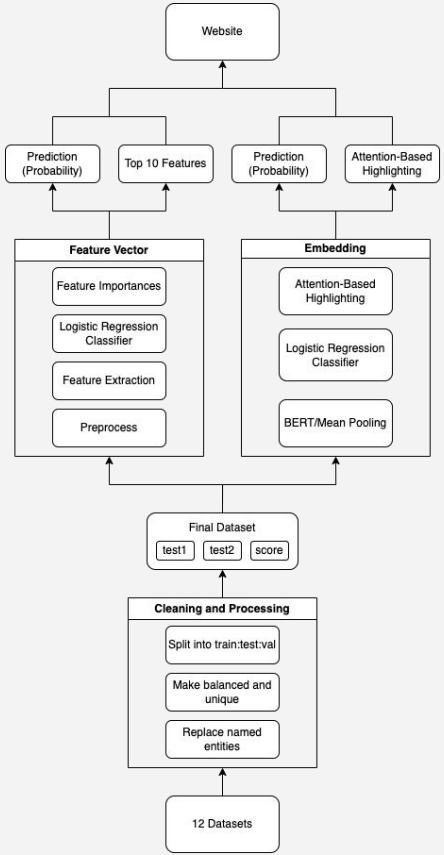
- Lexical/Linguistics
  - Language
  - Words
  - Morphology, syntax, phonetics, and semantics
  - Transformers/LLM
- Stylometry
  - Statistical variations in writing style
    - Features
      - We'll get back to this!

Topical information :(

Authorship Verification

~~Topical information :(~~

Training data diversity!


Website
Prediction (Probability)
Top 10 Features
Prediction (Probability)
Attention-Based Highlighting
Feature Vector
Feature Importances
Logistic Regression Classifier
Feature Extraction
Preprocess
Embedding
Attention-Based Highlighting
Logistic Regression Classifier
BERT/Mean Pooling
Final Dataset
test1
test2
score
Cleaning and Processing
Split into train:test:val
Make balanced and unique
Replace named entities
12 Datasets

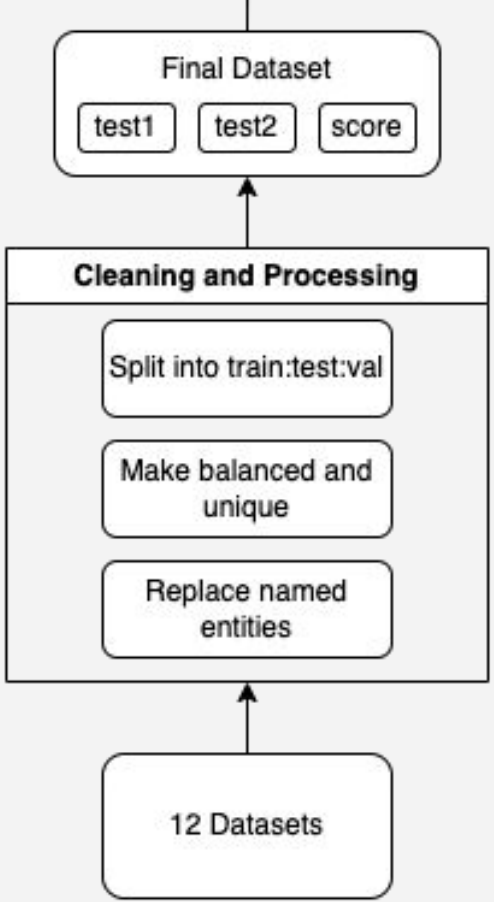
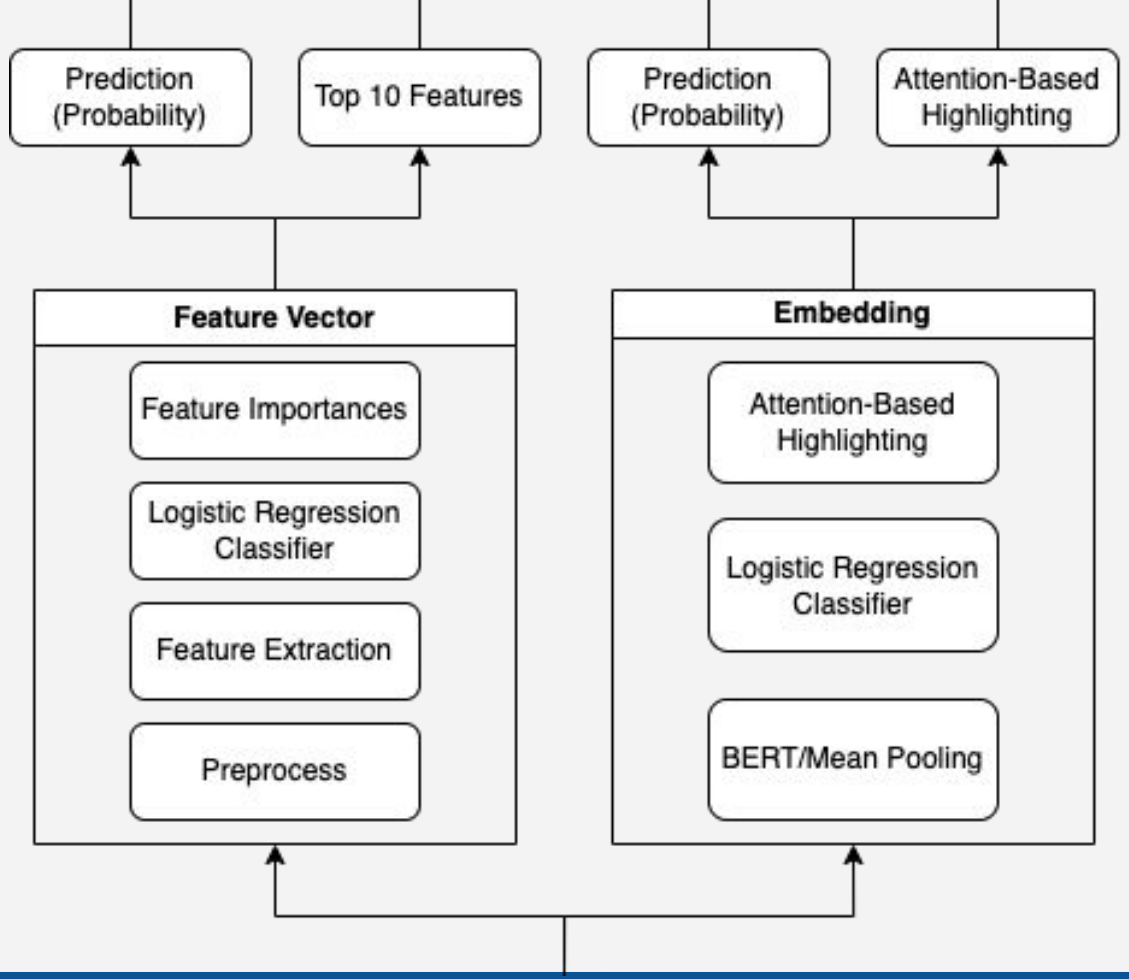

Final Dataset
test1
test2
score
Cleaning and Processing
Split into train:test:val
Make balanced and unique
Replace named entities
12 Datasets

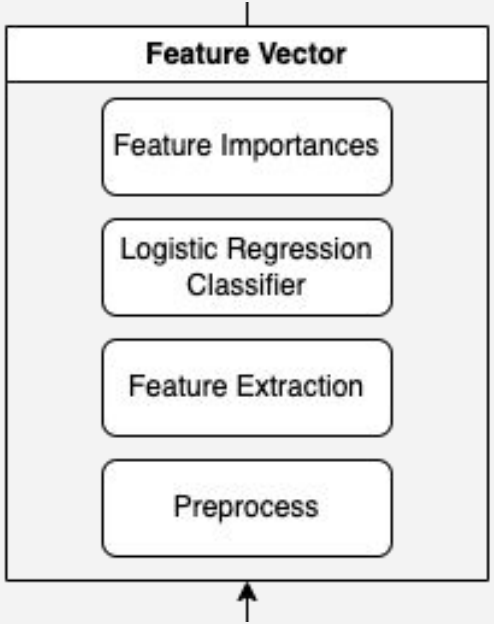
Table 1: Summary of AV and AA datasets used

Dataset	Text Form	Pairs Used*	Average Chars**	Length	Formality	Named Entities
Reuters	News Articles	1202	2770	Medium	Formal	Replaced
Blogs	Blog Posts	58930	1086	Short	Informal	Replaced
Victorian	Book Excerpts	10718	4923	Long	Formal	Replaced
arXiv	Paper Abstracts	704	803	Short	Formal	Replaced
DarkReddit	Reddit Comments	1028	2751	Medium	Informal	Replaced
BAWE	Student Writing	1150	14702	Long	Formal	Replaced
IMDB62	Movie Reviews	30982	1668	Short	Informal	Replaced
PAN11	Enron Emails	4650	300	Short	Mix	Replaced
PAN13	Various	120	7143	Long	Mix	Replaced
PAN14	Novels and Essays	900	15843	Long	Formal	Not Replaced
PAN15	Various	1265	3167	Medium	Mix	Not Replaced
PAN20	Fanfiction	275409	21473	Long	Informal	Not Replaced

*The amount of pairs of text from this dataset that were in the final compiled dataset
*Average amount of characters in each text used in the final compiled dataset
**General length of each text in the dataset

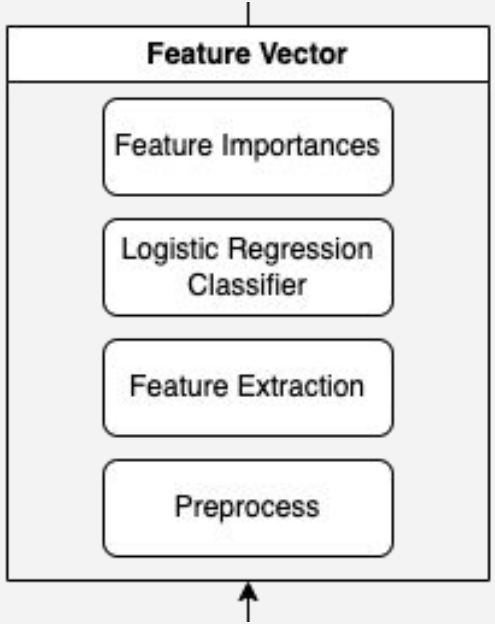

Prediction (Probability)
Top 10 Features
Prediction (Probability)
Attention-Based Highlighting
Feature Vector
Feature Importances
Logistic Regression Classifier
Feature Extraction
Preprocess
Embedding
Attention-Based Highlighting
Logistic Regression Classifier
BERT/Mean Pooling

Feature Vector



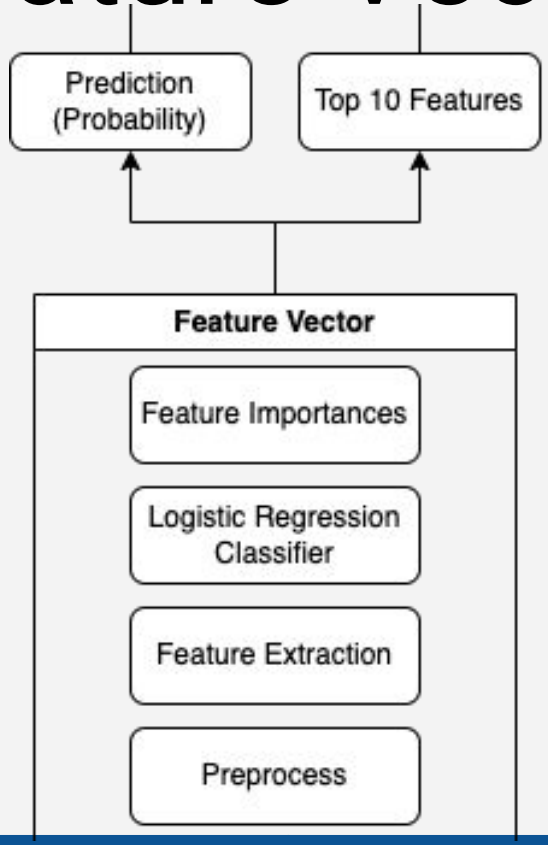
- PAN2021 - Weerasinghe et al.
- (Large - 3rd, Small - 1st)
- Binary classification
  - Same or not
- Input to classifier is a feature vector

Feature Vector



- Preprocessing
  - Tokenization
  - Part-of-Speech (POS) Tagging
  - POS Tag Chunking
- Features Extracted
  - Stylometric
    - Character n-grams
    - Average number of characters per word
    - Distribution of word-lengths, Vocabulary Richness
    - Spelling

Feature Vector



Feature Vector

happy birthday to you happy birthday to you happy birthday!!!

twinkle twinkle little star / how i wonder what you are up / above the world so high / like a diamond in the sky

Feature Vector

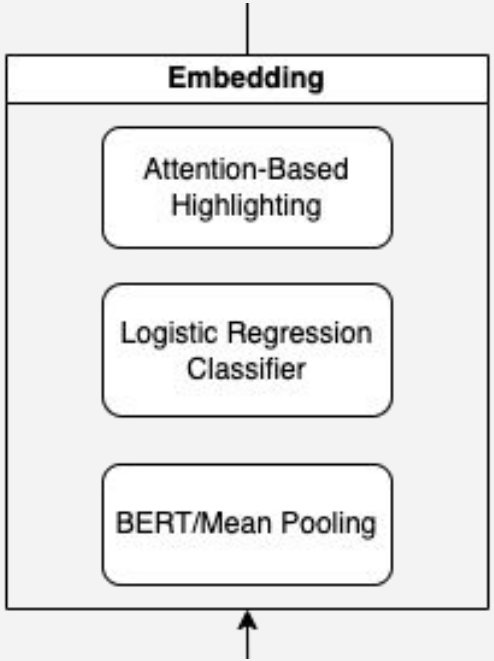
Top 10 Features

- above: 31.6957
- hap: 22.8684
- bov: 16.7269
- ppy: 16.5476
- across the NN: 13.8071
- NP JJ NP: 12.6481
- PRP JJ: 12.6132
- app: 11.7819
- rld: 10.9437
- py: 10.5696

Probability Same Author (Feature Vector)

0.8171

Embedding


Embedding
Attention-Based Highlighting
Logistic Regression Classifier
BERT/Mean Pooling

Embedding

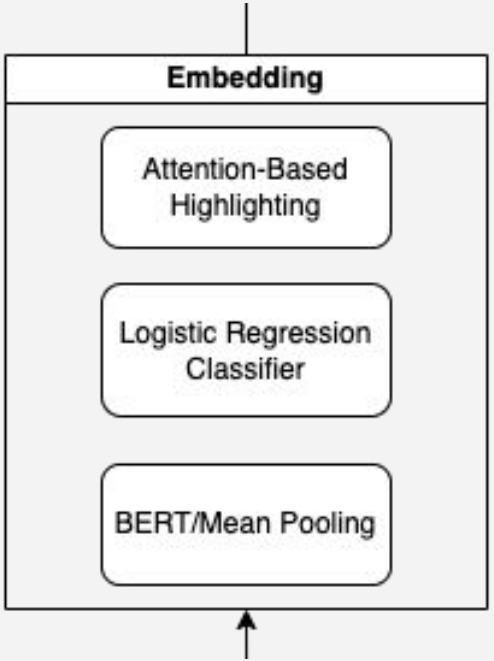
BERT

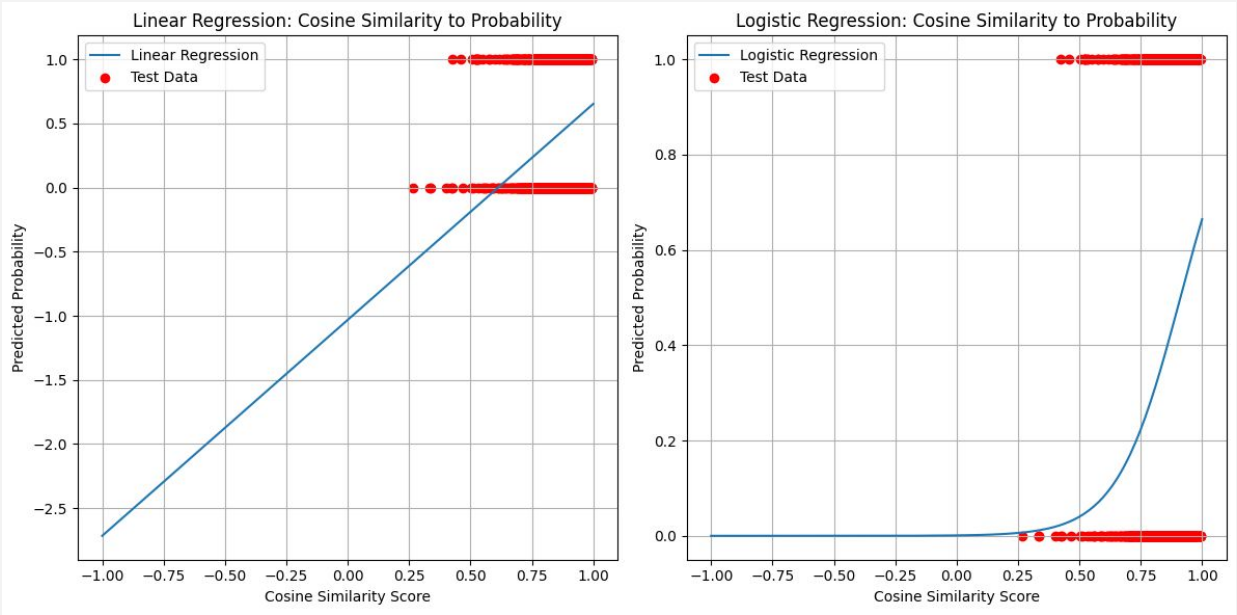
- Prasad and Chakkaravarthy (2022)
- Siamese network
  - Sub-networks
  - neural network that uses the same weights while working on two different input vectors
- Contrastive Loss

Embedding

~~BERT~~ -> mean pooling

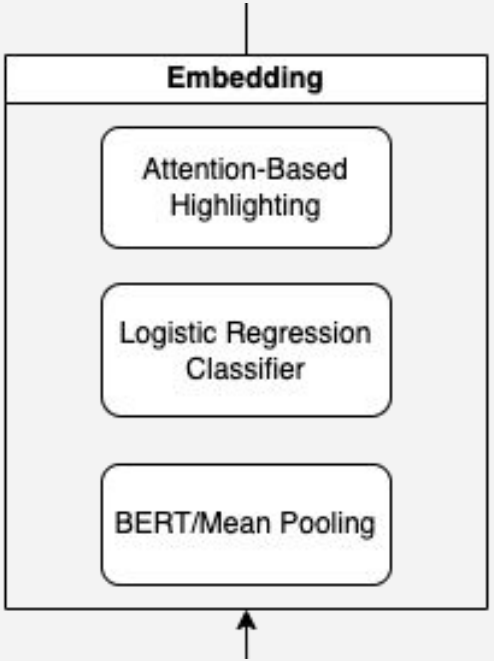
Embedding





Cosine score: 0.9169589281082153
Probability that the output is 1: 0.5107

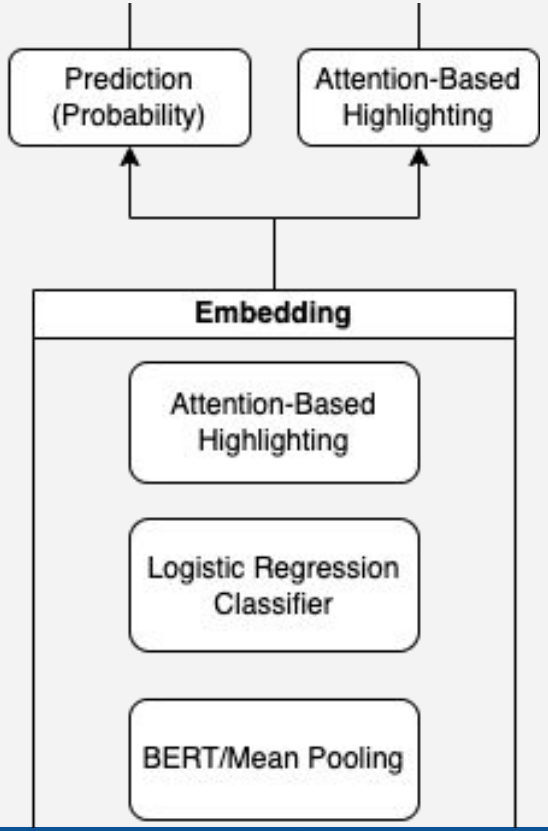
Embedding



The Cauchy problem for a coupled PRODUCT and PERSON system is shown to be globally well-posed for a class of data without finite energy. The proof uses the I-method introduced by ORG, Keel, PERSON, GPE, and PERSON.

The C ##au ##chy problem for a coupled PR ##OD ##UC ##T and P ##ER ##SO ##N system is shown to be globally well - posed for a class of data without finite energy . The proof uses the I - method introduced by OR ##G , Ke ##el , P ##ER ##SO ##N , GP ##E , and P ##ER ##SO ##N .

Embedding


Prediction
(Probability)
Attention-Based
Highlighting
Embedding
Attention-Based
Highlighting
Logistic Regression
Classifier
BERT/Mean Pooling

Embedding

happy birthday to you happy birthday to you happy birthday!!!

twinkle twinkle little star / how i wonder what you are up / above the world so high / like a diamond in the sky

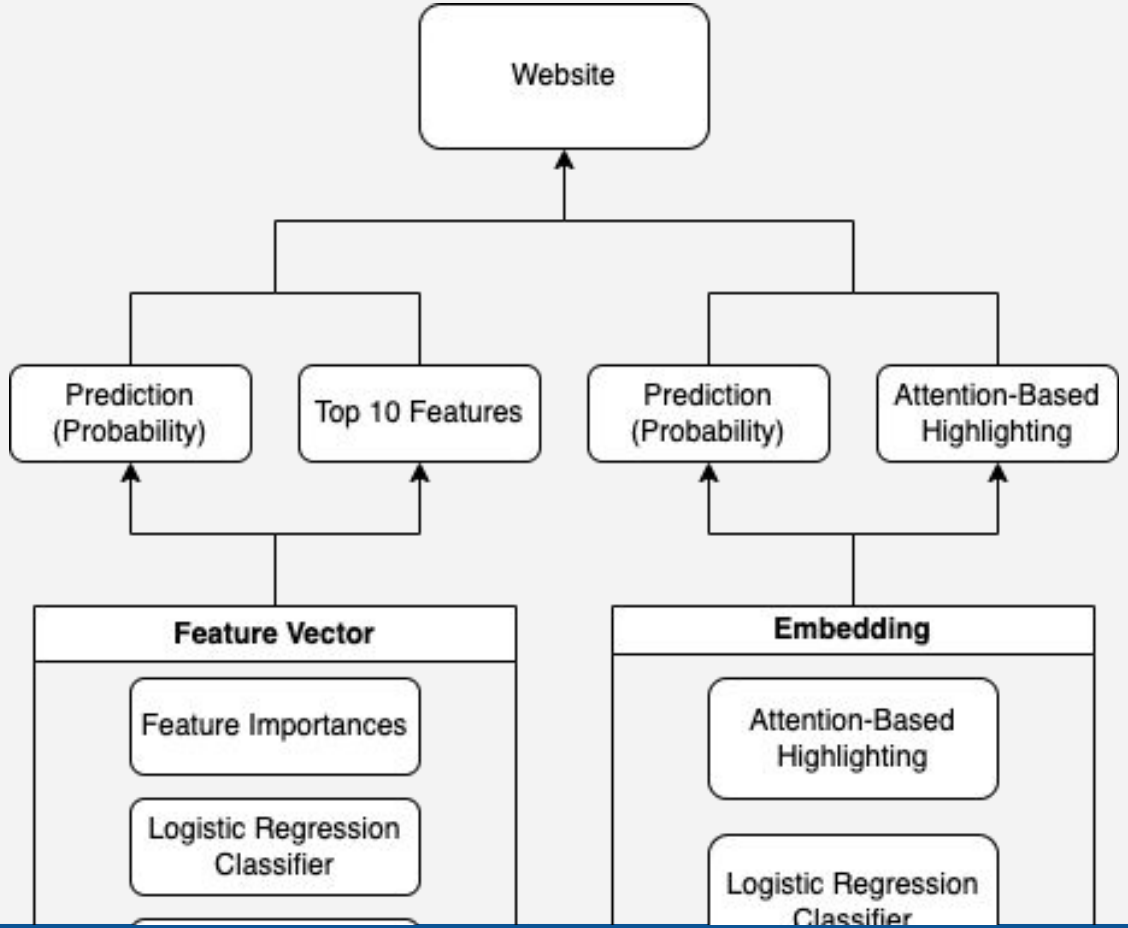
Embedding

Attention Highlighting

happy birthday to you happy birthday to you happy birthday ! ! ! twin ##kle twin ##kle little star / how i wonder what you are up / above the world so high / like a diamond in the sky

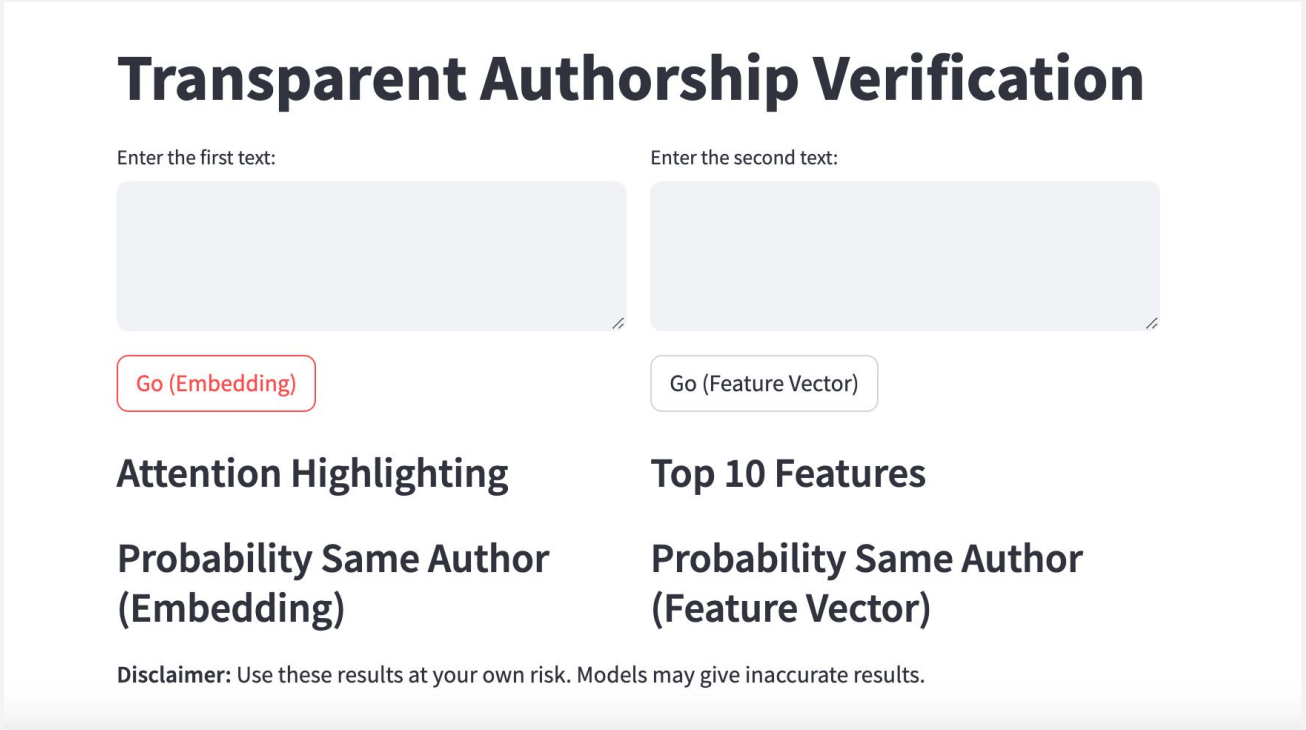
Probability Same Author (Embedding)

0.2615


Website
Prediction (Probability)
Top 10 Features
Prediction (Probability)
Attention-Based Highlighting
Feature Vector
Feature Importances
Logistic Regression Classifier
Embedding
Attention-Based Highlighting
Logistic Regression Classifier

Website

https://same-writer-detector.streamlit.app/



Discussion

Ethics

- lack of privacy and anonymity
- Repression
- AV and AA

Limitations

- Compute
- Time
- Overfitting & Accuracy
- Datasets

Bonus

Students (24) **Teachers (22)** **GPT 4-o** **Claude-3 Sonnet**

Bonus

Students (24)	Teachers (22)	GPT 4-o	Claude-3 Sonnet
0.616	0.767	0.7	0.9

Acknowledgements

- My family
- Mentors Gil Alterovitz and Ning Xie
- Dr. Slava Gerovitch, Prof. Srini Devadas, and the MIT PRIMES program
- Dr. Manesh Gani and Joanna Gilberti
- Trelis AI Grants

Main References

- Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., Potthast, M., & Stein, B. (2020). Overview of the Cross-Domain Authorship Verification Task at PAN 2020. CLEF (Working Notes).
- PAN. (n.d.). Pan.webis.de. Retrieved November 3, 2022, from https://pan.webis.de/
- Prasad, R. S., & Chakkaravarthy, M. (2022). State of the Art in Authorship Attribution With Impact Analysis of Stylometric Features on Style Breach Prediction. Journal of Cases on Information Technology, 24(4), 1–12. https://doi.org/10.4018/jcit.296716