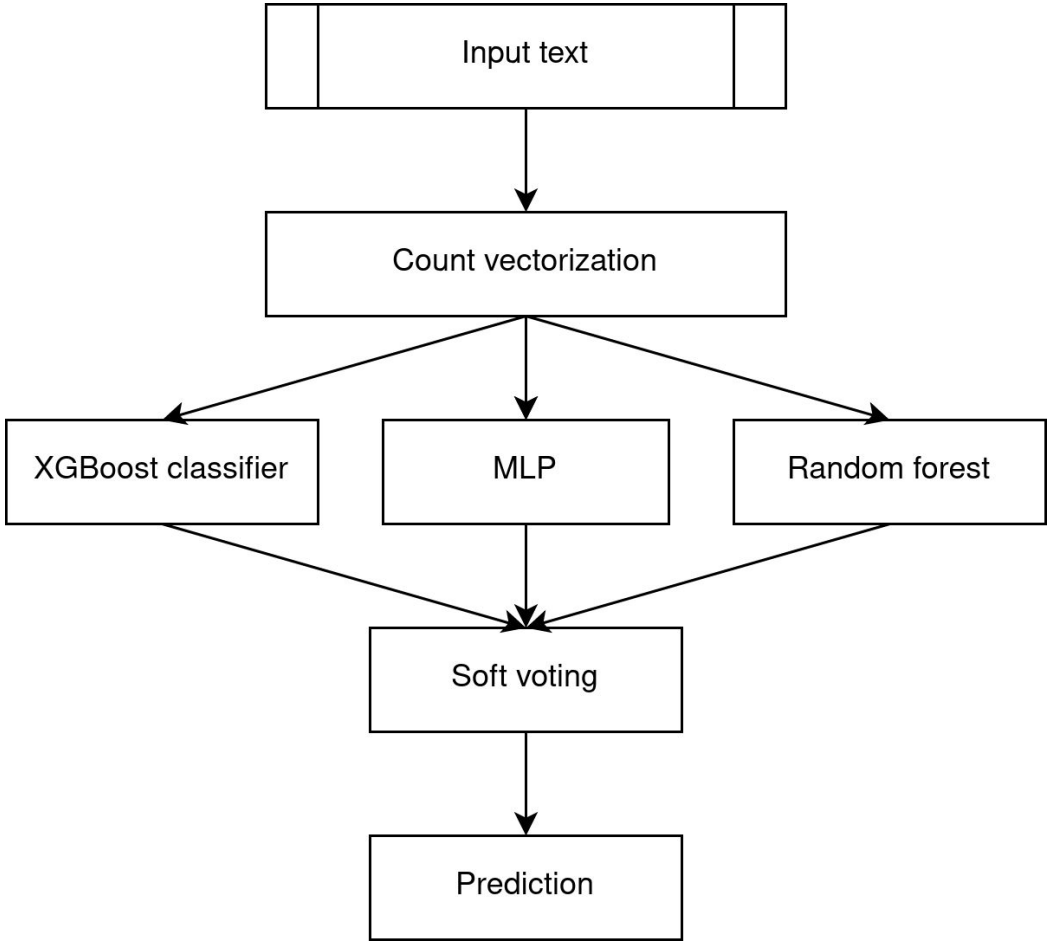
MUNI
FI

Authorship identification - comparison of selected algorithms

Adam Karásek

Ensemble model

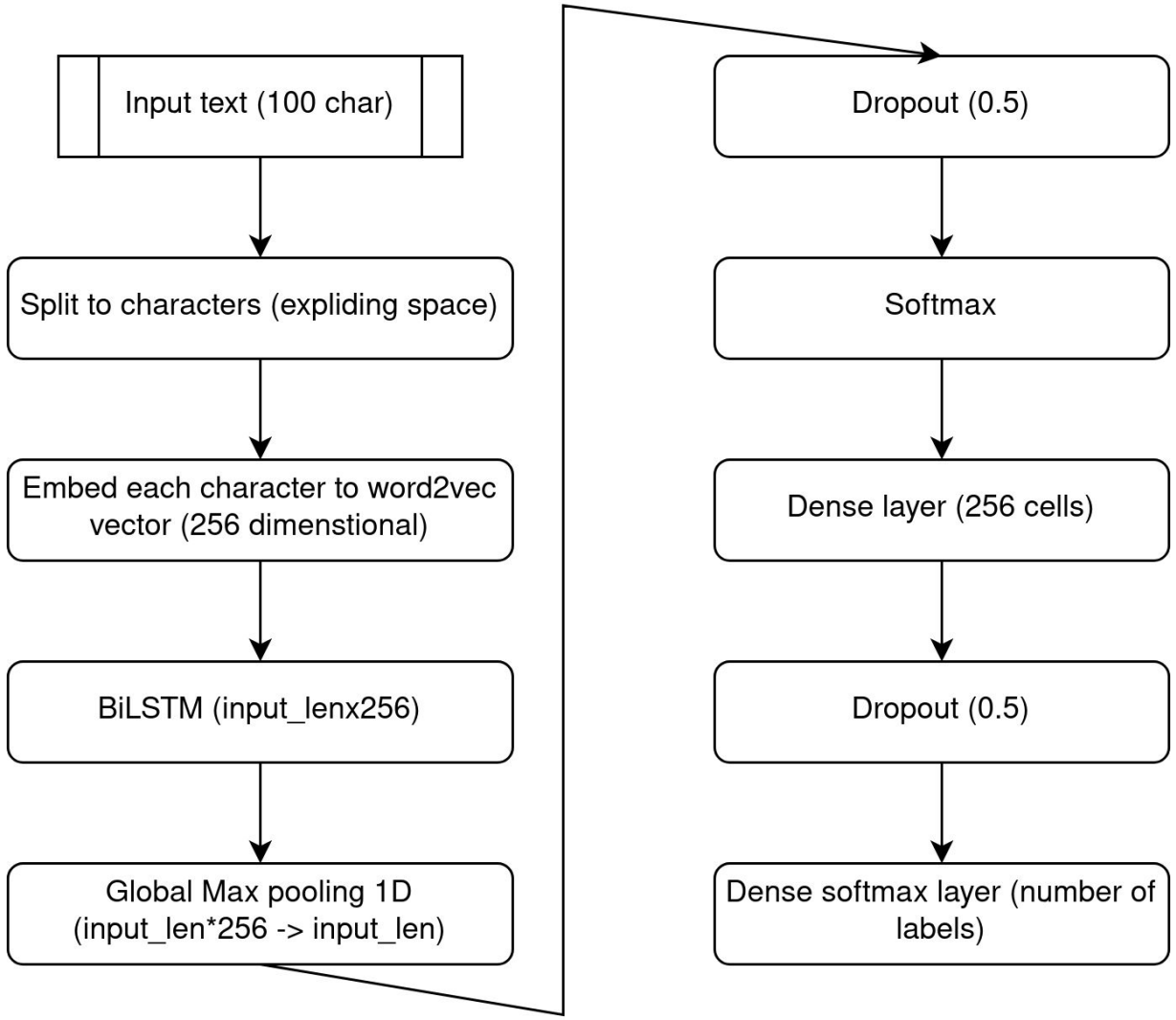
- Authorship identification using ensemble learning
- Ahmed Abbasi et al. [1]
- 97% on 10 authors



MUNI FI

Email detective

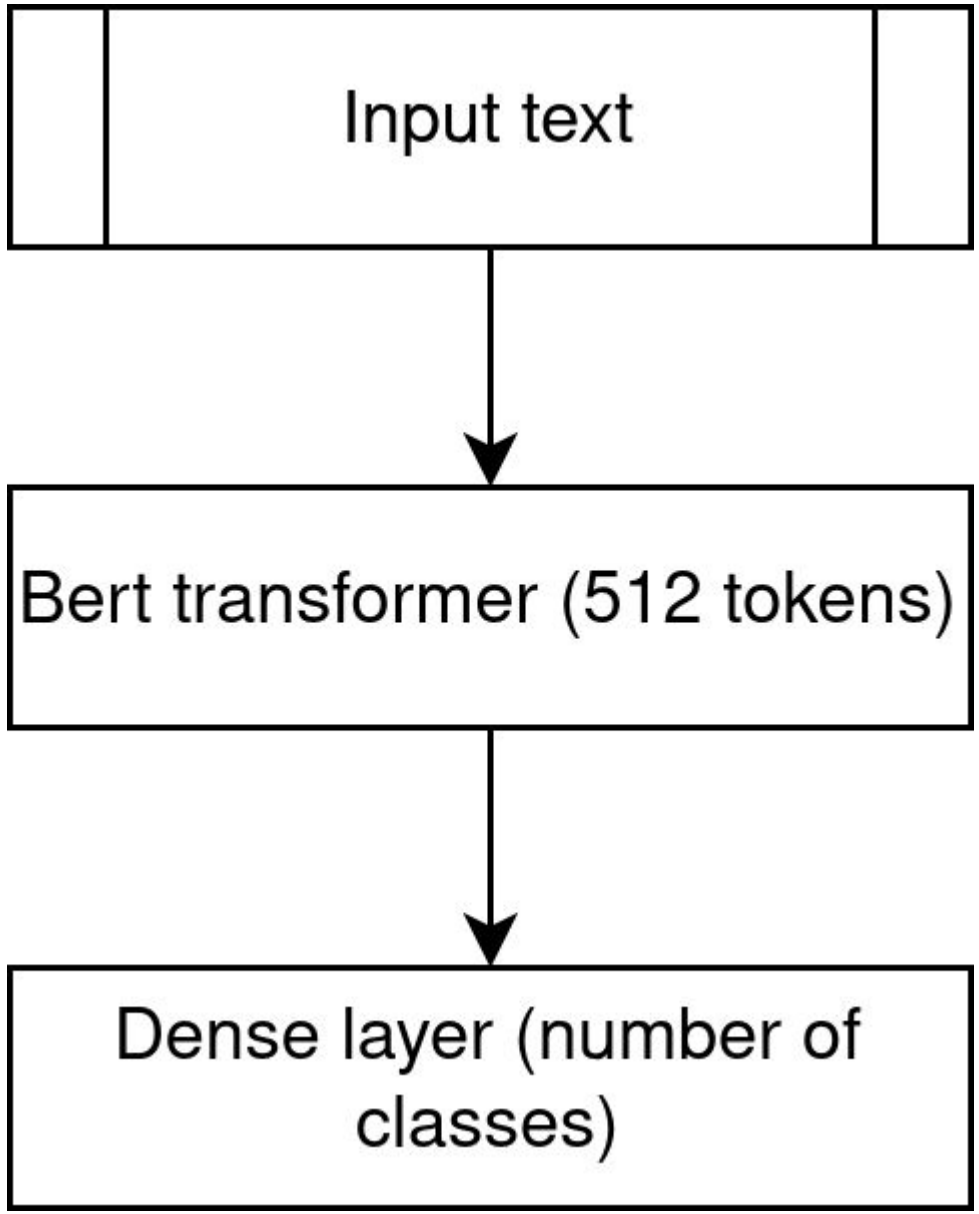
- Email Detective: An Email Authorship Identification And Verification Model
- Yong Fang et al. [2]
- 98.9% on 10 authors
- 92.9% on 25 authors



MUNI FI

BertAA

- BertAA: BERT fine-tuning for Authorship Attribution
- Maël Fabien et al. [3]
- 99.1 % on 10 authors
- 98.7% on 25 authors



MUNI FI

Datasets

- 5, 10 and 25 authors
- Enron, Techcrunch and crypto telegram

Where does the project name come from? What does it mean for you and why did you choose that name for your project??

Sign up & Trade: https://www.okex.com/academy/en/start-trading-cryptocurrency-on-okex?channelFlag=ACEAP4502833
OKEx 101 New User Tutorial database: https://telegra.ph/OKEx-101-New-User-Tutorial-10-12

Hi you could refer to the video to better the understanding of OKEx unified account operation https://twitter.com/i/status/1336627448155230209

Win Tesla Model Y 🚘 iPhone 12 📱 and 40,000 USDT 💰💰

To celebrate the launch of OKExChain mainnet, OKEx distribute rewards to participants who trade OKT.

During the promotion period, participants who register and trade OKT have a chance to win a Tesla Model Y 🚘 iPhone 12 📱 and 40,000 USDT 💰💰 Prize Pool

🕰Promotion period: 19 Jan 2021, 08:00 - 28 January 2021, 16:00 (UTC)

🏵 Trade $OKT to win prizes

👉 How to participate?
1. Join OKEx Global English
2. Fill up the form.
3. Participant register here

⁉No Account yet? Sign up here with $80 rewards

For $OKT trading campaign info, click here

MUNI FI

Datasets

Table 1. Number of documents per author in experiment set

Dataset	$k = 5$	$k = 10$	$k = 25$
Enron	$l = 4000$	$l = 2000$	$l = 800$
Telegram	$l = 1000$	$l = 650$	$l = 470$
Techcrunch	$l = 2500$	$l = 1200$	$l = 250$

Results - Ensemble

- Ensemble model

Techcrunch	$k = 5$	$k = 10$	$k = 25$
Ensemble	0.9624	0.9275	0.7568
Random Forest	0.904	0.8517	0.6096
XGB classifier	0.9616	0.915	0.7536
MLP	0.9728	0.943	0.7504
Training time	894s	1229s	1142s

Telegram	$k = 5$	$k = 10$	$k = 25$
Ensemble	0.34	0.2062	0.0896
Random Forest	0.344	0.2062	0.0902
XGB classifier	0.308	0.1985	0.0885
MLP	0.31	0.2338	0.0766
Training time	52s	78s	190s

Enron	$k = 5$	$k = 10$	$k = 25$
Ensemble	0.9675	0.9339	0.8417
Random Forest	0.9625	0.9228	0.8171
XGB classifier	0.9395	0.8961	0.8142
MLP	0.968	0.9234	0.8057
Training time	902s	1323s	1591s

Results - Ensemble model

- Number of features:
  - Telegram - 4500-6000
  - Enron - 26000
  - Techcrunch - 44000-55000
- Ensemble is more robust on more authors
- Random forest is better with less features
- MLP can be better than ensemble with lower number of authors and bigger number of features

Results - Email Detective

- Email detective - input set to 100 yields better results than larger input

Table 4.4: Email detective experiment results

Email detective	$k = 5$	$k = 10$	$k = 25$
Enron acc	0.9413	0.8719	0.7407
Enron time	678s	704s	827s
Techcrunch acc	0.7131	0.5695	0.2123
Techcrunch time	382s	492s	252s
Telegram acc	0.2667	0.2046	0.0948
Telegram time	60s	378s	560s

Results - BertAA

Table 4.5: BertAA experiment results

BertAA	$k = 5$	$k = 10$	$k = 25$
Enron acc	0.98	0.9555	0.899
Enron time			
Techcrunch acc	0.9587	0.9142	0.7296
Techcrunch time			
Telegram acc	0.3153	0.2236	0.0936
Telegram time			

Sources

- [1] ABBASI, Ahmed; CHEN, Hsinchun. Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. ACM Trans. Inf. Syst. 2008, vol. 26, no. 2. Issn 1046-8188. Available from doi: 10.1145/1344411.1344413.

- [2] FANG, Yong; YANG, Yue; HUANG, Cheng. EmailDetective: An Email Authorship Identification And Verification Model. The Computer Journal. 2020, vol. 63, no. 11, pp. 1775–1787. issn 0010-4620. Available from doi: 10.1093/comjnl/bxaa059.
- [3] FABIEN, Maël; VILLATORO-TELLO, Esau; MOTLICEK, Petr; PARIDA, Shantipriya. BertAA : BERT fine-tuning for Authorship Attribution. In: BHATTACHARYYA, Pushpak; SHARMA, Dipti Misra; SANGAL, Rajeev (eds.). Proceedings of the 17th International Conference on Natural Language Processing (ICON) [online]. Indian Institute of Technology Patna, Patna, India: NLP Association of India (NLPAI), 2020, pp. 127–137 [visited on 2023-11-26]. Available from: https://aclanthology.org/2020.icon-main.16.

MUNI FI