MUNI
FI

# Authorship identification
# Comparison of selected algorithms

Adam Karásek, RNDr. Zuzana Nevěřilová Ph.D.

# The Task

Classify a text to find the most probable author among known authors.

- Process several texts from each potential author
- Learn features for each author

Fair classification

- No author name (removed signatures from emails)
- How can we be sure the author really created the text?

# Datasets - Enron

- preprocessing

Jesus,

Friday, April 28, works for me. I am free between 8:00 - 10:30 to meet with
Aram.
Would you like to meet him for lunch or dinner?

Mike,

This one is very good. I am afraid we shall have a few employees left after this test.

http://www.google.com/search?q=cache:www.essential.org/monitor/hyper/mm0997.04
.h tml+GE+%2B+Enron+%2B+collusion&hl=en



MUNI
FI

# Datasets - Techcrunch

- articles

Pricing for the U.K.'s first 4G/LTE network has been revealed. The new 4GEE tariffs start at £36 per month for a plan with a mere 500MB of data, and rise to £56 for an 8GB plan. The 4GEE network, a joint venture between the Orange and T-Mobile carrier brands — which merged to create the Everything Everywhere joint venture (now EE) back in 2010, is due to go live on October 30. On that date all Orange and T-Mobile stores will also be rebranded EE.

While EE's field tests have achieved download speed tests on the 4GEE network of more than 50Mbps the average real-world download speeds for users are likely to be a much more modest 8Mbps to 12Mbps — albeit still around five times faster than the average 3G download rate, according to EE. That said, the company isn't giving 4GEE users any speed guarantees but a company spokesman told TechCrunch they are aiming for a "desired minimum" of 10Mbps. As EE builds its network footprint out in the coming years it will also be investing in network capacity to ensure it can maintain this informal speed floor, he added.

4GEE consumer price plans and services

All 4GEE phone plans are two-year contracts (update: EE has now announced 12-month plans and handset prices too); include unlimited texts and calls; access to BT Wi-Fi hotspots; and do allow VoIP and tethering. There are five consumer tariffs for 4GEE phone plans. Here's the full breakdown

Consumers who exceed their data limit can buy data add-ons, costing £3 for 50MB; £6 for 500MB; £16 for 2GB; and £20 for 4GB. EE says it will warn customers when they have used up 80 percent of their limit by sending a text message. A second warning will be sent when they have used up all that month's data — along with pop-ups giving them the option to purchase a data add-on.

One way to gobble up your data quickly will be to make a lot of use of EE's new Film Store service via 4G. The Film Store will offer films for download or streaming, plus trailers, cinema listings and a new two for one cinema ticket offer (duplicating the Orange Wednesdays offer), with a catalogue of more than 700 films in total. EE is offering new customers one free film download per week for the first four months — which won't eat into their data allowance. Any additional films will eat data, and also cost money to download/stream (with prices starting from 79p). The cost of downloads can be charged to customers' mobile bills or credit/debit cards.



MUNI FI

# Datasets - Telegram

TomoChain | TOMO

USD : $1.26000000
BTC : ฿0.0000334100
ETH : ♢0.00113270

Mkt Cap: $96,459,191
Volume : $13,906,795
1hr % : 0.36%
24hr % : 11.6%
7d % : 28.54%
30d % : 67.79%
1y % : 150.7%

What about your think

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

banking system in crypto mostly have problem with lack of efficient and fees, how does Latam Cash fix this issue?



MUNI FI

# Datasets - Experiment samples

- k = 5, 10 and 25 authors
- l - number of examples (at least 100 characters)

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

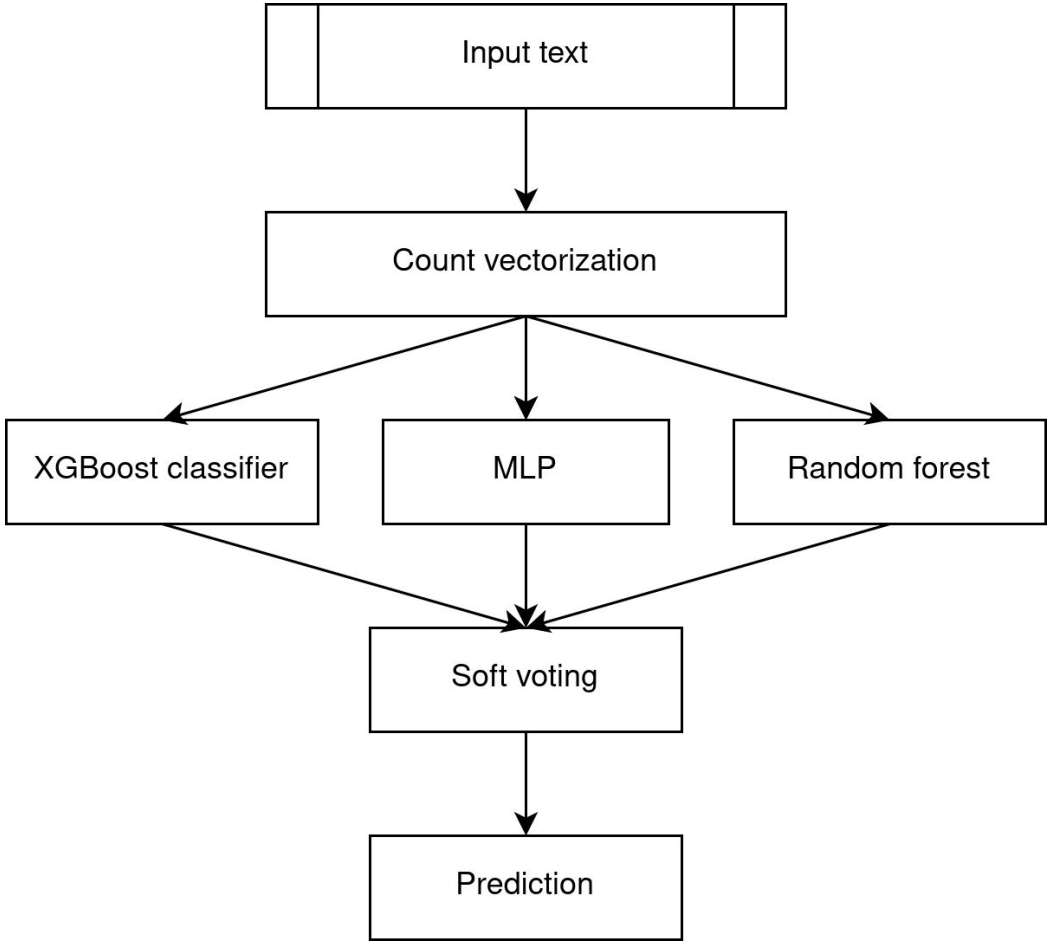
# Datasets - Experiment samples

Average example length (characters)

| Dataset | $k = 5$ | $k = 10$ | $k = 25$ |
|---|---|---|---|
| Enron | 478 | 486 | 427 |
| Techcrunch | 3461 | 3361 | 2927 |
| Telegram | 214 | 214 | 206 |

# Ensemble model

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

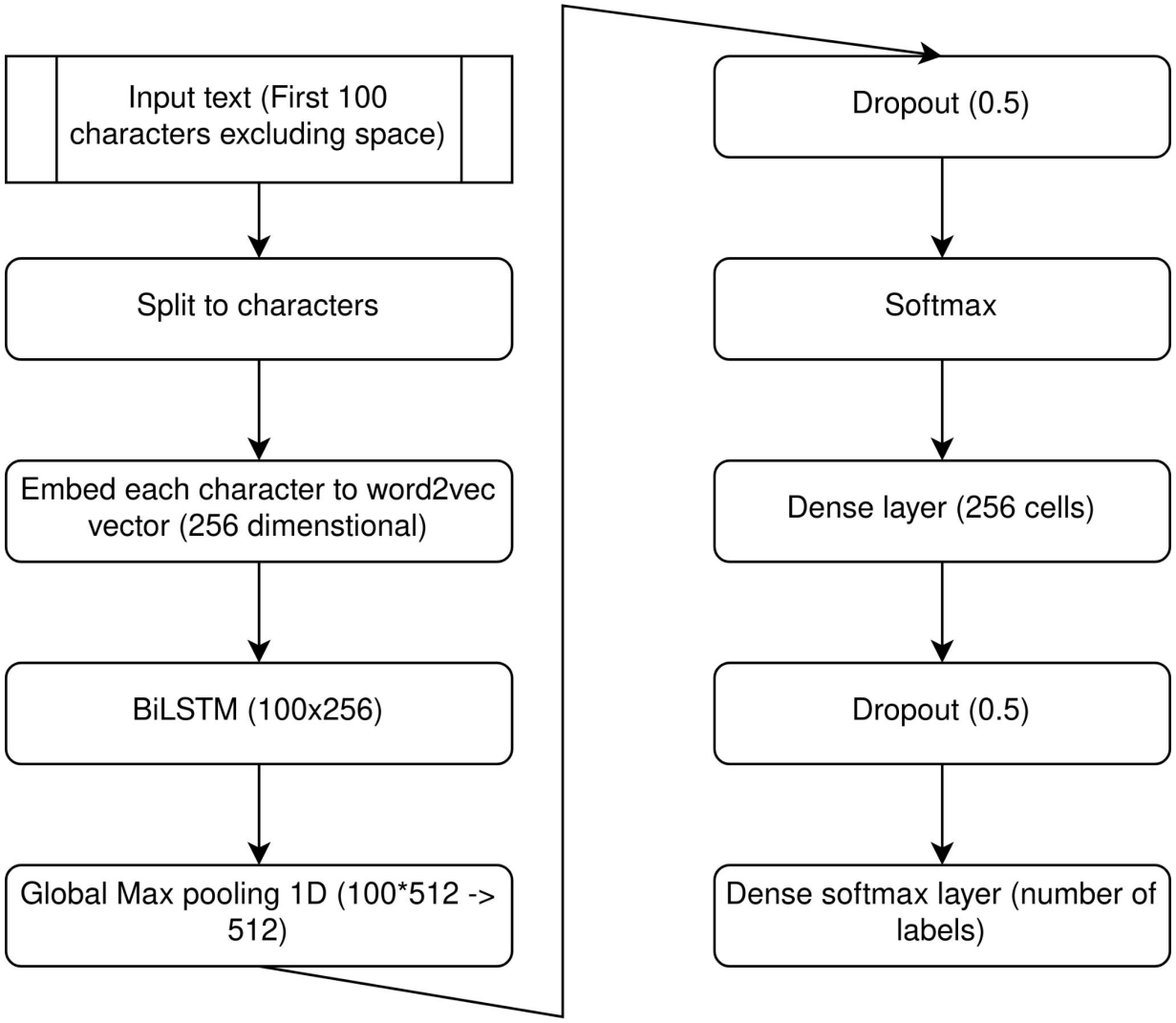





MUNI FI

# Email detective

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

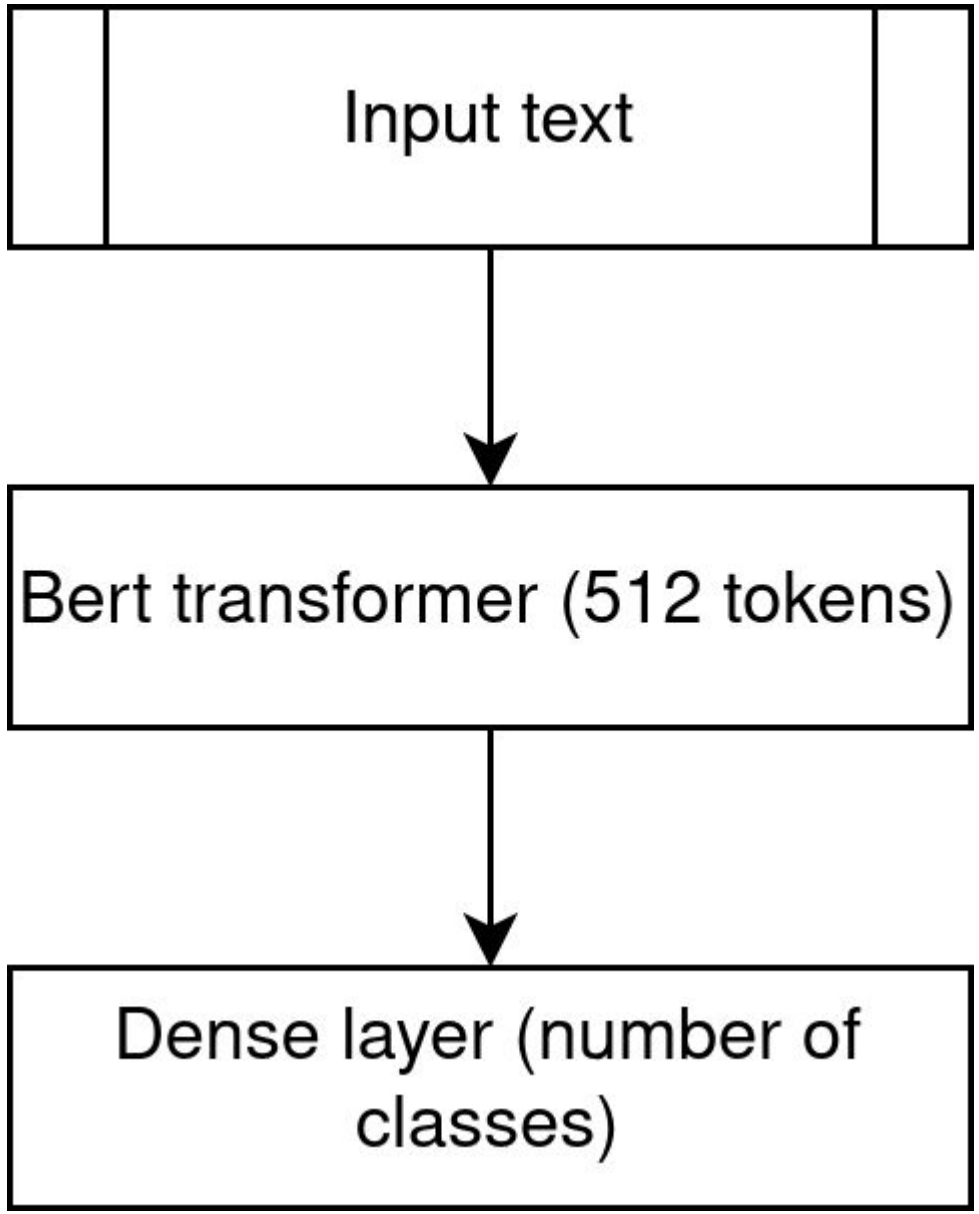





MUNI FI

# BertAA

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






MUNI FI

# Results - Ensemble

| Ensemble-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 | 71.52s | 102.42s | 182.72s |

| Ensemble-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 | 45.1s | 66.15s | 79.55s |

| Ensemble-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 | 10.4s | 12s | 16.17s |

# What did the Ensemble Model Learn

| Importance order | Enron | Techcrunch | Telegram |
|---|---|---|---|
| 1. | best | said | klaytn |
| 2. | shall | explains | frontier |
| 3. | hi | company | lon |
| 4. | counterparty | says | project |
| 5. | deal | like | tokenlon |
| 6. | ces | today | lord_beerus_samma |
| 7. | ge | apple | okex |
| 8. | ll | services | does |
| 9. | agreement | million | token |
| 10. | gas | hardware | defi |

# Ensemble Model: Author Characteristics

| Feature | Kaminski | Dasovich | Germany | Mann | Jones |
|---|---|---|---|---|---|
| best | 173 | 2729 | 97 | 55 | 28 |
| shall | 1627 | 55 | 1 | 361 | 51 |
| hi | 47 | 173 | 79 | 732 | 0 |
| counterparty | 0 | 7 | 57 | 24 | 1218 |
| deal | 38 | 505 | 2178 | 384 | 133 |
| ces | 2 | 0 | 1131 | 0 | 4 |
| ge | 9 | 5 | 0 | 626 | 0 |
| ll | 31 | 1048 | 214 | 572 | 314 |
| agreement | 35 | 209 | 119 | 854 | 1364 |
| gas | 124 | 931 | 1506 | 111 | 335 |

# Results - Ensemble

- 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
- The model learned reasonable author characteristics

# Results - Email Detective

- Email detective - input set to 100 yields better results

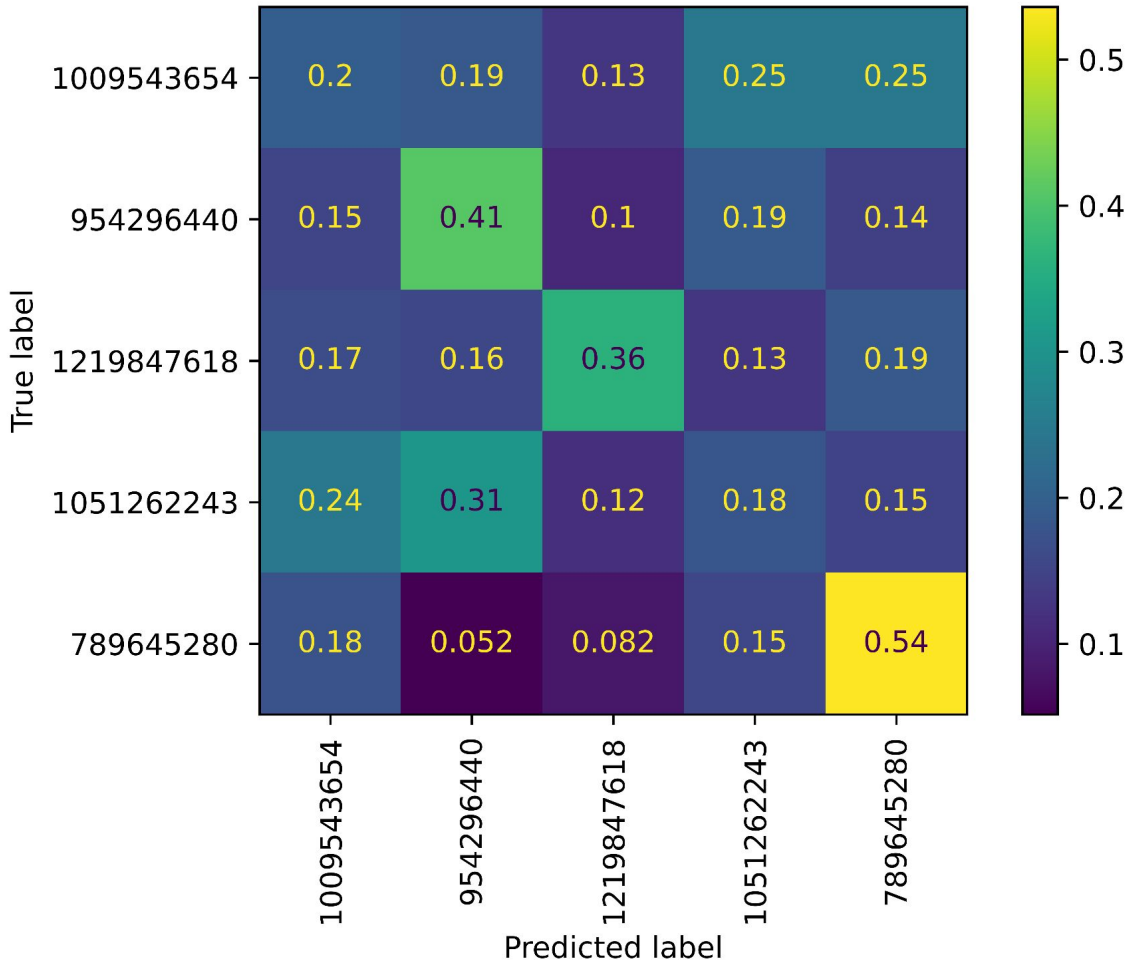
| Email detective | $k = 5$ | $k = 10$ | $k = 25$ |
|---|---|---|---|
| Enron acc | 0.9413 | 0.8719 | 0.7407 |
| Enron time | 33.9s | 35.2s | 41.35s |
| Techcrunch acc | 0.7131 | 0.5695 | 0.2123 |
| Techcrunch time | 30.56s | 41s | 40.32s |
| Telegram acc | 0.2667 | 0.2046 | 0.0948 |
| Telegram time | 10.4s | 58.15s | 47.66s |

# Results - BertAA

- Only a few epochs is enough

| BertAA | $k = 5$ | $k = 10$ | $k = 25$ |
|---|---|---|---|
| Enron acc | 0.9805 | 0.9633 | 0.9006 |
| Enron time | 327.85 | 342.95 | 250.25 |
| Techcrunch acc | 0.9587 | 0.9142 | 0.7296 |
| Techcrunch time | 308.16 | 621 | 424.64 |
| Telegram acc | 0.338 | 0.2236 | 0.0953 |
| Telegram time | 259 | 234.77 | 185.27 |

# Results - BertAA






MUNI FI

# Summary

- Some datasets are very difficult when compared to others
- Even simple models (Ensemble) can work well and learn reasonable things
- Even advanced models (BertAA) cannot work well on difficult data
- Reproducibility of quite recent papers is still poor.

# Sources


- 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.

- 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.
- 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