DeepFake Tweet Detection

Team 15: TextTechTitans

ADRIAN KAMIŃSKI, ADAM FREJ,
PIOTR MARCINIAK, SZYMON SZMAJDZIŃSKI

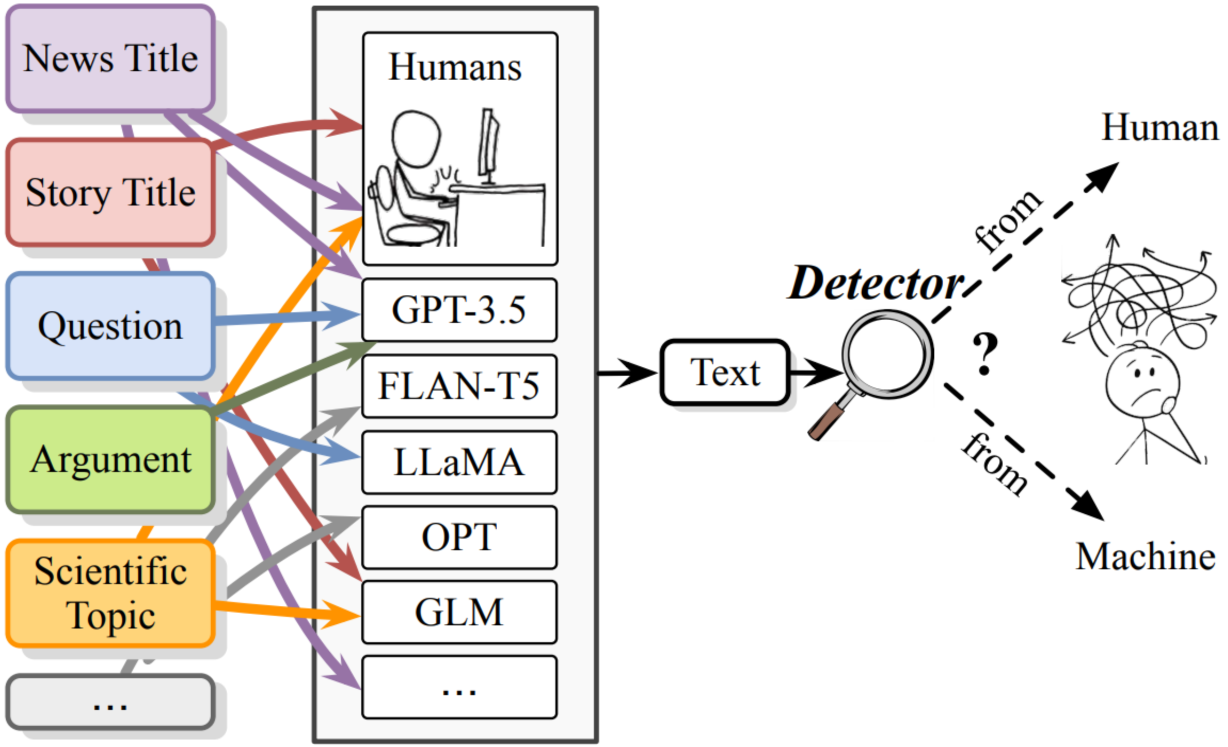
NOVEMBER 2023

Plan of presentation

1. Introduction
2. Datasets description
3. Results
4. Concept and work plan

Deepfake

- ➢Deep learning + fakes.
- ➢Tweets – short texts without context used in social media interactions.
- ➢Humans' performance on this task.
- ➢Why is it a problem? Why do we need detection tools?



Source: Deepfake Text Detection in the Wild, Yafu Li and Qintong Li and Leyang Cui and Wei Bi and Longyue Wang and Linyi Yang

Our research questions

- Can we build a reliable deepfake detection algorithm? By reliable algorithm, meaning detecting generated tweets while avoiding assigning false positives.
- What are the most effective features for deep-fake detection in tweets?
- Are there any patterns that indicate the model-generated tweet content?

Hypothesis

- The use of emoticons may be higher in human generated content. [H0]
- The use of mentions of other users may be higher in human-generated content. [H1]
- There will be more misspelled words in content generated by bots. [H2]
- The impact of different URL encoding, e.g., encoding all URLs to a single token vs extracting the basepath of the URLs. [H3]

TweepFake - main dataset

- Contains 25,572 tweets.
- Equal split between human-generated and bot-generated tweets.
- 17 human accounts as the basis for imitation.
- 23 bot accounts that mimic the behavior of these human accounts.

dril	this is every thing and its only 11:am, https://t.co/XOioXnwFQh	human	human
nsp_gpt2	guy, you have a pretty amazing dick, you're awesome, I appreciate that	bot	gpt2

GPT-2 output dataset - secondary dataset

- Contains 500,000 text from web.
- Equal split between human-generated and bot-generated text.
- Those text are longer than tweets.

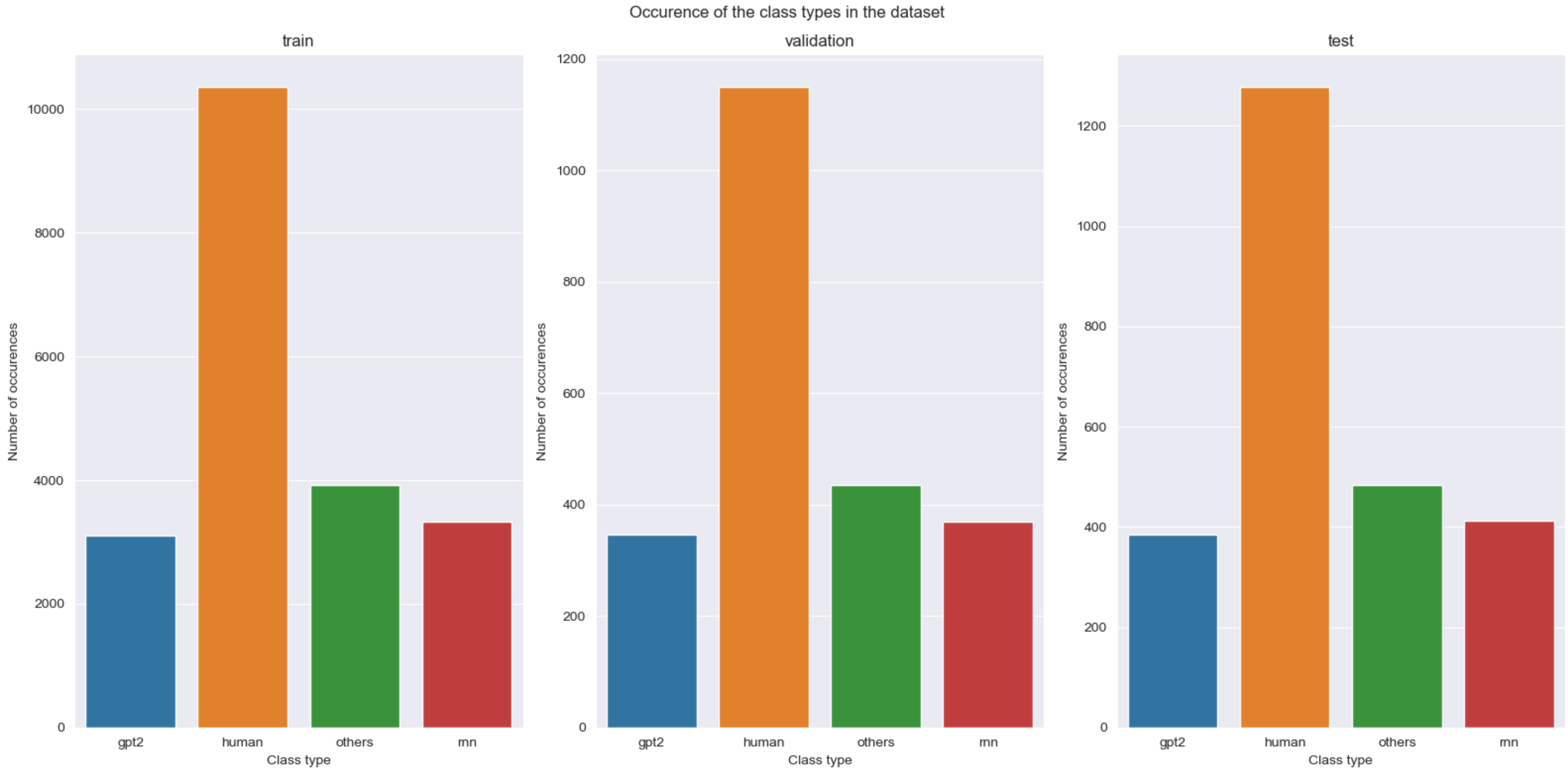
```
Real :: These girlfriends deserves a special mention for going that extra mile,
        hopefully doesn't set too many guys off on the path towards outrageous
        demands.\n\n1. She knows the severity of man-flu. ...

Fake :: As the final day of the presidential campaign approached Wednesday morning,
        polls suggested the race was closer than many had expected.
        Polls have shown Hillary Clinton leading Donald Trump by 9 percentage points
        since late September, with the Republican leading by 7 percent. ...
```

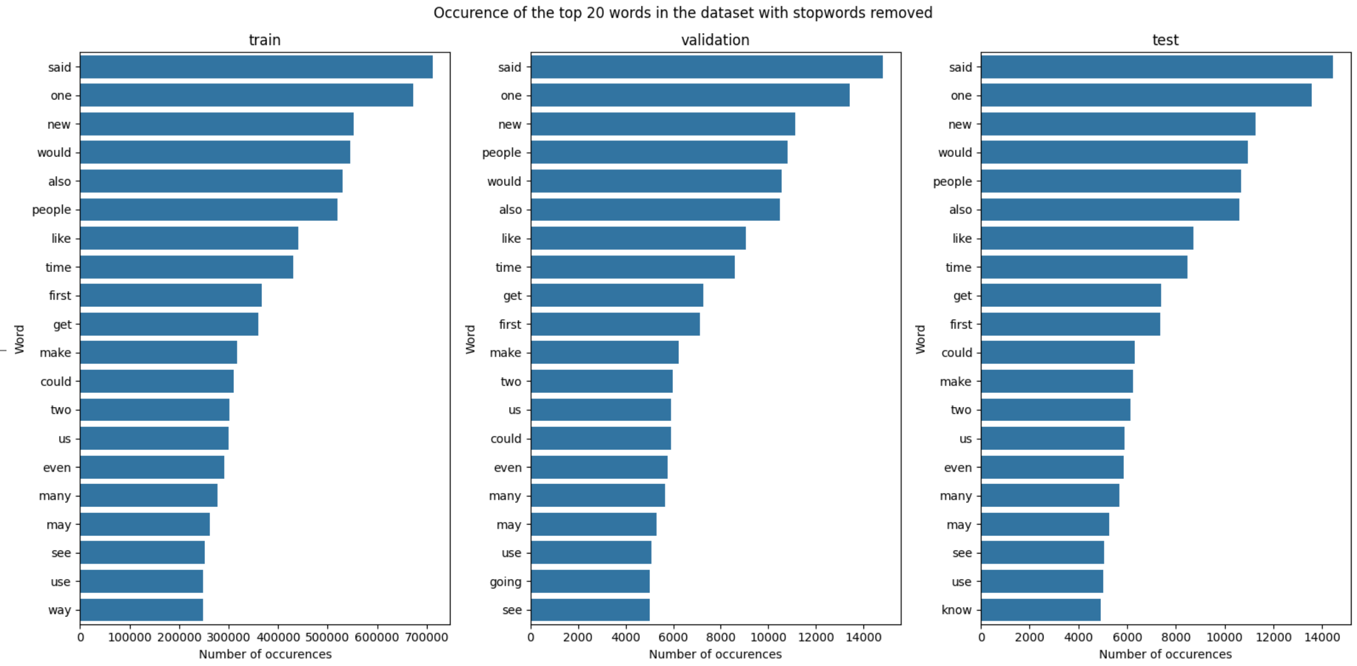
Methods outline

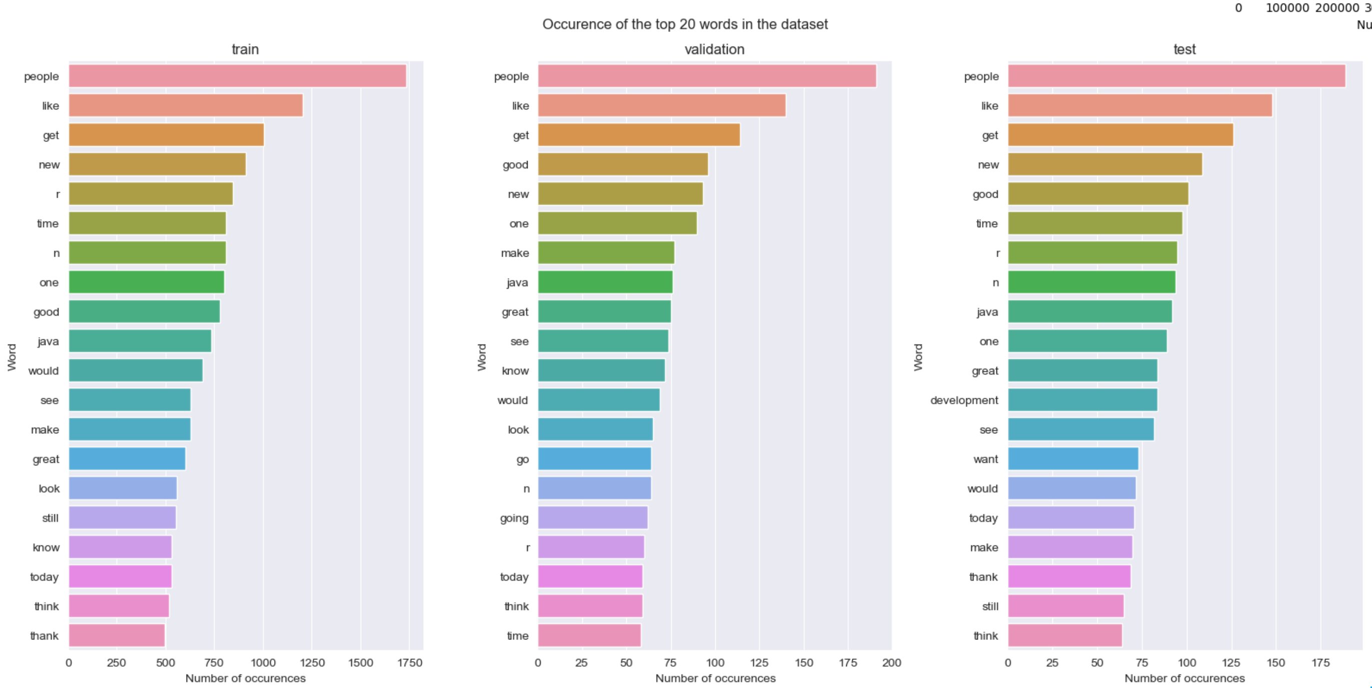
- General EDA
- Preprocessing: tagging hyperlinks, mentions, retweets, stemming, lemmatization
- Embedding:
  - TF-IDF
  - Bert
- Modeling:
  - ML methods: SVC, LGBM, RF, LR, XGB
  - DL methods: CNN, GRU, CNN + GRU
- Optuna optimizer for ML
- Char or word tokenizing for DL

Datasets/EDA

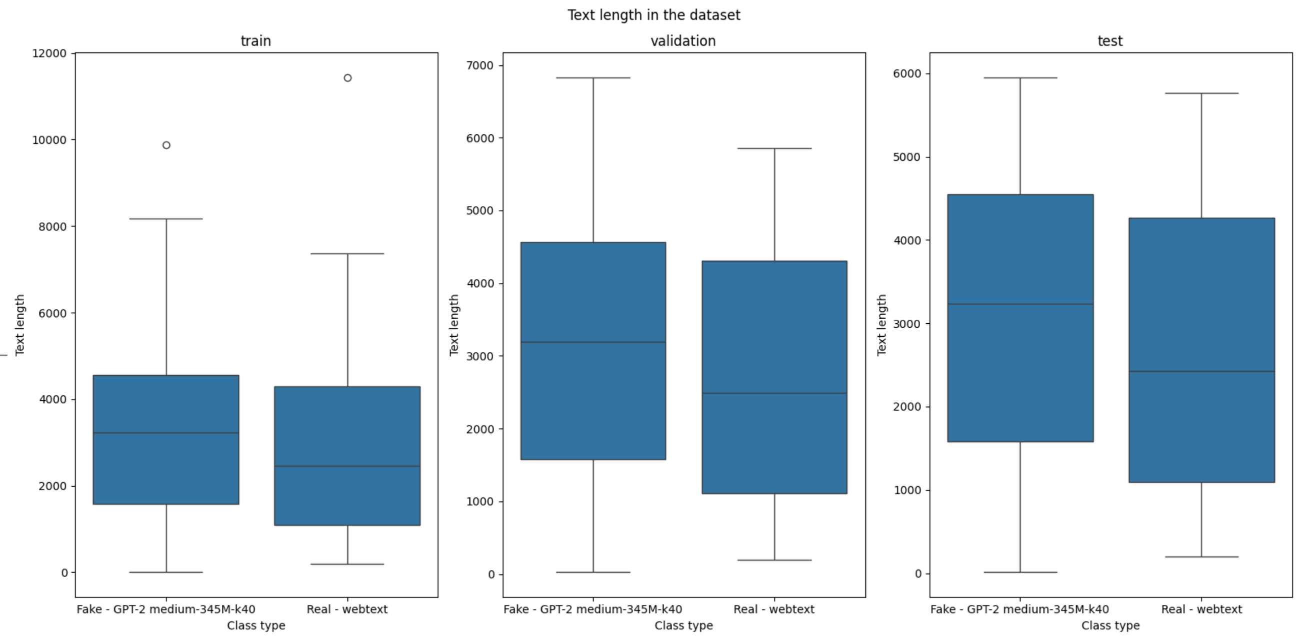


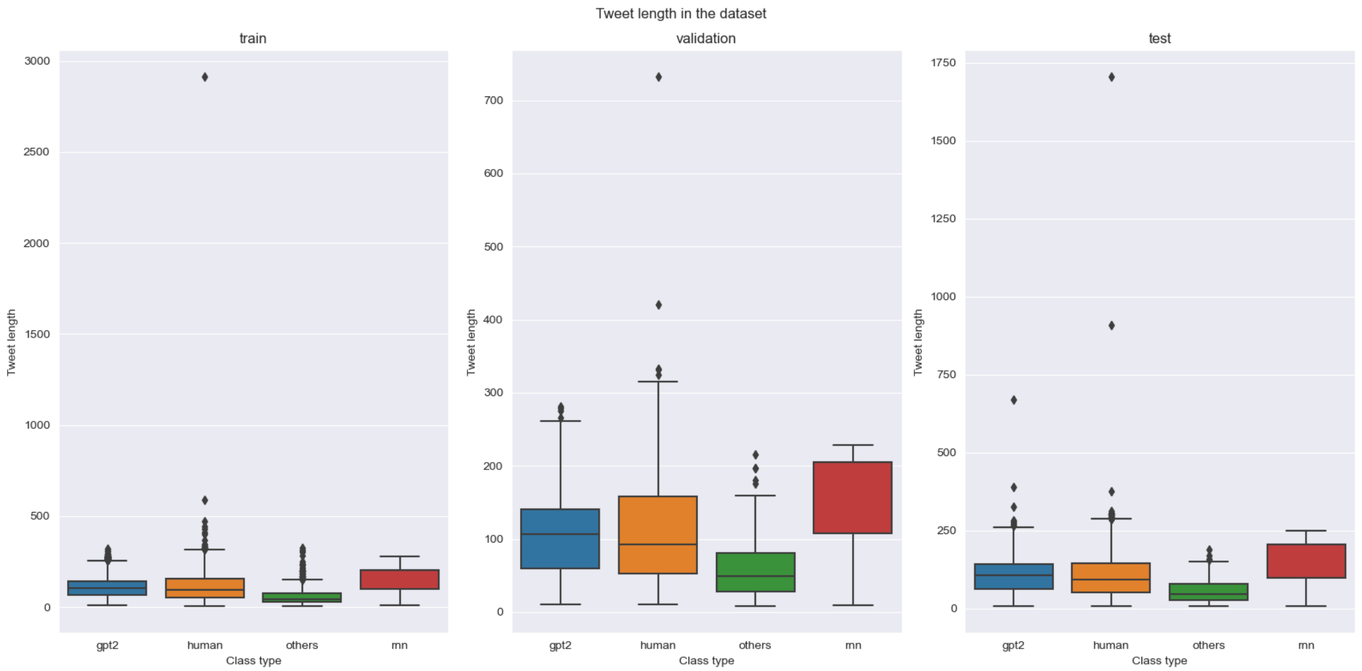
Datasets/EDA



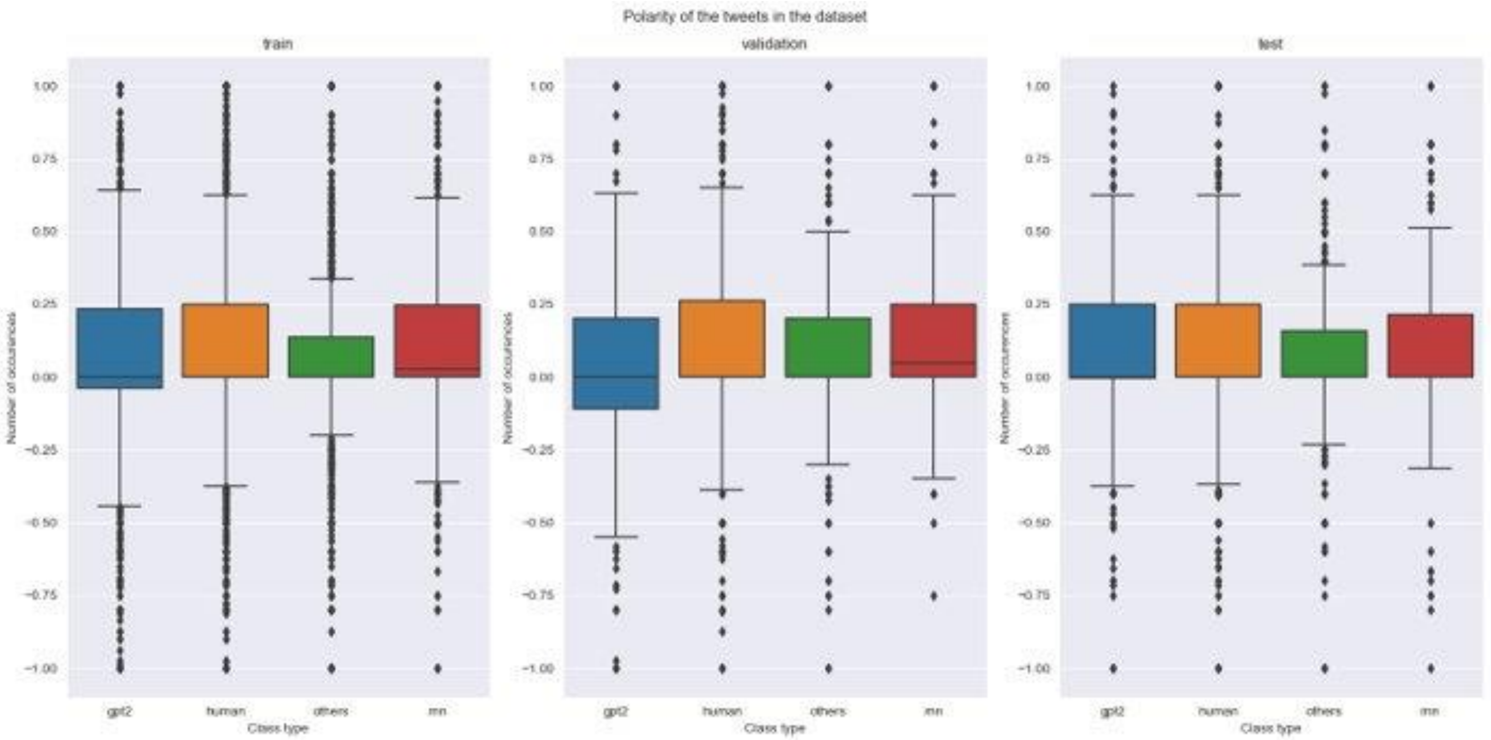


Datasets/EDA

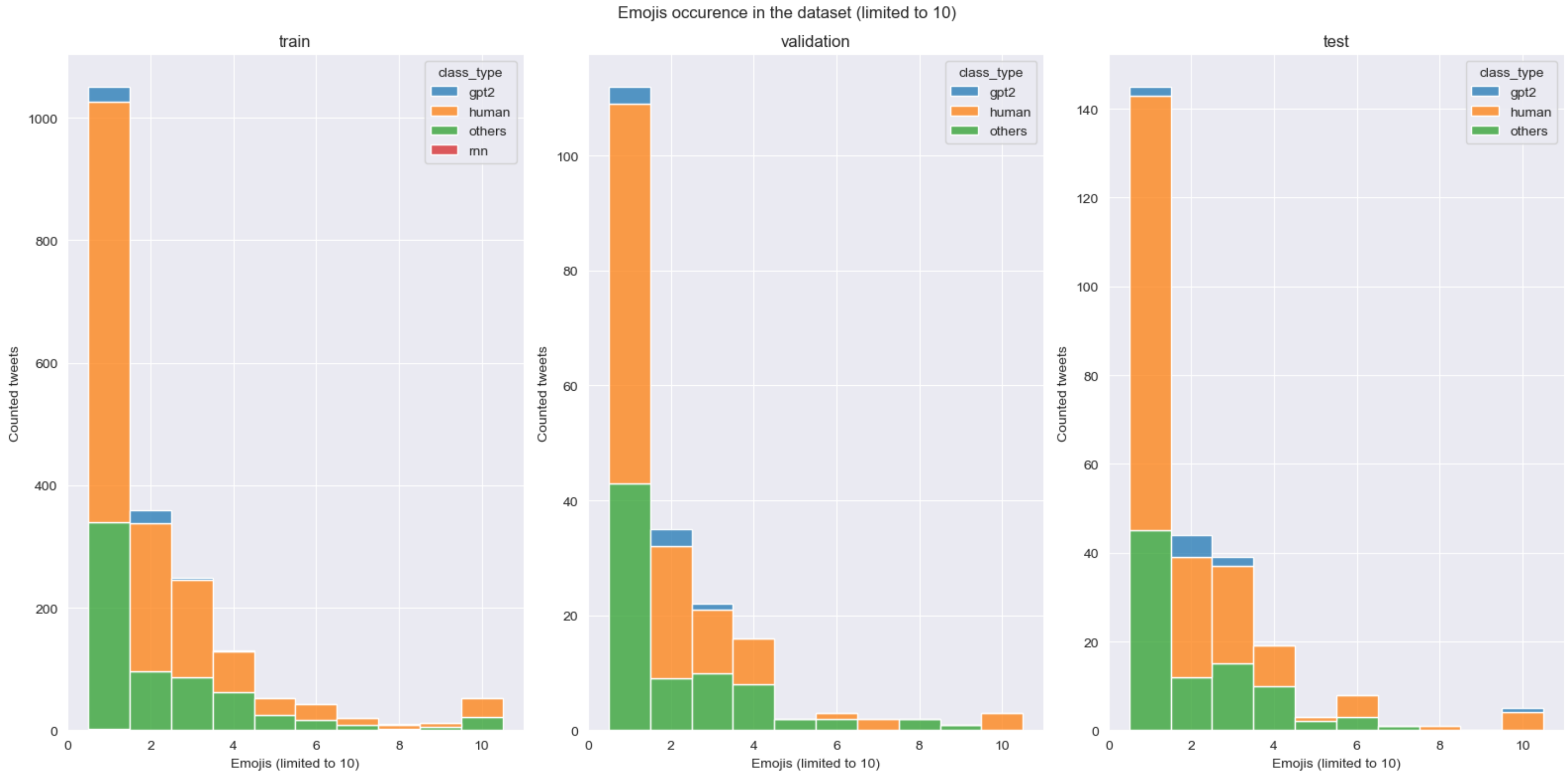




Datasets/EDA

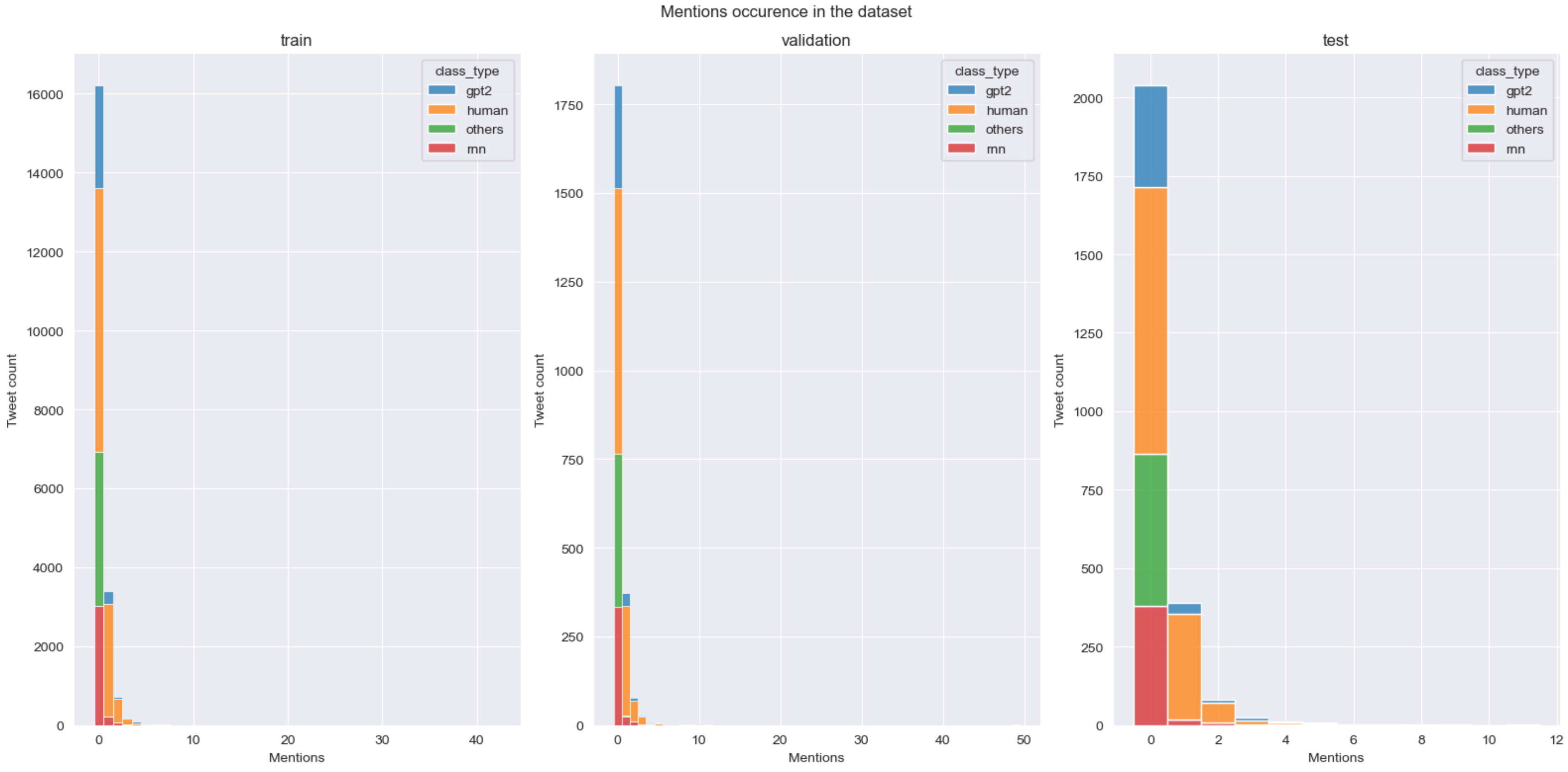


Datasets/EDA - verification of hypothesis



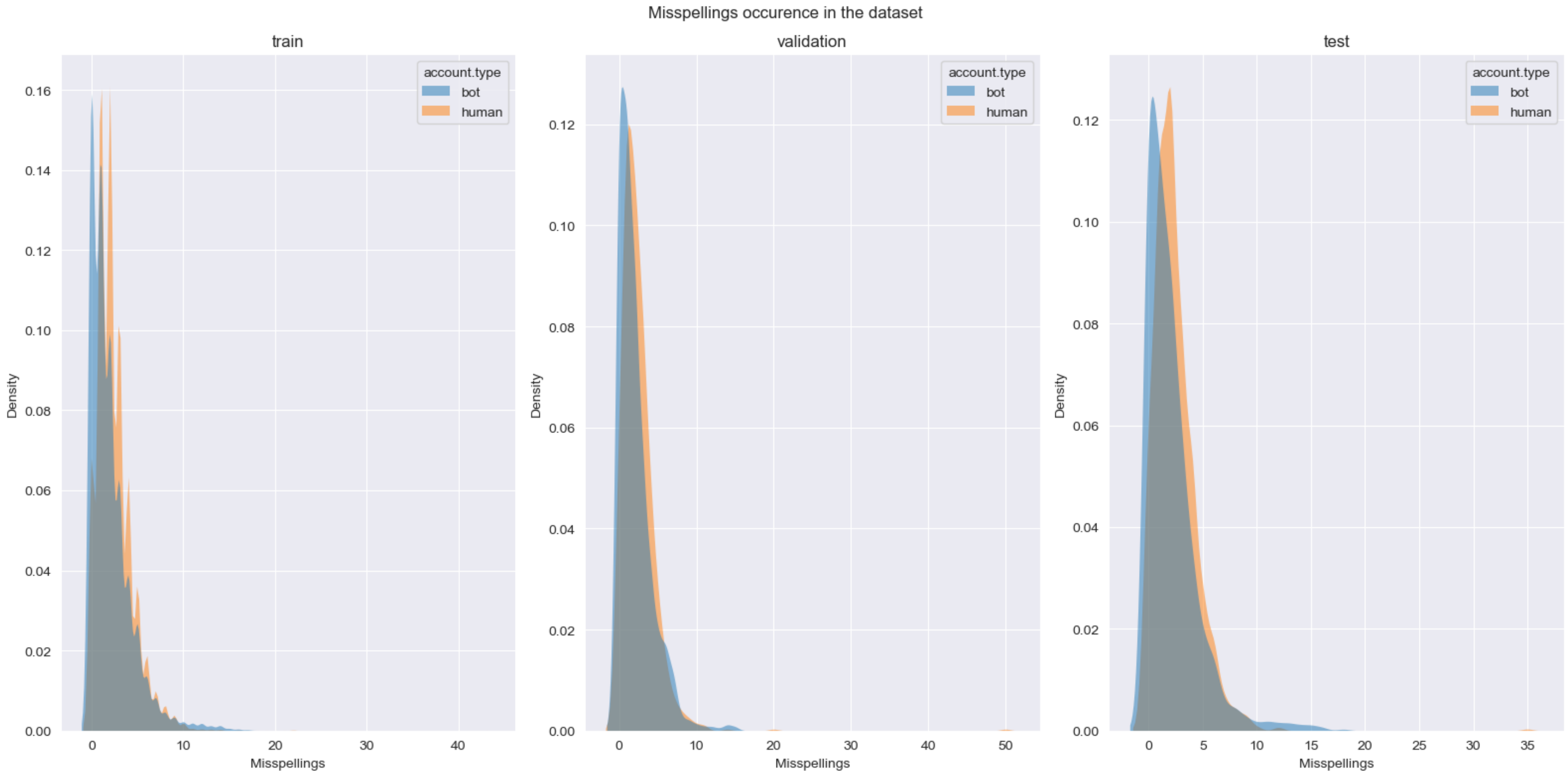
H0: The use of emoticons may be higher in human generated content.

Datasets/EDA - verification of hypothesis



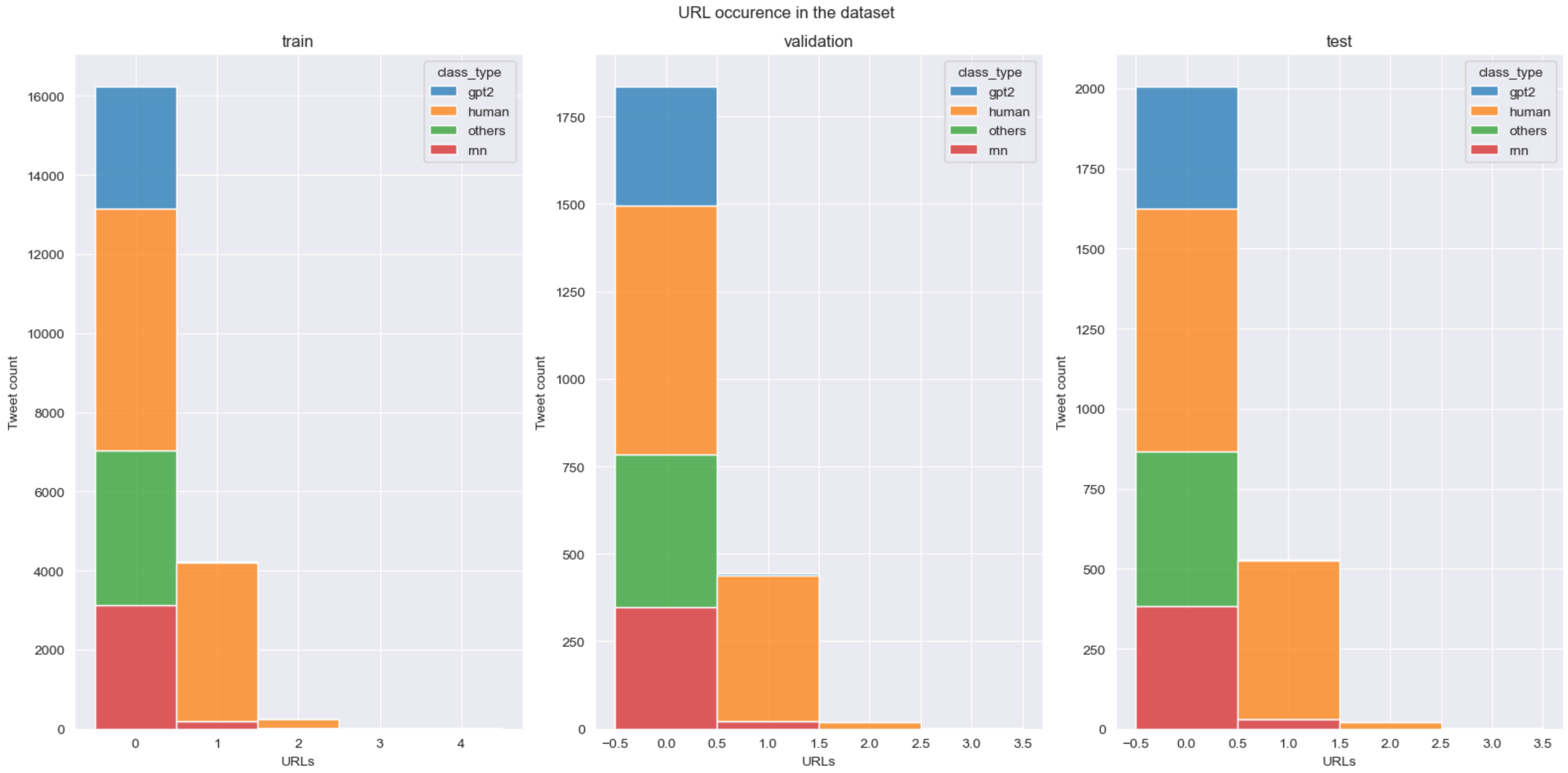
H1: The use of mentions of other users may be higher in human-generated content.

Datasets/EDA - verification of hypothesis



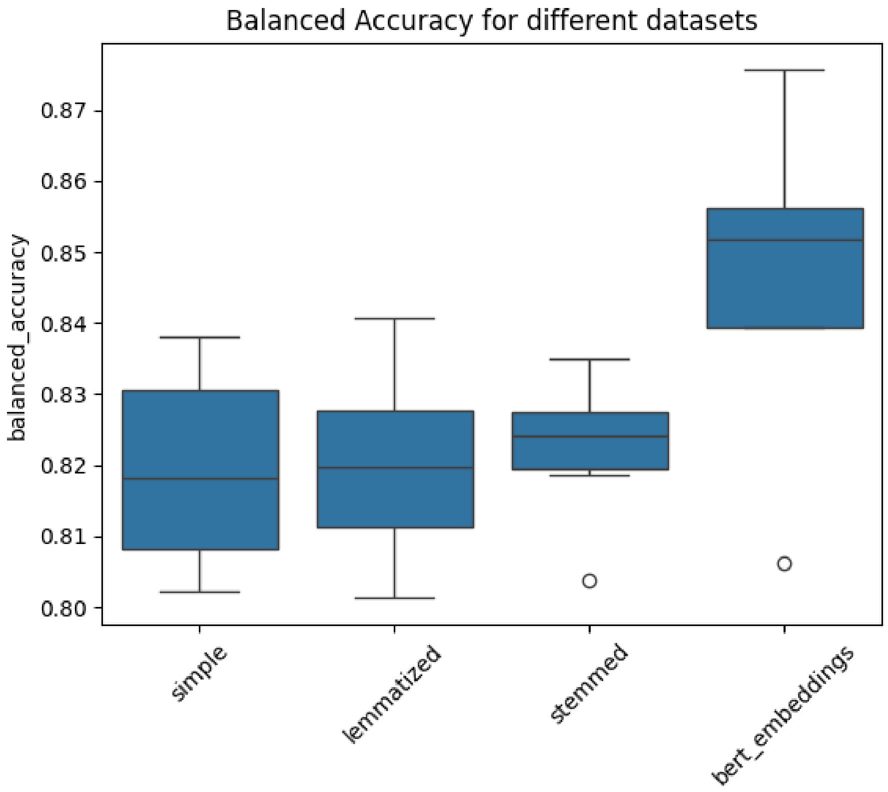
H2: There will be more misspelled words in content generated by bots.

Datasets/EDA - verification of hypothesis



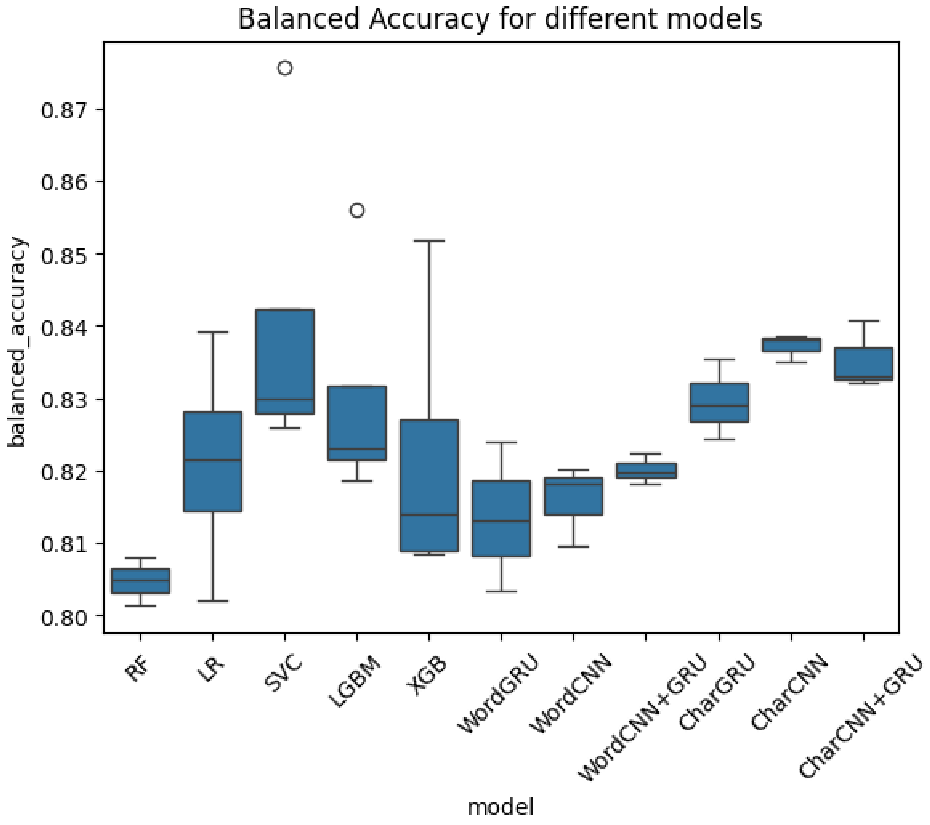
H3: The impact of different URL encoding, e.g., encoding all URLs to a single token vs extracting the basepath of the URLs.

Results



Balanced Accuracy for different datasets
balanced_accuracy
0.80
0.81
0.82
0.83
0.84
0.85
0.86
0.87
simple
lemmatized
stemmed
bert_embeddings
dataset

Results



Balanced Accuracy for different models
balanced_accuracy
0.87
0.86
0.85
0.84
0.83
0.82
0.81
0.80
RF
LR
SVC
LGBM
XGB
WordGRU
WordCNN
WordCNN+GRU
CharGRU
CharCNN
CharCNN+GRU
model

Results

model	dataset	balanced_accuracy	f1_score	precision	recall
SVC	bert_embeddings	**0.876**	**0.876**	**0.873**	0.880
LGBM	bert_embeddings	0.856	0.859	0.843	0.876
XGB	bert_embeddings	0.852	0.855	0.839	0.870
CharCNN+GRU	lemmatized	0.841	0.852	0.798	**0.914**
LR	bert_embeddings	0.839	0.842	0.830	0.853
CharCNN	lemmatized	0.839	0.845	0.813	0.880
CharCNN	simple	0.838	0.843	0.818	0.870
CharGRU	simple	0.835	0.842	0.810	0.877
CharCNN	stemmed	0.835	0.837	0.826	0.849
CharCNN+GRU	stemmed	0.833	0.844	0.792	0.904

Future works

- Fine-tuning transformers
  - With/without GPT-2 output dataset
- XAI

Thank you

Bibliography

- David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection. CoRR, abs/1907.09177.
- Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. CoRR, abs/1906.03351.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language under-standing.
- Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. PLOS ONE, 16(5):1–16, 05
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.
- Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. CoRR, abs/1602.01585.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1808–1822, Online, July. Association for Computational Linguistics.

Bibliography

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.
- Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1–47, mar.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203.
- James Vincent. 2018. Why we need a better definition of ‘deepfake’ / let’s not make deepfakes the next fake news. https://www.theverge.com/2018/ 5/22/17380306/deepfake-definition-ai-manipulation-fake-news, May.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019a. Defending against neural fake news. CoRR, abs/1905.12616.
- Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. CoRR, abs/1509.01626.