DeepFake Tweet Detection

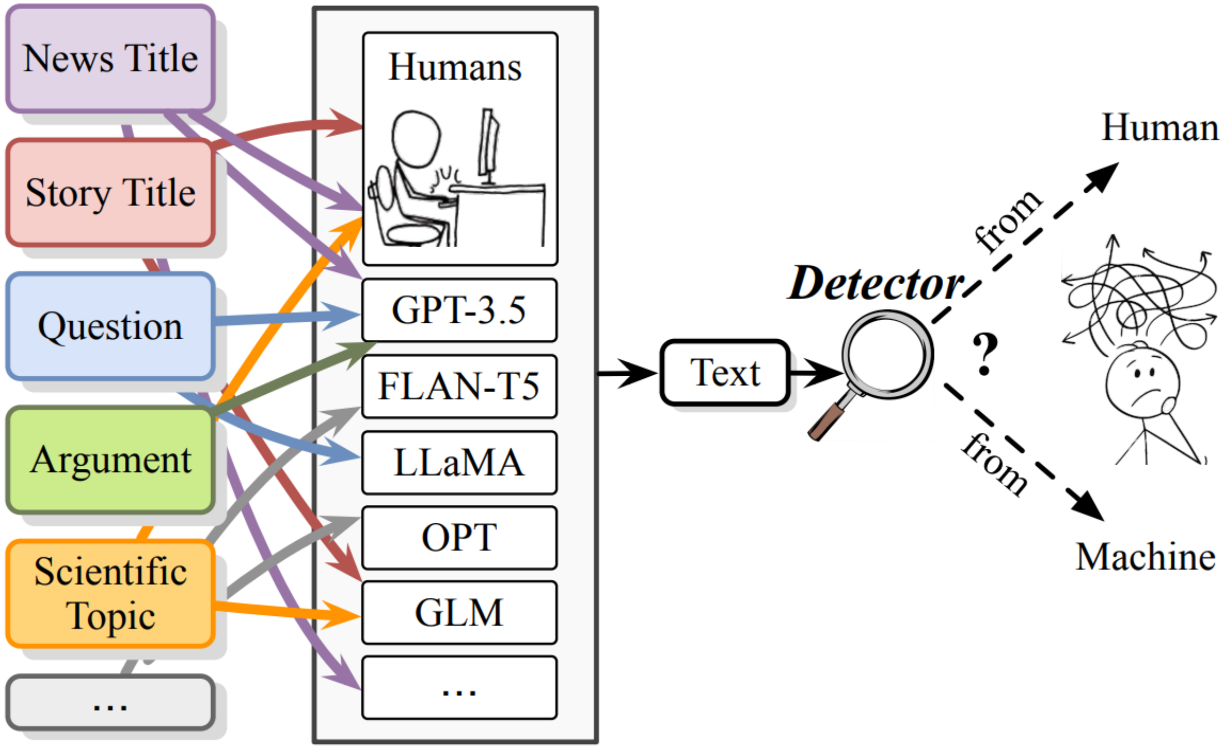
Team 15: TextTechTitans

ADRIAN KAMIŃSKI, ADAM FREJ,
PIOTR MARCINIAK, SZYMON SZMAJDZIŃSKI

JANUARY 2024

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

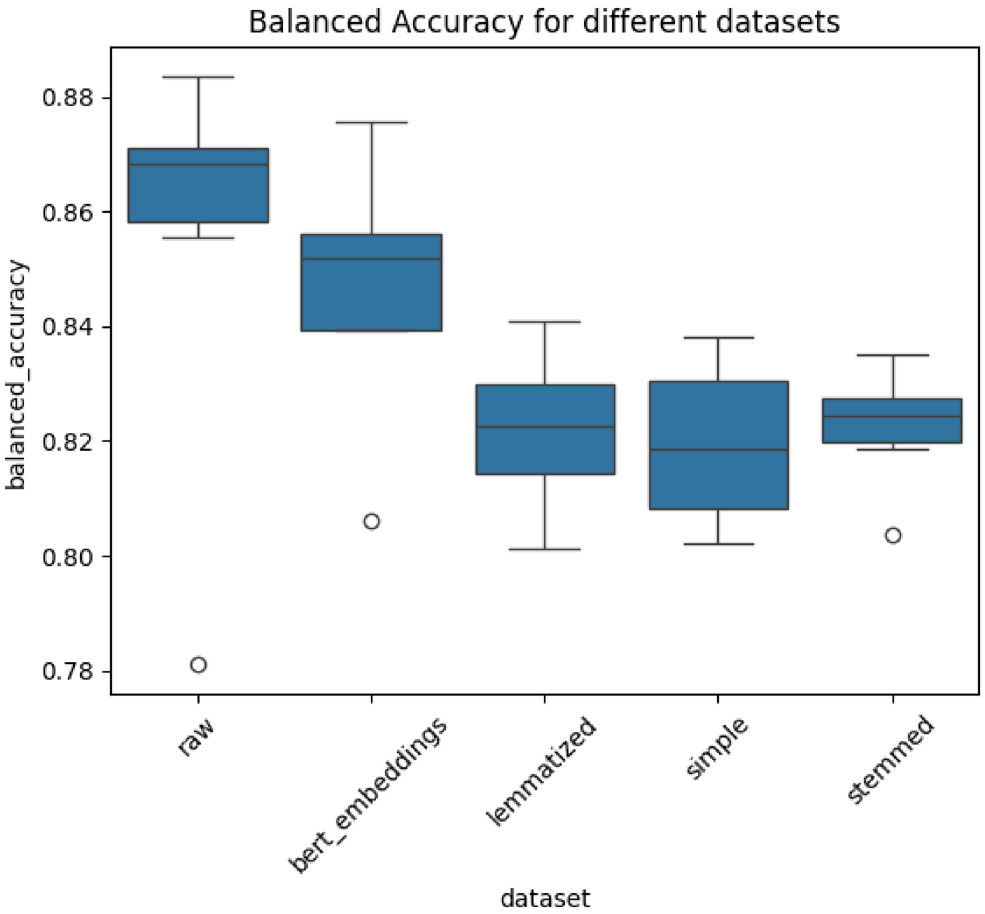
Hypotheses

- The use of emoticons may be higher in human-generated content.
  Roughly true
- The use of mentions of other users may be higher in human-generated content.
  True
- There will be more misspelt words in content generated by bots.
  False
- The impact of different URL encoding, e.g., encoding all URLs to a single token vs ex-tracting the basepath of the URLs
  Unrelated

Conclusions:

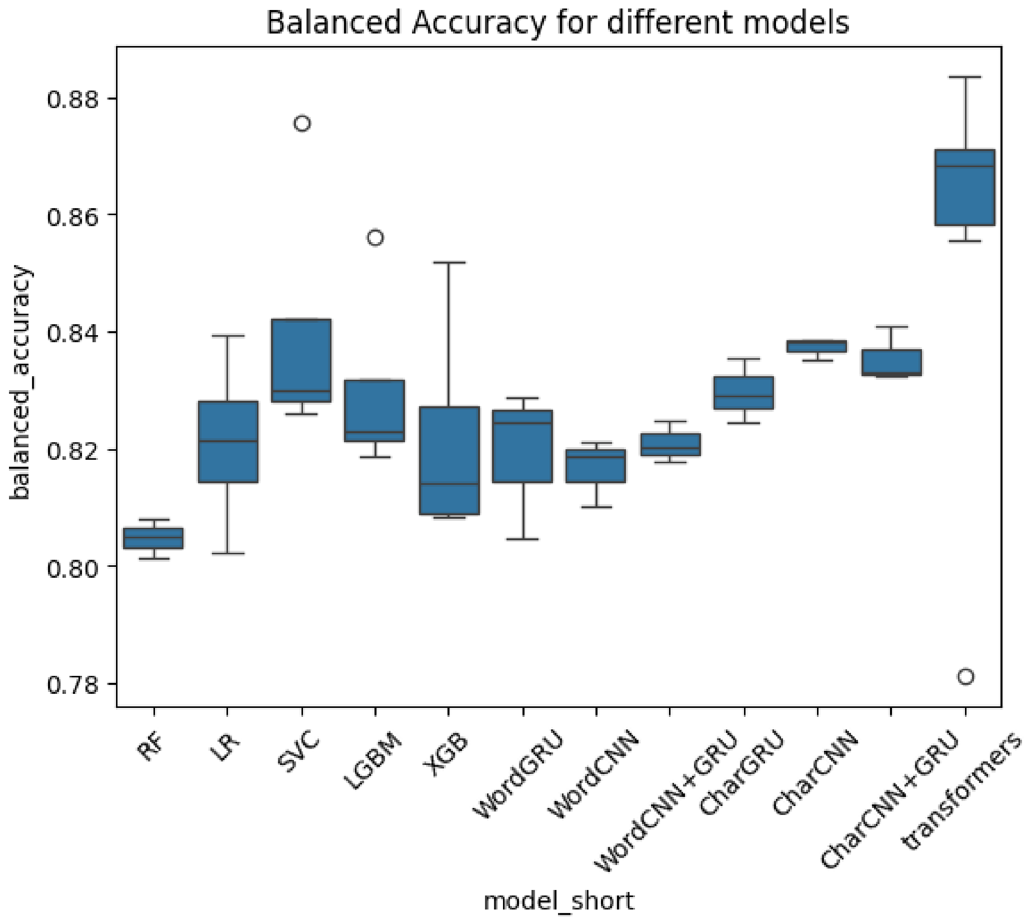
- Hypotheses show the direction of the research
- They are only a suggestion
- Not correct ≠ useless
- Hard to find good hypotheses at the beginning of the research

Results



Balanced Accuracy for different datasets
balanced_accuracy
0.88
0.86
0.84
0.82
0.80
0.78
raw
bert_embeddings
lemmatized
simple
stemmed
dataset

Results


Balanced Accuracy for different models
balanced_accuracy
0.78
0.80
0.82
0.84
0.86
0.88
RF
LR
SVC
LGBM
XGB
WordGRU
WordCNN
WordCNN+GRU
CharGRU
CharCNN
CharCNN+GRU
transformers
model_short

Results – accuracy in different categories

model name	TWEET CREATOR (CATEGORY)				
	ALL	GPT2	HUMAN	OTHERS	RNN
ROBERTA_FT (TweepFake)	**0.896**	0.74	0.89	0.95	**1.00**
XLM_ROBERTA2_raw	0.8835	0.6953	**0.8959**	0.9153	0.9830
SVC_bert_embeddings	0.8757	0.6927	0.8717	0.9442	0.9782
XLM_ROBERTA1_raw	0.8714	**0.8307**	0.8130	0.9607	0.9854
DisitlBERT0_raw	0.8698	0.6589	0.8795	0.9112	0.9879
GPT2_raw	0.8671	0.6693	0.8560	0.9587	0.9782
LGBM_bert_embeddings	0.8561	0.6745	0.8365	0.9483	0.9782
DistilBERT1_raw	0.8554	0.6849	0.8725	0.8471	0.9709
XGB_bert_embeddings	0.8518	0.6562	0.8333	0.9525	0.9733
CharCNN_GRU_lemmatized	0.8409	0.7760	0.7676	**0.9628**	0.9854
LR_bert_embeddings	0.8393	0.6380	0.8255	0.9236	0.9709

Results – with more details

model	**dataset**	**Accuracy**	**F1**	**Precision**	**Recall**
ROBERTA (TweepFake)	raw	**0.896**	**0.897**	0.891	0.902
XLM_ROBERTA2	raw	0.8835	0.8821	**0.8934**	0.8711
SVC	bert	0.8757	0.8763	0.8729	0.8797
XLM_ROBERTA1	raw	0.8713	0.8786	0.8328	**0.9297**
DISTIL_BERT0	raw	0.8698	0.8686	0.8773	0.8602
GPT2	raw	0.8671	0.8686	0.8593	0.8781
LGBM	bert	0.8561	0.8590	0.8429	0.8758
DISTIL_BERT (merged)	raw	0.8554	0.8529	0.8681	0.8383
XGB	bert	0.8518	0.8546	0.8395	0.8703
CharCNN+GRU	lemmatized	0.8408	0.8518	0.7975	0.9141
LR	bert	0.8393	0.8416	0.8304	0.8531

Lesson learnt

- The value of feedback
- Addition of more data does not necessary mean a better results.
- It is challenging to estimate workload of project in its early stages
- The literature review helps with future research
- Detection of text deepfakes – uphill battle

Contribution

- Promising results
- Research towards detection of GPT texts
- Overview of detection algorithms in different generative settings

Real world applications:

- Misinformation, fake news prevention
- Impersonation, privacy violations and identity theft prevention
- Increasing positive online experience (trust, safety and confidence in online interactions)

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.