DeepFake Tweet Detection

Adrian Kamiński, Adam Frej, Piotr Marciniak, Szymon Szmajdziński

November 2023

Presentation Outline

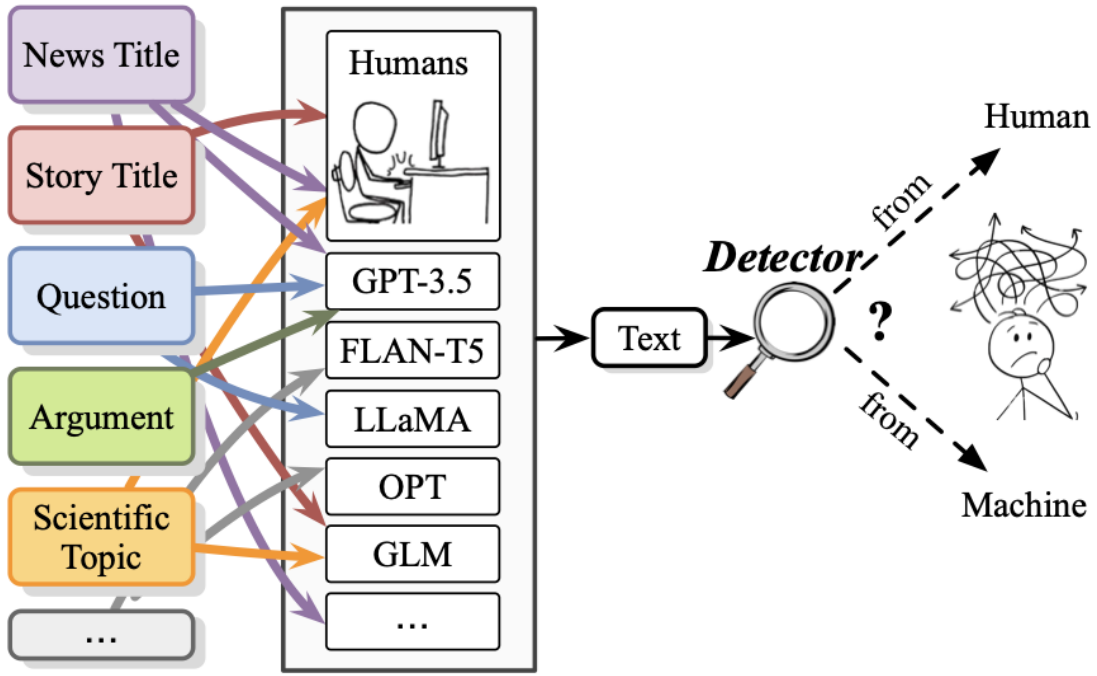
1. Introduction
2. Datasets description
3. Current methods for deepfake detection
4. Concept and work plan

Deepfake tweet detection

- there is nothing else as whaley lovely as a whale!!
- ....BUT MAYBE I STILL WILL. WITCH HUNT!

Deepfake

- Deep learning + fakes
- Tweets – short texts without context used in social media interactions
- Why is it a problem, why do we need detection tools. (humans low accuracy in this task)



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?

Other hypothesis

- The use of emoticons may be higher in human-generated content
- The use of mentions of other users may be higher in human-generated content
- There will be more misspelled words in content generated by bots

TweepFake - Twitter deep Fake text 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

Other datasets

- GPT-2 output datasets
  - Group of several datasets
  - Different decoding strategy settings and models sizes
  - Datasets based on excerpts of web texts
- HC3 – English (Human ChatGPT Comparison Corpus)
  - Group of datasets from different sources
  - Datasets consisting of questions and answers
  - For each question, human expert and ChatGPT3.5 answers are provided

Some popular methods

Bag-of-words + ML

Character level encoding + DL

BERT + ML

Pre-trained models

Bag-of-words + ML

TweepFake:

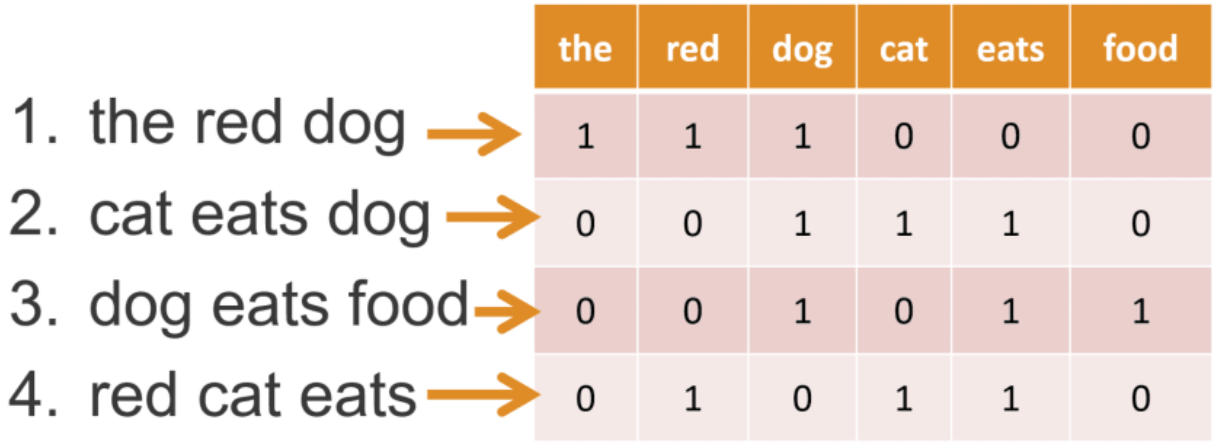
- TF-IDF function
- Logistic regression, random forest or SVM

Other works:

- GPT-2's 50,000 token vocabulary

Drawbacks:

- Curse of dimensionality
- Lack of semantic context



	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Source: https://www.ronaldjamesgroup.com/blog/grab-your-wine-its-time-to-demystify-ml-and-nlp

Character level encoding + DL

- Character as input instead of words or tokens.
- Encoding
- Deep Learning models, e. g. CNN, RNN
- Surprisingly good, fit to situations without pre-trained solutions

TweepFake:

- CHAR_CNN, CHAR_GRU, CHAR_CNNGRU

Other works:

- Zhang X, Zhao J, LeCun Y. Character-level Convolutional Networks for Text Classification (NIPS 2015)

BERT + ML

- Using BERT to provide embedded text with context
- Fixed-size vector representations
- Vectors depend on context
- Merging vectors to single per text (e.g. average)
- ML classifiers to predict labels

Pre-trained models

- Fine-tuned to specific dataset
- Take raw input and produce labels
- Complex sequence-based understanding level
- BERT, XLNet, RoBERTa, DistilBERT

- Most effective
- Well balanced in terms of precision - recall

Concept and work plan

- Exploratory Data Analysis (EDA)
- Data preprocessing
  - Removal of stop words
  - Stemming, Lemmatization
  - TF-IDF, Bag Of Words, Vector Word Representation
- Training ML models
  - Logistic Regression, SVC, Random Forest, XGBoost, LGBM
  - Usage of Bayes Optimization
- Inspection of ML models
  - Identification of important variables
  - Checking the results on different generative categories (RNN, GPT-2, others)

Concept and work plan

- Training simple DL models
  - Character-based
  - Word-based
- Investigation of DL models

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 understanding.
- 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.
- 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.