Deepfake tweets detection
Project Proposal for NLP Course, Winter 2022

Adam Frej
Warsaw University of Technology
01151392@pw.edu.pl

Adrian Kamiński
Warsaw University of Technology
01151387@pw.edu.pl

Piotr Marciniak
Warsaw University of Technology
01151428@pw.edu.pl

Szymon Szmajdziński
Warsaw University of Technology
01151438@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This research project aims to address the growing concern of deepfake text generation, specifically focusing on detecting deepfake tweets. Deepfakes, generated using advanced machine learning techniques, have the potential to spread misinformation, impersonate individuals, and facilitate malicious activities. Detecting such deepfake content is crucial to protect society from deception and harm. The project leverages the TweepFake dataset, which contains tweets from humans and bots. It also explores various text representations and preprocessing techniques.

The research questions revolve around the development of a reliable deepfake detection algorithm and the identification of effective features for tweet-based deepfake detection. The project hypotheses involve exploring patterns, such as the use of emoticons, mentions, and misspelled words, that may indicate machine-generated tweets. Different machine learning models will maximize detection accuracy while maintaining precision and recall balance.

Additionally, the project explores deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to capture the complexities of tweet structures. The research involves comparing and fine-tuning models for optimal performance.

The proposed work intends to improve upon state-of-the-art in deepfake tweet detection and contribute to more trustworthy online interactions. Ultimately, the project seeks to enhance the safety, trust, and confidence of users in their online experiences.

1 Introduction

The goal of our project is to determine if the content of a tweet is a deepfake. The term "deepfake" is a portmanteau of "deep learning" and "fakes" (Vincent, 2018). It refers to a type of synthetic content that is created using advanced machine learning techniques, particularly deep learning algorithms. Deepfakes are typically associated with manipulated videos, but they can also involve audio, images, and text. The core characteristic of a deepfake is that it convincingly alters or generates content to make it appear as if it is authentic, even when it is not. This work focuses on deepfake related to text corpora, in particular, to tweet data. This depicts human interaction in a short, dynamic form. Texts tend to be shorter and contain less context, which poses more challenges.

There are several reasons to tackle the problem

of deepfake tweet detection. One of them can be protection against misinformation. Deepfake tweets can be used to spread misinformation, disinformation, and fake news, which can have serious real-world consequences, such as influencing public opinion, election outcomes, and even inciting violence. Detecting and mitigating deepfakes is crucial to protect society from being misled by false narratives. Moreover, an increasing number of synthetic tweets can lead to a lack of trust in information shared on social media platforms.

Another reason for tackling the problem of deepfake tweet detection is the protection of individuals against impersonation, privacy violations, and identity theft. Deepfake tweets can be used to impersonate both public figures and private citizens, and by addressing this issue, we can safeguard people's rights and personal information.

Deepfakes can also be exploited for malicious purposes, including extortion, fraud, and harassment. They can be used to create a network of fake users who encourage other users to use some services. Detecting deepfakes is necessary to prevent these harmful actions.

Detection of deepfake corpora can also lead to exposure of flaws in existing text-generating algorithms, which can help to improve the language models in the future. This means creating texts that are more exciting and semantically plausible to read. Furthermore, comparing the models to human benchmarks can identify the types of errors in generated sequences that humans tend to notice and make text appear unnatural.

The proper detection of deepfake tweets can increase the trust, safety, and confidence of users in online interactions and content consumption. It can contribute to a more positive online experience.

2 Related works

The TweepFake dataset was created in (Fagni et al., 2021), where authors compared multiple approaches in both corpora encoding and algorithms to detect fake tweets. What is unique about this dataset is that the data used there was generated by various algorithms and was extracted directly from social media. The following setup can provide more generic results when evaluating models. Their results suggest that transformer-based models offer better results. They also found out that all approaches struggled with tweets generated by GPT-2, and the best results were obtained using an RNN decoder using character encoding.

One of the main difficulties in natural language processing is text representation. Related works investigate several solutions to this problem in the context of feeding data models detecting machine-generated corpora. The TweepFake (Fagni et al., 2021) article proposes 4 options. One of them is a popular method, bag-of-words (BoW), with features weighted using TF-IDF function (Sebastiani, 2002). The output of this methodology was processed by either logistic regression, random forest, or SVM. However, this approach suffers from the curse of dimensionality, as the features are very sparse and require a lot of data. Another drawback is that the algorithm misses the semantic context of the words. The article got the worst result here, hovering around 0.80 accuracy.

Another option is to encode text using a high-level language model, which overcomes these limitations. The TweepFake used BERT (Devlin et al., 2019) to provide contextual embeddings that include words context and can be encoded into a vector representing specific text. Yet again, this representation is later processed by classifiers.

The third approach operates on the character level. A vocabulary of characters is mapped to the internal embedding matrix, which is passed to the selected deep learning networks. This results in a surprisingly good score, up to 0.85 accuracy, and can be useful in cases without pretrained models available, for example, another language.

However, the most successful approach is utilizing pre-trained models. They take raw input directly and, with fine-tuning on a given dataset, solve the classification problem of labeling a sentence as human or machine-generated. This way, the models operate on the complex sequence-based understanding level. The TweepFake tested several language models, all related to BERT. Besides the original BERT, XLNet (Yang et al., 2020) and RoBERTa (Liu et al., 2019) were utilized as they can produce 15% better results thanks to architecture modifications and a bigger training dataset. The article also tested DistilBERT (Liu et al., 2019), which tries to keep performance while simplifying the architecture and halving the number of parameters. Generally, the results get up to 0.90 accuracy, the best being RoBERTa, which indicates that the most complex representations are most effective. They are also well-balanced in

terms of precision and recall.

One of the works tried to detect generated text by exploiting situations when humans are fooled by it (Ippolito et al., 2020). They also investigated several text representations. Similarly, the primary one is a fine-tuned BERT (Devlin et al., 2019). Again, this approach far surpassed other methods depicted in the article.

Another representation in this work is a simple BoW. This time, the GPT-2's $50{,}000$ token vocabulary (Sennrich et al., 2016) is used to count the occurrences of tokens in text sequences. This embedding allowed for training logistic regression binary classifier, which achieved the next best result.

The article also proposed Histogram-of-Likelihood Ranks. As in GLTR (Gehrmann et al., 2019), they created an energy-based deepfake text decoder by calculating the probability distribution of the next word given the previous words in the sequence according to GPT-2 language model. They ranked the words by likelihood and then binned them either into 4 groups or uniformly over the whole vocabulary. Such histograms served as input for logistic regression binary classifiers. Unfortunately, this approach resulted in worse scores despite successes reported in GLTR, which can be explained by training data selection.

Another deepfake detection performed on social media texts was conducted on Amazon reviews in (Adelani et al., 2019). Fake news were generated using the GPT-2 text generation model (Radford et al., 2019). Authors, in order to adjust the model to new corpora, adapted the original GPT-2 model to Amazon (He and McAuley, 2016) and Yelp reviews (Zhang et al., 2015). As for detection algorithms, they used Grover (Zellers et al., 2019a), GLTR (Gehrmann et al., 2019), and OpenAI GPT-2 (Solaiman et al., 2019). They also tried combining those models using logistic regression at the score level. One of the experiments they performed was selecting a real review out of 4, of which 3 were fake. Human participants tended to randomly guess which review was real since they were right about 25% of the time. However, the models were not much better. The best configuration achieved 20% accuracy on this task.

In (Guo et al., 2023), authors proposed HC3 dataset consisting of questions and their corresponding human/ChatGPT answers. Based on this dataset, they conducted a comprehensive human evaluation and linguistic analysis as well as developed several detecting models. They used the GLTR model with logistic regression, RoBERTa, and RoBERTa-QA - a Question Answering version of the model that supports a text pair input format, where a separating token is used to join a question and its corresponding answer. In their work, they came to the following conclusions:

- The robustness of the RoBERTa-based-detector is better than GLTR.
- RoBERTa is not affected by indicating words (characteristic words for ChatGPT).
- RoBERTa is effective in handling Out-Of-Distribution scenarios, whereas we can observe a significant decrease in performance on GLTR's when testing on data in first-seen format.
- Detecting ChatGPT-generated texts is more difficult in a single sentence than in a full text.

3 Datasets

Several datasets were created to build a solution to detect content created by deep learning algorithms. Some of them are presented below.

1. The TweepFake dataset (Fagni et al., 2021) - it is a dataset which contains $25{,}572$ tweets half human and half bots generated. They used 17 human accounts, which were imitated by the 23 bots. Some of the fake accounts imitated the same human profile. These bots were using different technologies, and for almost all of them (except one bot account), the used technology is known.

2. The GPT-2 output datasets (Ippolito et al., 2020) - it is a group of several datasets in which deepfakes are generated by GPT-2 models (Radford et al., 2019). The datasets differ because different decoding strategy settings and different sizes of models are applied during generation. Each dataset contains $500{,}000$ training and $5{,}000$ validation and test samples, which are evenly spread across classes (human-generated excerpts of web texts or GPT-2 generated).

3. The HC3-English datasets (Guo et al., 2023) - it is a group of datasets from different sources, in which for each question, there

is provided at least one human, and ChatGPT3.5 answer. The questions and answers by the human experts come mainly from publicly available question-answering datasets. There is also an additional source in which Wikipedia is treated as a human expert who is asked questions based on concepts in crawled data.

In most papers (Zellers et al., 2019b; Bakhtin et al., 2019), datasets containing generated and human texts are not provided because big corpora are used to build a generative model in which descriptors act as detectors of deepfakes. Later, they test their descriptors on unused parts of corpora and text produced by generators to see if the descriptors are working in the correct way.

4 Concept and work plan

In our project, we would like to investigate several research questions, which are:

- Can we build a reliable deepfake detection algorithm? By reliable algorithm, we mean the model that will maximize accuracy on a balanced dataset while remaining well-balanced in terms of precision and recall. That means detecting generated tweets while avoiding assigning false positives.
- What are the most effective features for deepfake detection in tweets?
- Are there any patterns that indicate the model-generated tweet content?

There are also some hypotheses which we would like to test:

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

Our investigation of these questions and hypotheses is going to start with explanatory data analysis (EDA) of the Tweep-Fake dataset. We hope that the results of this analysis will give us some insight into text structure and allow us to preprocess it accordingly.

Table 1: The timetable of project

deadline	description of task
10.11.23	Exploratory Data Analysis (EDA)
12.11.23	Data preprocessing
16.11.23	Initial building of ML models
20.11.23	Inspection of ML models
21.11.23	Report + presentation
22.11.23	Project 1 - PoC
02.12.23	Post review code adjustments
04.12.23	Post review report adjustments
08.12.23	Initial building of simple DL models
10.12.23	Inspection of DL models
12.12.23	Report + presentation
13.12.23	Final report
20.12.23	Post review code adjustments
20.12.23	Post review final report adjustments
20.12.23	Deadline, project 1

We are going to try different text encodings, i.e., different tf-idf functions, a bag of words, word vectors, as well as text preprocessing (URL, user mentions replacements). We would also like to check the impact of stemming and lemmatization of words on machine learning effectiveness. The usage of stemming and lemmatization was not mentioned in the article (Fagni et al., 2021).

In this work, we will focus on training the following machine learning models on the Tweep-Fake dataset: logistic regression, SVC, Random Forest, as well as models that weren't considered in (Fagni et al., 2021), XGBoost, and LGBM. For these machine-learning models, the most optimal hyperparameters will be chosen by Bayesian optimization instead of grid search with cross-validation that was used in the original TweepFake paper. Later, we will investigate which features occurred to be the most important according to SHAP values or in-built model methods and which approaches achieved the best results. We are going to check which category of the model caused the most problems for our classification models. We will try to improve the results of models presented in article (Fagni et al., 2021).

As well as in (Fagni et al., 2021), we will lever-

age another effective way to encode textual contents by working at the character level. We will try different simple CNN/RNN models in such a configuration. The simple RNN models mean the models consisting of 1-2 layers of LSTM or GRU (bidirectional or not). The simple CNN models mean a network composed of a couple of convolutional and pooling layers. Additionally, we would like to check simple CNN/RNN models working at the word level with custom (trained) and Bert embeddings.

The schedule of the project with the deadlines is in the Table 1.

References

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.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019b. 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.