Looking for Traces of Textual Deepfakes in Bulgarian on Social Media

Irina Temnikova[1] Iva Marinova[2] Silvia Gargova[1] Ruslana Margova[1] Ivan Koychev[3]
[1]Big Data for Smart Society Institute (GATE), Bulgaria,
[2]Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria,
[3]Sofia University "St. Kliment Ohridski", Bulgaria
{irina.temnikova, silvia.gargova,ruslana.margova}@gate-ai.eu
iva.marinova@identrics.ai
koychev@fmi.uni-sofia.bg

Abstract

Textual deepfakes can cause harm, especially on social media. At the moment, there are models trained to detect deepfake messages mainly for the English language, but no research or datasets currently exist for detecting them in most low-resource languages, such as Bulgarian. To address this gap, we explore three approaches. First, we machine translate an English-language social media dataset with bot messages into Bulgarian. However, the translation quality is unsatisfactory, leading us to create a new Bulgarian-language dataset with real social media messages and those generated by two language models (a new Bulgarian GPT-2 model – GPT-WEB-BG [1], and ChatGPT). We machine translate it into English and test existing English GPT-2 and ChatGPT detectors on it, achieving only 0.44-0.51 accuracy. Next, we train our own classifiers on the Bulgarian dataset, obtaining an accuracy of 0.97. Additionally, we apply the classifier with the highest results to a recently released Bulgarian social media dataset with manually fact-checked messages, which successfully identifies some of the messages as generated by Language Models (LM). Our results show that the use of machine translation is not suitable for textual deepfakes detection. We conclude that combining LM text detection with fact-checking is the most appropriate method for this task, and that identifying Bulgarian textual deepfakes is indeed possible.

1 Introduction

The term "deepfake", comes from "deep learning" and "fake" and indicates (potentially) fake texts, images, or videos, generated using deep learning models (Gambini, 2020). Among them, "Textual DeepFakes" (TDF) refer to texts generated automatically with the help of Generative Models (GMs, and lately with Large Language Models - LLMs), which may also contain fake or untrue content. This makes those of them, which are spread with the intention to deceive, the automatic variant of disinformation (as defined by the European Commission (EC)[2]. According to this EC's definition, "disinformation" is *"false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm"*. There exist useful GMs and LLMs applications (Kasneci et al., 2023). However, textual deepfakes can be a serious problem when spread on official information channels, as they can reach a large number of people. They are also problematic because, when fluent, they are hard to recognize by humans (Crothers et al., 2022). TDFs can be used by politicians in their political fights and destroy a person's reputation, or to influence a large number of people about sensitive topics such as a war or health. TDFs can be especially problematic on social media, as anybody can have access to such platforms and freely post information, which can be easily spread to a larger number of population subgroups including those who do not usually follow the official media channels (such as teenagers).

There is Natural Language Processing (NLP) research on detecting LM-generated texts and TDFs for English and other languages (Jawahar et al., 2020; Fagni et al., 2021; Kowalczyk et al., 2022; Gambini et al., 2022; Stiff and Johansson, 2022; Sadiq and Ullah; Shamardina et al., 2022; Chen et al., 2022b). However, to the best of our knowledge, there is no such research for Bulgarian.

Differently from most previous works, we consider Textual DeepFakes (TDFs) not just as any texts generated by LMs, but specifically those LM-

[1] https://huggingface.co/usmiva/gpt-web-bg.

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423. Last accessed on April 7th, 2023.

Proceedings of Recent Advances in Natural Language Processing, pages 1151–1161
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_122

generated texts that contain fake information. However, we also consider detecting LM-generatedness as an important aspect in detecting TDFs.

Detecting textual deepfakes for Bulgarian is a challenging task, as there are no LM-generated and textual deepfakes datasets in Bulgarian. While some English-language methods can use LM-generated messages actually posted on Twitter by self-proclaimed bots (Fagni et al., 2021), we could not identify such in Bulgarian.

We test three approaches to detect LM-generated texts. Two of them use Machine Translation (MT), as this is a very frequent method in lower-resourced settings. While we suspect that we might not get good results in translating already broken LM-generated texts, we experiment with MT due to the lack of appropriate Bulgarian datasets and LM-generated text detectors for Bulgarian.

Approach 1 (described in Section 4.1) tests MT for translating into Bulgarian an existing English-language dataset of actually occurring LMs-generated tweets (TweepFake[3] (Fagni et al., 2021)). Our subsequent plan is to build classifiers on the machine-translated messages.

As the results of machine translating the TweepFake messages into Bulgarian are not satisfactory, we test Approach 2 (explained in Section 4.2). We generate a Bulgarian language dataset composed of human-written messages and those generated by ChatGPT and GPT-WEB-BG. Next, we machine translate this dataset into English. We do this to test existing LM detectors for English, which are already trained on much more data.

As the English-language LM detectors in Approach 2 show a low accuracy, we apply Approach 3 by training classifiers (see Section 4.3) on our Bulgarian LM-generated dataset.

Finally, in order to add the "fakeness" aspect of textual deepfakes to them being generated by an LM, we run a final experiment (described in Section 4.4). We apply the classifier with the highest test results from Approach 3 on a recently published Bulgarian social media dataset manually fact-checked and annotated for containing untrue information and disinformation. We do this to check if the classifier would recognize any untrue messages or such containing disinformation as LM-generated.

The rest of the article is structured as follows: Section 2 discusses the Related Work. Section 3 introduces the existing datasets used. Section 4 presents each approach with its results. Section 5 provides a Discussion, Conclusions, and Future Work, and the following unnumbered sections contain the Limitations of this work, the Ethical and Legal statements, the Broader Impact Assessment, and the Acknowledgments.

2 Related Work

In comparison with detecting deepfake images and videos, until recently there was a limited number of Natural Language Processing (NLP) works on detecting textual deepfakes, and efforts were focused mostly on English (Fagni et al., 2021). With the recent appearance of several Large Language Models (LLMs), including the freely available for many languages ChatGPT[4], the amount of NLP works detecting LLMs-generated and deepfake texts has increased (Orenstrakh et al., 2023). Detectors for new languages[5] have also appeared (Antoun et al., 2023).

The work on detecting textual deepfakes usually checks if the texts have been generated by one or more LMs, generally training classifiers on LM-generated and human texts (Fagni et al., 2021; Gambini, 2020; Gambini et al., 2022), with recent zero-shot approaches appearing too (Mitchell et al., 2023).

The most recent **language models** are the deep learning ones: Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), the Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019), GPT-3 (Brown et al., 2020), ChatGPT[6], the Google's Pathways Language Model 2 (PaLM 2), used in Bard[7] and BLOOM (Scao et al., 2022). LM detectors check also for texts, generated with older neural LMs, such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN).

Most LMs can generate content in English, but there are also models for generating text in other languages such as Chinese, Bengali, Arabic, Russian, Korean, Slovak, Spanish, Czech, German, French, and Macedonian. Pre-trained Large Language Models (LLMs) can be found online (e.g.

[3]https://www.kaggle.com/datasets/mtesconi/twitter-deepfake-text. Last accessed on April 7th, 2023.

[4]https://chat.openai.com/. Last accessed on July 27, 2023.

[5]For example https://detector.dng.ai/ - a ChatGPT detector for English and French. Last accessed on July 27th, 2023.

[6]https://openai.com/blog/chatgpt

[7]https://bard.google.com/. Last accessed on July 27, 2023.

on the Hugging Face platform). **The newest LMs for Bulgarian** are OpenAI's GPT-3.5 and GPT-4, Google's Bard, and the GPT-WEB-BG[8] (GPT-2-based) model, which we use together with ChatGPT (model 3.5) in this article. There are also 3 older pre-trained Bulgarian GPT-2 LMs (all by the same author) - 2 small and one medium[9]. These models were trained on Bulgarian data from books, Wikipedia, and the Oscar corpus. We considered them unsuitable for our task, as they generated texts with unsatisfactory quality.

With the fast advances in text generation models came the need for **synthetic text detectors**. LMs detection methods fall into three major categories: 1) simple classifiers; 2) zero-shot classifiers and 3) fine-tuning neural LMs (NLMs) (Jawahar et al., 2020). Simple classifiers use classical machine learning methods to train models from scratch to discriminate between synthetic text generated from LMs and human-written texts. Zero-shot classifiers use a pre-trained generative model (e.g. GPT-2 output detector, DetectGPT (Mitchell et al., 2023)) to detect if a text has been generated by the model used or by a similar model. These detectors do not require further training. In the NLM fine-tuning method a pre-trained LM (e.g., BERT, RoBERTa) is fine-tuned to detect text generated from itself or similar models. These detectors *do require* additional training. Several **pre-trained models for synthetic text detection** (mostly for English) are available online (for example in Hugging Face) - BERT, CLTR, GROVER, Open-AI GPT-2, AI Text Classifier, DetectGPT[10] and RoFT[11](human detector in the form of a game). Until our work, to the best of our knowledge, there weren't any synthetic text detectors that could work with Bulgarian.

Although there are generators for different languages, the **existing datasets for detectors training** are mostly in English (Fagni et al., 2021; Liyanage et al., 2022), with Chinese (Chen et al., 2022a) and Russian[12] (Posokhov et al., 2022; Shamardina et al., 2022) also available. However, there are no datasets with Bulgarian LM-generated texts, especially social-media-like.

Among the works, which are the most similar to ours are those on detecting LM-generated texts in other languages (e.g. Russian, Chinese, French) (Shamardina et al., 2022; Chen et al., 2022a; Antoun et al., 2023), but they only detect LM-generated texts, and ignore any fakeness of their content. Similar to ours is also the new research on detecting ChatGPT. An example is (Pegoraro et al., 2023), which tests a large number of available English LMs detectors and discovers that they are all not good at detecting ChatGPT (achieving <50 in True Positives Rate). However, this research detects English ChatGPT only and does not work with Bulgarian, nor does it detect textual deepfakes.

Finally, there are also approaches that detect (usually human-written) fake texts in social media, without taking into account the LM-generation aspect of textual deepfakes. These methods are usually based on detecting a specific style or analyzing the behavior of source accounts, comparing the messages with external news sources, and performing various types of (semi-)automatic fact-checking (Ghadiri et al., 2022; Krishnan and Chen, 2018).

3 Datasets Used

This section describes the existing datasets used in our experiments.

3.1 English TweepFake Dataset

The TweepFake dataset[13] (Fagni et al., 2021) contains 25,836 tweets in English (half of which are human-written and half are bot-generated), with each tweet actually published on Twitter. The data comes from 23 bots, imitating 17 human accounts, and the respective human accounts that the bots are imitating. The bots use different text generation models, such as Markov Chains, RNN, RNN+Markov, LSTM and GPT-2. We use TweepFake in our Approach 1 in Section 4.1.

3.2 Bulgarian Social Media Datasets

We have used five recently released (Temnikova et al., 2023) datasets of social media messages, posted on Twitter and Telegram between 1 January 2020 and the end of June 2022. Among them, 4 datasets (of a total of 118,570 messages) contain non-fact-checked social media texts. However, these datasets are on topics, related to Covid-19, lies and manipulation, and famous Bulgarian cases

[8]https://huggingface.co/usmiva/gpt-web-bg. Last accessed on July 27, 2023.

[9]https://huggingface.co/rmihaylov/

[10]https://detectgpt.ericmitchell.ai/

[11]https://roft.io/

[12]https://www.kaggle.com/competitions/ruatd-2022-multi-task/data

[13]https://www.kaggle.com/datasets/mtesconi/twitter-deep-fake-text. Last accessed on March 3rd, 2023.

when Bulgarian politicians were accused of lying. We selected exactly these datasets because they are more likely to contain untrue information or disinformation, given the nature of the topics (e.g., Covid-19, political statements), and because they are more recent than the previous ones (e.g. Nakov et al. (2021)). We used messages from these 4 datasets to generate our own LM texts for Approach 2 (Section 4.2).

The fifth dataset is a subset of these 4 datasets, containing 4083 messages[14]. Each message of it was fact-checked using external sources and manually annotated by 3 Bulgarian journalists for containing or not "Untrue information" and "Disinformation". This dataset is used in our Approach 4 (Section 4.4).

To these 5 datasets, we have added our own 104,138-messages Facebook dataset[15]. The Facebook dataset contains messages collected from official pages and public groups of Bulgarian media, parties, politicians, and political influencers from June 2021 to June 2022 using CrowdTangle[16], as well as from a historical search for the keyword "избори" (meaning in English "elections") in Bulgarian from 2006 until now. We selected this keyword, as according to our observations, many accusations of lying are published during elections. The Facebook dataset has been pre-processed similarly to what is described for the five publicly available datasets (Temnikova et al., 2023) in order to ensure compatibility: we removed duplicates, messages with fewer than 5 words, and non-Bulgarian messages using FastText's language identification tool.

In total, we used 222,708 Bulgarian social media messages.

4 Experiments

4.1 Approach 1: Machine Translating TweepFake into Bulgarian

First, we used the existing English Tweepfake dataset, due to the unavailability of Bulgarian-language bots on social media.

4.1.1 Methods

We performed experiments in which we tested the results of using Machine Translation (MT) to translate only the Tweepfake bot messages into Bulgarian. We suspected that we might obtain low-quality machine translation results; nevertheless, we tested this approach due to its common use in lower-resourced settings. We selected 16 messages, generated with different LMs, such as GPT-2 and RNN. We run them through 5 publicly available MT engines that were known to work well with the English-Bulgarian language pair: 1) Google Translate's free User Interface (UI)[17] 2) Google Translate's Google Spreadsheets function[18], 3) DeepL Translator UI[19], 4) GoURMET project's demo[20]; and 5) ChatGPT interface. Manual evaluation was done by two Bulgarian linguists, both with Natural Language Processing (NLP) and professional translation expertise. Each translation by all 5 MT engines of each message was evaluated for two categories: a) Is the meaning preserved? b) Are the characteristics of the message preserved (e.g. broken syntax, specific formatting, etc.)? Both categories had a 3-point scale (1 - not preserved; 2 - partially preserved; 3 - preserved).

4.1.2 Results from machine translating TweepFake dataset into Bulgarian

None of the engines performed satisfactorily for this task. The average score of both human evaluators was around 1 ("not preserved") for both categories. Google Translate and DeepL performed slightly better (1.5). IAA varied per engine and question, with higher agreement on the first question.

The analysis has revealed that all the engines encountered difficulties with translating the bot messages. This is due to the fact that the bot messages either contained slang or were almost completely incomprehensible with broken English syntax. The MT engines were either adding noise (Google Translate and GoURMET) or making the translated messages more fluent and human-like (DeepL). ChatGPT either corrected the bot's messages or commented that the messages were incomplete and could not provide a translation. Due to the aforementioned translation problems, we decided not to use this dataset for training the classifiers, and to instead create a new dataset for this purpose.

[14] https://zenodo.org/record/7702054. Last accessed on April 16th, 2023.

[15] This dataset cannot be shared due to Facebook's requirements.

[16] https://www.crowdtangle.com/

[17] https://translate.google.com/

[18] =GOOGLETRANSLATE

[19] https://www.deepl.com/translator.

[20] https://translate.gourmet.newslabs.co/

4.2 Approach 2: Bulgarian Dataset Generation and English LM Detectors Testing

We created our own dataset with real and LM-generated "social-media"-like messages. We then test existing English LMs detectors, which are supposed to work well because they are trained on much larger datasets.

4.2.1 Methods

Dataset Generation

We created a Bulgarian language dataset (from now on referred to as Deepfake-BG[21]), containing 9824 messages. Half of the messages (4912) were randomly chosen from the larger existing datasets, described in Section 3.2 in a way to have a higher probability that they were written by humans. For example, they were selected from Covid-19 disease and travel mutual help Facebook and Telegram public channels and groups, as well as from politicians' and political influencers' Facebook pages. The other 4912 messages contained an equal number of "social-media"-like messages, generated by two LMs - a new Bulgarian-language GPT-2 model (called GPT-WEB-BG) (Marinova et al., 2023) and ChatGPT for Bulgarian.

Generating messages with GPT-WEB-BG

GPT-WEB-BG[22] was trained on a dataset containing scraped content from major Bulgarian online media providers. The model is a part of an active development of a suite of LLMs for Bulgarian and the authors are incorporating more data from various domains such as social media, Wikipedia, books, and scientific literature. A specialized procedure was followed for source filtering, topic selection, and lexicon-based removal of inappropriate language for Bulgarian in order to prevent gender, race, and political bias, toxicity, or discrimination practices. GPT-WEB-BG generated messages by completion, starting from randomly selected Twitter and Facebook messages from the datasets, described in 3.2, which were different from those included in the "human" part of this dataset. The Deepfake-BG messages were generated using two methods: 1) 5 words from the original message, completed with 200 characters, and 2) 10 words from the original message, completed with 250 characters. Such generation produced properly looking messages, but also messages, containing repeated phrases or sentences, and truncated (interrupted) sentences. We removed the last two types of messages to make the classifiers' task harder. Next, we selected a random sample of the messages generated by GPT-WEB-BG. If there were two generated versions of an original message (one from both methods), we took randomly only one of them. Duplicates were removed, which led to the final number of 2456 messages on the following topics: 482 from Facebook public pages of Bulgarian media and political parties, 172 generated from Twitter messages on the "Covid-19" topic, and 1802 generated from Twitter messages on the "lies and manipulation" topic.

Bulgarian ChatGPT Generation

We also generated 2,456 ChatGPT messages on the same topics and in the same quantity per topic as the GPT-WEB-BG messages. The ChatGPT messages were generated by typing manually instructions into the UI in two ways: 1) Copy-pasting examples of human messages with the instructions: "Generate (5 or 10) social media messages (with emoticons and hashtags) like this one:...". The number (5 or 10) varied, according to the speed of generation and the necessary amount of messages. The instructions were written half of the time in Bulgarian, and half in English. 2) In 10 cases, and to generate more variety, we experimented with giving this instruction: "Write (5 or 10) social media messages (with emoticons and hashtags) on this topic:...". As in the previous cases, we cleaned the obtained messages from duplicates.

Testing English LM detectors

Next, we translated Deepfakes-BG dataset into English using three widely used and freely available MT engines - DeepL, Google Translate UI and the GOOGLETRANSLATE() function in Google Sheets. Upon reviewing the existing English LM detectors, we identified several problems. Firstly, freely available tools are usually trained to recognize either GPT-2 or ChatGPT, but not both (excluding zero-shot approaches). Among the available tools, only GPTZero is trained to recognize GPT-2, GPT-3, and ChatGPT, but it is a paid tool. Additionally, the majority of classifiers require longer texts, typically at least 40 words or a minimum of 2000 characters, while our texts are approximately 250 characters in length.

Another challenge is that each detector produces a different type of output. Some return only binary

[21]This dataset will be partially shared upon publication of this paper, and in compliance with social media platforms' requirements.

[22]https://huggingface.co/usmiva/gpt-web-bg.

labels (e.g., "Human" and "Machine"), while others also provide label probabilities. Additionally, some detectors only return a probability value (e.g., "52.63% AI-generated content"). This variability in output types makes comparative evaluation difficult.

We selected four detectors based on the following criteria: (1) freely available and (2) trained to recognize both GPT-2 and/or ChatGPT. The first detectors we selected are *roberta-base-openai-detector* detector [23] which is a RoBERTa base model for detection of GPT-2 generated texts, *chatgpt-detector-single* detector [24] for ChatGPT detection which uses pretrained large models based classifiers. We tested two more detectors *ChatGPT-Detection* [25] and *baykenney/bert-base-gpt2detector-topp96* [26]. However, the authors of these detectors do not provide information about them.

For our experiments, we used the binary version of our dataset as most detectors return a binary output. However, *ChatGPT-Detection* only returns probabilities. Consequently, we evaluated the output of the other detectors that provide both labels and probabilities and observed that the minimum probability for automatically generated texts was 50%. Based on this, we classified texts with a probability greater than 50% as "automatically generated" and those with a probability equal to or lower than 50% as "human texts".

After processing the translated texts using the detectors, we compared their results with the original labels and evaluated their accuracy.

4.2.2 Results from Deepfakes-BG Generation and Testing English LM-detectors

Comments on the Deepfakes-BG Generation Results

We observed that ChatGPT tended to generate advertisement-like short texts, and it needed several reminders, in order to change its style to be more social media-like. Since the original datasets contained messages both pro- and against official Covid-19 measures, we tried to generate messages about the adverse effects of Covid-19 vaccines. ChatGPT either refused to generate such messages, or generated messages, always ending with "however, it is better to get vaccinated". Our observations also reveal that ChatGPT's bias towards officially accepted positions can generate highly inaccurate statements. In fact, the model may attribute to a public figure, who has typically expressed opposing views to widely accepted beliefs, words that this individual never actually uttered.

Results from Testing the English LM Detectors on the Translated Deepfakes-BG Dataset

The experimental results are presented in Table 1. The tested detectors show an accuracy of approximately 50%. However, the results reveal that some detectors perform poorly on one of the classes, which may be attributed to two factors: (1) the translation of the text into English affects the outcome, and (2) the dataset is balanced, so even if the model predicts only one label for the entire test dataset (as in one of the cases), it will still achieve approximately 50% accuracy.

The length of the texts may also impact the results. As previously mentioned, many detectors require longer texts to accurately determine whether the text is automatically generated or human-written. This approach may not be practical, and there is a need to develop tools that can work with shorter texts.

We evaluated additional detectors beyond those previously described, however, the results obtained were similar to those already reported. Therefore, we have opted not to include them in the table.

4.3 Approach 3: Building Bulgarian-Language Classifiers on the Bulgarian Dataset

Due to the low accuracy results of Approach 2, we trained our own classifiers on the Deepfake-BG dataset.

4.3.1 Methods

We have trained several classifiers: Naive Bayes, Logistic Regression, K-Nearest Neighbors, Support Vector Machines (SVM), Decision Trees, Random Forests, and we have fine-tuned the newly released BERT-WEB-BG[27], obtaining BERT-Deepfake-BG[28]. We developed 2 models from the dataset- binary (human vs. LM) and multi class (human, GPT-WEB-BG, and ChatGPT). The dataset was split into train, validation, and test in

[23] https://huggingface.co/roberta-base-openai-detector

[24] https://huggingface.co/spaces/Hello-SimpleAI/chatgpt-detector-single

[25] https://huggingface.co/spaces/imseldrith/ChatGPT-Detection

[26] https://huggingface.co/baykenney/bert-base-gpt2detector-topp96

[27] https://huggingface.co/usmiva/bert-web-bg.

[28] https://huggingface.co/usmiva/bert-deepfake-bg, https://huggingface.co/usmiva/bert-deepfake-bg-multiclass.

Det.	Class	Google Sheets Function				Google Translate				DeepL			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
D1	Human		0	0	0		0	0	0		1	0	0
	LM		0.49	1	0.66		0.49	1	0.66		0.5	1	0.66
	Total	0.49	0.25	0.49	0.33	0.49	0.25	0.49	0.33	0.5	0.75	0.5	0.33
D2	Human		0.43	0.34	0.38		0.49	0.53	0.51		0.5	0.55	0.52
	LM		0.44	0.53	0.48		0.48	0.44	0.46		0.49	0.44	0.46
	Total	0.43	0.43	0.43	0.43	0.49	0.49	0.49	0.49	0.5	0.5	0.5	0.49
D3	Human		0.47	0.8	0.59		0.48	0.83	0.6		0.48	0.83	0.61
	LM		0.29	0.08	0.13		0.29	0.07	0.12		0.31	0.08	0.13
	Total	0.45	0.38	0.45	0.36	0.45	0.38	0.45	0.36	0.46	0.4	0.45	0.37
D4	Human		0.51	1	0.67		0.38	0.45	0.36		0.51	1	0.67
	LM		0.75	0.01	0.01		0.39	0.45	0.36		0.8	0.02	0.03
	Total	0.51	0.63	0.51	0.35	0.51	0.65	0.51	0.36	0.51	0.65	0.51	0.36

Table 1: The table presents the outcomes of an experiment on translating Bulgarian texts into English and the subsequent testing of third-party LM detectors. In the first column, the selected detectors are listed as follows: **D1**, which is *bert-base-gpt2detector*; **D2**, which is *roberta-base-openai-detector*; **D3**, which is *chatgpt-detection*; and **D4**, which is *SimpleAI-chatgpt*.

Model	Class	Acc.	Prec.	Rec.	F1
BdB	LM		0.96	0.98	**0.97**
	Human		0.98	0.96	**0.97**
	Total	0.97	0.97	0.97	**0.97**
SVM	LM		0.90	0.92	0.91
	Human		0.92	0.90	0.91
	Total	0.91	0.91	0.91	0.91
Logist.	LM		0.89	0.90	0.90
Regr.	Human		0.90	0.89	0.90
	Total	0.90	0.90	0.90	0.90

Table 2: Best models for Human vs. LM-generated text classification. **Total** is the macro average, as the dataset is balanced. *BdB* stands for BERT-Deepfake-BG.

Model	Class	Acc.	Prec.	Rec.	F1
BdB	cGPT		0.93	0.94	**0.93**
	BwB		0.94	0.95	**0.95**
	Human		0.95	0.94	**0.95**
	Total	0.94	0.94	0.94	**0.94**
SVM	cGPT		0.88	0.82	0.85
	BwB		0.89	0.83	0.86
	Human		0.87	0.92	0.89
	Total	0.88	0.88	0.88	0.87
Logist.	cGPT		0.80	0.80	0.80
Regr.	BwB		0.85	0.85	0.85
	Human		0.87	0.87	0.87
	Total	0.85	0.85	0.85	0.85

Table 3: Best models for Human vs. ChatGPT vs. GPT-WEB-BG classification. **Total** is the weighted average, as the dataset is unbalanced for the 3 classes. *BdB* stands for BERT-Deepfake-BG. *cGPT* stands for ChatGPT.

this way: 80:10:10. We used v. 0.24.2 of the Python library *sklearn*[29].

4.3.2 Results

Table 2 shows the results of the three classifiers, which obtained at least 0.90 F1-Score for human vs. LM (bot) classification. Table 3 shows the results of the classifiers, which achieved at least 0.90 F1-Score for the three-class classification (human, ChatGPT, BERT-Deepfake-BG).

As expected, BERT-Deepfake-BG shows the highest results for both binary and 3-class classification. Figure 1 shows the confusion matrix of BERT-Deepfake-BG's human vs. LM (bot) classification and Figure 2 shows the confusion matrix of BERT-Deepfake-BG's human vs. GPT-WEB-BG vs. ChatGPT classification.

4.4 Applying the Bulgarian Classifier with the Highest Results on a Manually-Fact-Checked Bulgarian Dataset

The three previous approaches worked on recognizing LM-generated texts. In order to account for the fact that textual deepfakes may potentially contain also fake information, we applied BERT-Deepfake-BG on the 4083-messages dataset manually annotated by journalists, mentioned as the 5th subset

[29]https://scikit-learn.org/stable/. Last accessed on April 16, 2023.

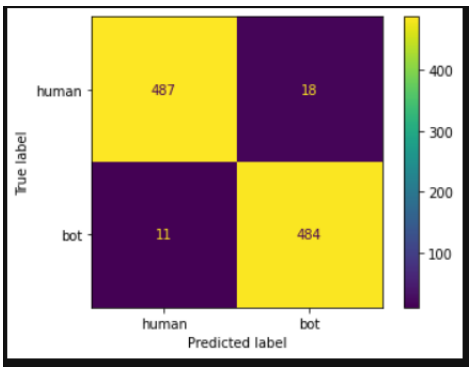

Figure 1: Confusion matrix of BERT-Deepfake-BG's binary classification.

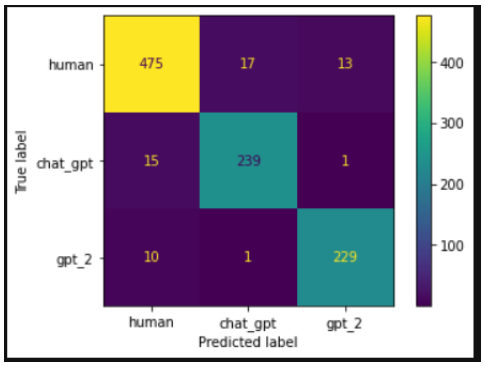

Figure 2: Confusion matrix of BERT-Deepfake-BG's 3-class classification.

dataset in Section 3.2.

4.4.1 Methods

We selected a subset of the messages from the dataset annotated by journalists, aiming to achieve the highest possible confidence that the messages recognized by BERT-Deepfake-BG are fake. To achieve this, we selected only the messages, which have been annotated by all 3 annotators as containing "Untrue information", and at the same time annotated by all 3 annotators as containing "Disinformation". We considered both the responses "yes" and "partially". We have removed the messages that are simultaneously present in the dataset annotated by journalists, as well as in the larger Twitter and Telegram datasets, in order to avoid any overlap with the messages used for building BERT-Deepfake-BG.

We applied both the binary (human vs. LM) and the multiclass (human, GPT-WEB-BG, and ChatGPT) versions of BERT-Deepfake-BG on the manually annotated dataset. We decided to experiment with both models, even if we realize that it is not technically correct to attempt to identify instances of ChatGPT among social media messages posted before it was made publicly accessible (1 January 2020 to 27 June 2022). Differently from that, the binary (human vs. LM) BERT-Deepfake-BG model could potentially identify messages, generated by other similar GPT models.

Category	Untrue	Untrue+Disinf.
LM	42	28
ChatGPT	16	9
GPT-WEB-BG	26	14

Table 4: Number of messages recognized by BERT-WEB-BG as LM-generated, ChatGPT-, and GPT-WEB-BG-generated in the 4083 messages dataset.

4.4.2 Results

BERT-Deepfake-BG recognized several messages as LM-generated. Specifically, among the messages, annotated by three annotators as containing untrue information 42 were recognized as being LM-generated, out of which 16 as ChatGPT and 26 as GPT-WEB-BG-generated. The number of messages, annotated by three annotators as untrue and by three annotators as containing disinformation, and recognized by BERT-Deepfake-BG as LM-generated represented 50-60% of each of the above categories (see Table 4 for more details). Our observations show that the messages, labeled by BERT-Deepfake-BG as ChatGPT resemble propaganda style, contain groups of words entirely written in capital letters, and sound more dramatic. This could be related to the fact that ChatGPT tended to generate advertisement-like texts, as we mentioned in Section 4.2.2. We show below an example of a message, labeled by BERT-Deepfake-BG as *ChatGPT*, and by three human annotators as both containing "untrue information" and "disinformation":

In Bulgarian: "ВОЙНАТА СЕ РАЗГАРЯ: Радев обвинява "Има такъв народ" в корупция!"

(In English: THE WAR IS IN FULL SWING: Radev accuses "There is such a nation" in corruption.)

The messages labeled by BERT-Deepfake-BG as GPT-WEB-BG exhibit more frequently broken syntax or unusual punctuation. What follows is an example of a message, manually annotated both as "untrue information" and "disinformation" and as a *GPT-WEB-BG-generated* one by BERT-Deepfake-BG.

"НЕЩО ИНТЕРЕСНО НЕДОСЕГАЕМИТЕ

ХУНТАТА Гешев иска Борисов и здравните власти да затегнат на мерките срещу COVID-19”
(In English, with broken syntax preserved: SOMETHING INTERESTING: UNTOUCHABLES JUNTA - Geshev wants Borisov and the health authorities to tighten the measures against COVID-19)

5 Discussion, Conclusions and Future Work

This article presents the first experiments aiming to find a solution to answer the challenging question of whether textual deepfakes in Bulgarian can be found in social media. We tested three approaches for detecting the “LM-generatedness” and one for the fakeness of textual deepfakes. The results indicate that utilizing machine translation (MT) in either language pair direction is not a viable solution, as textual deepfakes style may get lost in the process and the accuracy of English LM detectors is low. We conclude that the most appropriate approach for detecting textual deepfakes in Bulgarian should be one involving creating our own LM-generated dataset, in combination with fact-checking. In future work, we plan to generate more data with more models and on more topics. Applying the classifier with the highest accuracy on Bulgarian fact-checked social media texts posted after ChatGPT’s release is also a possible future work.

Limitations

- We have experimented with messages generated by only two language models. Testing with more LMs is desirable.
- We have also used a manually fact-checked real social media dataset with messages posted prior to the public release of either of the two language models. While this is motivated by the lack of a Bulgarian language fact-checked social media dataset released in 2023, it is desirable to experiment with newer fact-checked social media messages.
- Having a pre-trained GPT-2 model including social media texts in Bulgarian in the data could also enhance the results.

Ethics and Legal Statement

The research presented in this article has been conducted according to the Ethical Code of Sofia University “St. Kliment Ohridski” and after frequent consultations with lawyers specialized in Bulgarian and European Union’s laws.

Broader Impact Assessment

This article presents the first known to us effort to automatically recognize textual deepfakes in Bulgarian in social media. For this reason, it paves the way to building better working automatic tools, which will be able to recognize textual deepfakes in Bulgarian. This would benefit Bulgarian society as a whole, Bulgarian journalists, and fact-checkers, and may also contribute to the work of Natural Language Processing researchers and developers in other languages.

Acknowledgments

This article presents the joint work of several researchers. 1) Part of it took place during project TRACES[30], and indirectly received funding from the European Union’s Horizon 2020 Research and Innovation Action Programme, via the AI4Media Open Call #1 issued and executed under the AI4Media project (Grant Agreement no. 951911)[31]. 2) Additionally, the work has been also supported by the GATE project, funded by Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002-C01. 3) Furthermore, the models GPT-WEB-BG and BERT-WEB-BG are kindly provided by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, Grant number DO1-301/17.12.21. 4) Finally, the work was partially funded by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

The authors express their gratitude to the RANLP 2023 reviewers and conference organizers.

[30] https://traces.gate-ai.eu/.

[31] The article reflects only the authors’ view.

References

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a robust detection of language model generated text: Is chatgpt that easy to detect? *arXiv preprint arXiv:2306.05871*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020-December.

Xingyuan Chen, Peng Jin, Siyuan Jing, and Chunming Xie. 2022a. Automatic detection of chinese generated essays based on pre-trained bert. *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*.

Xingyuan Chen, Peng Jin, Siyuan Jing, and Chunming Xie. 2022b. Automatic detection of chinese generated essayss based on pre-trained bert. In *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, volume 10, pages 2257–2260. IEEE.

Evan Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2022. Machine generated text: A comprehensive survey of threat models and detection methods. *ArXiv*, abs/2210.07321.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415.

Margherita Gambini. 2020. Developing and experimenting approaches for deepfake text detection on social media.

Margherita Gambini, Tiziano Fagni, Fabrizio Falchi, and Maurizio Tesconi. 2022. On pushing deepfake tweet detection capabilities to the limits. In *14th ACM Web Science Conference 2022*, pages 154–163.

Zahra Ghadiri, Milad Ranjbar, Fakhteh Ghanbarnejad, and Sadegh Raeisi. 2022. Automated fake news detection using cross-checking with reliable sources. *arXiv preprint arXiv:2201.00083*.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Peter Kowalczyk, Marco Röder, Alexander Dürr, and Frédéric Thiesse. 2022. Detecting and understanding textual deepfakes in online reviews.

Saranya Krishnan and Min Chen. 2018. Identifying tweets with fake news. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 460–464.

Vijini Liyanage, Davide Buscaldi, and Adeline Nazarenko. 2022. A benchmark corpus for the detection of automatically generated text in academic publications. *arXiv.org*.

Iva Marinova, Kiril Simov, and Petya Osenova. 2023. Transformer-based language models for bulgarian. In *Proceedings of the International RANLP Conference 2023*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. Covid-19 in bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009.

Michael Sheinman Orenstrakh, Oscar Karnalim, Carlos Anibal Suarez, and Michael Liut. 2023. Detecting llm-generated text in computing education: A comparative study for chatgpt cases. *arXiv preprint arXiv:2307.07411*.

Alessandro Pegoraro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. To chatgpt, or not to chatgpt: That is the question! *arXiv preprint arXiv:2304.01487*.

Pavel Posokhov, Stepan Skrylnikov, and Olesia Makhnytkina. 2022. Artificial text detection in russian language: a bert-based approach. *Computational Linguistics and Intellectual Technologies*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Saima Sadiq and Saleem Ullah. Unmasking deepfake tweets: Leveraging deep learning and word embeddings for accurate classification of machine-generated text on social media. *Available at SSRN 4494619*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, and Niklas Muennighoff. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv.org*.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, A. E. Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and E. Artemova. 2022. Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian. *ArXiv*, abs/2206.01583.

Harald Stiff and Fredrik Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13(4):363–383.

Irina Temnikova, Silvia Gargova, Ruslana Margova, Veneta Kireva, Ivo Dzhumerov, Tsvetelina Stefanova, and Hristiana Krasteva. 2023. New bulgarian resources for studying deception and detecting disinformation. *10th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Poznań, Poland.*