

TweepFake: about detecting deepfake tweets

Tiziano Fagni¹, Fabrizio Falchi^{2*}, Margherita Gambini³, Antonio Martella⁴, Maurizio Tesconi¹

1 Istituto di Informatica e Telematica - CNR, Pisa, Italy

2 Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - CNR, Pisa, Italy

3 University of Pisa, Italy

4 University of Trento, Italy

The authors contributed equally to this work.

* fabrizio.falchi@cnr.it

Abstract

The recent advances in language modeling significantly improved the generative capabilities of deep neural models: in 2019 OpenAI released GPT-2, a pre-trained language model that can autonomously generate coherent, non-trivial and human-like text samples. Since then, ever more powerful text generative models have been developed. Adversaries can exploit these tremendous generative capabilities to enhance social bots that will have the ability to write plausible *deepfake* messages, hoping to contaminate public debate. To prevent this, it is crucial to develop deepfake social media messages detection systems. However, to the best of our knowledge no one has ever addressed the detection of machine-generated texts on social networks like Twitter or Facebook. With the aim of helping the research in this detection field, we collected the first dataset of *real* deepfake tweets, *TweepFake*. It is *real* in the sense that each deepfake tweet was actually posted on Twitter. We collected tweets from a total of 23 bots, imitating 17 human accounts. The bots are based on various generation techniques, i.e., Markov Chains, RNN, RNN+Markov, LSTM, GPT-2. We also randomly selected tweets from the humans imitated by the bots to have an overall balanced dataset of 25,572 tweets (half human and half bots generated). The dataset is publicly available on Kaggle. Lastly, we evaluated 13 deepfake text detection methods (based on various state-of-the-art approaches) to both demonstrate the challenges that Tweepfake poses and create a solid baseline of detection techniques. We hope that TweepFake can offer the opportunity to tackle the deepfake detection on social media messages as well.

Introduction

During the last decade, the social media platforms - developed to connect people and make them share their ideas and opinions through multimedia contents (like images, video, audio, and texts) - have also been used to manipulate and alter the public opinion thanks to *bots*, i.e., computer programs that control a fake social media account as a legitimate human user would do: by “liking”, sharing and posting old or new media which could be real, forged through simple techniques (e.g., editing of a video, use of gap-filling texts and search-and-replace methods) or deepfake.

Deep learning is a family of machine-learning methods that use artificial neural networks to learn a hierarchy of representations, from low to high non-linear features representation, of the input data. The term *deepfake* is a portmanteau of *deep learning* and *fake*; it refers to AI-generated multimedia (images, videos, audios, and texts) that are potentially deceptive [1], although good usages of deepfakes can be found [2]. The generation and sharing of deepfake multimedia over social media, tricking people into believing that they are human-generated, have already caused distress in several fields (such as politics and science). Therefore, it is necessary to continuously probe the generative model’s ability to produce deceiving multimedia by enhancing and developing appropriate detectors. This is more than ever necessary for the text generation field: in 2019, for the first time, a generative model (GPT-2 language model [3]) showed incredible text generation

capabilities that deeply worries the research community: [4] and [5] proved that humans seem unable to identify automatically generated text (their accuracy is near random guessing, i.e. 54%). Deepfake social media texts (GPT-2 samples included) can already be found, though there is still no misuse episode on them.

Deepfake detecting strategies are continuously developed, from deepfake video [6–8] to audio [9] and text detection methods. Under the hood, current automatic neural text detectors tend to learn not to discriminate between neural text and human-written text, but rather decide what is characteristic and uncharacteristic of neural text [10] (i.e., statistics of the language for machine-generated texts); however, it emerged that some strategies (substituting homoglyphs to characters or adding some common misspelled words) can alter the statistical characteristics of the generated text making the detection task ever more difficult [10]. Moreover, nowadays scientific works focus on *ad-hoc* generated texts only; also, deep fake detectors usually run knowing the adversarial generative model. This is a white-box approach; [11] studied the black-box approach (pretending not knowing the text generator), but the text samples were always *ad-hoc* generated. Besides, the majority of the studies deal with long deepfake texts like news articles or stories: according to [12], “For both human raters and automatic discriminators, the longer the provided text excerpt is, the more easily its provenance can be identified”. Having said that, there is a lack of knowledge on how state-of-the-art detection techniques perform in a *real social media setting*, in which the machine-generated text samples are the ones actually posted on a social media, the social media content is often short (above all on Twitter) and the generative model is not known (also, the text samples can be altered to make difficult the automatic detection).

Additionally, to the best of our knowledge, a properly labeled dataset containing *only* human-written and *real* deepfake social media messages still doesn’t exist. [13] and [14] tried to detect auto-generated tweets over a dataset of tweets produced by a large variety of bots [15] (spam bots, social bots, sockpuppet, cyborgs), meaning that the detection task is *not* focused *only* on *real* deepfake messages. Furthermore, those tweets are *human*-labelled at the account level, i.e., by examining the messages produced by a user, but [16] proved that a human is not reliable on this labeling task.

Our work provides the *first* properly labeled dataset of human and real machine-generated social media messages (coming from Twitter in particular): *TweepFake - A Twitter Deep Fake Dataset*. TweepFake contains *real* deepfake tweets that we collected with the goal of testing existing and future detection approaches. The dataset is *real* in the sense that each deepfake tweet was actually posted on Twitter. We collected tweets from a total of 23 bots, imitating 17 human accounts. The bot accounts are based on various generation techniques, including Markov Chains, RNN, RNN+Markov, LSTM, GPT-2. We randomly selected tweets from the humans imitated by the bots to have an overall balanced dataset of 25,572 tweets (half human and half bots generated). We made TweepFake publicly available on Kaggle [17]. More information can be found in the Subsection *The TweepFake dataset*. With the aim of showing the challenges that TweepFake poses and providing a solid baseline of detection techniques, we also evaluated 13 different deepfake text detection methods: some of them exploiting text representations as inputs to machine-learning classifiers, others based on deep learning networks, and others relying on the fine-tuning of transformer-based classifiers. The code used in the experiments is publicly available on GitHub [18].

Related work

Deepfake technologies have first risen in the computer vision field [19–22], followed by effective attempts on audio manipulation [23, 24] and text generation [3]. Deepfakes in computer vision usually deal with face manipulation - such as entire face synthesis, identity swap, attribute manipulation, and expression swap [6] – and body re-enacting [22]. Recently, audio deepfakes involved the generation of speech audio from a text corpus by using the voice of different speakers after five seconds of listening time [24]. In 2017, the development of the *self-attention* mechanism and the [25]’s transformer led to the improvement of the language models. Language modeling refers to the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. The subsequent transformer-based language models (GPT [26], BERT [27], GPT2 [3] etc.) did not only enhance the natural language understanding tasks, but language generation as well. In 2019, [3] developed GPT-2, a pre-trained language model that can *autonomously* generate coherent human-like paragraphs of text by having in input just a short sentence; in the same year, [28] contributed to text generation with GROVER, a new approach

for efficient and effective learning and generation of multi-field documents such as journal articles. Soon after, [29] released CTRL, a conditional language model that uses control codes to generate text having a specific style, content, and task-specific behavior. Last but not least, [30] presented OPTIMUS, putting the variational autoencoder in the game for text generation.

Currently, the approaches to automatic deepfake text detection roughly fall into three categories, here listed in order of complexity:

Simple classifier: a machine-learning or deep-learning binary classifier trained from scratch.

Zero-shot detection: using the output from a pre-trained language model as features for the subsequent classifier. This classifier may be a machine-learning based one or a simple feed forward neural network.

Fine-tuning based detection: jointly fine-tuning a pre-trained language model with a final simple neural network (consisting of one or two layers).

The GPT-2's research group made an in-house detection research [31] on GPT-2 generated text samples: first, they evaluated a standard machine-learning approach that trains a logistic regression discriminator on *tf-idf* unigram and bigram features. Then, they tested a simple zero-shot baseline using a threshold on the *total probability*: a text excerpt is predicted as machine-generated if its likelihood according to GPT-2 is closer to the mean likelihood over all machine-generated texts than to the mean of the human-written ones.

[4] provided GLTR (Giant Language model Test Room), a visual tool that helps humans to detect deepfake texts. Generated text is sampled word by word from a next token distribution (several sampling techniques can be used [32], the simplest way is to take the most probable token): this distribution usually differs from the one that humans subconsciously use when they write or speak. GLTR tries to show these statistical language differences to aid people in discriminating human-written text samples from machine-generated ones.

GROVER's authors [28] followed the fine-tuning based detection approach by using BERT, GPT2 and GROVER itself as the pre-trained language model. GROVER was the best, suggesting that maybe the best defense against the transformer-based text generators is a detector based on the same kind of architecture. However, OpenAI [31] proved it wrong on GPT-2 generated texts: they showed that fine-tuning a RoBERTa-based detector achieved consistently higher accuracy than fine-tuning a GPT-2 based detector with equivalent capacity.

[11] developed an energy-based deepfake text detector: unlike auto-regressive language models (e.g. GPT-2 [3], XLNET [33]), which are defined in terms of a sequence of conditional distributions, an energy-based model is defined in terms of single scalar energy function, representing the joint compatibility between all input variables. Thus, the deepfake discriminator is an energy function that scores the joint compatibility of an input sequence of tokens given some context (e.g. a text sample, some keywords, a bag of words, a title) and a set of network parameters. These authors tried also to generalize the experimental setting, where the generator architectures and the text corpora are different between training and test time.

The only research on the detection of deepfake social media texts was conducted by [5] on Amazon reviews written by GPT-2. They evaluated several human-machine discriminators: the Grover-based detector, GLTR, RoBERTa-based detector from OpenAI and a simple ensemble that fused these detectors using logistic regression at the score level.

The above deepfake text detection methods have got two flaws: except for [5]'s research, they dealt with generated news articles, having a longer length with respect to social media messages; then, just a single known adversarial generative model is usually used to generate the deepfake text samples (usually GPT-2 or GROVER). In a real-setting scenario, we don't know how many and what generative architectures are used. Our Tweepfake dataset provides a set of tweets produced by several generative models, hoping to help the research community in detecting shorter deepfake texts written by heterogeneous generative techniques.

DeepFake tweets generation

There exist several methods to generate a text. What follows is a short description of the generative methods used to produce the machine-generated tweets contained in our dataset.

Generation techniques

First and foremost, the training set of text corpora is tokenized (punctuation included), and then one of the following methods can be applied. Notice that the following techniques write a text token-by-token (a token could be a word, a char, a byte pair, a Unicode code point) until a stop token is encountered or a pre-defined maximum length is reached. *RNN*, *LSTM*, *GPT2* are language models. Therefore, at each token generation, they always produce a multinomial distribution - in which a category is a token of the vocabulary derived from a set of human-written texts - from which the next token is sampled with a specific sampling technique (e.g., max probability, top-k, nucleus sampling [34]). A special start token is given in input to the generative model to prime the text generation; with language models, a short sentence can work as a priming text as well: each token of the start sentence is processed without computing the multinomial distribution, just to *condition* the generative model.

Markov Chains is a stochastic model that describes a sequence of states by moving from a state to another with a probability which depends on the current state only. For the text generation a state is identified as a token: the next token/state is randomly selected from a list of tokens following the current one. The probability of a token t to be chosen is proportional to the frequency of the appearance of t after the current token.

RNN, helped by its loop structure, stores in its *accumulated* memory the information on the previously encountered tokens and computes the multinomial distribution from which the next token is chosen. The selected token is given back in input so that the RNN can produce the following one.

RNN+Markov method *may* employ the Markov Chain's next token selection as a sampling technique. In practice, the next token is randomly sampled from the multinomial distribution produced by RNN, with the tokens having the highest probability value being the most likely to be chosen. However, no reference was found to confirm our hypothesis on RNN+Markov mechanism.

LSTM generates text as RNN does. However, it is smarter than the latter because of its more complicated structure: it can learn to selectively keep track of only the relevant information of the already seen piece of text *while* also minimizing the vanishing gradient problem that affects a RNN. LSTM's memory is "longer" than RNN's.

GPT-2 is a generative pre-trained transformer language model relying on the *Attention* mechanism of [25]: by employing the *Attention*, a language model pre-trained on millions of sentences/texts learns how each token/word relates to every other in every possible context. This is the trick to generate more coherent and non-trivial paragraphs of text. Anyhow, being a language model, GPT-2's text generation steps are the same as RNN and LSTM: generation of a multinomial distribution at each step and then selection of the next token from it by using a specific sampling technique.

CharRNN employs RNN at *char level* to generate a text char-by-char.

The TweepFake dataset

In this section we describe the process of building the novel *TweepFake - A Twitter Deep Fake Dataset* together with the results of the experimentation on the deepfake detection task. Twitter accounts have been searched heuristically on the web, GitHub and Twitter looking for keywords related to automatic or AI text generation, deepfake text/tweets, or to specific technologies as well as GPT-2, RNN, etc. in order to collect a sample of Twitter profiles as huge as possible. We selected only accounts referring to automated generated text technologies in Twitter descriptions, profile URLs, or related GitHub. From this sample, we selected a subset of accounts mimicking (often fine-tuned on) human Twitter profiles. Thus, we obtained 23 bots and 17 human accounts because some fake accounts imitate the same human profile (see Table 1). Then we downloaded timelines of both deep fake accounts and their corresponding humans via Twitter REST API. In order to get a data set balanced on both categories (human and bots) we randomly sampled tweets for each accounts' couple (human and bot/s) based on the less productive. For example, after the download, we had

3,193 tweets by human#11 and 1,245 by the corresponding bot#12, thus we random sampled 1,245 tweets by the human account timeline to get the same amount of data. In total, we had 25,572 tweets half human and half bots generated. In Table 1, we report, for each fake account we considered, the human account imitated, the technology used for generating the tweets, and the number of tweets we collected from both the fake and the human account. In Table 2, we grouped the fake accounts by technology reporting, together with the number of collected tweets, the citation of the information we found about the bot (i.e., more technical information, code, news, etc.). Please note that in our detection experiments we grouped the technologies in three main groups: GPT-2, RNN, others (see Sections *Results* and *Discussion*).

Table 1. The proposed *TweepFake* dataset tweets grouped by imitated human account.

fake account	tweets	technology	human account	tweets
bot#1	946	RNN + Markov	human#1	946
bot#2	348	GPT-2	human#2	348
bot#3	132	GPT-2	human#3	132
bot#4	1792	GPT-2	human#4	1803
bot#5	11	RNN		
bot#6	38	LSTM	human#5	56
bot#7	18	Torch RNN		
bot#8	217	GPT-2	human#6	217
bot#9	1289	RNN + Markov	human#7	1289
bot#10	1030	Markov Chains	human#8	515
			human#9	515
bot#11	2409	RNN	human#10	2409
bot#12	1245	RNN	human#11	1245
bot#13	228	GPT-2	human#12	228
bot#14	355	GPT-2		
bot#15	33	GPT-2		
bot#16	286	RNN	human#13	1293
bot#17	549	GPT-2		
bot#18	18	GPT-2		
bot#19	52	unknown		
bot#20	100	CharRNN	human#14	100
bot#21	39	GPT-2	human#15	39
bot#22	128	OpenAI	human#16	128
bot#23	1523	unknown	human#17	1523

DeepFake tweets detection

Detection techniques

To verify the difficulty level in the detection task of automatically generated natural language contents, we used the built dataset to measure the effectiveness of a set of ML and DL methods of increasing complexity. The results obtained allow us to fix some baseline configurations in terms of performance and give an idea on which approaches are most promising in solving this specific problem.

In Table 3, we report all the methods that have been tested in this work. We explored the usage of four main approaches to model the solutions to this specific task. The first scenario uses a text representation based on bag-of-words (BoW) [53] with encoded feature weighted according to TF-IDF function [53]. The tweets encoded in this way have been next processed by a statistical ML algorithm able to produce a suitable classifier to solve the specific problem. In this work, we have chosen to implement three popular classifiers: logistic regression, random forest, and SVM.

Table 2. The proposed *TweepFake* dataset tweets grouped by technology.

technology	fake acc.	human acc.	info	code	tweets	tweets
GPT-2	bot#2	human#2	[3]	[35]	348	3711
	bot#3	human#3	[36]	[37]	132	
	bot#4	human#4	-	[38]	1792	
	bot#8	human#6	-	[39]	217	
	bot#13	human#12	228	-	-	
	bot#14	human#13	-	[40]	355	
	bot#15	human#13	-	-	33	
	bot#17	human#13	-	-	549	
	bot#18	human#13	-	-	18	
bot#21	human#15	-	-	39		
RNN	bot#5	human#4	-	-	11	3969
	bot#11	human#10	[41]	-	2409	
	bot#12	human#11	[42]	[43]	1245	
	bot#16	human#13	[44]	[45]	286	
	bot#7	human#5	[46]	[47]	18	
RNN + Markov	bot#1	human#1	-	-	946	2235
	bot#9	human#7	-	-	1289	
Markov Chains	bot#10	human#8 human#9	[48]	[49]	1030	1030
LSTM	bot#6	human#5	[50]	[51]	38	38
OpenAI	bot#22	human#16	-	-	128	128
CharRNN	bot#20	human#14	-	[52]	100	100
unknown	bot#19	human#13	-	-	52	52
unknown	bot#23	human#17	-	-	1523	1523

Table 3. Description of the methods used in the experimentation.

Encoding	Method name	Algorithm
BoW+TF-IDF	LOG_REG_BOW	Logistic regression classifier [54]
	RAND_FOREST_BOW	Random forest classifier [55]
	SVC_BOW	Support vector machine classifier [56]
BERT	LOG_REG_BERT	Logistic regression classifier
	RAND_FOREST_BERT	Random forest classifier
	SVC_BERT	Support vector machine classifier
Characters	CHAR_CNN	Single CNN network [57] using internal char embeddings representation
	CHAR_GRU	Single GRU network [58] using internal char embeddings representation
	CHAR_CNNGRU	Combined CNN and GRU networks using internal char embeddings representation
Native LM	BERT_FT	BERT language model with fine-tuning
	DISTILBERT_FT	DistilBERT language model with fine-tuning
	ROBERTA_FT	RoBERTa language model with fine-tuning
	XLNET_FT	XLNet language model with fine-tuning

The approach based on BoW+TF-IDF, although being very popular and used for many years as the primary methodology to vectorize texts, suffers from two main drawbacks. The first problem is related to the curse of dimensionality [59], i.e., the feature space is very sparse, and the amount of data required to produce statistically significant models is very high. The second issue of BoW is that it ignores the information about word order, and thus it misses completely any information about the semantic context on which the words

occur. To overcome these limitations, on the second approach, we encoded texts using BERT [27], a recent pre-trained language model that contributed to improving state-of-the-art results on many NLP problems. BERT provides contextual embeddings, fixed-size vector representations of words which depend not only by the words itself but also by the context on which the words occur: for example, the word *bank*, depending on being near to word *economy* or *river*, will assume a different meaning and consequently a different contextual vector. Therefore, these contextual representations can be merged together to obtain a contextualized fixed-size vector of a specific text (e.g., by averaging the vectors of all words composing a specific text). As in the previous scenario, the tweets encoded through BERT has been processed using the same set of classifiers.

On the third approach, we leverage another effective way to encode textual contents by working at the character level instead of words or tokens [60]. This methodology has the advantage of not requiring access to any external resource, but it only exploits the dataset used to learn the model. The encoding process is summarized in Fig 1.

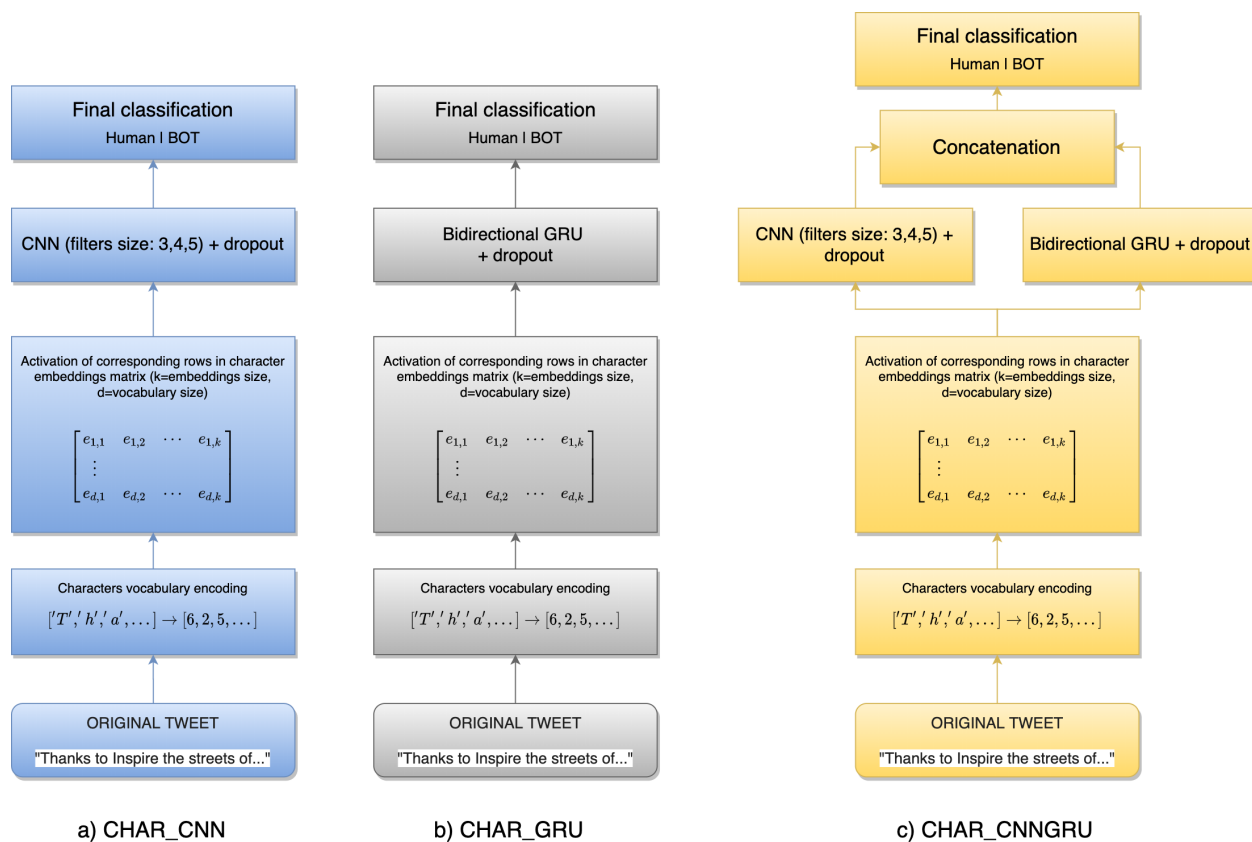


Fig 1. Architecture of the tested deep neural networks based on character encoding. Three different architectures were tested: a) a CNN sub-network using three different kernel sizes (3,4, and 5) combined together and followed by a dropout layer, b) a bidirectional GRU followed by a dropout layer, and c) a network exploiting both CNN and GRU to extract spatial and temporal features from data in order to try to improve the effectiveness of the solution.

Each tweet is encoded as a set of contiguous characters IDs obtained from a fixed vocabulary of characters. This mapping allows us to use the internal embeddings matrix (learned during training phase) to select, at each time step in the text, only the row vector corresponding to the current analyzed character, thus contributing to building a proper matrix representation of current text. The resulting text's embedding matrix is next passed as input to the successive layers in the tested deep learning networks.

As the final and most effective approach, we used several pre-trained language models by fine-tuning them directly on the built dataset. This process just consists of taking a trained language model, integrate

its original architecture with a final dense classification layer, and perform training on a specific dataset (typically small) for very few epochs [27]. This step of fine-tuning allows us to customize and adapt the native language model’s network weights to the considered use case, maximizing the encoding power of the model to solve that specific classification problem. As reported in Table 3, in this work, we tested four different language models, all based on transformer architecture [25], which have provided state of the art results on many text processing benchmarks. BERT [27] was presented in 2018, and thanks to the innovative transformer-based architecture with dual prediction tasks (Masked Language Model and Next Sentence Prediction) and much data, it was able to basically outperform all other methods on many text processing benchmarks. XLNet [33] and RoBERTa [61] tried to increase BERT effectiveness by slightly varying and optimizing its original architecture, and using a lot more data on training step, resulting in improvements on prediction powers on the same benchmarks up to 15%. DistilBERT [61], on the other hand, tried to keep the performance of the original BERT model (97% of original ones) but greatly simplifying the network architecture and halving the number of parameters to be learned.

Experimental setup

The main parameters of each algorithm (except for those based on deep learning models where, for computational reasons, we used default parameters) have been optimized using the validation set.

Baselines built on standard machine learning methods (with both BoW and BERT representations) have been implemented using `scikit-learn` Python library [62]. In BoW experiments, we performed tweets tokenization by splitting texts into words, removing all hashtags, replacing all user mentions with the token `__user_mention__`, replacing all URLs with token `__url__`, and leaving all found emoticons as separated tokens. During the encoding phase, to minimize computational cost, we only left the most frequent 25,000 tokens, and we weighted each token inside tweets according to tf-idf method [53]. In BERT experiments we encoded tweets using `bert-base-cased` pre-trained model from `transformers` Python library [63]. In SVC configurations we tried different kernels (`linear` and `rbf`), and a range of values for `C` and `gamma` parameters. The `C` misclassification cost has also been optimized on logistic regression configurations. On random forest baselines we have chosen the best setting varying these parameters: `max_depth`, `min_samples_leaf`, `min_samples_split`, and `n_estimators`.

Solutions based on characters deep learning networks have been implemented using `Keras` Python library [64]. We used a fixed window of length 280 (on Twitter, 280 is the maximum length of a tweet in terms of the number of characters) to represent input tweets and `tanh` activation function at every level of hidden layers. In all three configurations of chars neural networks, the first hidden layer is an embedding layer of size 32. At the second level, `CHAR_CNN` is characterized by three independent CNN subnetworks (CNN layer composed by 128 filters and followed by a global max pooling layer) with different kernel sizes (3,4, and 5) which are next concatenated and filtered by a dropout layer before performing final classification. `CHAR_GRU` configuration is more simple, composed at the second level by a bidirectional GRU layer followed by dropout and a final classification layer. `CHAR_CNNGRU` configuration adds to the first hidden layer two different subnetworks (one CNN-based and one GRU-based with the same architecture as defined before), concatenates them, applies a dropout, and performs final classification.

We used `simpletransformers` Python library [65] to implement all models in fine-tuned configurations. In agreement with other works in literature and for computational reasons, we decided to limit the number of epochs to just three complete iterations over training data.

A summary of the customized parameter values used in the final configurations is reported in ?? All the other unspecified parameters used by tested algorithms are left to their default values, as defined in the software libraries providing their implementation.

Our experiments are reported in a GitHub repository [18].

Results

As evaluation measures, we used the canonical adopted metrics in text classification contexts: precision, recall, F1, and accuracy [53]. In this context, given that the analyzed dataset is balanced in terms of

examples, the accuracy seems the most reasonable measure to capture the effectiveness of a method. On Table 4, we report the results obtained on test set using the proposed detection baselines.

Table 4. Experimental results on test set obtained with the proposed baselines.

Method	HUMAN			BOT			GLOBALLY Accuracy
	Precision	Recall	F1	Precision	Recall	F1	
LOG_REG_BOW	0.841	0.749	0.792	0.774	0.859	0.814	0.804
RAND_FOREST_BOW	0.759	0.798	0.778	0.787	0.747	0.767	0.772
SVC_BOW	0.851	0.754	0.800	0.779	0.869	0.822	0.811
LOG_REG_BERT	0.846	0.820	0.833	0.826	0.851	0.838	0.835
RAND_FOREST_BERT	0.864	0.776	0.818	0.797	0.878	0.836	0.827
SVC_BERT	0.860	0.818	0.838	0.827	0.867	0.846	0.842
CHAR_CNN	0.896	0.794	0.842	0.815	0.908	0.859	0.851
CHAR_GRU	0.899	0.743	0.814	0.781	0.916	0.844	0.830
CHAR_CNNGRU	0.848	0.820	0.834	0.826	0.853	0.839	0.837
BERT_FT	0.899	0.882	0.890	0.884	0.901	0.892	0.891
DISTILBERT_FT	0.894	0.880	0.886	0.882	0.895	0.888	0.887
ROBERTA_FT	0.901	0.890	0.895	0.891	0.902	0.897	0.896
XLNET_FT	0.914	0.832	0.871	0.846	0.922	0.882	0.877

To have a better understanding on how the tested baselines behave at detection time, we split all available accounts on the dataset into four different categories:

human The set of Twitter accounts having contents produced only by a human.

gpt2 The set of Twitter accounts having contents produced only by GPT2-based generative algorithms.

rnn The set of Twitter accounts having contents produced only by RNN-based generative algorithms.

others The set of Twitter accounts having contents produced only by generative algorithms using mixed (e.g., RNN + Hidden Markov models) or unknown approaches.

Each account has been assigned to one of those categories according to the specific information found in the corresponding Twitter account’s description or in a linked Web page describing its purpose. For some accounts, we were not able to find any information provided by the author about the technology used to implement the BOT, so in that case we assigned the account to *others* category.

On Fig 2 we show a qualitative evaluation of the accuracy of the proposed baselines in relation to the category of accounts and the “global” performance over all categories.

To obtain a fair comparison between *human* and the other types of categories, giving that the *human* class has more examples than the other categories alone, we performed a random undersampling of *humans* to match the maximum size in terms of examples given by one of the other three categories. The resulting distribution in terms of examples has been the following: *humans* (484), *GPT-2* (384), *RNN* (412), and “others” (484).

Discussion

Globally, in terms of accuracy, we can observe (Table 4) that the methods based on BoW representations have the worst performance (around 0.80), followed by those using both BERT (around 0.83) and character encodings (up to 0.85), and remarkably outperformed by methods using native language modeling encoding (0.90 for ROBERTA_FT). A high level view of the results thus indicates that the most complex text representations, especially those based on big amount of external data, provide evident advantages in terms of effectiveness. An interesting exception is the character encoding in deep learning methods which is simple, able to generally provides good performance, and be useful in cases where no pretrained models are available (e.g., for non-english languages).

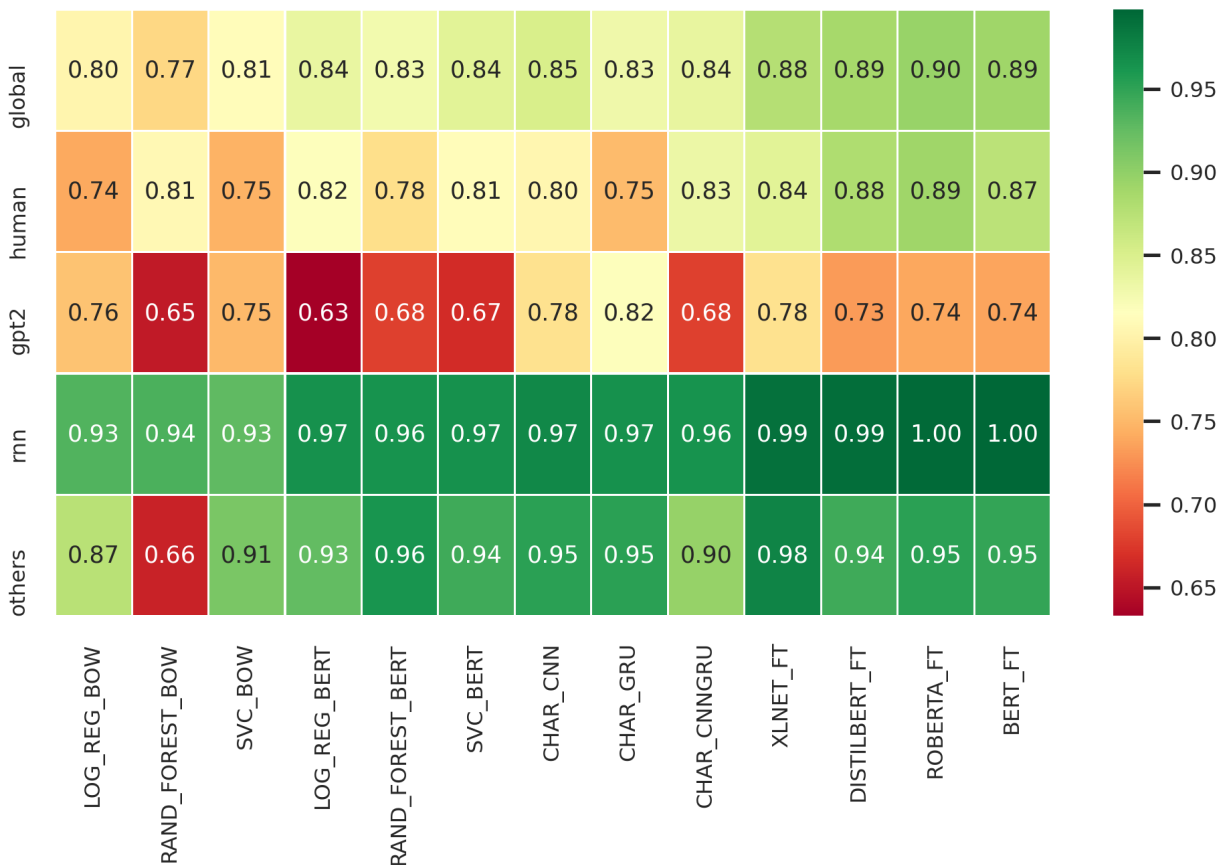


Fig 2. Accuracy heat-map over fake account type.

Going more on details, the baselines based on fine tuning (except for XLNET) show very well balanced performance in terms of precision and recall, differently from the other configurations where one the two measures is a lot higher than the other. Another observation is that all methods provide a) higher precision on human label examples than on bot examples ones and b) higher recall on bot label examples than on human ones. This translates into having more difficulties to detect correctly a tweet as written by a bot instead than a human, although the algorithms have more troubles finding all relevant human label examples.

The qualitative analysis of the accuracy in relation to the accounts' categories highlights some interesting facts (Fig 2): a) all methods (except RAND_FOREST_BOW) perform extremely well in identifying tweets as BOT on both on *RNN* and *others* accounts; b) tweets from *human* accounts are easily identifiable by methods based on fine tuning but not from the others; and c) all methods have difficulties in identifying correctly tweets produced by *GPT-2* accounts. In particular, on this last point it is interesting to note that all complex fine tuned LM methods perform remarkably worst than some character based methods like CHAR_GRU. This could indicate that RNN networks maintain slight advantages in temporal representations for short contexts respect to newer transformer networks, an important aspect to be investigated in the future.

To sum up, these findings suggest that a wide variety of detectors (text representation-based using machine-learning or deep-learning methods and transformer-based using transfer-learning) have greater difficulties in detecting correctly a deepfake tweet rather than a human-written one; this is especially true for GPT-2 generated tweets, insinuating that the newest and more sophisticated generative methods based on the transformer architecture (here GPT-2) can produce more human-like short texts than old generative methods like RNNs. We manually verified several GPT-2 and RNN tweets: the former were harder to label as bot-generated. In any case, a future work could deeply investigate the humanness of tweets coming from

several generative methods by questioning people.

Conclusion

Deepfake text detection is increasingly critical due to the development of highly sophisticated generative methods like GPT-2. However, to the best of our knowledge no deepfake detection has been conducted over social media texts yet. Therefore, the aim of this paper was to present the first *real* deepfake tweets dataset (TweepFake) to help the research community to develop techniques fighting the deepfake threat on social media. The proposed *real* deepfake tweets are publicly available on the well-known Kaggle platform. The dataset is composed of 25,572 tweets, half human and half bots generated, posted on Twitter in the last few months. We collected them from 23 bots and from the 17 human accounts they imitate. The bot accounts are based on various generative techniques, including GPT-2, RNN, LSTM and Markov Chain. We tested the difficulty in discriminating human-written tweets from machine-generated ones by evaluating 13 detectors: some of them exploiting text representations as inputs to machine-learning classifiers, others based on deep learning networks, and others relying on the fine-tuning of transformer-based classifiers.

Our detection results suggest that the newest and more sophisticated generative methods based on the transformer architecture (e.g., GPT-2) can produce high-quality short texts, difficult to unmask also for expert human annotators. This finding is in line with the ones in [3] covering long texts (news, articles, etc.). Additionally, the transformer-based language models provide very good word representations for both text representation-based and fine-tuning based detection techniques. The latter provide a better accuracy (nearly 90% for RoBERTa-based detector) than the former.

We recommend to further investigate the RNN-based detectors, as the CHAR_GRU-based detector was the best at correctly labelling GPT2-tweets as bots. Moreover, a study of the capability of humans to discriminate human-written tweets from machine-generated ones is necessary; also, the humanness of tweets produced by different generative methods could be assessed. Of course, different detection techniques are appreciated.

Acknowledgments

This work has been fully supported by the EU H2020 Program under the scheme INFRAIA-01-2018-2019: Research and Innovation action grant agreement number 871042 SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics. Fabrizio Falchi has been also supported by AI4Media - A European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant 95191.

References

1. Vincent J. Why we need a better definition of ‘deepfake’. The Verge; 2018. Available from: <https://www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news> [cited 2021 March 09].
2. Museum D. Behind the Scenes: Dali Lives.; 2019. In: Youtube [Video]. Available from: <https://www.youtube.com/watch?v=BIDax14xqJ4&t=35s> [cited 2021 March 09].
3. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners; 2019. In: OpenAI Blog [Internet]. Available from: <https://openai.com/blog/better-language-models/> [cited 2021 March 09].
4. Gehrmann S, Strobelt H, Rush A. GLTR: Statistical Detection and Visualization of Generated Text. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics; 2019. p. 111–116.

5. Adelani DI, Mai H, Fang F, Nguyen HH, Yamagishi J, Echizen I. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In: International Conference on Advanced Information Networking and Applications. Springer; 2019. p. 1341–1354.
6. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Information Fusion*. 2020;64:131–148.
7. Lyu S. Deepfake Detection: Current Challenges and Next Steps. In: 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW); 2020. p. 1–6.
8. Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S. Deep Learning for Deepfakes Creation and Detection; 2019. arXiv:1909.11573 [Preprint]. Available from: <https://arxiv.org/abs/1909.11573> [cited 2021 March 09].
9. Chen T, Kumar A, Nagarsheth P, Sivaraman G, Khoury E. Generalization of Audio Deepfake Detection. In: Proc. Odyssey 2020 The Speaker and Language Recognition Workshop; 2020. p. 132–137.
10. Wolff M, Wolff S. Attacking Neural Text Detectors; 2020. arXiv:2002.11768 [Preprint]. Available from: <https://ui.adsabs.harvard.edu/abs/2020arXiv200211768W> [cited 2021 March 09].
11. Bakhtin A, Gross S, Ott M, Deng Y, Ranzato M, Szlam A. Real or fake? learning to discriminate machine from human generated text; 2019. arXiv:1906.03351 [Preprint]. Available from: <https://arxiv.org/abs/1906.03351> [cited 2021 March 09].
12. Ippolito D, Duckworth D, Callison-Burch C, Eck D. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 1808–1822.
13. Garcia-Silva A, Berrio C, Gómez-Pérez JM. An Empirical Study on Pre-trained Embeddings and Language Models for Bot Detection. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). Association for Computational Linguistics; 2019. p. 148–155.
14. Lundberg J, Nordqvist J, Matosevic A. On-the-fly Detection of Autogenerated Tweets; 2018. arXiv:1802.01197 [Preprint]. Available from: <https://arxiv.org/pdf/1802.01197.pdf> [cited 2021 March 09].
15. Gorwa R, Guilbeault D. Unpacking the social media bot: A typology to guide research and policy. In: Special Issue:Cleaning Up Social Media. vol. 2. Policy & Internet; 2020. p. 225–248.
16. Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In: Proceedings of the 26th international conference on world wide web companion. International World Wide Web Conferences Steering Committee; 2017. p. 963–972.
17. Fagni T, Falchi F, Gambini M, Martella A, Tesconi M. TweepFake dataset; 2020. Repository: Kaggle [Internet]. Available from: <https://www.kaggle.com/mtesconi/twitter-deep-fake-text> [cited 2021 March 09].
18. Fagni T. Scripts used in “TweepFake: about Detecting Deepfake Tweets”; 2021. Repository: GitHub [Internet]. Available from: https://github.com/tizfa/tweepfake_deepfake_text_detection [cited 2021 March 16].
19. Bregler C, Covell M, Slaney M. Video rewrite: Driving visual speech with audio. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co.; 1997. p. 353–360.

20. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Trans Graph*. 2017;doi:10.1145/3072959.3073640.
21. Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. *Commun ACM*. 2018;62(1):96–104.
22. Chan C, Ginosar S, Zhou T, Efros AA. Everybody dance now. In: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE; 2019. p. 5933–5942.
23. Lyons K. FTC says the tech behind audio deepfakes is getting better. *The Verge*; 2020. Available from: <https://www.theverge.com/2020/1/29/21080553/ftc-deepfakes-audio-cloning-joe-rogan-phone-scams> [cited 2021 March 09].
24. Jia Y, Zhang Y, Weiss R, Wang Q, Shen J, Ren F, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates Inc.; 2018. p. 4480–4490.
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc.; 2017. p. 6000–6010.
26. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training; 2018. In: *OpenAI Blog* [Internet]. Available from: <https://openai.com/blog/language-unsupervised/> [cited 2021 March 09].
27. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. vol. 1. Association for Computational Linguistics; 2019. p. 4171–4186.
28. Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F, et al. Defending Against Neural Fake News. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.; 2019. p. 9054–9065.
29. Keskar NS, McCann B, Varshney LR, Xiong C, Socher R. Ctrl: A conditional transformer language model for controllable generation; 2019. arXiv:1909.05858 [Preprint]. Available from: <https://arxiv.org/abs/1909.05858> [cited 2021 March 09].
30. Li C, Gao X, Li Y, Li X, Peng B, Zhang Y, et al. Optimus: Organizing sentences via pre-trained modeling of a latent space. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics; 2020. p. 4678–4699.
31. Solaiman I, Brundage M, Clark J, Askill A, Herbert-Voss A, Wu J, et al. Release strategies and the social impacts of language models; 2019. arXiv:1908.09203 [Preprint]. Available from: <https://arxiv.org/abs/1908.09203> [cited 2021 March 09].
32. Platen PV. How to generate text: using different decoding methods for language generation with Transformers.; 2020. In: *Huggingface Blog* [Internet]. Available from: <https://huggingface.co/blog/how-to-generate> [cited 2021 March 09].
33. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. *Advances in Neural Information Processing Systems*. vol. 32. Curran Associates, Inc.; 2019. p. 5754–5764.

34. Mann B. How to sample from language models; 2019. In: Towards Data Science [Internet]. Available from: <https://towardsdatascience.com/how-to-sample-from-language-models-682bceb97277> [cited 2021 March 09].
35. Macbeth S. Whalefakes code; 2019. Repository: Github [Internet]. Available from: <https://github.com/sorenmacbeth/gpt-2-whalefakes> [cited 2021 March 09].
36. Woolf M. How To Make Custom AI-Generated Text With GPT-2; 2019. In: Max Woolf's blog [Internet]. Available from: <https://minimaxir.com/2019/09/howto-gpt2/> [cited 2021 March 09].
37. Woolf M. Train a GPT-2 Text-Generating Model w/ GPU; 2020. In: Google Colab Notebook [Internet]. Available from: <https://colab.research.google.com/drive/1VLG8e7YSEwypxU-noRNhsv5dW4NftGce> [cited 2021 March 09].
38. Cohn B. drillbot: GPT-2 finetuned on dril tweets; 2019. Repository: Github [Internet]. Available from: <https://github.com/o1brad/drillbot> [cited 2021 March 09].
39. Woolf M. twitter-cloud-run; 2020. Repository: Github [Internet]. Available from: <https://github.com/minimaxir/twitter-cloud-run> [cited 2021 March 09].
40. Woolf M. botustrump; 2020. Repository: Github [Internet]. Available from: <https://github.com/osirisguitar/botus-twitter> [cited 2021 March 09].
41. Hooke K. Generating Tweets Using a Recurrent Neural Net (torch-rnn); 2018. In: DZone [Internet]. Available from: <https://dzone.com/articles/generating-tweets-using-a-recurrent-neural-net-tor> [cited 2021 March 09].
42. Aggarwal R. Develop and publish Text Generating Twitter bot; 2019. In: Medium [Internet]. Available from: <https://medium.com/@aggarwal.rohan.me/develop-and-publish-text-generating-twitter-bot-9da60cfd20b8> [cited 2021 March 09].
43. Aggarwal R. AI-NarendraModi-twitter-Bot; 2019. Repository: Github [Internet]. Available from: <https://github.com/aggrowal/AI-NarendraModi-twitter-bot> [cited 2021 March 09].
44. Hayes B. Postdoc develops Twitterbot that uses AI to sound like Donald Trump; 2016. In: MIT's blog [Internet]. Available from: <https://www.csail.mit.edu/news/postdoc-develops-twitterbot-uses-ai-sound-donald-trump> [cited 2021 March 09].
45. Hayes B. DeepDrumpf; 2016. Repository: Github [Internet]. Available from: <https://github.com/deepdrumpf/deepdrumpf.github.io> [cited 2021 March 09].
46. Deep Elon: AI Tweets from Mars; 2018. In: Twitter [Internet]. Available from: https://twitter.com/musk_from_mars [cited 2021 March 09].
47. Smith D. deep-elon-tweet-generator; 2018. Repository: Github [Internet]. Available from: <https://github.com/DaveSmith227/deep-elon-tweet-generator> [cited 2021 March 09].
48. Meyer S. Create a Twitter Politician Bot with Markov Chains, Node.js and StdLib; 2017. In: Hackernoon [Internet]. Available from: <https://hackernoon.com/create-a-twitter-politician-bot-with-markov-chains-node-js-and-stdlib-14df8cc1c68a> [cited 2021 March 09].
49. Meyer S. The Jaden Trudeau Code: twitter-markov-chain;. In: Autocode [Internet]. Available from: <https://autocode.com/lib/steve/twitter-markov-chain/> [cited 2021 March 09].

50. Deep Elon Musk: IA entraînée pour tweeter comme Elon Musk; 2017. In: OpenClassrooms [Internet]. Available from: <https://openclassrooms.com/forum/sujet/ia-deep-elon-musk> [cited 2021 March 09].
51. Elon Musk Transcripts Dataset; 2017. Repository: Github [Internet]. Available from: <https://github.com/berpj/elon-musk-dataset> [cited 2021 March 09].
52. Mete J. deepThorin; 2017. Repository: Github [Internet]. Available from: <https://github.com/Jmete/deepThorin> [cited 2021 March 09].
53. Sebastiani F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*. 2002;34(1):1–47.
54. Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification. In: *IJCAI-99 workshop on machine learning for information filtering*. vol. 1. Stockholm, Sweden; 1999. p. 61–67.
55. Breiman L. Random Forests. *Mach Learn*. 2001;doi:10.1023/A:1010933404324.
56. Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Nédellec C, Rouveirol C, editors. *Machine Learning: ECML-98*. Springer; 1998. p. 137–142.
57. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998;86(11):2278–2324.
58. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2014. p. 1724–1734.
59. Verleysen M, François D. The curse of dimensionality in data mining and time series prediction. In: *International Work-Conference on Artificial Neural Networks*. Springer; 2005. p. 758–770.
60. Zhang X, Zhao J, LeCun Y. Character-level Convolutional Networks for Text Classification. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc.; 2015. p. 649–657.
61. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al.. Roberta: A robustly optimized bert pretraining approach; 2019. arXiv:1907.11692 [Preprint]. Available from: <https://arxiv.org/abs/1907.11692> [cited 2021 March 09].
62. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*; 2013. p. 108–122.
63. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics; 2020. p. 38–45.
64. Chollet F. Keras; 2015. Repository: Github [Internet]. Available from: <https://github.com/keras-team/keras> [cited 2021 March 09].
65. Rajapakse TC. Simple Transformers; 2019. Repository: Github [Internet]. Available from: <https://github.com/ThilinaRajapakse/simpletransformers> [cited 2021 March 09].

Method	Parameters
LOG_REG_BOW	C=1
RAND_FOREST_BOW	max_depth=30,min_samples_leaf=2, min_samples_split=2, n_estimators=500
SVC_BOW	kernel='linear', C=1
LOG_REG_BERT	C=1
RAND_FOREST_BERT	max_depth=30,min_samples_leaf=2, min_samples_split=15, n_estimators=100
SVC_BERT	kernel='linear', C=1
CHAR_CNN	batch_size=256, optimizer='adam', epochs=25
CHAR_GRU	batch_size=256, optimizer='adam', epochs=50
CHAR_CNNGRU	batch_size=256, optimizer='adam', epochs=25
BERT_FT	num_train_epochs=3, model_type='bert', model_name='bert-base-cased'
DISTILBERT_FT	num_train_epochs=3, model_type='distilbert', model_name='distilbert-base-cased'
ROBERTA_FT	num_train_epochs=3, model_type='roberta', model_name='roberta-base'
XLNET_FT	num_train_epochs=3, model_type='xlnet', model_name='xlnet-base-cased'

Fig 3. Supp. info 1: Parameter values. Parameter values used in the final experimentation on the test set.

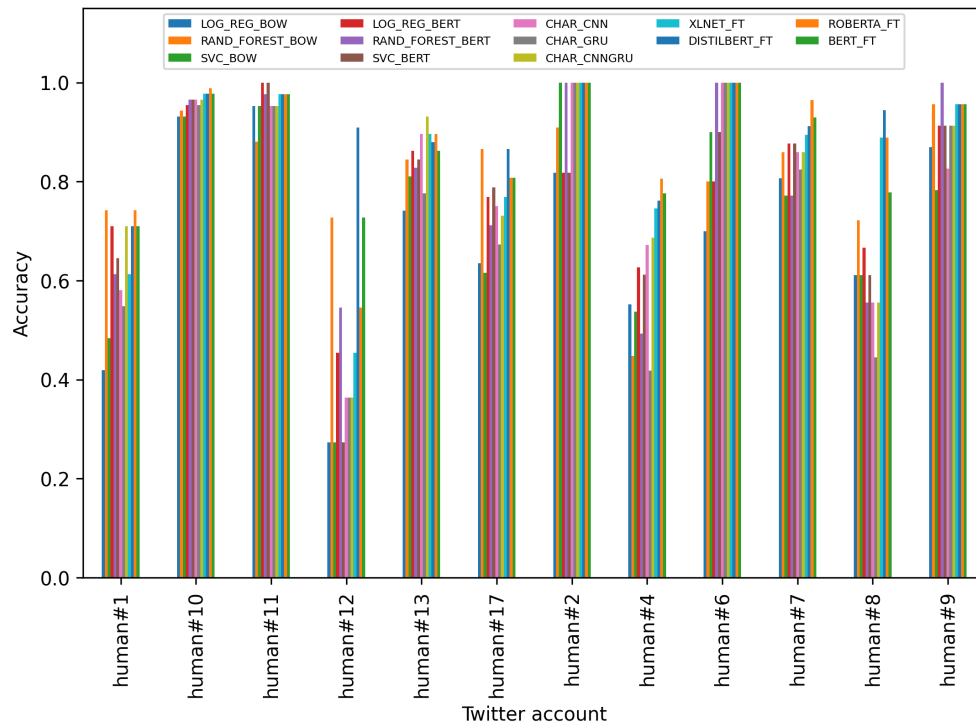


Fig 4. Supp. info 2: Acc. over humans. Detection Accuracy of tested methods on 'human' accounts with at least 10 examples.

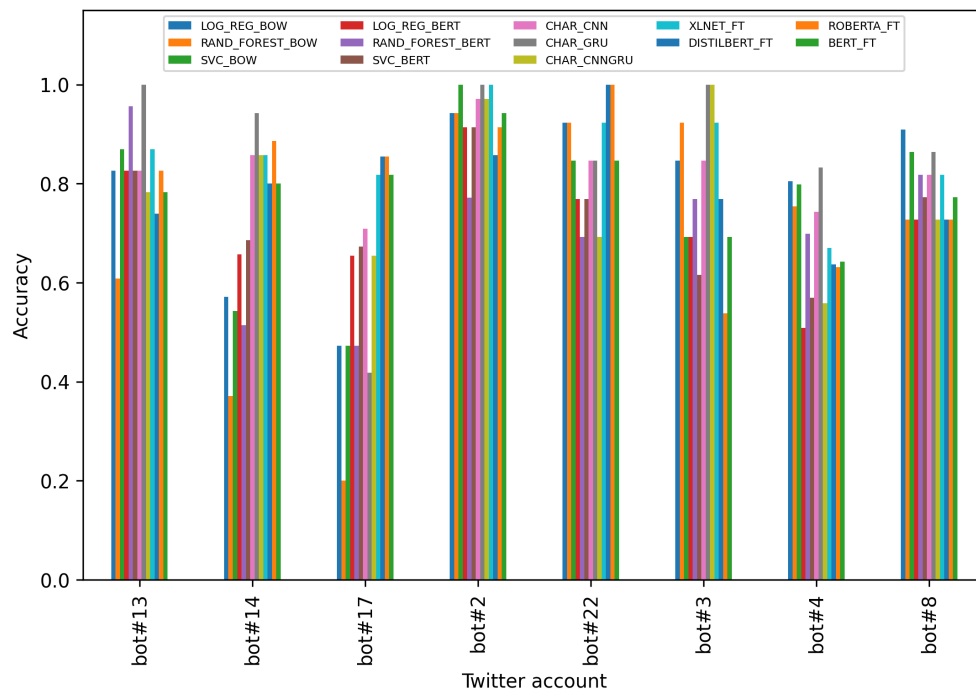


Fig 5. Supp. info 3: Acc. over GPT-2. Detection Accuracy of tested methods on 'gpt2' accounts with at least 10 examples.

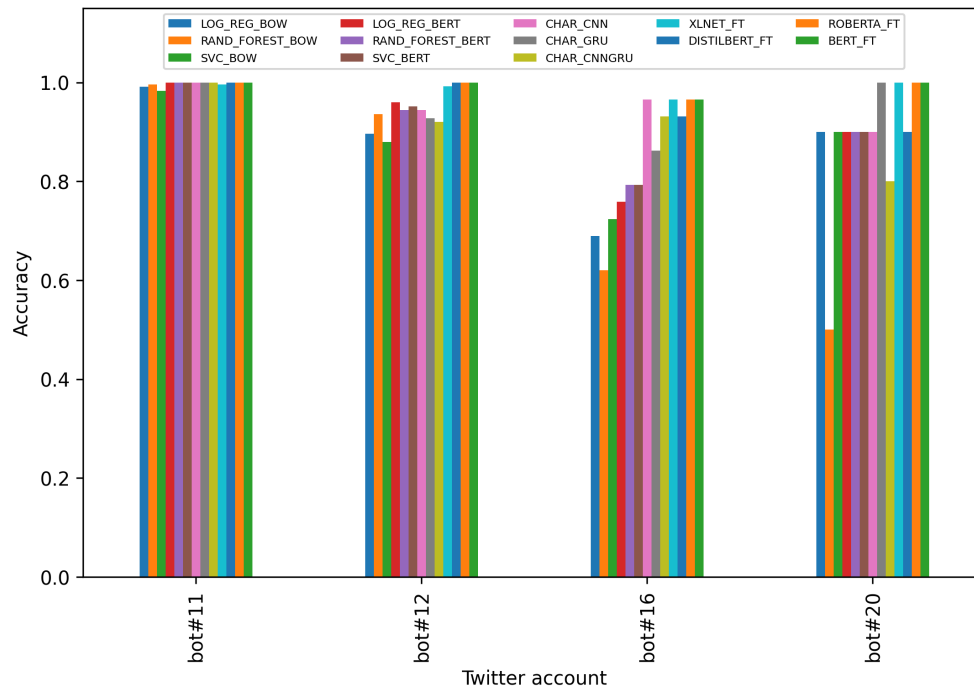


Fig 6. Supp. info 4: Acc. over RNN. Detection Accuracy of tested methods on 'rnn' accounts with at least 5 examples.

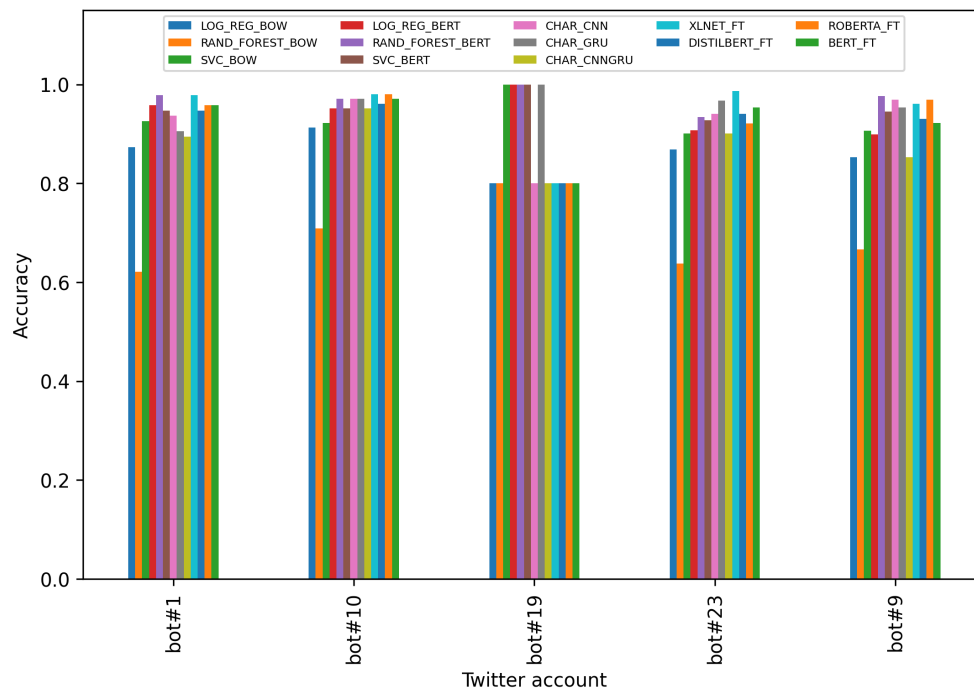


Fig 7. Supp. info 5: Acc. over others. Detection Accuracy of tested methods on 'others' accounts with at least 5 examples.