proceedings

MDPI

Abstract

Machine Learning for Dissimulating Reality †

Andrea Giussani

Department of Decision Sciences and Bocconi Institute for Data Science and Analytics, Bocconi University, 20136 Milan, Italy; andrea.giussani@unibocconi.it

† Presented at the Global Safety Evaluation Workshop, Online, 1 July–31 December 2020.

Abstract: In the last decade, advances in statistical modeling and computer science have boosted the production of machine-produced contents in different fields: from language to image generation, the quality of the generated outputs is remarkably high, sometimes better than those produced by a human being. Modern technological advances such as OpenAI's GPT-2 (and recently GPT-3) permit automated systems to dramatically alter reality with synthetic outputs so that humans are not able to distinguish the real copy from its counteracts. An example is given by an article entirely written by GPT-2, but many other examples exist. In the field of computer vision, Nvidia's Generative Adversarial Network, commonly known as StyleGAN (Karras et al. 2018), has become the de facto reference point for the production of a huge amount of fake human face portraits; additionally, recent algorithms were developed to create both musical scores and mathematical formulas. This presentation aims to stimulate participants on the state-of-the-art results in this field: we will cover both GANs and language modeling with recent applications. The novelty here is that we apply a transformer-based machine learning technique, namely RoBerta (Liu et al. 2019), to the detection of human-produced versus machine-produced text concerning fake news detection. RoBerta is a recent algorithm that is based on the well-known Bidirectional Encoder Representations from Transformers algorithm, known as BERT (Devlin et al. 2018); this is a bi-directional transformer used for natural language processing developed by Google and pre-trained over a huge amount of unlabeled textual data to learn embeddings. We will then use these representations as an input of our classifier to detect real vs. machine-produced text. The application is demonstrated in the presentation.

Keywords: machine learning; natural language processing; supervised learning; classification task

Citation: Giussani, A. Machine Learning for Dissimulating Reality . *Proceedings* **2021**, *77*, 17. https://doi.org/10.3390/proceedings2021077017

Published: 27 April 2021

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available upon request directly from the author.