Mapping Risks in AI Generated Texts

Joseph Bullock[1 2 3], Miguel Luengo-Oroz[1]

[1] UN Global Pulse, [2] Institute for Data Science, Durham University
[3]Institute for Particle Physics Phenomenology, Durham University

Introduction

Although AI bring many positive benefits to society, several major risks to peace and political stability have been identified, including: the use of automated surveillance platforms to suppress dissent; fake news reports with realistic fabricated video and audio; and the manipulation of information availability [1].

Of specific focus here will be language models, which are employed in a variety of useful way, but can also be used for malicious intent. Being able to synthesise and publish text in a particular style could have detrimental consequences augmenting those witnessed from the dissemination fake news articles and generated videos, or 'deep fakes'. Furthermore, by utilising social media platforms, such textual content can now be disseminated widely and rapidly, and used for propaganda, disinformation and personal harm on a large scale.

Here, we train a language model on political speeches and generate highly realistic samples given little resources, proving the ease with which this can be done and highlighting some of the potential risks.

Results

The model is seeded with the statements in **bold** and autocompletes the paragraphs.

Example 1: 'Regular' speech-style text

Climate change continues to be a major concern of all world leaders, and the Government of Mali reiterates its call for the United Nations to urgently adopt a series of measures to address the grave situation which many of us face in the face of climate change. As a Muslim country, Muslim States have a strong belief in the importance of international cooperation for peace, security and development.

Nuclear disarmament has been one of the basic objectives of the United Nations. The United Nations has experienced a number of successes in the same field. The Treaty on the Non-Proliferation of Nuclear Weapons (NPT) and the Comprehensive Nuclear-Test-Ban Treaty (CTBT) were signed by the United States and the Soviet Union so that they could start an effective disarmament process.

Example 2: Inflammatory speech

Refugees are terrorists and are taking the lives of their citizens. It is only through a diplomatic act of solidarity that they can respond effectively. It is a humanitarian duty. Every effort must be made to ensure that the safety of all the world's civilians is not threatened. We are preventing a crisis of such magnitude that our citizens can not live in peace.

Implications

1. **Availability**. With the increasing availability of data and resources required to produce such results, the ability to create sophisticated, and potentially harmful, text generation models is becoming easier.

2. **Easier disinformation and fake news dissemination**. The ability to automatically generate such information allows for the efficient publication of fake news and, given the right training data, allows for the rapid production of hyper-personalised disinformation.

3. **Automated generation of hate speech** (see Example 2 in Results) presents a critical challenge to human rights. This is of particular importance in situations where groups are targeted on the basis of discrimination or to incite political instability.

4. **Impersonation**. Text can be generated in a variety of styles, allowing for convincingly attributable text to a given person or group. Fake remarks can be further reinforced through the use of other technologies such as 'deep fakes'.

Data & Methodology

As a proof-of-concept experiment we train a language model on a dataset of 7,507 speeches made by world leaders at the UN General Assembly (UNGA) between 1970-2015 [2].

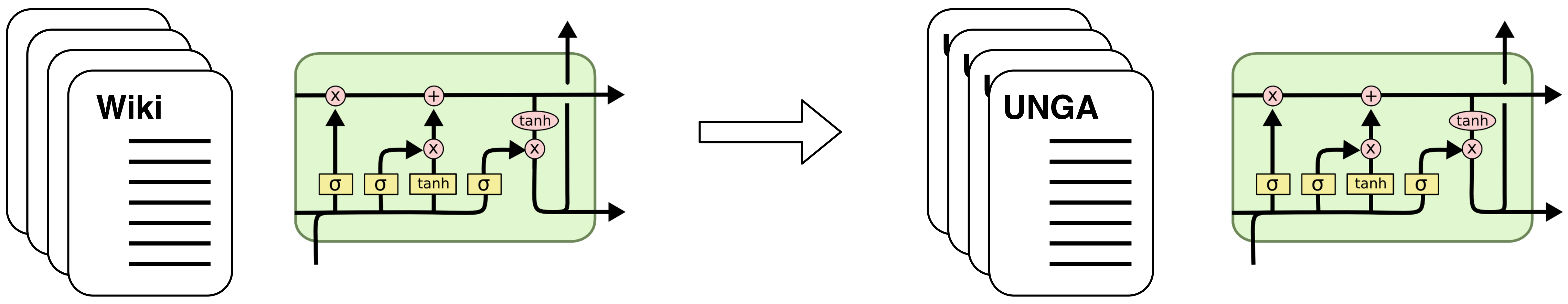

Figure 1: AWD-LSTM pre-trained on Wikitext-103 is fine-tuned dataset of UNGA speeches. LSTM image taken from [3].

In training the language model we follow the methodology as laid out by Howard and Ruder [4]: we first download an AWD-LSTM [5] language model pretrained on the Wikitext-103 dataset and then fine-tune the model on the speeches dataset.

The model is fine-tuned in as little as 13 hours, costing ∼$7.80 using cloud computing resources.

Recommendations

1. **Mapping the potential human rights impacts of these technologies** - we must continue to assess impacts in specific contexts to enhance mitigation efforts and better understand the struggles of potential victims.

2. **Development of tools for systematically and continuously monitoring AI generated content** - there needs to be greater awareness and ownership across institutions outside of the technology sector. Public and private institutions should work together to implement relevant monitoring systems, adapting them to the different evolving cultural and societal contexts.

3. **Setting up strategies for countermeasures and scenario planning for critical situations** - preemptive strategies, such as better societal education on identifying fake reports, along with relevant countermeasures, can help lessen the impact of disinformation attacks.

4. **Building alliances including civil society, international organisations and governments with technology providers, platforms and researchers** - for a coherent and proactive global strategy, a multidisciplinary approach to tackling the risks should be recognised.

Acknowledgements

JB and MLO are with the United Nations Global Pulse innovation initiative supported by the Governments of Sweden, Netherlands and Germany and the William and Flora Hewlett Foundation. JB also is supported by the UK Science and Technology Facilities Council (STFC) grant number ST/P006744/1.

Correspondence: joseph@unglobalpulse.org

References

[1] M. Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.*

[2] A. Baturo, N. Dasandi, S.J. Mikhaylov, *Understanding state preferences with text as data: Introducing the UN General Debate corpus.*

[3] C. Olah, https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[4] J. Howard, S. Ruder, *Universal Language Model Fine-tuning for Text Classification.*

[5] S. Merity, N.S. Keskar, R. Socher, *Regularizing and Optimizing LSTM Language Models.*