Public Policy for Responsible Language Model Deployment

Phillip Wang (pkwang) and Jeremy Ong (tto)

Carnegie Mellon University

Introduction

Recent advancements in Natural Language Processing have created language models that can generate human-like text which could be used maliciously to for example generate large amounts of fake news or perform astroturfing. We explore the implications of the inevitable deployment of convincing language models and what kinds of policies we should enact to mitigate the negative consequences.

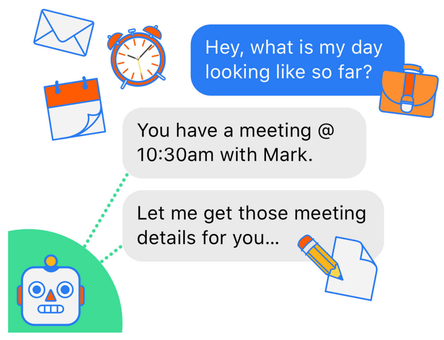
.

Background

- OpenAI created a language model trained on 40GB of text from 8 million web pages, capable of generating convincing human-like samples of text. Beats state of the art in all NLP benchmarks [1]
- An RNN was trained to generate product reviews [2].
- OpenAI's recent language models are results that could conceivably be developed by a number of research groups
  - Researchers from AllenNLP and Google have released work that is very similar to the approach that OpenAI did, but OpenAI used more data

Positive Implications

- AI writing assistants
- More capable dialogue agents
- Automated systems for translation
- Better speech recognition



Negative Implications

- Generate misleading articles
- Impersonating others online
- Automate the production of abusive/faked content on social media
- Decreases trust of content that you see on the internet
- Astroturfing – creating fake grassroots movements

Possible Policies

We don't need any policies

- We've seen the effects of fake news and fake reviews and the existence of these language models will exacerbate these problems, so policy is necessary.

Per-institution policies on the development and release of strong language models

- Only effective under the assumption that the institutions capable of developing such systems are few and easy to control which is not the case.
- Would also halt the research progress on language models.

Internet-level policies

- De-anonymizing people on the internet to make it harder to publish large amounts of fake content or perform astroturfing.
  - Requires publishers to identify themselves
  - Social media sites that require you to enter SSN-like ID
- Watermark on anything generated by deep models and shown to the public
  - Hard to enforce, using current technology
  - How to create a watermark that can't be modified
- Controlling who has access to powerful text-generation networks similar to control of cryptography technology[3].
  - Hard to enforce because deep learning is pretty accessible
- Make software that detects computer generated text publicly accessible [4,5]
  - More research and development required to create systems that can detect generated text. It is harder to detect generated text than generated videos.

SYSTEM PROMPT (HUMAN-WRITTEN)	*Recycling is good for the world.* *NO! YOU COULD NOT BE MORE WRONG!!*
MODEL COMPLETION (MACHINE-WRITTEN, 25 TRIES)	Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources. And THAT is why we need to get back to basics and get back to basics in our recycling efforts. One of the

Figure 1: Example of generated text from the OpenAI language model

Our Proposed Policy

With our current internet structure, it is easy to use language models maliciously because there is a lack of accountability when publishing internet content.

International Internet Association issues internet identification number (IIN) to each person. IINs have associated internet credit scores.

No one institution (such as governments) can trace back to the person

- Still allows free speech on the internet
- Possibly a decentralized system (like blockchain)

Social media and news websites will require you to provide your IIN upon creating an account. This way, publishers of fake news can be held accountable and be suspended if detected, and it will be hard to astroturf without being able to easily make a lot of fake accounts.

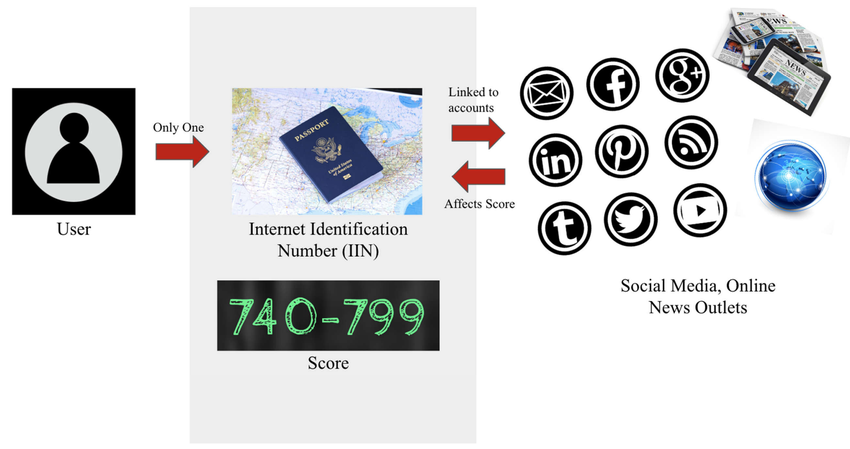


Figure 2: Our Proposed Policy Visualized

Conclusion

Widespread accessibility to language models that can generate human-like text is inevitable given the rate of progress and nature of the technology. Internet Identification Numbers will mitigate societal concerns while still allowing us to pursue research in the field by providing greater accountability for the publishers of content on the internet. A lot of overhead is required to implement the system, but we believe this is the best long term solution to handle highly capable language models.

Bibliography

1. https://openai.com/blog/better-language-models
2. [2015, Lipton et. al] "Generative Concatenative Nets Jointly Learn to Write and Classify Reviews," arxiv.org
3. https://en.wikipedia.org/wiki/Export_of_cryptography_from_the_United_States
4. https://ieeexplore.ieee.org/document/8282270
5. https://applymagicsauce.com/about-us