Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (34^{th}cycle)

Dissecting Deep Language Models

The Explainability and Bias Perspective

By

Giuseppe Attanasio

******

Supervisor(s):
Professor Elena Baralis

Doctoral Examination Committee:
Professor Giuseppe Rizzo, Links Foundation, Torino, Italy
Professor Sara Tonelli, Fondazione Bruno Kessler, Trento, Italy

Politecnico di Torino
2022

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Giuseppe Attanasio
2022

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Dissecting Deep Language Models

Giuseppe Attanasio

Language models are statistical representations of language that allow AI systems to work with text. They are increasingly ubiquitous, powering language technologies such as social networks, chatbots, writing assistants, translation tools, and more.In recent years, we have seen the release of larger and more complex models – we call them Large Language Models (LLMs) – to accommodate diverse tasks and contexts.

However, recent studies have shown that language models can learn social biases from training data. Production-ready systems that subsequently use these models often harm underrepresented groups and categories. For example, a language model for hate speech detection would classify the sentence "Girl, I adore you" as misogynous because the word "Girl" tends to appear in misogynous utterances. Moreover, as the complexity of LLMs increases, this undesirable behavior becomes harder to detect or control. Studying models' learning dynamics and explaining their predictions would help detect and mitigate harmful outputs.

This work provides a critical overview of common pitfalls in the sensitive task of automatic hate speech detection and presents practical techniques to detect and mitigate unintended bias. First, we study sentence embeddings for misogyny detection. Results demonstrate that peculiar social media language confounds models that fail to generalize. Next, we propose a novel regularization technique to reduce lexical overfitting and mitigate bias. Entropy-based Attention Regularization (EAR) acts on self-attention weights to improve the representations of words. Finally, we tackle the issue of explainability in language modeling by benchmarking four post-hoc feature attribution methods on the misogyny identification task.

Our results highlight issues in both pre-trained and fine-tuned language models. However, this thesis demonstrates how intentional training choices and improved model transparency can help detect and mitigate biased outcomes. Furthermore, our findings open future avenues for understanding large language models' learning and inference dynamics.

Nomenclature

The next list describes several symbols that will be later used within the body of the document

LLMs Large Language Models

LMs Language Models

NLP Natural Language Processing