1859

Politecnico
di Torino

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

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (34thcycle)

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).

To my family, my safe harbor.

Acknowledgements

Having a Ph.D. during a global pandemic is different. Friendships and relationships grow stronger than usual.

I have crossed roads with many in the last three years. I want to thank every one of you; I genuinely feel that everybody has made me grow in some sense. You are the true essence of this achievement.

Elena Baralis, my advisor, for helping weigh every choice, discussing what I thought was one way but wasn't, and the freedom she gave me in pursuing the research I liked.

Luca Cagliero and Paolo Garza, for guiding my first steps as an inexperienced researcher.

Evelina, Stefano, and Elena for having eased my way in and made my days joyful.

Flavio Giobergia, for being a friend, a colleague, and the best flatmate during a pandemic.

Francesco Ventura, for having told me once: "This Ph.D. thing has moments. You start with high hopes, get knocked hard, and eventually find a way."

Federico Manuri, for all you did when I was in need. Our long list of tabletop nights is and will be unparalleled.

Antonio Cipolletta, for the true friendship and countless advice. Your passion for science has inspired and guided me through these years.

Gian Pietro Bellocca, for sharing most of the path together, with all highs and lows that come.

C*zi, my long-standing inner circle. We have changed, but we did that together, and, despite the distance, I know we are growing tighter.

Dirk Hovy, Federico Bianchi, and Amanda Curry, for your incredible mentorship, advice, and collaboration we had in this late period. All of that was a catalyst to my maturation as a researcher.

Debora Nozza, for being an incredible mentor, collaborator, and friend, all together. I would not be here if it were not for you.

Eliana, my guiding light, especially in difficult times.

I have likely forgotten someone. It is said it is common but sorry for that. Thank you to you as well.

Abstract

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.

Contents

List of Figures

List of Tables

Nomenclature

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

AMI	Automatic Misogyny Identification
AUC	Area Under the Curve
BERT	Bidirectional Encored Representations from Transformers
BoW	Bag-of-Words
CAGE	Commonsense Auto-Generated Explanation
CDA	Counterfactual Data Augmentation
CLM	Casual Language Modeling
EAR	Entropy-based Attention Regularization
FNED	False Negative Equality Difference
FPED	False Positive Equality Difference
GPT	Generative Pre-trained Transformer
HTA	Hidden Token Attribution
IG	Integrated Gradients
LLMs	Large Language Models
LMs	Language Models

MLM	Masked Language Modeling
NLP	Natural Language Processing
NN	Neural Network
PTLM	Pre-Trained Language Model
RNN	Recurrent Neural Network
SBERT	Sentence-BERT
SEAT	Sentence-Embedding Association Test
SHAP	SHapley Additive exPlanations
TF-IDF	Term Frequency-Inverse Document Frequency
USE	Universal Sentence Encoder
WEAT	Word-Embedding Association Test
XAI	eXplainable AI

Preamble

Disclaimer

The manuscript contains sentences with slurs, stereotypical language, and passages some readers might find offensive. Although we do not entirely censor them (Jane, 2014), **they do not reflect the views of the authors.**

Chapter 1

Introduction

1.1 Motivation

Natural Language Processing (NLP) systems have become extremely popular. Production tools filter out spam e-mails, moderate hateful content online, [1] [2] and translate hundreds of languages. [3] [4] Successful consumer-oriented products span from writing assistants[5] to realistic text-based role-playing games.[6]

With very few exceptions, system designers build NLP applications around a mathematical abstraction of the human language, commonly known as a **language model**.[7] Language models are statistical tools: when fitted to some training text, they can model word distribution and syntactic rules, and learn how to combine them to solve language-related tasks.

Facilitated by more efficient computing, language models have increased in size and complexity. Starting from small and non-parametric, they evolved to intricate neural networks (NNs).

[1] https://techcrunch.com/2022/01/24/spectrum-labs-b/

[2] https://techcrunch.com/2020/06/11/sentropy-emerges-from-stealth-with-an-ai-platform-to-tackle-online-abuse-backed-by-13m-from-initialized-and-more/

[3] https://techcrunch.com/2022/05/11/google-translate-adds-24-new-languages-including-its-first-indigenous-languages-of-the-americas/

[4] https://ai.facebook.com/research/no-language-left-behind/

[5] https://app.grammarly.com/

[6] https://aidungeon.io/

[7] In modern NLP, a language model is, technically speaking, a mathematical model used to predict the next word given a sequence of previous words. We detail this definition in Section 2.1.

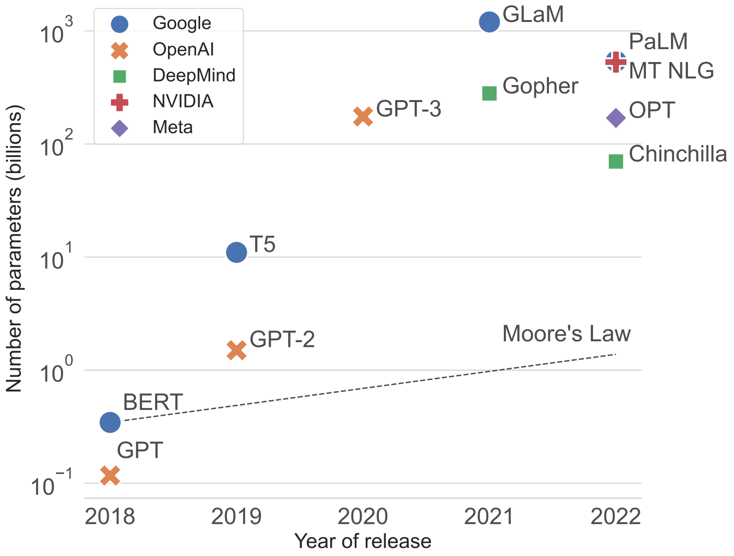


Fig. 1.1 Recent LLMs. This non-exhaustive list includes BERT (Devlin et al., 2019), Generative Pre-trained Transformer (GPT) (Radford et al., 2018), GPT-2 (Radford et al.), GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), GLaM (Du et al., 2021), Gopher (Rae et al., 2021), Megatron Turing Natural Language Generation (MT NLG) (Smith et al., 2022), Pathways LM (PaLM) (Chowdhery et al., 2022), Open Pretrained Transformer (OPT) (Zhang et al., 2022), Chinchilla (Hoffmann et al., 2022).

Recently, the flexibility of the **Transformer** (Vaswani et al., 2017b), a new type of new neural network, along with empirical evidence on scaling laws (Kaplan et al., 2020) has fueled the run for larger models: since the publication of *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019) size has increased quicker than Moore's law (Figure 1.1).[8] This thesis refers to these models as **Large Language Models** (LLMs).

As LLMs find increasing public adoption, the societal impact of the application they enable has become paramount. Recent studies have shown that LLMs can encode social biases if those are present in the training data.

Gender bias, for example, is a well-known issue. Trained models learn stereotypical associations between gender and occupation (e.g., "man" is to "CEO" as "women" to "nurse") leading to biased pronoun resolution (de Vassimon Manela et al., 2021; Zhao et al., 2019), machine translation (Stanovsky et al., 2019), sentiment

[8] Moore conceived his theory in the blooming age of transistors. Although no evidence suggests that transistors and model parameters follow the same scaling laws, the parallelism that sees them as the "base unit" of a larger computational system holds.

analysis (Basta et al., 2019), and more. Further work has shown that LLMs can learn a biased representation of words (Bhaskaran and Bhallamudi, 2019) and sentences (May et al., 2019). We cover in detail bias in language models in Section 2.2.

In this thesis, we study the interplay between learning dynamics and regularization to mitigate such biases in Transformer-based hate speech classifiers. Further, we analyze the strengths and weaknesses of state-of-the-art explainability methods with such classifiers and provide guidelines for their use. Our work is motivated by two core research questions:

RQ1 Are hate speech classifiers biased by lexical features, such as trigger words and language-specific constructs? If it is the case, what mitigation strategies can we adopt that do not require *a-priori* access to these words or phrases?

RQ2 Can explainability approaches shed light on biases caused by lexical features? If it is the case, what method provides *better* explanations?

1.2 Contribution

This work focuses on bias in hate speech detection systems based on large language models and proposes effective mitigation techniques. Our contribution is three-folded.

First, we focus on state-of-the-art sentence encoders, i.e., systems that use a language model to encode text into a dense, semantically rich vector. Here, we probe these vectors to understand whether a given tweet contains misogynous speech or not (Attanasio and Pastor, 2020). Our results show that off-the-shelf models do not provide robust representations and perform best when paired with TF-IDF, curated lexicons, and semantic parsing. Finally, we identified confounding factors in social media text that fool the model, like the presence of particular words. The leading cause we found is poor generalization capabilities.

In light of such findings, we propose Entropy-based Attention Regularization (EAR) (Attanasio et al., 2022b), a novel regularization approach to mitigate bias caused by lexical overfitting. EAR builds on the idea that tokens learned with a narrow self-attention induce overfitting as they bring small meaning from the surrounding context. Therefore, we mitigate bias by constructing tokens that **observe more context**. We report technical details in Chapter 4. Training BERT with EAR on English and Italian hate speech detection datasets improves performance and bias metrics. Moreover, we propose an automatic procedure to extract overfitting terms.

Third, we address the topic of explainability in Transformer-based language models for misogyny detection (Attanasio et al., 2022c). We benchmark four state-of-the-art post-hoc feature attribution explainable approaches on LLMs and discovered that not all methods provide faithfull and plausible explanations. Finally, we compared Attention and Hidden Token Attribution and demonstrated that Attention could not offer any interpretability insight.

1.3 Outline

We organize the rest of the manuscript as follows.

- **Chapter 2: Background.** We introduce introductory notions on modern language models. Further, we provide a thorough overview of recent research on social bias in NLP, encompassing intrinsic and extrinsic bias characteristics. Finally, we introduce the topic of Explainable AI and describe recent advances in the field of NLP.

- **Chapter 3: Improving Sentence Embedding with Misogyny Lexicons.** We report our study on probing sentence embeddings for automatic misogyny detection. Most of the content reflects work done and published in Attanasio and Pastor (2020).

- **Chapter 4: Entropy-based Attention Regularization for Unintended Bias Mitigation.** We describe our novel regularization technique to reduce lexical overfitting in language models. Most of the content reflects work done and published in Attanasio et al. (2022b).

- **Chapter 5: Benchmarking Post-Hoc Interpretability Approaches for Misogyny Detection.** We detail our benchmarking study on explainability approaches applied to LLMs for the task of misogyny detection. Most of the content reflects work done and published in Attanasio et al. (2022c).

- **Chapter 6: Conclusion.** We review the main contribution of this thesis and discuss future work and open directions.

Chapter 2

Background

This chapter introduces the core background notions of topics covered later in the document. It also presents relevant literature on bias and explainability in modern language models.

First, we introduce the concept of **language model** across a historical excursus.[1] We then overview the **Transformer** architecture (Vaswani et al., 2017b) which has become the standard de facto to learn word representations and relationships. Transformer builds on the **attention** mechanism (Bahdanau et al., 2015; Graves, 2013) a novel solution to learn alignment between entities and build new representations based on it. Unsurprisingly, in language models, entities are words and word representations. We then provide key intuitions to distinguish *autoregressive* from *bidirectional* models, the two most common variants. Finally, we review the difference between *pre-training* and *fine-tuning*.

Next, we discuss the topic of **social bias in language models**. We disambiguate the classic definition of “bias” known in the Machine Learning from “bias” seen as the issue of a model discriminating against minorities, with the latter being the main focus of this work. Further, we situate discriminatory “bias” in NLP, surveying discussions and results in recent literature, focusing on the distinction between intrinsic and extrinsic biases.

Finally, we overview recent advances in the field of **eXplainable AI** (XAI) (Lipton, 2018), a broad, established area that addresses the problem of interpreting

[1] As widely accepted, most advances in computational linguistics have been achieved in the last 20 years, starting from the first neural network-based language model in Bengio et al. (2000).

models and the models' predictions. Following chronologically the development of the field, we first introduce foundational concepts such as the different ways to explain an outcome, the distinction between *local* and *global* explanations. Next, we present novel XAI methodologies for NLP.

The three sections of our literature review are tightly interconnected. How we train modern language models – i.e., with Transformer and Attention – is far from perfect. Many examples using off-the-shelves tools have shown undesired discriminatory bias (Chapter 3). One solution would be working on the model, for instance, mitigating bias due to lexical overfitting on training words, such as slurs or terms identifying minority groups (Chapter 4). Another approach entails explaining the model and its predictions to detect and debug harmful associations before deployment. However, reliability and quality assessment of NLP explanation is still an open research field (Chapter 5).

2.1 Anatomy of a Language Model

A language model is a probabilistic model designed to assign a probability to an arbitrary sequence of words, i.e.,

$$\text{LM} := P(w_0, w_1, ..., w_S) \tag{2.1}$$

where w_i is a word and S the length of the sequence. Although simple, this formulation highlights at least three crucial aspects.

First, words are the basic unit of the model. Reflecting this view, many reformulate the problem as predicting the probability of a word given its *context*, i.e., all or a subset of the sentence. For example, modern *autoregressive* models learn a conditional word probability on the preceding context, i.e.,

$$P(w_i | w_0, ..., w_{i-1}) \tag{2.2}$$

Second, the formulation of P defines the complexity of the model. Classic (and outdated) models compute statistics based on word occurrences. Most modern models are parametric neural networks that learn conditional word probability using gradient optimization.

Third, coding and operationalizing a model as in Equation 2.1 entails defining a representation of words a computer can understand. Indeed, recent research has put a great effort into learning language via *understanding the meaning of its units*, i.e., words. An established approach in computational linguistics is representing words as dense vectors. Ideally, the vector acts as a briefcase where all the relevant information is squoze.

We briefly discuss how models have changed in the past two decades (the P), discussing advances that involve the representation of word vectors (the w_i), and how these ideas blend into the Transformer (Vaswani et al., 2017a), the building block of modern language models.

2.1.1 A Historical Perspective

Pre-neural language models solved linguistics tasks using count-based approaches. The idea is to count word occurrences from a corpus and use them to estimate word probabilities. The well-known *n-gram* model (Manning et al., 2010) is a representative of this class of models. It approximates Equation 2.1 under the Markov assumption: word probability depends only on a fixed-width context (n) preceding it, i.e.,

$$P(w_0, w_1, ..., w_S) = \prod_{i=0}^{S} P(w_i | w_{i-n+1}, ..., w_{i-1}) \tag{2.3}$$

Refreshing ideas dated back to the beginning of the century (Bengio et al., 2000), in the early 2010s, count-based word representations were superseded by prediction models (Baroni et al., 2014). In Mikolov et al. (2013) the authors learn dense vector representations - what we commonly call today *word embeddings* – using neural networks and gradient descent. An efficient learning algorithm and large corpora allowed training word representations that capture syntax and semantics.[2]

Building on new, semantically rich word representations, the NLP community shifted the focus toward the probability model itself, i.e., finding better Ps. But, again, the community borrowed ideas from the past, and Recurrent Neural Networks

[2] Algebraic operations in the vector space show that v[“biggest”] - v[“big”] + v[“small”] = v[“smallest”] and that v[“France”] is to v[“Paris”] as v[“Germany”] is to v[“Berlin”], with v being the dictionary of words.

(RNNs) (Elman, 1990) found new life in practical applications.[3] An RNN is a model designed to process sequences as input data. It differentiates from other neural models in two main aspects: it ingests one item at a time and uses an internal status to keep track of "experience" from previous steps. This formulation appeared to fit nicely with modeling language. Leveraging word embeddings as inputs led to successful applications of RNN-based language models in various tasks, such as sentiment classification, summarization, image captioning, and machine translation (Cho et al., 2014; Chopra et al., 2016; Sutskever et al., 2014; Xu et al., 2015, *inter alia*).

The path that led to the Transformer required one last step. Around 2014, RNN variants (e.g., Long-Short Term Memory (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Unit (Cho et al., 2014)) found successful applications in sequence-to-sequence tasks. One example is machine translation, i.e., modeling the probability of a sentence in a target language (say, Italian) conditioned on an input in a source language (say, English). Early approaches involving an encoder-decoder network had two major issues: non-parallelizable computation – a known limitation of RNNs that impose sequential computation – and the information passage between the networks. The decoder RNN accessed the source sentence only via the last status update of the encoder, which severely limited long-range interactions (e.g., a decoding term translating a word early in the source sentence).

Although some training choices mitigated the latter issue, systems achieved the most considerable improvement by introducing **Attention** to align source and target sentences (Bahdanau et al., 2015). The Attention mechanism elegantly formulates this alignment problem as a learning algorithm; it does not require sequential computation and solves long-range dependencies. These characteristics set the stage for the Transformer.

The Attention Mechanism

It is safe to say that the attention mechanism lies at the core of *all* best-performing language models. This simple alignment algorithm is the foundation of how we model natural language today.

[3] In parallel, several works (Kim, 2014) explored the use of Convolutional Neural Networks (CNNs) as well.

Attention was introduced in Bahdanau et al. (2015) as an *alignment* mechanism in a encoder-decoder translation network. The idea was to connect every target word with every source word and learn *attention weights* as part of the training, all of that under a parallelizable implementation. Before reviewing Attention in Transformer (Section 2.1.2), we provide the intuition using influencers and dress styles.

Fashion trends change rapidly. Harry knows that and tries to keep his wardrobe ready. Every season he goes over the social profiles of his favorite fashion influencers to look for ideas. Harry finds nice shirts in profile 1, suitable shoes in profile 2, nothing exciting in profile 3, and so on. From each influencer, he chooses part of the outfit for the upcoming season. In a sense, he *aligns* his preferences with social profiles and *mixes* different styles following his intuition on what is best for his final goal – we do not know Harry. Maybe he is trying to be a famous influencer himself.

Transformers learn word representations similarly. Each word is a *query* (Harry's outfit) whose representation is updated in alignment with a set of other words (the influencers' profiles), the *keys*, mixing some of their *values* (the influencers' products). Again, some training objective (Harry's dream of becoming an influencer) drives the process.

2.1.2 The Transformer Model

The Transformer (Vaswani et al., 2017b) is an encoder-decoder neural network originally devised for sequence-to-sequence tasks. The encoder and the decoder use attention to learn and align word representations. Notably, computations in the system involve only attention and fully-connected layers, requiring no sequential computation.

Encoder The transformer encoder (Figure 2.1, left) mixes input words using attention, then feds the results to a fully-connected feed-forward block with point-wise non-linear activation. Both the operations apply residual connection and layer normalization. This computation is repeated N times by identical, stacked replicas to compute the final word representations. The first unit of the encoder applies a *multi-headed self-attention*, meaning that i) words "mix and align to the sentence itself" and ii) multiple, different alignments are learned at once (see Section 2.1.2 for details) – each alignment is inputed to one *attention head*. This simple learning paradigm – based on mixing and aligning words in sentences – paired with a linguistically

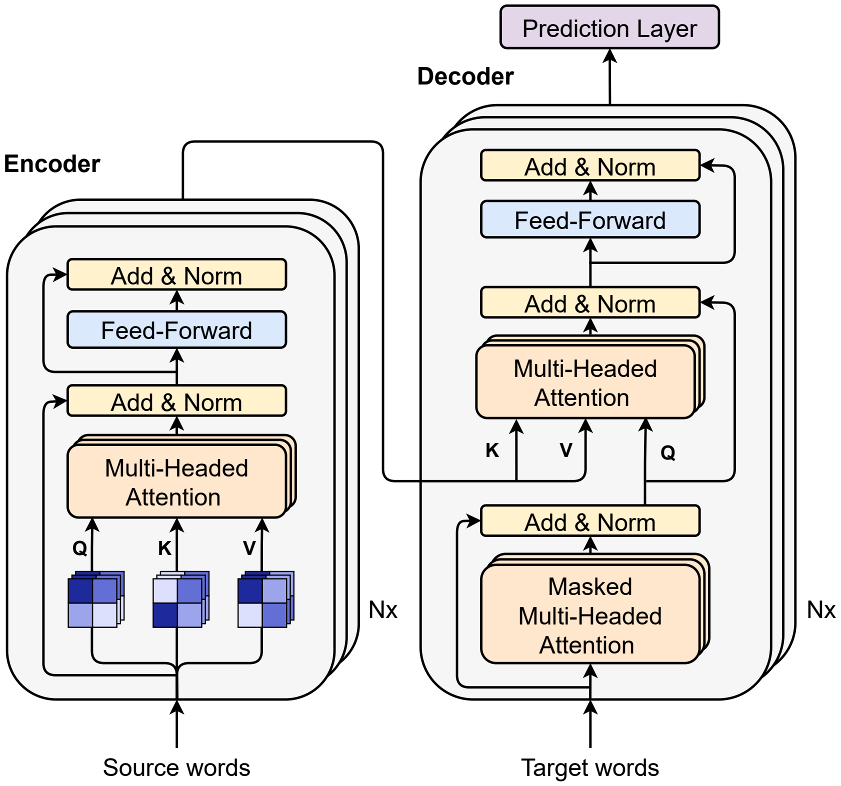


Fig. 2.1 Transformer model. In the encoder (gray box, left), linear projections (blue squares) generate queries (**Q**), keys (**K**), and values (**V**) from the input sequence. In the decoder (gray box, right), a simplified view of the masked attention and the cross-attention block. A prediction layer maps the decoder output to logits.

founded training objective (see Section 2.1.3), enables the best performing language models.

Decoder Similarly, the decoder (Figure 2.1, right) mixes words from the target sentence using masked self-attention.[4] However, before the fully-connected unit, an additional *cross-attention* block computes the alignment with the source words. Recall the example on dress styles: the decoded word is the query (Harry looking for ideas), while words from the encoder are the keys (the influencers' profiles) *and* the values (the products used by Harry next season).

[4] Masked self-attention allows words to express attention weights only to the left context, i.e., all preceding, already decoded words.

Forward pass in the Transformer encoder

We provide details for a standard forward pass in the encoder. In attention blocks, the multi-head output is computed with Scaled Dot-Product Attention between a set of queries and keys of dimension d_k, and a set of values of dimension d_v. Let Q, K and V be the respective matrix representations. The attention is then computed as

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The network replicates the operation on NH different, independent linear projections of the same queries, keys, and values in the so-called attention *heads*. The heads are then concatenated, projected back to the original input space, and finally fed through the fully connected neural network to produce the next layer embeddings. Let $E = [e_0, ..., e_{d_s}]$ be the sequence of input embeddings[5], with $e_i \in \mathbb{R}^{d_m}$. In the specific case of a transformer encoder, queries, keys and values correspond to the input embeddings - i.e. $Q = K = V = E$. Since the values are projections of the tokens themselves, each weight in self-attention measures the contribution of its token to the attention head and, in turn, to the new token representation. The output of the multi-head self-attention block is computed applying the previously presented Equation to the N token projections, concatenating and projecting back to the original space:

$$\text{MultiHead}(Q,K,V) = (\text{o}_0||\dots||\text{o}_N)W^O$$

where

$$\text{o}_h = \text{Attention}\left(QW_h^Q, KW_h^K, VW_h^V\right)$$

and W^O and each W_h^Q, W_h^K, W_h^V are projection matrices.

2.1.3 Large Language Models

The Transformer has quickly become the standard de-facto for language modeling, with final models showing impressive transfer learning abilities to downstream tasks. It allows direct communication between every pair of words through self-attention

[5] The input embeddings for the first layer are the static token embeddings plus their position encoding.

(Section 2.1.2), it has different learning units – or attention heads – and computation is parallelizable. But one more essential aspect led to widespread adoption.

The architecture enables easy scaling of the learnable parameters – e.g., increasing N, the number of stacked layers or NH, the number of attention heads – effectively improving the model's capacity. Large-scale training corpora and curated training choices have enabled increasingly performant models – the examples in Figure 1.1 are best in many NLP tasks and linguistic benchmarks (Wang et al., 2018, 2019a).[6]

This thesis refers to these Transformer-based models as Large Language Models (LLMs). But *how* do these models learn from Transformer? The key lies in size and training objectives, and data.

Architecture

Primarily, language models are bigger variants of one among the transformer encoder and decoder. Architectures are designed by stacking more layers (N), increasing the number of attention heads (NH), or the dimensionality of word embeddings (d). As a reference, the recently introduced Open Pretrained Transformer (Zhang et al., 2022) counts 96 stacked decoder layers, attention with 96 heads and 12288-dimensional hidden vectors in its largest variant.[7]

Pre-training procedure

Training LLMs is, most commonly, a self-supervised procedure. First, the model is fed with a raw textual corpus - which eventually undergoes pre-processing, cleansing, de-duplication, etc. - and some form of pre-training objective guides learning. This phase is commonly known as "pre-training" as the model is learning general word representations that account for syntax, grammar, and word-in-context meaning, and the resulting models **Pre-Trained Language Models** (PTLMs).

Two families of pre-training objectives have prevailed in recent years: **Masked Language Modeling** (MLM) and **Casual Language Modeling**.

[6] The flexibility attracted other fields as well. We have seen new models in computer vision (Dosovitskiy et al., 2021), reinforcement learning (Chen et al., 2021), computational biology (Avsec et al., 2021), *inter alia*, and settings, e.g., vision-text multi-modality (Radford et al., 2021).

[7] To study the effect of size and improve accessibility in low-resource contexts, authors often release smaller versions that share the training procedure with their bigger counterparts.

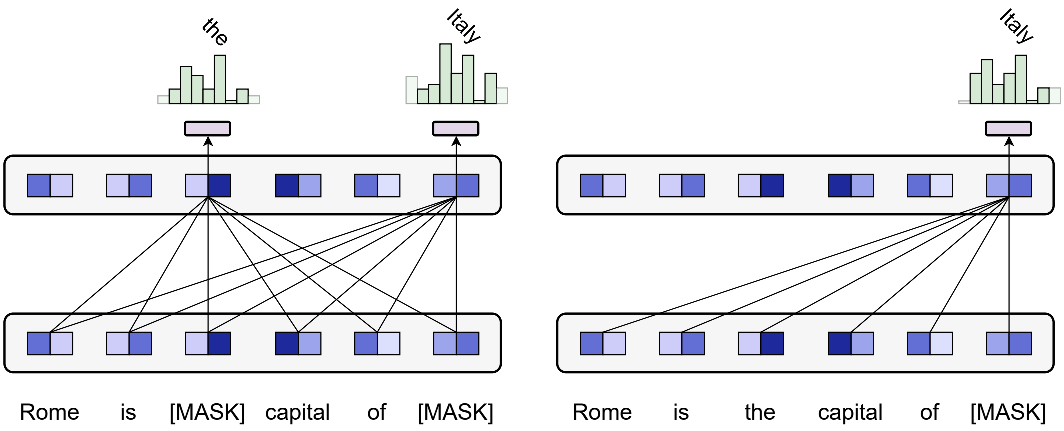


Fig. 2.2 Simplified view of Masked (left) and Casual (right) Language Modeling pre-training in a two-layers Transformer language model. Attention weights are shown only for masked and target tokens. Gray boxes are encoders or decoders; blue squares are token embeddings. In MLM, tokens attend left and right contexts, whereas CLM allows left-context attention only. Classification heads compute a distribution probability over the words in the vocabulary.

In MLM (Figure 2.2, left), a subset of words is masked, and the goal is to predict what is missing. For each masking, a classification head denotes a probability distribution over the vocabulary of allowed tokens, and standard cross-entropy loss is applied. Most models pre-trained with MLM use a stack of transform encoders that allows for *bi-directional* self-attention: tokens express attention weights to both right and left contexts. The first representative of this class of models is BERT (Devlin et al., 2019). In this thesis, we will focus mainly on this type of model.

In CLM (Figure 2.2, right), the goal entails predicting the next word given a preceding context. Models that use CLM are typically decoder-only. Given the nature of the pre-training task, these models – also known as *autoregressive* language models – learn to *generate* text given a context. Recently, large-scale models have become state-of-the-art in tasks such as question-answering, summarization, or original storytelling.

As a further remark, we prompted[8] GPT-3 (Brown et al., 2020) to explain the difference between MLM and CLM. The unedited result follows.

> Prompt: *Can you explain the difference between the Masked and Casual Language Modeling pre-training tasks?*

[8] `text-davinci-002` model from https://beta.openai.com/playground. Temperature 0.7, maximum length 2048, top P 1, frequency penalty 0.05, presence penalty 0, best of 1.

> Generated text: *The Masked Language Modeling pre-training task randomly masks a percentage of the input tokens and then asks the model to predict the masked tokens. The Casual Language Modeling pre-training task predicts the next token in the sequence, without masking any tokens.*

Pre-training objectives are not, however, the only aspect involved. The tokenization strategy, positional encoding schemes, tied input and output weights, optimizer, learning rate schedule, complementary pre-training tasks, and more play a role in the learning dynamics. As these settings depend on the type of application and training data, we leave them out of this chapter and provide details when needed in central chapters (3-5).

Note that pre-training involves a massive amount of data commonly collected from the Web. Models hence **reflect the views of groups most represented online and fail to express ideas from underrepresented groups** (Bender et al., 2021). As one might expect, PTLMs internalize stereotypical and prejudicial misconceptions – which can easily be undercovered via sentence completion (Nozza et al., 2021), coreference resolution (de Vassimon Manela et al., 2021), or template filling (Bartl et al., 2020) tests – that those transfer even after fine-tuning (Steed et al., 2022).

Downstream Fine-Tuning

As empirical results and probing setups have shown, pre-training models syntax and semantics of words (Hewitt and Manning, 2019; Jawahar et al., 2019) while learning linguistically aware representations (Conneau et al., 2018; Ettinger, 2020).

However, one cannot apply the model to any specific task without some *specialization*. Therefore, the current trend is to run a second training step – commonly known as **fine-tuning** – and tune the model for a specific NLP task such as Sentiment Classification, Natural Language Inference, Question Answering, and more (Devlin et al., 2019).

In this thesis, we leverage pre-trained sentence embedding (Chapter 3) and language models (Chapter 4-5) and fine-tune them on task-specific datasets.

2.1.4 Pre-Trained Language Models for Sentence Embeddings

Sentence Embedding is one of the successful applications of Transformer-based language models. A sentence embedding model encodes a sentence into a multi-

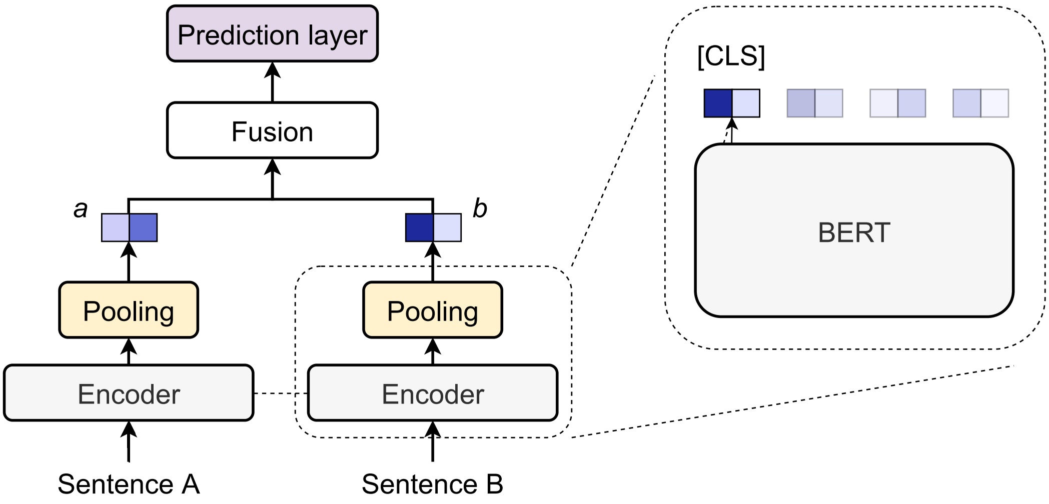


Fig. 2.3 General sentence embedding training architecture (left). Gray boxes are the shared encoder, yellow boxes are pooling layers. Example of BERT as the encoder as in Reimers and Gurevych (2019) and selection of "[CLS]" token as pooling strategy (right). Fusion layer is $(a, b, |a-b|, a*b)$ in Conneau et al. (2017) and $(a, b, |a-b|)$ in Reimers and Gurevych (2019). Prediction layer is a three-way classifier when training on NLI data.

dimensional vector and has a crucial property: the model maps *similar* sentences close in the vector space.

Sentence Embedding models are trained by feeding pairs of sentences. If the two texts are related (i.e., a *positive* pair), the model learns to get the two embeddings closer[9] in the vector space. Conversely, if they are unrelated (i.e., a *negative* pair), the model pushes them apart. Sentence embedders have found application in semantic text similarity, retrieval tasks (e.g., query-product retrieval), topic modeling (Bianchi et al., 2021), clustering, and more.

Figure 2.3 shows a generic sentence embedding architecture at training time. A shared encoder network encodes the sentences, and a fusion strategy combines the resulting embeddings. Notably, most models are trained on NLI data (but not only), such as SNLI (Bowman et al., 2015): the authors construct these datasets explicitly to test semantics.

Seminal work (Conneau et al., 2017) used LSTM or GRU as encoders using as the sentence embedding either the last hidden state of the network or some mean/max pooling over all hidden states.

[9] In sentence embedding training and applications, one commonly uses cosine similarity or Euclidean distance.

After the release of Transformer, Cer et al. (2018) used a stacked transformer encoder to extract sentence embedding by summing element-wise last word representations. Harvesting the representational power of PTLMs, Reimers and Gurevych (2019) train Sentence-BERT (SBERT), a sentence embedding model using pre-trained BERT (Devlin et al., 2019) as the encoder.

In Chapter 3, we fine-tune pre-trained SBERT sentence embedding models for misogyny detection on Twitter. Although we achieve promising results over several baselines, posterior error analysis has highlighted how these embeddings overfit to words or cannot embed complex behaviors such as self-mocking references or quoted speech.

2.2 Social Bias in Language Models

"Bias" has undergone different definitions depending on the field. Therefore, we need to disambiguate the terminology first and provide the reader with clear scope and description for the remainder of the work.

At the intersection between statistics and classic Machine Learning, "bias" is associated with measuring how well a point estimator over observed data (e.g., training data) approximates the true value (Goodfellow et al., 2016). We often tradeoff it with the variance of the estimator.

In this work, instead, we align with the research field and literature on the **social bias** an NLP system may express (Blodgett et al., 2020). With "bias" in NLP, we refer to the case of methods that show harmful behaviors against one or more social categories. The conceptualization of critical concepts in this definition, such as the type of discriminatory behavior identified or targeted groups, differs from paper to paper – sometimes to an inconsistent extent (Blodgett et al., 2020) – and has led to several related efforts in measuring and mitigating bias.

2.2.1 The Social Impact of NLP Systems

Investigating the social impact of NLP systems started long before the rise of LLMs. Hovy and Spruit (2016) discuss the **situatedness of language**, i.e., its property of i) happening in a specific time and place and ii) carrying the individual traits of the speaker: language is, by all means, something that fits each person's essential qualities and habits. Consequently, training language technologies can lead to **demo-**

graphic underrepresentation (e.g., we discussed the specificity of views a PTLMs learns training on data from the internet in Section 2.1.3) or over-generalization, i.e., failing at identifying groups.

More recently, Bender et al. (2021) have highlighted how large-scale training data crawling and cleansing procedures retain hegemonic viewpoints (Section 4 in the paper). Most data is collected from the Internet, whose usage is dominated by users in developed countries,[10] and social networks, accessed mainly by younger people.[11] Further, moderation policies exacerbate demographic unbalance: while reacting to hatred and aggressiveness online, they can censor minorities due to faulty or poorly calibrated automatic moderation tools.[12] Further, several studies on language models (Bolukbasi et al., 2016; Brunet et al., 2019; Dinan et al., 2020a; Kaneko and Bollegala, 2021; Stanovsky et al., 2019, inter alia) have established that models encode societal bias, suggesting that current datasets are from being free from unwanted discourse such as stereotypical speech or harmful content against minorities.

In sum:

1. language is a powerful proxy of one's attitudes and individual traits, and thus language technologies have a strong social impact once deployed;

2. Internet-crawled datasets reflect the views of a narrow part of the world population, mainly younger and from developed countries;

3. PTLMs learn discriminatory biases based on stereotypical misconceptions that are hard to remove from data entirely.

In the following, we review relevant literature on bias identification, evaluation, and mitigation and discuss recent advances in the specific case of language models fine-tuned for hate speech detectors.

Ultimately, we refer the reader to Blodgett et al. (2020) for a thorough survey on "bias" in NLP systems.

[10] https://ourworldindata.org/grapher/correlation-between-internet-users-as-a-share-of-the-population-and-gdp-per-capita

[11] https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

[12] For instance, in Chapter 4, we show that hate-speech classifiers become overly-reliant to the presence of specific features in text and result in misclassifying those mentioning them.

2.2.2 Measuring Bias in Language Models

We organize this section along the distinction of intrinsic and extrinsic bias evaluation in language models (Czarnowska et al., 2021; Goldfarb-Tarrant et al., 2021).

Intrinsic evaluation and metrics study bias encoded in pre-trained representations of a language model. As it is natural from a historical viewpoint, seminal works on the topic focused on bias in static pre-trained word embeddings (e.g., word2vec (Mikolov et al., 2013), fastText (Bojanowski et al., 2017a), or GloVe (Pennington et al., 2014)). More recent work studies bias in Transformer-based large language models used either for sentence embedding (May et al., 2019) or for extracting contextualized word representations. For context, most of the intrinsic tests studied gender bias.

Extrinsic fairness metrics test models for *group equality* in downstream tasks (Czarnowska et al., 2021). These metrics are numerous and capture different facets of social bias. Still, all lie on the same foundation: models should express no evident difference in performance across groups, i.e., **no demographic group should be penalized against others**.

In Goldfarb-Tarrant et al. (2021), extensive empirical evaluation across different models, metrics, and tasks shows no reliable correlation between intrinsic and extrinsic metrics exists. Following the authors, we suggest a line of research that focuses on extrinsic evaluation as it matches closely with production scenarios and propose a regularization technique – introduced in Chapter 4 – that we assess using different extrinsic bias metrics.

2.2.3 Intrinsic Bias Assessment

Seminal work on intrinsic bias in language models focused on representational bias in embedding spaces. Bolukbasi et al. (2016) have found gender bias in word2vec embeddings (Mikolov et al., 2013) trained on Google News articles. Specifically, the authors showed how embeddings encode gender stereotypes in the space geometry: they discovered that the embedding of “doctor” is closer to “man” than “woman”, or, using the analogy framework of Mikolov et al. (2013), that “man” is to “computer programmer” as “woman” is to “homemaker”.

Similar evidence of gender bias was also found in multilingual fastText embeddings (Sabbaghi and Caliskan, 2022) and monolingual contextualized embeddings

(Basta et al., 2019). Gender-biased representations affect also coreference resolution (Zhao et al., 2018), machine translation (Stanovsky et al., 2019), relation extraction (Gaut et al., 2020), dialogue generation (Dinan et al., 2020a), and sentiment analysis (Bhaskaran and Bhallamudi, 2019).

Evaluation Benchmarks

Recently, related research has constructed evaluation benchmarks for intrinsic bias.

Caliskan et al. (2017) introduces the Word-Embedding Association Test (WEAT) to test relationships between sets of words. Testing GloVE embeddings (Pennington et al., 2014) on WEAT shows that female-related words (e.g, "she", "woman") are more associated with arts-related words than science-related ones.

May et al. (2019) extends the WEAT test to several state-of-the-art sentence embedding models. Using templates to turn words into sentences, the Sentence-Embedding Association Test (SEAT) tests for spurious associations between sets.

In Nadeem et al. (2021) the authors collect StereoSet, a large-scale English dataset to evaluate stereotypical bias across four categories: gender, profession, race, and religion. Then, mimicking the MLM pre-training objective, they collect *fill-the-blank* templates specific to demographic groups. Finally, they test if models fill templates with more or less stereotypical alternatives (e.g., in the template *My housekeeper is* ___, the words "Mexican" and "American" should have an equal probability).

2.2.4 Extrinsic Bias Evaluation

Most of the intrinsic biases arise from pre-training on online-crawled training data. However, Steed et al. (2022) introduce and prove the *bias transfer hypothesys*: once models have internalized social biases during pre-training, mitigation techniques have little effect even if applied *before* fine-tuning.[13] Therefore, it is crucial to assess social bias in fine-tuned models.

Recently, we have seen the development of extrinsic fairness metrics. These metrics are computed on the model's prediction (or the scores) and measure performance

[13] In Chapter 4, we present encouraging results opposite to this theory, showing how regularization on how models distribute self-attention leads to reduce lexical overfitting and, in turn, mitigates bias.

separately by demographic groups. Ideally, an unbiased model should show similar performance across all sub-groups.

In the following, we provide a non-comprehensive list of metrics to acquaint the reader with the intuition of *performance-based group parity*. Please refer to Czarnowska et al. (2021) for a complete overview on fairness metrics.

Hardt et al. (2016) introduce the notion of Equality of Odds. A prediction model shows equalized odds on two or more demographic groups if it has equal true-positive and false-positive rates across the groups.

Later, Dixon et al. (2018) introduced the notion of "unintended bias" for a text classifier. Following the authors' definition, a model shows unintended bias "if it performs better for some demographic groups than others". Inspired by Equality of Odds, they introduce two measures of Error Rate Equality Difference:

- **False Positive Equality Difference**: measured as

$$FPED = \sum_{t \in T} |FPR - FPR_t|$$

- **False Negative Equality Difference**: measured as

$$FNED = \sum_{t \in T} |FNR - FNR_t|$$

FPR and FNR are the false-positive and false-negative rates on the entire test set. FPR_t and FNR_t are the same measures computed considering only the samples mentioning a specific demographic group. The author use lists of identity terms (e.g., "woman", "Muslim", etc.) to identify subgroups. In the best case, rates are similar across subgroups (e.g., $FPR \simeq FPR_t, \forall t$).

Similarly, Borkan et al. (2019) designed three Area Under the Curve (AUC)-based metrics to assess subgroup performance. AUC is measured directly on prediction scores and avoids defining a threshold to get prediction labels. **SubgroupAUC** measures AUC limitedly to data mentioning the group. **Background Positive Subgroup Negative (BPSN) AUC** and **Background Negative Subgroup Positive (BNSP) AUC** measure AUC on samples mentioning the group and those from the *background* set, i.e., the rest of the samples. Low BPSN-AUC and BNSP-AUC values indicate the model will likely misclassify subgroup samples as false positives and false negatives, respectively. Finally, the authors propose **Positive Average Equality Gap** and **Negative Average Equality Gap** as threshold-agnostic extensions of FPED and FNED.

We refer the reader to Czarnowska et al. (2021) for a broader perspective on extrinsic bias metrics. In the paper, the authors survey most of the papers on social bias in NLP and propose three generalized bias metric formulations to unify the different metrics found in the literature.

- **Pairwise Comparison Metric** quantifies the difference in performance between two groups of interest;
- **Background Comparison Metric** measures the difference in performance between a group and its *background*. FPED and FNED are BCM metrics since they compare error rates on subgroups with the entire test set.
- **Multi-group Comparison Metric** quantifies bias considering all subgroups of a protected category (e.g., to measure how subgroups within the *Gender* protected category impact performance).

2.2.5 Debiasing Language Models

Debiasing techniques encompass both intrinsic representational bias and extrinsic group parity. Here, we present debiasing techniques based on vector re-embedding, data augmentation, and debiasing-oriented training procedures.

Vector re-embedding

Bolukbasi et al. (2016) propose a debiasing technique to remove gender bias in word embeddings, e.g., making gender-neutral words such as “doctor” equally distant from a set of gendered pairs of words, e.g., “he”-“she”. The proposed approach learns a *gender subspace* (or *direction*) from gender pairs and subtracts it from word embedding to *neutralize* the *gender component*.[14] Similarly, Liang et al. (2020) propose to remove gender bias in sentence embeddings by finding a linear gender subspace, projecting the embedding, and removing the resulting vector from the original one.

[14] The authors call this approach **hard debiasing** and discuss an alternative *soft bias correction* which we do not cover here.

Founded on similar ideas, Ravfogel et al. (2020) propose to train a linear classifier on the word (or sentence representations) to predict a target property (e.g., *Gender*). The idea is the following: if the classifier learns to project embeddings such that they are easily separable for a classification task, its **nullspace** will do the opposite, i.e., it will remove every information useful to identify the property. The authors show how nullspace re-projection mitigates bias in word embeddings.

In a follow-up paper to Bolukbasi et al. (2016), Gonen and Goldberg (2019) criticizes the neutralization approach showing that re-embedded vectors retain bias. These results have opened the discussion: can representational bias be removed by re-projection and linear algebra in the latent space?

Recently, original work (Brunet et al., 2019) proposed to track gender bias back to training documents and remove them from training to mitigate gender bias in word embeddings.

Data Augmentation

We divide data augmentation solutions into class-based resampling and counterfactual augmentation.

Several works have explored the possibility of reducing bias in language models by augmenting training data using resampling. Let us use a hate speech detection setup as an example. One of the causes of the lack of group parity in performance (*extrinsic* evaluation) is a training dataset that contains a polarized view of a demographic group (e.g., a collection of tweets where the topic of homosexuality is mainly associated with hateful tweets). A classifier learning this distribution will likely replicate the polarization found in the data. A way out would be collecting new samples concerning the demographic group in question, but with a different sentiment (in our example, it will be tweets dealing with homosexuality in non-hateful contexts).

Dixon et al. (2018) define a set of identity terms of specific protected attributes – e.g., “Muslim” for *Religion* – and observe that they are skewed toward the positive class (hateful content) in a Wikipedia Talk Pages corpus. They use identity terms to sample additional data from Wikipedia and rebalance the distribution. Models trained on the new corpus show lower unintended bias measured by extrinsic metrics. Similar results (de Vassimon Manela et al., 2021; Nozza et al., 2019; Sharma et al., 2020, *inter alia*) motivate the effectiveness of identity terms rebalancing in training data for bias mitigation.

Counterfactual Data Augmentation (CDA) (Zmigrod et al., 2019) uses syntax parsing, feature engineering, or lists of identity terms to select and modify parts of a sentence to generate counterfactual examples (e.g. from Zmigrod et al. (2019), *Los ingenieros son expertos* (the male engineers are skilled) is turned into *Las ingenieras son expertas* (the female engineers are skilled) as a counter-stereotype). Models further trained on counterfactually augmented datasets have shown less bias in both intrinsic and extrinsic evaluation (Dinan et al., 2020a; Lauscher et al., 2021; Meade et al., 2022; Park et al., 2018; Zmigrod et al., 2019).

Debiasing-Oriented Training Procedure

Other approaches to debiasing do not imply changing data. For example, some tested adversarial training, some others regularization.

Kennedy et al. (2020) compute an importance score for identity terms using post-hoc explanations (see Section 2.3 for an introductory definition). Then, the authors debias a BERT-based text classifier with an additional regularization term to the training loss: the new term minimizes the importance assigned by the model to the identity term. The idea is that reducing the importance of identity terms would also reduce the false positive rate of that group.

Both Dixon et al. (2018) and Kennedy et al. (2020) assume that data collection about protected demographic groups retrieves much more hateful texts than non-hateful. Zhang et al. (2020) formalizes this assumption as a type of **selection bias from the non-discriminatory distribution to the discriminatory one**. In other words, two distributions exist, but training datasets sample only from the discriminatory one. The authors obtain the non-discriminatory loss by re-weighting samples from the observed discriminatory distribution. They compute the weights once using the prior distribution of specific identity terms (e.g., “gay”, “Muslim”) and hence do not impact training.

Finally, debiasing often involves changing model parameters with the risk of catastrophic forgetting (Goodfellow et al., 2013). Lauscher et al. (2021) updates only injected adapter modules (Pfeiffer et al., 2020) while debiasing on a counterfactually augmented corpus, leaving model parameters unchanged. The technique, dubbed ADELE, effectively mitigates gender bias for monolingual and multilingual BERT.

2.2.6 Identity Terms-Free Debiasing

Note that data augmentation and identity term-based regularization require a list of identity terms. However, lists are typically built for a specific domain and language and may be incomplete.

In Chapter 4, we introduce a novel regularization – called Entropy-based Attention Regularization (EAR) – that builds on ideas similar to Dixon et al. (2018) and Kennedy et al. (2020), i.e. that some terms drive the classification outcome, i.e., they induce *lexical overfitting*. However, crucially, **our approach works with no predefined identity terms**. Testing on three datasets in English and Italian, we show how EAR generalizes well to different domains, languages, and tasks (Attanasio et al., 2022b).

We relate lexical overfitting to token contextualization (the lower, the higher it will induce overfitting) measured by the attention entropy. We then propose a novel regularization term to *widen* overall token attention, increasing entropy and mitigating lexical overfitting.

2.3 Explainable AI

EXplainable AI (XAI) is a well-established research today. In this section, we provide the key concepts to get acquainted with the field: we cover the basic notions and motivations and the established taxonomy of XAI techniques. Finally, we close the section by briefly reviewing recent advances in XAI techniques applied to NLP systems.

In other words, why provide basic answers to the following questions:

- *Why* is interpretability needed?
- What does model interpretability *mean*?
- What does *distinguish* interpretability approaches?

The breadth and complexity of the topic make it hard to provide a self-concluded overview. We hence restrict our scope to supervised learning and model interpretability. For the complete picture of the state-of-the-art, we refer the reader to comprehensive surveys such as Guidotti et al. (2019) or Danilevsky et al. (2020) for NLP specifically.

2.3.1 Motivating eXplainable AI

As pioneered by Lipton (2018), the need for interpretability is tightly coupled with a misalignment between the final goal of a supervised learning algorithm and its real-world cost once in production. For example, a supervised model can optimize accuracy, while in production, one would also require a **socially unbiased** model even though it did not directly optimize for it. In all these cases, interpretability can serve as an aid to bridge the two requirements.

More broadly, interpretability can serve:

- **Trust**, in the sense of human trust and confidence in the prediction of a model (e.g., think of the importance of *trustworthy hate speech detection models*);
- **Causality**, i.e., discovering new causal relationships in supervised learning algorithms by looking at explanations (e.g., one might discover that quoted speech is often self-mocking reference and therefore is not necessarily a hateful content);
- **Transferability**, i.e., suggesting how the model can behave if the environment changes, shifting away from the training distribution (e.g., in NLP, this is the case of new slang in social media);
- **Informativeness**, i.e., to provide *additional information* to the model prediction, e.g., by providing analogous cases (Koh and Liang, 2017);
- **Fair Decision-Making**, i.e., providing further proof that models do not make decisions based on social biases (Section 2.2). This goal is paramount whenever individuals are affected by algorithmic decisions and have the *right to an explanation* (Goodman and Flaxman, 2017). Recent works (Pastor et al., 2021; Sagadeeva and Boehm, 2021) *slice* model performance along demographic groups to discover discriminated categories.

2.3.2 On the Notion of Interpretability

Roughly, interpreting a machine learning model entails **producing an *explanation***, an additional artifact that helps human to understand part or all of the computation pipeline. [15] For instance, the model can be inherently interpretable (or self-interpretable) (Guidotti et al., 2019)). Rule-based systems are an example within this category. Otherwise, we might be interested in understanding how it conceptualizes one of the class labels (e.g., how a vision system *represents* the concept of a "horse" (Wu et al., 2020)).

In the following, we focus on the **outcome explanation** (Guidotti et al., 2019; Lipton, 2018) in the context of supervised learning: an explanation is an artifact (e.g., a visualization, some scores, a generated text, or else) that enables humans to understand the rationale behind a given outcome.

2.3.3 A Taxonomy or Interpretability Approaches

Recent literature (Danilevsky et al., 2020; Guidotti et al., 2019) has settled on a two-fold categorization structure.

1. **Local vs. Global Methods** As we said above, we are focusing on the explanation of a model outcome. Interpretability can disclose either the global properties of the model or its behavior for a specific instance.

   Global explanations highlight any general behavior of the model. Wu et al. (2020) explains a vision model globally discovering class-related *concepts* (e.g., black and white stripes related to the class "zebra"). Looking at an NLP case, Pryzant et al. (2018) induce lexicons – which we can loosely relate to concepts – that are predictive of the positive class. Pastor et al. (2021) and Sagadeeva and Boehm (2021) highlight dataset slices where model performance significantly degrades. These global explanations provide insights for model debugging and fairness considerations.

   Global features and behaviors are hard to define and often depend on the domain. These issues have led to limited development of global methods compared to local ones.

[15] We will use the terms *interpretability* and *exaplainability* interchangeably in the remainder of the document. Following related work (Miller, 2019; Molnar, 2020), we consider a model interpretable as soon as one can produce an explanation of its outcome readable and understandable by a human. Hence, the two properties often come together: XAI techniques make models explainable (we have an algorithm to explain the rationale) and interpretable (we interpret the outcome *through* the explanation). Some work also relates interpretable models to white-box algorithms (e.g., Decision Trees, or Rule-based Algorithms). Here, we use the notion of *inherently interpretable models*.

Local explanations provide artifacts that **disclose the rationale behind the prediction** of the specific instance (e.g., why the predicted label is "Positive" for the movie review *I love movies with both pathos and action*). Many local explainability methods have been introduced lately with different underlying ideas and goals in mind. We review the most common types in Section 2.3.4.

2. **Inherently Interpretable vs. Post-Hoc Interpretability**

   Inherently interpretable models (or *self-explaining* model) are by construction human interpretable and does not require specific post-processing after a prediction. In other words, the model itself discloses the rationale. Classic examples are rule-based models or methods that learn sequential splits of the input space, such as decision trees. Attanasio et al. (2020) use this property to characterize the predictions of a quantitative trading system. Other examples are works that infer object properties and concepts from weights and activations of neural layers (Olah et al., 2017, 2018).

   Post-Hoc Interpretability algorithms leverage part of the pipeline (either the dataset, the model itself, an external model, the features of the computational graph such as the gradients, or a combination of those) to produce the explanation artifact. Post-hoc interpretability is versatile (i.e., *model-agnostic*) and highly customizable (i.e., well-suited to modern predictors). For these reasons, it has attracted most of the research efforts in the field.

Figure 2.4 represents graphically the described taxonomy.

2.3.4 Post-Hoc Local Explanation Methods

The surge of post-hoc local explanation methods in recent years mainly involved **feature attribution** – also known as **saliency methods** – and **explanation-by-example** methods. Lately, novel solutions involve text generation and counterfactual edits. We review two of them in Section 2.3.5.

Feature Attribution Methods

Feature attribution entails finding a contribution score to the prediction.[16] We review four of the most influential feature attribution methods: input occlusion, surrogate

[16] It is common to explain the predicted class to understand the rationale behind the prediction. However, one can technically explain other classes.

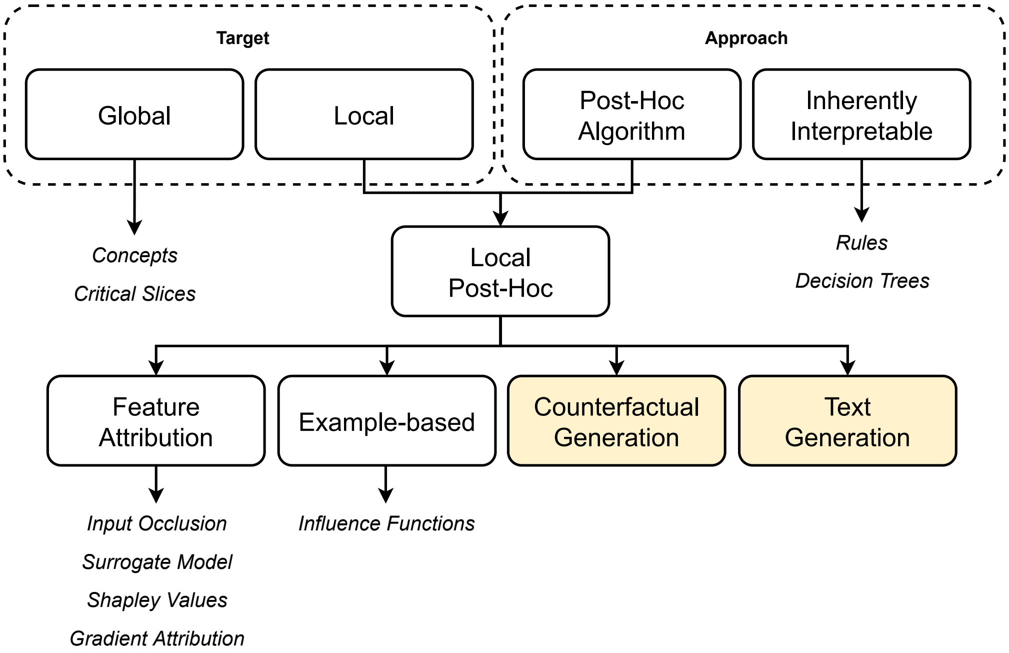


Fig. 2.4 Taxonomy of interpretability approaches introduced in this thesis. Yellow boxes are novel categories for NLP systems. Instances within each category are in *italic*.

models, Shapley Values and gradient attribution, and two explanation-by-example methods.

1. **Input Occlusion.** These methods measure feature importance by removing it and computing the difference in the prediction of the class. The idea is that influential features would cause a large change in the score. Zeiler and Fergus (2014) uses patches to occlude image portions and compute importance. Li et al. (2016) runs a similar analysis by removing one word at a time for a text classifier.

2. **Surrogate Models.** These methods learn a simpler model, a *surrogate*, in the locality of the instance. Ribeiro et al. (2016) sample neighbors using input perturbation and learn a linear model under constrained optimization on the generated neighborhood. The weights of the linear model measure express the explanation in the form of **feature importance scores** (e.g., while explaining a text classifier, each word receives an importance score).

   Relatedly, other works build simpler surrogate models via rule mining (Guidotti et al., 2018; Pastor and Baralis, 2019; Ribeiro et al., 2018).

3. **Shapley Values.** Lundberg and Lee (2017) proposes SHapley Additive exPlanations (SHAP), a novel framework that assigns feature attribution via Shapley Values estimation. Shapley values (Shapley, 1997) come from game theory and allow to estimate the contribution of each *player* to a given result. Here, the result is the prediction, and the players are the input features. The authors demonstrate that SHAP is a general case for many related feature attribution methods such as LIME (Ribeiro et al., 2016).

4. **Gradient Attribution.** Gradient attribution methods measure input contribution directly to the prediction score. Roughly, these methods compute the gradient of the loss (or the logits) with respect to input features. Intuitively, the gradient relates to feature importance as it measures how likely the prediction will change for a small feature variation. Highly influential features would impact the prediction the most. Gradient-based methods have been first applied in Computer Vision (Selvaraju et al., 2020; Simonyan et al., 2014a) and then in NLP (Han et al., 2020; Sanyal and Ren, 2021).

Finally, a relevant line of research questions the reliability of saliency methods (Kindermans et al., 2019; Wang et al., 2020), but the discussion is out of the scope of this thesis, so we leave it to the curious reader.

Explanation-by-example

Explanation-by-examples methods provide the most influential (but not necessarily the most *similar*) samples to the instance to explain. In Koh and Liang (2017), the authors propose a framework to select these samples using **Influence Functions**. Roughly, the algorithm proceeds as follows. First, they compute the loss $\mathscr{L}$ of the model over the sample $\mathbf{x}$. Then, for each training sample, they estimate a new loss $\hat{\mathscr{L}}$ but as if the training dataset did not include the training sample. Training examples that lead to the largest change ($|\hat{\mathscr{L}} - \mathscr{L}|$) are the most influential training examples for $\mathbf{x}$.

Koh and Liang (2017) tested influence functions on Support Vector Machine and Convolutional Neural Networks. Recently, Han et al. (2020) extended the framework to BERT-based classifier.

2.3.5 Novel Approaches for NLP Systems

We summarize some of the novel ideas to explain language models specifically. Some of these approaches build on related ideas such as input perturbation and gradient-based search.

Sanyal and Ren (2021) propose a generalization of Integrated Gradients (IG) (Sundararajan et al., 2017) to word embedding spaces. IG requires to interpolate gradients following a continuous "straight line" from a baseline point in the space and the sample explained. However, the authors propose to interpolate over discrete steps that pick points in the space close to actual word embeddings.

Other works find adversarial examples using input perturbation on either characters (Ebrahimi et al., 2018) or words (Wallace et al., 2019). Even though they do not provide a formal explanation, they are used to improve robustness. However, these methods can shed light on the model's inner workings, such as extreme sensitivity to specific input phrases.

Originally, Rajani et al. (2019) collected human-annotated explanations and used them to train a text generation language model. They then propose Commonsense Auto-Generated Explanation (CAGE) to provide explanations using natural language for the CommonsenseQA task (Talmor et al., 2019).

Another recent line of research explores **counterfactual generation**. In the case of binary text classification, the goal is finding a counterfactual example, i.e., a text similar to the original one but classified differently. In Ross et al. (2021) the authors generate the counterfactuals as follows. First, they train a T5 (Raffel et al., 2020) sequence-to-sequence model using as input any sentence with masked tokens concatenated to the gold label (e.g., "Text: That book was [MASK] Label: Positive") and as output the masked tokens (e.g., "awesome!"). To generate a counterfactual for a given sample, they i) mask the most important tokens using gradient attribution, ii) fill the template with the label opposite to the class to explain, and iii) feed the fine-tuned T5 with the new template. They repeat the steps above using beam search until the predicted class does not change.

2.4 Summary

In this section, we have introduced the background notions propaedeutic to follow along with the remainder of the thesis.

We have provided details on the Transformer architecture and modern deep language models, covering concepts such as self-attention and pre-training objectives (Section 2.1). We have also explained how we currently use language models for sentence embeddings.

We have then introduced the issue of Social Bias in LLMs and how it is related to training data and dynamics (Section 2.2). Further, we have described bias evaluation in intrinsic and extrinsic metrics and listed recent advances in debiasing techniques.

Finally, we have established the notion of eXplainable AI and introduced part of the taxonomy of existing methods (Section 2.3).

Chapter 3

Improving Sentence Embedding Models with Misogyny Lexicons

Misogynous speech in social media is a frequent phenomenon that companies and institutions fight with strict policies and dedicated teams. But size and diversity of posts require automatic strategies to facilitate filtering and moderation. Recent work has proposed using NLP systems for automatic misogyny identification as a countermeasure. In this most straightforward formulation (i.e., without considering contextual information like user demographics, image, audio, or the conversation history), the task entails predicting whether a given text contains misogynous speech.

Standard NLP approaches use a two-step pipeline. First, a text encoding model – or a *sentence embedder* – transforms raw text into an informational rich representation. Classic examples of this encoding are Bag-of-Words (BoW), n-grams, or Term Frequency-Inverse Document Frequency (TF-IDF). Modern encoders span from Recurrent Neural Networks (RNNs) to Transformer-based models (e.g., BERT) (see Section 2.1.3). Second, a classification model uses text encodings to perform the actual task. If embedders produce meaningful representations, those will be easy to separate and classify.

Recent literature (Section 2.1.4) has supported the idea that modern models produce rich embeddings, versatile for many tasks.

In this chapter, we challenge this claim using such embeddings for misogyny identification in Italian tweets. We present a multi-agent classification approach that uses a Sentence Embedding model, TF-IDF vectorization, and hand-crafted Misogyny Lexicons. A first agent uses a BERT-based (Devlin et al., 2019) model to encode tweets and a Support Vector Machine to produce initial labels. A second agent, based on TF-IDF, Misogyny Italian Lexicons, and semantic parsing, contributes a second set of predictions to collaborate with the first agent on uncertain predictions. We evaluate our approach in the Automatic Misogyny Identification Shared Task (Fersini et al., 2020b) of the EVALITA 2020 campaign (Basile et al., 2020). Results show that TF-IDF and lexicons effectively improve the supervised agent trained on sentence embeddings.

Remarkably, error analysis on sentence embedding models has shown brittle generalization capabilities and overfitting to specific terms.

3.1 Motivation

Many public forums exist for people's opinions, such as blogs and social networks. While pursuing free speech is a positive endeavor, it also creates fundamental challenges as not all intentions come for good. In these platforms, where access cannot – and should not – be restricted to anyone, hatred is a critical issue, and protecting minorities based on gender, ethnicity, religion, and sexual orientation is paramount.

Violence against women manifests in social networks whenever the offensive language targets women directly or indirectly (Ellsberg et al., 2005). Although platform owners frequently update regulatory terms – often as a result of renewed misconducts[1] – quantity and diversity of posts pose critical challenges to monitoring systems.

Many recent works in the NLP community show effective results in online monitoring of hate speech (Badjatiya et al., 2017; Fortuna and Nunes, 2018; Waseem and Hovy, 2016) and misogynous content specifically (Anzovino et al., 2018; Frenda et al., 2019; Pamungkas et al., 2020). Furthermore, as a part of the collective effort to build better systems, research communities propose evaluation campaigns (Bosco

[1] https://www.theverge.com/2020/3/5/21166940/twitter-hate-speech-ban-age-disability-disease-dehumanize,
https://www.theverge.com/2020/8/11/21363890/facebook-blackface-antisemitic-stereotypes-ban-misinformation,
https://edition.cnn.com/2020/10/12/tech/facebook-holocaust-denial-hate-speech/index.html,
https://www.theguardian.com/technology/2020/jun/29/reddit-the-donald-twitch-social-media-hate-speech

et al., 2018) presenting challenging shared tasks (Basile et al., 2019; Fersini et al., 2018, 2020b) and new datasets.

The Automatic Misogyny Identification (AMI) shared task (Fersini et al., 2020b) proposed at EVALITA 2020 focuses on the automatic identification of misogynous content on Twitter in Italian. The challenge has two subtasks.

- Subtask A (Misogyny and Aggressive Behaviour Identification) entails identifying misogynous speech in tweets, i.e., it is a binary classification task. In the case of misogyny, it also requires predicting if the tweet has an aggressive tone.
- Subtask B (Unbiased Misogyny Identification) entails classifying misogynous speech while guaranteeing the model's fairness on a synthetic dataset.[2] The task authors measure fairness under extrinsic bias metrics.

The task admits *constrained* and *unconstrained* submissions. The former allows the system to train only on the provided data, the latter on additional external data.

In the following, we describe our solution to the AMI task and the most interesting insights from our error analysis.

3.2 Methodology

We adopt a multi-agent classification procedure to address each proposed subtask.

We build a first agent as follows. First, we encode tweets to their sentence embeddings using a pre-trained multi-lingual sentence encoder. Next, we train a Support Vector Machine (SVM) on the latent embedding space.

Similarly, we build a second agent from two different text representations: 1) the smoothed TF-IDF of the tweet; 2) a set of features extracted from semantic parsing and misogynous lexicons. We feed this representation to an additional SVM.

Finally, we propose a classification schema that substitutes uncertain predictions from the first agent (sentence embedding and SVM) with certain ones from the second agent. Figure 3.1 shows an overview of the system.

[2] The task organizer provided pre-defined train and test sets.

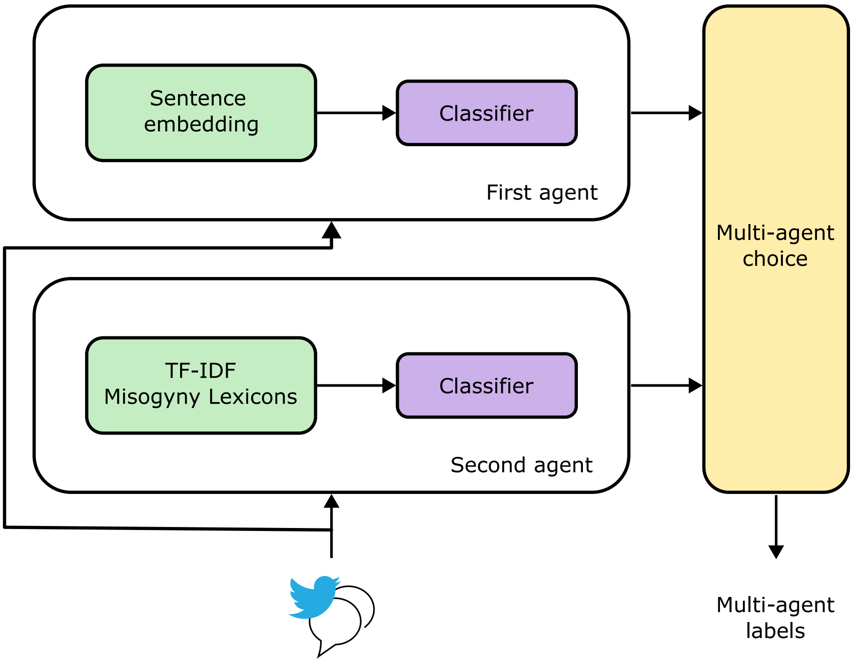


Fig. 3.1 Multi-agent system overview. Color encodes functionality. Both agents extract a dense representation using an encoder (green) from raw tweets and use it to train a supervised classifier (purple). A multi-agent decision module (yellow) pools labels from the two agents and produces a final classification label.

The following paragraphs describe the data preprocessing step, expand the description of the classification system, and provide insights into its application to subtasks A and B.

3.2.1 Sentence Embedding Models

We build sentence embeddings using two models.

We test the monolingual BERT-based model introduced by Aluru et al. (2020). The authors fine-tuned it from a multilingual BERT on an Italian corpus for hate-speech detection tasks. We fine-tune Aluru et al. (2020) to our specific subtasks.

Second, we use a multilingual Sentence-BERT (Reimers and Gurevych, 2019) embedding model.[3] Since results for the monolingual BERT were not encouraging from the beginning, in both the subtasks, we will focus the discussion on multilingual Sentence-BERT.

[3] We use the implementation found in https://github.com/UKPLab/sentence-transformers

Lexicon	Number of Words	Type of Words
Sexist	138	Misogynous and sexist
Profanity	4	Vulgar and swear
Sexuality	7	Sexual references
Female body	6	Feminine body

Table 3.1 Statistics on lexicons used in this study.

The final agent is then a supervised classifier trained on multilingual sentence embeddings (referred to as the *SE* agent). We use a Support Vector Machine (SVM) with a Radial Basis Function kernel, which achieves the best results on our validation set. Please refer to Section 3.3 for more details on parameter configuration and performance.

3.2.2 TF-IDF and Misogyny Lexicons

We build a second classification agent on TF-IDF and additional features extracted from misogyny lexicons and semantic parsing.

As for the first agent, we use an SVM. We train the model by concatenating all the features extracted from text. We refer the reader to Section 3.3 for details on the experimental setting.

We replace every URL found in tweets with the token *LINK*. Next, we tokenize and lemmatize using a pre-trained Italian model in spaCy.[4] Finally, we vectorize the pre-processed tweet using smoothed TF-IDF.

Features from Lexicons

Misogynous tweets often contain sexist attacks, swear words, or sexual references. We include specific lexicons as input features for dealing with hate and misogynous speech (Frenda et al., 2018). We proceed as follows.

We collect Italian lexicons from multiple online sources. We divide them into four categories: sexists, profanity, sexuality, and female body as described. We report statistics in Table 3.1. The complete list and sources are available at our repository.[5]

[4] https://spacy.io/models/it#it_core_news_lg

[5] https://github.com/g8a9/ami20-improving-embedding

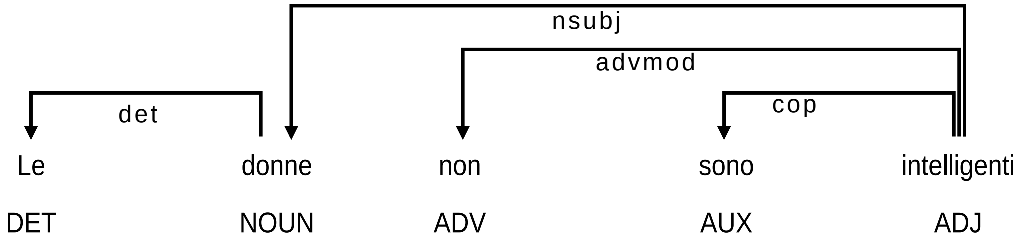


Fig. 3.2 Example of dependency-based parse tree with sentiment polarity inversion. Sentence: "Le donne non sono intelligenti" (Eng: *Women are not smart*).

We lemmatize lexicons using spaCy and derive four integer features, one per category. Each feature will show the count of words belonging to the category found in the tweet.

Features from Semantic Parsing

We use a sentiment lexicon to characterize the polarity of tweets. We use the OpeNER Italian Sentiment Lexicon (Russo et al., 2016) to assign a sentiment to words. This sentiment lexicon consists of 24.293 lexical entries annotated with positive, negative, and neutral polarity. In our analysis, we consider only positive and negative polarity.

However, the model attributes polarity without considering the word's context. We augment our polarity detection using a simple strategy: search the parse tree to consider negation. Specifically, we search for words affected by negation. For these words, we invert the polarity if available.

For example, consider the phrase "le donne non sono intelligenti" (eng: *women are not intelligent*). Figure 3.2 shows the extracted parse tree. The polarity of the word "intelligenti" (intelligent) is inverted, from positive to negative, since it is affected by negation.

We represent the overall tweet polarity with two features, one counting the positive words and one the negative ones. Finally, we normalize the polarity counts by the number of words in the tweet.

Additional Features

Tweets may contain quotations of misogynous content without being misogynous themselves. We hence consider as an additional feature the relative frequency of quotation marks. We also include the tweet's length.

3.2.3 Multi-Agent Prediction

We propose a multi-agent system that maximizes the prediction confidence. Specifically, we deem a prediction confident if its associated probability score is above a given threshold.

We produce the final classification label by combining the outcomes of the two agents. We score the tweet using the sentence embedding agent. If the prediction is not confident (i.e., above a set threshold), we probe the second agent. If the latter has a confident prediction, we use it. Otherwise, we fall back to the first agent output.

We applied the multi-agent classification procedure for both subtasks.

Subtask A: Misogyny and Aggressiveness Detection

In this subtask, participants have to assign a label indicating whether a tweet is misogynous or not. Then, limited to the misogynous ones, a second label should tell if the tweet is also aggressive.

We apply our multi-agent classification in a chained fashion. Specifically, we train a first instance of the system on the binary misogyny problem and label every tweet. In this step, we use the complete corpus.

Next, we train a second instance on the binary aggressiveness problem. Here, we score only the tweets that the first instance deemed misogynous. Finally, we label all the (predicted) non-misogynous tweets as non-aggressive.

This strategy presents advantages and drawbacks since the predictions are chained. On the one hand, the two models are independent and can learn a simpler problem separately. On the other hand, this design propagates errors from the misogyny task to the aggressiveness one. We discuss the issue in Section 3.3.

Subtask B: Unbiased Misogyny Identification

We apply our multi-agent model (*SE*+*Lex* agents) with no modifications. We use both training sets (real and synthetic) to let the model learn the structure of synthetic templates.

Run	Misogyny ↑	Aggressiveness ↑	Task Score ↑
SE	76.88	**59.31**	68.10
Lex	72.22	57.24	64.73
SE+Lex	**77.50**	59.20	**68.35**
Avg.	75.53	58.58	-

Table 3.2 F1 score (macro) for subtask A (test set).

3.3 Experimental setting

We held a validation set (20%) from the training set using stratified sampling on the misogyny and aggressiveness labels.

We tested manually different classifiers: Support Vector Machines, Feed-Forward Neural Network, Random Forest, and Logistic Regression. In addition, we used the F1 score (macro) of the *SE* agent as a reference. SVM with RBF kernel with *gamma="scale"* and *C=10* achieved highest performance. We used this configuration for the supervised classifier of the second agent.

For the TF-IDF, we tuned the n-grams from *n=1* to *n=3*, and the number of maximum tokens from 5.000 to 10.000. We achieved the highest F1 score with unigrams and 10.000 tokens as maximum vocabulary size.

Finally, we optimized the confidence threshold value in $[0.6, 0.95]$ with steps of 0.05 and selected 0.9.

3.3.1 Misogyny and Aggressiveness Identification

Table 3.2 reports the performance on subtask A (test set).

Our multi-agent system achieved a 77.50 F1 on misogyny identification but a much worse 59.20 on aggressiveness identification (note that the final score for subtask A is the arithmetic mean between *misogyny* and *aggressiveness*). We track this subpar performance on aggressiveness detection back to our training choices. Specifically:

1. we trained the aggressiveness detector on misogynous tweets only. This subset is smaller and has a heavy class imbalance, with a prevalence of aggressive tweets. We did not rebalance the dataset nor used non-misogynous tweets, leading to a skewed aggressiveness detector;

Run	**Test ↑**	**Synthetic ↑**
SE	76.93	85.58
Lex	72.41	65.39
SE+Lex	**77.46**	**85.92**

Table 3.3 F1 score (macro) for subtask B (test and synthetic sets).

2. relatedly, non-misogynous tweets are out-of-distribution samples for the aggressiveness detector. The issue arises whenever the misogyny classifier produces a false positive;

3. we based our hyperparameter validation on the misogyny identification performance. We expect other choices could improve the aggressiveness classifier.

Notably, the number of false negatives in misogyny identification is low (16 out of 365 total errors). Further, while the strategy led to subpar performance in subtask A, it resulted in the best system in the *constrained* for subtask B (see Section 3.3.2).

TF-IDF and features extracted from lexicons and semantic parsing (*Lex*) achieve the worst results. On the other hand, sentence embeddings alone (*SE*) improve it on average by two F1 points, reinforcing the better expressiveness of dense vector representations.

Our multi-agent system (*SE+Lex*) achieves our highest result (68.35). This result shows that TF-IDF and features from lexicons and semantic parsing effectively improve plain sentence embeddings.

Official Ranking. Our system ranked 12th out of 20 teams, considering *all* submissions, and 7th considering only the *constrained* ones.

3.3.2 Unbiased Misogyny Classification

The score for subtask B is the weighted combination between AUC on the test set and three AUC-based extrinsic bias metrics on the synthetic set. We refer the reader to Fersini et al. (2020b) for the complete description of the evaluation metrics.

Table 3.2 reports classification performance on subtask B. Similarly to subtask A, *SE* outperforms *Lex*. In addition, our multi-agent *SE+Lex* achieves best performance, further motivating our approach.

Rank	Team	Task Score ↑
1	Jigsaw (unconstrained)	88.26
2	***SE+Lex* (constrained)**	**81.80**
3	*SE* (constrained)	81.37
6	*Lex* (constrained)	69.40
11	MDD	60.13

Table 3.4 Partial leaderboard for subtask B. Our placement is in bold. Unconstrained runs use additional training data.

Official Ranking. Our system ranked 2nd out of 11 teams, considering *all* submissions, and 1st considering only the *constrained* ones. We report the leaderboard in Table 3.4.

3.4 Related Work

Several approaches to the automatic misogyny identification task show two traits in common with our system, i.e., (i) some form of encoding of the text followed by a supervised classifier, and (ii) the use of ensemble systems and pooling of different predictions. In the following, we show related encoding and techniques.

Lees et al. (2020) use an ensemble of BERT-based classifiers (Devlin et al., 2019) and majority voting on the predictions. The proposed system achieved state-of-the-art performance on both subtask A and B.[6] In contrast, we use BERT-based sentence embedding models to encode tweets: our worse results suggest that pre-trained sentence encoders provide less meaningful representations in this particular task. Lees et al. (2020) also address the issue of bias mitigation explicitly, leading to the best result in subtask B (Unbiased Misogyny Identification). The authors collect a set of Wikipedia articles that mention the identity terms provided by the task and include them in the training data as new non-misogynous, non-aggressive data points.

Proving valuable representation extraction from static word embeddings, Fabrizi (2020) achieved state-of-the-art results on a constrained setup, i.e., using only the provided training dataset. The author used Word2Vec embeddings trained on an Italian corpus (Cimino et al., 2018) and one-dimensional convolutional filters to encode them. Similar approaches leverage FastText word embeddings (Bojanowski et al., 2017b) to extract base word representations.

[6] Please note that both the results were achieved in an unconstrained setup, meaning that the authors leveraged additional data to train their system.

3.5 Error Analysis

In subtask B, *SE+Lex* predicts 72 false negatives and 157 (x2.2) false positives. We run a posterior error analysis to dive into the positive misclassifications. We describe four recurrent types of error.[7]

- **Body parts.** Our system misclassifies tweets containing body parts that can have a sexual reference based on the context. These words polarize the assignment to the misogynous class. For example, 15% of false positives contain the word "gola" (throat). This behavior somewhat mimics the bias of models towards specific identity terms.
- **Self-mocking reference.** Another category hard to model is self-referencing text containing misogynous speech. Although the tone of these tweets is primarily ironic, the model decontextualizes and predicts the positive class.
- **Targeted gender.** In these tweets, the model correctly detects the hateful tone of voice but fails to identify the gender of the target. Mostly, our system predicts tweets attacking males as misogynous. The problem is even more challenging when the gender derives from prior knowledge (e.g., some aggressive tweets mention *@bonucci_leo19*, a male Italian football player).
- **Reported speech.** Reported misogynous speech provides another complex scenario. Frequently, users quote an unpleasant, misogynous passage while trying to support precisely the opposite. It can happen directly using quotation marks or indirectly by citing the original speaker. In both cases, our model fails to recognize the reported speech.

Our analysis highlights that sentence embedding models hardly generalize to these types of language. In the following, we motivate why.

First, the training data might suffer from selection bias on specific body parts, i.e., the word is frequently associated with the positive class. Consequently, the model does not contextualize how the term is used within its context and reinforces the spurious correlation.

[7] We provide a list of tweets for each category at https://github.com/g8a9/ami20-improving-embedding

Self-mocking references and reported speech are types of language peculiar to social media. Here, sentence embedding models fail to encode irony from the text and the grammatical and linguistic patterns that characterize reported speech. Again, it is likely that the embedding is learning a decontextualized version of the passage.

Finally, the text often does not expose the gender of the targeted subject. Ideally, we would accept a system that i) identifies this scenario and ii) predicts a neutral class label, e.g., "Unknown," but we are limited to binary classification for the task. In some tweets, inferring gender requires some prior knowledge. However, general-purpose sentence embedding – and their underlying language models – have limited ontological knowledge about the world. This issue reflects in misogyny detection: they fail to understand whether the attacker is targeting a woman.

We argue the model has likely learned a precise relationship between words and labels that influences the prediction **regardless of the context**.

3.6 Conclusion

In this chapter, we have presented a multi-agent solution for misogyny and aggressiveness detection. The system employs jointly pre-trained SBERT embeddings, TF-IDF, and features from misogyny lexicons and semantic parsing.

We evaluated our system in the AMI shared task (Fersini et al., 2020b) at the EVALITA 2020 evaluation campaign (Basile et al., 2020), achieving encouraging results compared to *constrained* solutions in both subtasks.

Our contribution is two-folded. First, we observed weak lexical generalization capabilities of modern sentence embedding models through error analysis on the test set. We relate this behavior with the same underlying issue: sentence embedding models fail to contextualize words and passages. This behavior is strictly related to our first research question (**RQ1**): there are, indeed, specific words or phrases that steer the output with their *lexical presence* showing the brittle beyond-text generalization capabilities of these models.

Second, our experimental analysis has shown that hand-crafted misogyny lexicons and semantic parsing rules bring valuable representational information to the system, effectively improving state-of-the-art transformer-based sentence encoders. This is valuable to many practitioners who use today's off-the-shelf language models.

It is now an excellent time to recall our line of thought. Across the chapter, we discussed sentence embeddings based on BERT (Devlin et al., 2019) and error analysis has shown that embeddings i) become oversensitive to trigger words and ii) fail to generalize to particular types of language. We are then left with an open question: **how do these considerations translate to a generic Transfomer-based language model?**

In the following chapter, we will describe the tight interplay between the core learning paradigm of language models, i.e., self-attention, contextualization, and lexical overfitting. Then, using hate speech detection as a running example, we show how to identify overfitting words and how to **regularize** them. Additionally, we show how to achieve this with no *a-priori* list of words – e.g., lexicons: the proposed approach extracts biased terms as part of the training procedure.

Chapter 4

Entropy-based Attention Regularization for Unintended Bias Mitigation

Natural Language Processing (NLP) models risk overfitting to specific terms in the training data, thereby reducing their performance, fairness, and generalizability. E.g., neural hate speech detection models are strongly influenced by identity terms like *gay*, or *women*, resulting in false positives, severe unintended bias, and lower performance. Most mitigation techniques use lists of identity terms or samples from the target domain during training. However, this approach requires a-priori knowledge and introduces further bias if important terms are neglected. Instead, we propose a knowledge-free Entropy-based Attention Regularization (EAR) to discourage overfitting to training-specific terms. An additional objective function penalizes tokens with low self-attention entropy. We fine-tune BERT via EAR: the resulting model matches or exceeds state-of-the-art performance for hate speech classification and bias metrics on three benchmark corpora in English and Italian. EAR also reveals overfitting terms, i.e., terms most likely to induce bias, to help identify their effect on the model, task, and predictions.

4.1 Motivation

Online hate speech is growing at a rapid pace, with effects that can result in dangerous criminal acts offline. Due to its verbal nature, various Natural Language Processing

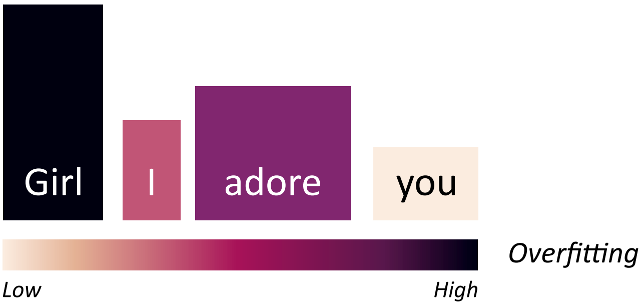


Fig. 4.1 False positive from BERT as a hate speech detector. The darker and taller the bar, the higher the overfitting on the term.

approaches have been proposed (Attanasio and Pastor, 2020; Indurthi et al., 2019; Kennedy et al., 2020; Qian et al., 2018; Vidgen et al., 2021, *inter alia*). Recently, detection performance has significantly improved with the use of large pre-trained language models based on Transformers (Vaswani et al., 2017a), such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). However, several works have shown that by fine-tuning neural language models on hate speech detection, the classifiers obtained contain severe *unintended bias* (Dixon et al., 2018), i.e. they perform better or worse when texts mention specific *identity terms* (such as *gay*, *Muslim*, or *woman*). As a result, a sentence like "As a Muslim woman, I agree" would be wrongly classified as hate speech, purely due to the presence of two identity terms, i.e., terms referring to specific groups based on their socio-demographic features. One cause of false positives is selection bias in the keyword-driven collection of corpora (Ousidhoum et al., 2020). Figure 4.1 shows a false positive example for a fine-tuned BERT model on hate speech detection. Ideally, the model should rely on the words *adore* and *you*. Instead, BERT overfitted to the word *Girl* and associated it with a hateful context. This unwanted effect demonstrates the issues of lexical overfitting, and how they cause unintended bias on identity terms.

Various methods have been proposed to mitigate and measure (unintended) bias (Dixon et al., 2018; Elazar and Goldberg, 2018; Kennedy et al., 2020; Nozza et al., 2019; Park et al., 2018; Vaidya et al., 2020). However, all those methods rely on the availability of a set of *identity terms*. This is a severe limitation, which hinders the generalizability and applicability of hate detection models to real-world contexts. For example, a model designed to reduce the unintended bias on gender-related terms (such as *woman*, *wife*) will not address unintended bias for religious affiliation.

So practitioners must decide a-priori *“which vulnerable groups are present in our data?”*

We propose an Entropy-based Attention Regularization (EAR) that forces the model to build token representations by attending to a wider context, i.e., consider a larger number of tokens from the rest of the sentence. We measure the attended context as the entropy of the self-attention weight distribution over the input sequence. We use EAR as a regularization term in the loss computation to maximize each token’s entropy. We apply EAR to BERT. The resulting model (BERT+EAR) significantly improves performance on unintended bias mitigation in English and Italian. In addition, it requires no a-priori knowledge (e.g., sets of identity terms), making it fairer and more general. The contextualized representations EAR induces avoid basing the classification on individual terms and, ultimately, mitigate lexical overfitting and intrinsic bias from pre-trained weights.

As a training by-product, EAR lets us extract the overfitting terms, i.e., terms accounting for narrower context that most likely induce unintended bias. These terms can highlight possible weaknesses in the model: from the over-sensitivity of pre-trained weights to specific words (Nangia et al., 2020; Sheng et al., 2019; Vig et al., 2020), to over-specialization of training corpora on the keywords used for collecting data (Ousidhoum et al., 2020).

Note that while we show results on BERT, EAR applies to any attention-based architecture.

4.2 Entropy-based Attention Regularization

Attention was originally designed for aligning target and source sequences in machine translation (Bahdanau et al., 2015; Graves, 2013). However, in the Transformer architecture (Vaswani et al., 2017a), it has become a means to account for lexical influence and long-range dependencies. It also provides useful information about the importance of a term for the output (Brunner et al., 2020; Sun and Marasović, 2021; Wiegreffe and Pinter, 2019). Here, we use the notion of attention entropy, and EAR’s use of it in BERT. Note, though, that EAR can be used with *any* attention-based architecture.

Attention entropy. Information *entropy* was first introduced in Shannon (1948), and measures the average information content of a random variable X with the set

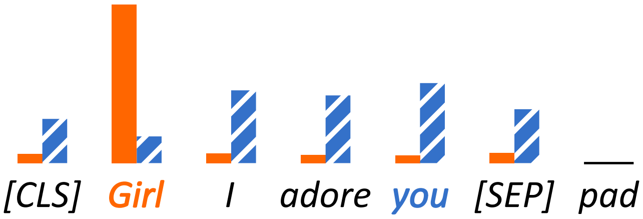


Fig. 4.2 Self-attention distribution on tokens *Girl* (solid orange) and *you* (shaded blue). Attention for *Girl* is concentrated on its representation: its entropy is low. Attention for *you* is spread: its entropy is high.

$[x_0, \ldots, x_n]$ of possible outcomes. It is defined as

$$H(X) = -\sum_i P(x_i) \log P(x_i) \tag{4.1}$$

Following Ghader and Monz (2017), we compute the entropy in the self-attention heads by interpreting each token's attention distribution as a probability mass function of a discrete random variable. The input embeddings are the possible outcomes, and the attention weights their probability.

For the sake of simplicity, we now discuss the computation of attention entropy of a single token in a standard transformer encoder. Attention weights are first averaged over heads by defining $a'_{i,j} = \frac{1}{h} \sum_h a_{h,i,j}$ as the mean attention that the token at position i pays to the token at position j. Then, we define a probability mass function by applying a softmax operator:

$$a_{i,j} = \frac{e^{a'_{i,j}}}{\sum_j e^{a'_{i,j}}} \tag{4.2}$$

We define the attention entropy as follows

$$H_i = -\sum_{j=0}^{d_s} a_{i,j} \log a_{i,j} \tag{4.3}$$

Intuitively, attention entropy measures the degree of contextualization while constructing the model's upper level's embedding. A large entropy suggests that a wider context contributes to the new embedding, while a small entropy tells the opposite: only a few tokens are deemed relevant. From a broader viewpoint, contextualized tokens improve the information passage between continuous layers by re-distributing the information content for every unit involved.

Figure 4.2 shows a toy example of self-attention distributions for two arbitrary tokens. Solid orange bars correspond to $a_{\text{Girl},j}$, while shaded blue bars correspond to $a_{\text{you},j}$. The toy example illustrates the correlation between attention distributions and entropy. The representation of *you* uses a wider context and, thus, it has a higher attention entropy. Note that, if present, we discard padding tokens from the attention entropy computation. Conversely, we include special tokens when required by the downstream task.

EAR in BERT. We introduced attention entropy as a proxy for the degree of contextualization of token representations above. Following this intuition, we propose BERT with EAR mitigation (BERT+EAR), a novel model trained to learn tokens with maximal self-attention entropy over the input sequence. We fine-tune BERT+EAR in the downstream task of hate speech detection. Note, though, that the approach is feasible for any classification task. In classification models, having more contextualized tokens avoids individual terms driving the classification outcome because they got over-attentioned.

Although EAR is applicable to any Transformer-based model, we base our approach here on the BERT (Devlin et al., 2019) base architecture. BERT provides an informative case study, given the number of architectures it has spawned and the recent interest in its attention patterns (Clark et al., 2019b; Kovaleva et al., 2019; Serrano and Smith, 2019). BERT consists of twelve stacked transformer encoders, each running self-attention on the output of the previous encoder. In BERT+EAR, we build new tokens with the maximal information content coming from the previous layer for every transformer layer in the architecture. Using Equation 4.3, we first compute the attention entropy of each token in the input sentence. We then take their mean and define the *average contextualization* for the ℓ*-th* layer as

$$H^{\ell} = \frac{1}{d_s} \sum_{i=0}^{d_s} H_i^{\ell} \tag{4.4}$$

where H_i^{ℓ} is the attention entropy of the token at position i, and d_s is the length of the input sequence (excluding the padding tokens but including the [CLS] and [SEP] special tokens). Finally, we introduce a new regularization term to the model loss to

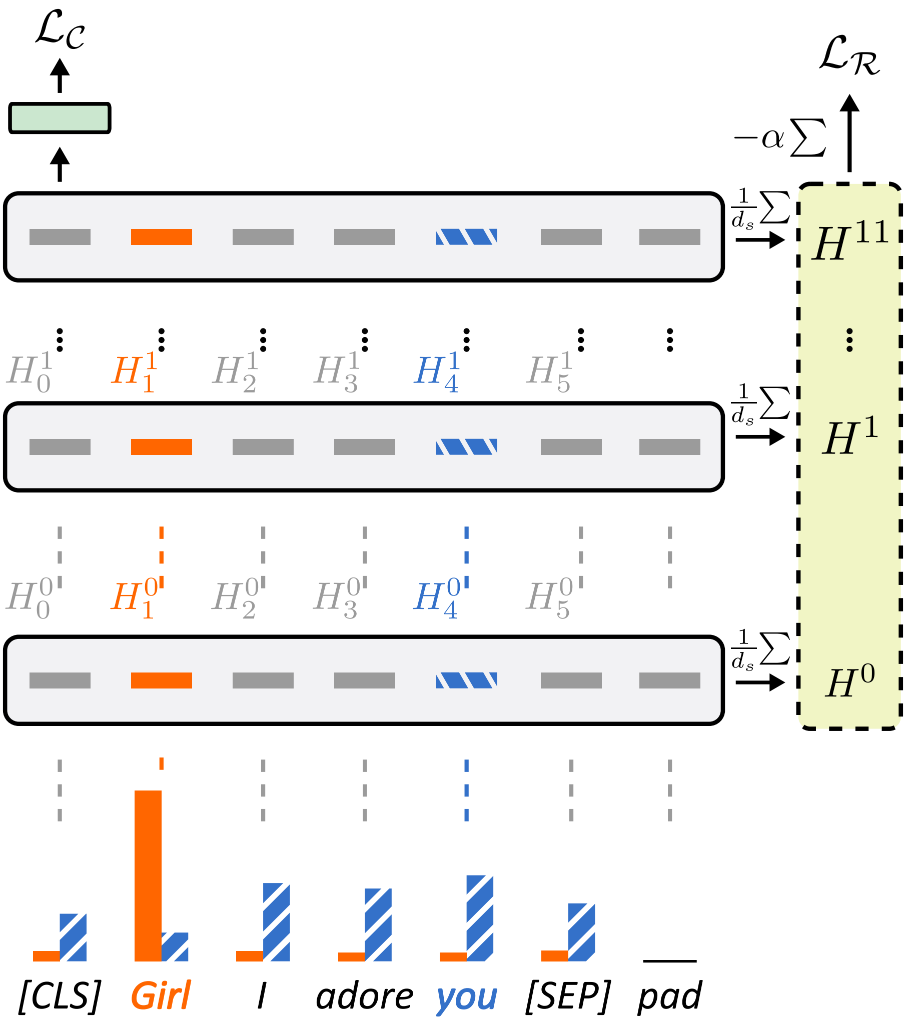

$\mathcal{L}_{\mathcal{C}}$
$\mathcal{L}_{\mathcal{R}}$
$-\alpha\sum$
$\frac{1}{d_s}\sum$
H^{11}
H^{1}
H^{0}
H_0^1 H_1^1 H_2^1 H_3^1 H_4^1 H_5^1
H_0^0 H_1^0 H_2^0 H_3^0 H_4^0 H_5^0
[CLS] Girl I adore you [SEP] pad

Fig. 4.3 Overview of BERT+EAR. Grey boxes are Transformer encoder layers. Each builds a token with attention entropy H_i^ℓ. Right green box pools layer-wise contextualization contributions and outputs regularization loss. First layer self-attention distribution (bottom) shown for “you” (shaded blue) and “Girl” (solid orange).

maximize the entropy at each layer:

$$\mathscr{L} = \mathscr{L}_C + \mathscr{L}_R, \quad \mathscr{L}_R = -\alpha \sum_l H^\ell \tag{4.5}$$

$\mathscr{L}$ is the total loss, $\mathscr{L}_C$ and $\mathscr{L}_R$ are the classification and regularization loss, respectively, and $\alpha \in \mathbb{R}$ is the regularization strength. As in previous work, $\mathscr{L}_C$ is the Cross Entropy loss obtained with a linear layer on top of the last encoder as a classification head. It receives the [CLS] embedding and outputs the probability of the positive class (Hate).

The new regularization term $\mathscr{L}_R$ frames the task of maximal contextualization learning in the network. This framing has several advantages over existing approaches. First, it is a sum of differentiable terms and is hence differentiable. We can thus optimize BERT+EAR with classical back-propagation updates. Second, the regularization is agnostic to specific identity terms. It instead induces the network to learn contextualized tokens globally. This induction is crucial to regularize biased terms that might not be known in advance. Finally, note that the $\mathscr{L}_R$ pools each layer's entropy-based contributions H^ℓ. Each term H^ℓ is in turn dependent on the sole attention entropy defined in Equation 4.3. This makes the setup a general framework not limited to BERT. $\mathscr{L}_R$ can be used to evaluate and maximize the token contextualization in any attention-based architecture.

Figure 4.3 shows a graphical overview of BERT+EAR. Each layer provides a contextualization contributing to the loss independently, where layers with a low average contextualization increase the loss the most. Note also that, similarly to He et al. (2016), $\mathscr{L}_R$ introduces skip connections between layers and the classification head, so shorter paths for the contextualization information to flow.

Insights from attention entropy. On the one hand, we use attention entropy maximization to train BERT+EAR and test its classification and bias mitigation performance. On the other hand, we can leverage attention entropy to automatically extract the tokens with the lowest contextualization, which are the most likely to induce unintended bias. When a sentence is fed through a model like BERT, we can inspect the attention distribution of its terms[1].

We propose to exploit entropy, and hence contextualization, to gain insights into any attention-based model. Given a corpus and a model we want to inspect, we repeatedly query the model with sentences from the corpus and collect each token's attention entropy. Finally, we take each token's mean to measure the impact it has on bias, where lower is worse. Note that the same term can impact bias differently depending on the sentence.

While our approach works for any attention-based model and data set, we test it on fine-tuned classifiers to extract the biased terms learned on the training data set. We discuss this functionality in Section 4.5.

[1] For complex terms, we average the attention entropy of their sub-words.

4.3 Experimental settings

Here, we consider the problem of *unintended bias* (Dixon et al., 2018): "*a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others*".

Datasets. Unintended bias is measured on synthetic test sets, artificially generated by filling manually defined contexts with identity terms (e.g., *I hate all* ___, *I love all* ___) . By construction, each identity term appears 50% of the time in hateful contexts and 50% in non-hateful ones. If a model then classifies the instances related to one identity term differently than the others, it means that the model contains unintended bias towards that term, e.g., if every instance containing the term *women* is labelled hateful, independently of the context. Synthetic test sets simulate new data, so a model that has low performance on this set demonstrates low generalization abilities and incapacity to be used in real-world contexts and applications.

We test BERT+EAR on hate speech datasets with associated synthetic test sets to measure unintended bias.

MISOGYNY (EN) (Fersini et al., 2018) is a state-of-the-art corpus for misogyny detection in English. The related synthetic test set (Nozza et al., 2019) was created via several manually defined templates and synonyms for "woman" as identity terms.

MISOGYNY (ITA) (Fersini et al., 2020b) is the benchmark corpus for misogyny detection in Italian. The synthetic test set has been generated similarly to the English one. This dataset allows us to study EAR's impact on cross-lingual adaptation.

MULTILINGUAL AND MULTI-ASPECT HATE SPEECH (MLMA) (Ousidhoum et al., 2019) consists of tweets with various hate speech targets. We choose to work on its English part. We use the synthetic test provided in Dixon et al. (2018), generated by slotting a wide range of identity terms into manually defined templates.

Table 4.1 reports statistics of the data sets. Alongside the size of train, test, and validation sets, we report also the percentage of hateful instances to show the class balance. Note that MLMA is highly unbalanced with 88% of instances associated with the hateful class. Note that the original MULTILINGUAL AND MULTI-ASPECT dataset comes in a multi-label, multiple class setting. Following Ousidhoum et al. (2021), we used the *Hostility* dimension of the dataset as target label and created a

	MISOGYNY (EN)	MISOGYNY (IT)	MLMA
# Train	4,000	5,000	5082
# Test	1,000	1,000	565
% Validation	10	10	10
% Hate (train, test)	45, 46	47, 53	88, 88
B_2	0.858	0.852	0.881
# Synthetic	1,464	1,908	77,000
# Identity terms	12	18	50
% Hate (Synthetic)	50	50	50

Table 4.1 Statistics of the data sets.

Hate binary from it as follows. We considered single-labeled "Normal" instances to be non-hate/non-toxic and all the other instances to be toxic.

To further characterize our data sets, we explore the aspect of selection bias, reporting the measure B_2 (Ousidhoum et al., 2020). The metric ranges from 0 to 1 and evaluates how likely topics of the data set are to contain keywords of the data collection. Values above 0.7 demonstrate high selection bias, implying the need for unbiasing procedures.

We report also the size and number of identity terms used in the synthetic test sets. The percentage of hateful content is perfectly balanced (50%) since each identity term should appear exactly in the same context as the others to measure the unintended bias. See Appendix A for the list of identity terms and further preprocessing details.

4.3.1 Metrics

We use the weighted and binary F1-score of the hateful class ($\mathbf{F1_w}$ and $\mathbf{F1_{hate}}$) as classification metrics. We consider both due to the class imbalance of test sets (see Table 4.1).

We compute the unintended bias metrics from Dixon et al. (2018) and Borkan et al. (2019). They are computed from differences in the score distributions between instances mentioning a specific identity-term (*subgroup distribution*) and the rest (*background distribution*). The three per-term AUC-based bias scores are:

1) $AUC_{subgroup}$ calculates AUC only on the data subset of a given identity term. A low value means the model performs poorly in distinguishing between hateful and non-hateful comments that mention the identity term.

2) *Background Positive Subgroup Negative* (AUC_{bpsn}) calculates AUC on the hateful background examples and the non-hateful subgroup examples. A low value means that the model confuses non-hateful examples that mention the identity term with hateful examples that do not.

3) *Background Negative Subgroup Positive* (AUC_{bnsp}) calculates AUC on the non-hateful background examples and the hateful subgroup examples. A low value means that the model confuses hateful examples that mention the identity with non-hateful examples that do not.

We report the averaged metrics across identity terms, i.e., $\mathbf{AUC_{subg}}$, $\mathbf{AUC_{bpsn}}$, and $\mathbf{AUC_{bnsp}}$.[2]

4.3.2 Baselines

We compare BERT+EAR against the following existing approaches: (1) *BERT* (Devlin et al., 2019), (2) *BERT+SOC mitigation* (Kennedy et al., 2020), where the authors modify BERT's loss to lower the importance weight of identity terms, computed with the Sampling-and-Occlusion (SOC) algorithm (Jin et al., 2019), (3) Nozza et al. (2019), a single-layer neural network architecture based on the Universal Sentence Encoder (USE) representation (Cer et al., 2018), (4) Lees et al. (2020), a multilingual BERT model fine-tuned on the training data, (5) Ousidhoum et al. (2021), a classifier based on TF-IDF and Logistic Regression, and (6) Zhang et al. (2020), a debiasing training framework based on instance weighting.

The *debiased* version proposed in Lees et al. (2020) is obtained by training the model on additional samples from Wikipedia articles (assumed to be non-hateful) to balance the distribution of specific identity terms. Nozza et al. (2019) extracted these additional non-hateful samples from an external Twitter corpus (Waseem and Hovy, 2016).

To address the impact of different term lists, we also consider two different versions of BERT+SOC mitigation, one where we test the effect of *missing identity*

[2] Statistical significance and results from Lees et al. (2020) on these metrics could not be computed due to data unavailability and label distribution assumptions.

	Unintended bias (synthetic)					test	
	AUC$_{\text{subg}}$	**AUC$_{\text{bnsp}}$**	**AUC$_{\text{bpsn}}$**	**F1$_{\text{w}}$**	**F1$_{\text{hate}}$**	**F1$_{\text{w}}$**	**F1$_{\text{hate}}$**
Nozza et al. (2019), no mitigation	49.83	49.83	49.83	49.97	51.33	**72.29**	**71.62**
Nozza et al. (2019), debiased	50.27	50.21	50.21	45.40	29.31	71.43	69.37
Zhang et al. (2020)	69.99	62.19	62.19	43.01	66.70	31.35	63.21
BERT, no mitigation	70.97	66.62	66.62	58.19	64.61	69.60	70.21
BERT+SOC mitigation	78.11	**76.60**	**76.60**	51.88	58.89	57.39	60.47
BERT+SOC mitigation, missing ITs	68.58	67.38	67.38	38.49	41.38	51.14	43.65
BERT+EAR	**80.08**	75.18	75.18	**62.59** •▲	**70.58** •▲	70.90 ▲	70.83 ▲
Lees et al. (2020), debiased	-	-	-	**47.00**	58.58	79.87	82.45
Zhang et al. (2020)	48.10	48.29	48.29	33.33	**66.66**	33.54	66.69
BERT, no mitigation	47.30	47.54	47.54	39.72	61.17	81.57	83.56
BERT+SOC mitigation, translated ITs	45.54	45.88	45.88	46.34	51.62	80.28	81.73
BERT+EAR	**48.59**	**48.65**	**48.65**	40.64	62.71 •▲	**83.29** •▲	**84.68** ○▲
Ousidhoum et al. (2021), no mitigation	63.87	60.80	61.10	33.33	66.66	82.84	**93.80**
Zhang et al. (2020)	74.14	64.74	65.76	33.33	66.66	82.84	93.79
BERT, no mitigation	69.38	67.12	67.12	**50.24**	39.65	64.70	70.14
BERT+SOC mitigation	56.15	55.83	55.58	33.79	59.89	76.49	86.24
BERT+EAR	**74.31**	**71.43**	**71.25**	40.09	**67.45** •▲	**83.05** •▲	91.88 •▲

Table 4.2 Results (in %) on MISOGYNY (EN) (top), MISOGYNY (ITA) (middle), and MLMA. Significance of BERT+EAR over BERT without mitigation (•: $p \leq$ 0.01) and BERT with SOC mitigation (▲: $p \leq 0.01$).

terms and the other where the identity terms are *translated* for adapting to a new language.

4.4 Experimental Results

Table 4.2 shows classification and bias metrics on both synthetic and test set for the three corpora, i.e., MISOGYNY (EN) (top), MISOGYNY (ITA) (middle), and MLMA (bottom). The top rows in each table section report the performance of hate speech detection models specifically proposed for the respective dataset. The lower rows show the results of baselines and BERT+EAR. BERT+SOC mitigation uses the identity terms from Kennedy et al. (2020) (see Appendix A), unless a different identity terms lists is specified (e.g., "BERT+SOC mitigation, translated ITs").

BERT+EAR obtains comparable and, in most cases, better performance on all three datasets than all state-of-the-art debiasing approaches, which are based on (i) the knowledge of identity terms and (ii) data augmentation techniques. However, identity terms are not always readily available, which severely limits the generalization of those approaches. Similarly, there are several drawbacks to data augmentation with (assumed) non-hateful samples containing the identity terms. 1) Data augmentation

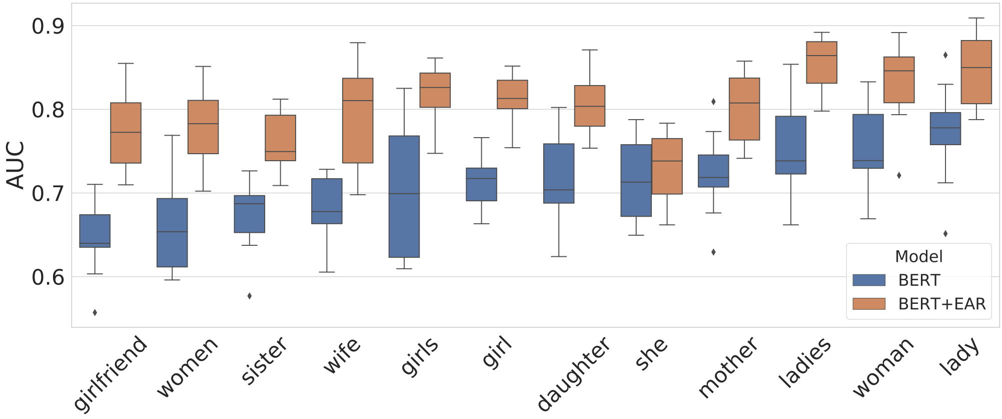


Fig. 4.4 $AUC_{subgroup}$ results broken down by identity term on MISOGYNY (EN).

is expensive. It requires filtering a large dataset (usually Wikipedia) and retraining the model with a much larger set of instances. 2) Data augmentation with task-specific identity terms requires prior knowledge of those terms, and is therefore limited by the authors' knowledge. 3) The overlap between identity terms in the evaluation set and the augmented data inevitably (but somewhat unfairly) improves the performance on the synthetic dataset.

BERT+EAR is overall the best debiasing model considering the proposed bias metrics. The only exception is MISOGYNY (EN), for which BERT+EAR has lower $\mathbf{AUC_{bnsp}}$ and $\mathbf{AUC_{bpsn}}$ than BERT+SOC mitigation. The latter's advantage, however, comes with high variability in the results. BERT+SOC mitigation seems more sensitive to random initialization. The standard deviation over 10 runs is 37%, compared to 13% of BERT+EAR.

Figure 4.4 shows the $AUC_{subgroup}$ metric separately by identity term on MISOGYNY (EN). We compare BERT and BERT+EAR over 10 different initialization runs. EAR improves BERT across all identity terms.

Most existing models and AUC-based metrics for unintended bias focus only on the false positives (i.e., hateful instances wrongly recognized as non-hateful). While correctly recognizing hateful instances is important, we believe that the problem of false negatives is equally important. Since BERT+EAR does not rely on identity term lists, it regularizes terms that impact *both* the positive and negative class. BERT+EAR obtains an average decrease of 15.04% in false negative rate compared to BERT and BERT+SOC mitigation. Indeed, the performance difference between BERT+EAR vs. BERT and BERT+SOC is mainly due to non-hateful instances

(∼95% of the time). Reducing the impact of overfitting terms like *f*ck* and *p*ssy* in MISOGYNY (EN) causes BERT+EAR to consider a larger context, and correctly labels them as non-hateful.

4.4.1 Error Analysis

Text	**BERT**	**BERT+SOC**	**BERT+EAR**
I'm just a sk*nk for understanding the basics of life!	1	1	0
You're such a f*cking hoe, I love it - the new Kanye and Lil Pump I kings make women feel comfortable about their sexuality.	1	1	0
GIRL, YOU'RE HYSTERICAL. I AM DANCING SO HAPPY FOR TODAY	0	1	0
#metoo I'm a victim of rape, abuse and harrassment. Every woman who had any these experiences.	1	1	0
some people at school drive me insane. like cool b*tch! im depressed too!! doesnt mean im a f*cking c*nt	1	1	0
@male_user And you are a hysterical k*nt.	0	1	0
@male_user F*ck you p*ssy	1	1	0

Table 4.3 Sample of non-misogynous tweets from MISOGYNY (EN). The tweets were misclassified by BERT, BERT+SOC, or both, and correctly classified by BERT+EAR. Next to the tweet we report prediction of each model (1 is misogynous). Exact phrasing changed to protect privacy.

Table 4.3 shows tweets from the MISOGYNY (EN) data set, which have been correctly predicted by BERT+EAR but misclassified by BERT or BERT+SOC. These tweets serve as qualitative examples of the effectiveness of forcing the model to attend to a wider context and not overfit to training-specific terms, exploiting the richness of information (Nozza et al., 2017). The examples are an excerpt of the most common cases where BERT+EAR classifies the non-hateful examples correctly: (1) when slurs or negative words (such as *sk*nk*) are used in a non-hateful context, like slang or lyrics, (2) when many words associated with misogyny appear in the sentence (e.g., *rape*, *abuse*) and (3) when the hateful target is male and the instance should not be classified as misogynous. The use of a wider context by BERT+EAR allows the model identify such non-misogynous instances compared to BERT and BERT+SOC. In particular, BERT+SOC is even more biased in these cases because its debiasing techniques overly rely on specific terms (e.g. *woman*) and increase overfitting to training-specific examples.

4.4.2 Impact of predefined identity terms

We also analyze the impact of predefined identity term lists on performance by evaluating the effect of (i) missing identity terms, and (ii) adapting to a new language where the list is unavailable.

First, we remove every identity term of BERT+SOC from MISOGYNY (EN) that appears at least once in the evaluation set, here *women* and *woman* out of 24 terms. This reflects the real-world case where the identity term list does not contain a specific group present in the data. The significant performance drop resulting from this case (Table 4.2, top, "missing ITs") highlights a strong weakness of term-based mitigation strategies.

Second, we analyze the case where identity terms need to be adapted to a new language, e.g., Italian. We translated the English identity terms from BERT+SOC to Italian via Google Translate.[3] Table 4.2 (middle, "translated ITs") shows that the performance is lower than BERT+EAR. A simple translation of predefined identity terms is therefore not an option for cross-lingual settings. This aligns with the findings by Nozza (2021), that demonstrated that cross-lingual hate speech detection is limited by the use of non-hateful, language-specific taboo interjections that are not directly translatable.

In sum, we demonstrated that relying on a predefined list of identity terms is a strong limitation for performance and generalizability of the model. In contrast, BERT+EAR's independence from any predefined terms makes it the ideal model in real-world scenarios.

[3] For gendered Italian words, we kept both the masculine and the feminine (e.g., *muslim* → *musulmana*, *musulmano*).

4.5 Extracting overfitting terms

Dataset	**Overfitting terms**
MISOGYNY (EN)	girls, womens*ck, hoes, c*ck, shut, stupid, hoe, p*ssy, trying, f*ck
MISOGYNY (ITA)	pezzo, bel, bellissima, scoperei, p*ttanona, zitta, sb*rro, t*ttona, bella, c*lone _(piece, nice, very nice, I’d f*ck, sl*t, shut up, c*m, b*sty, beautiful, fat*ss)_
MLMA	n*gger, n*gro, shut, chong, ching, d*ke, okay, sp*c, tw*t, f*ggot

Table 4.4 Terms with highest lexical overfitting identified using attention entropy.

While being the core of EAR, attention entropy serves another purpose. Once we finish standard fine-tuning (i.e., with no regularization involved), models have overfitted specific terms. We identify these terms using attention entropy.

To extract the most indicative terms, we replicate training conditions. Specifically, we run inference using all the training data using a fine-tuned checkpoint and a standard BERT tokenizer. We collect attention entropy values for each term and average them over all training instances. Terms with the lowest average entropy show the highest overfitting as the model learned them with a narrow context.[4]

Retrieving these terms after training allows us to gain insights into the domain and language-specific aspects driving the outcome.

Table 4.4 shows the top 10 terms with the highest lexical overfitting on the studied datasets extracted from the corresponding fine-tuned model. We extract terms strongly correlated with the positive class, e.g., _womens*ck_ (97%), _shut_ (96%), _n*gger_ (92%), _sb*rro_ (97%), _c*lone_ (95%). Note that these terms are _not_ frequent in the corpus. Overfitting terms appear with an average document frequency of only 4.7%, while the most frequent terms have 32.5% average document frequency across datasets. These results suggest that the higher the class polarization of a token, the narrower the context BERT will use to learn its representation and the higher the overfitting.

[4] To filter out noise, we report only words with a document frequency higher than 1%.

4.6 Related Work

The first works to study bias measurement and mitigation in neural representation removed intrinsic gender bias from word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Ravfogel et al., 2020; Romanov et al., 2019). More recently, researchers have focused on contextualized sentence representations and effective neural models for understanding the presence and resolution of bias (Nozza et al., 2021; Ousidhoum et al., 2021).

While the majority of proposed approaches focus on data augmentation (Bartl et al., 2020; de Vassimon Manela et al., 2021; Dixon et al., 2018; Nozza et al., 2019; Sharma et al., 2020), different approaches have been proposed for bias mitigation intervening directly in the objective function. Kennedy et al. (2020) proposed to apply regularization during training to the explanation-based importance of identity terms, obtained with Sampling-and-Occlusion (SOC) explanations (Jin et al., 2019). Kaneko and Bollegala (2021) proposed a method for debiasing pre-trained contextual representation by retaining the learned semantic information for gender-related words (e.g., *she*, *woman*, *he*, *man*) and simultaneously removing any stereotypical biases in the pre-trained model. Zhou et al. (2021) exploited debiasing methods for natural language understanding (Clark et al., 2019a) to explicitly determine how much to trust the bias given the input. Vaidya et al. (2020) proposed a multi-task learning model for predicting identity terms' presence alongside a sentence's toxicity.

The main drawback of all previous works is their strict reliance on predefined identity terms. This list can be either defined manually by experts or extracted a-priori from the data set. Either way, the subsequent debiasing models will be strongly affected by these biased terms, limiting the applicability of the trained model to new data. This requirement is a severe limitation since it is not always possible to retrain a model on new data to reduce bias, resulting in limited use in real-world cases.

4.7 Discussion

We introduce EAR, a regularization approach applicable to any attention-based model. EAR does not require any a-priori knowledge of identity terms, e.g., lists. This feature (i) allows us to generalize to different languages and contexts and (ii) avoids neglecting essential terms. As part of the training procedure, EAR also discovers the impact of relevant domain-specific terms. This automatic term extraction provides researchers with an analysis tool to improve data collection and bias mitigation approaches.

EAR, applied to BERT, reliably classifies data with competitive performance and substantially improves various bias metrics. BERT+EAR generalizes better to new domains and languages than similar methods.

4.7.1 Ethical Considerations

In this chapter, we propose term-attention entropy as a proxy for unintended bias in attention-based architectures. Our approach allows us to extract, for a given classifier and data set, a list of terms that induce most of the bias in the model. While this list is intuitive and easy to obtain, we would like to point out some ethical dual-use considerations.

Collecting the list is a data-driven approach, i.e., strongly dependent on the task, the corpus, the token frequencies, and the chosen model. Therefore, the list might lack specific terms or include terms that do not strictly perpetrate harm but are prevalent in the sample. Because of these twin issues, the resulting lists should *not* be read as complete or absolute. We discourage users from developing new models based solely on the extracted terms. We want, instead, the terms to stand as a starting point for debugging and searching for potential bias issues in the task at hand, be it in data collection or model development.

Further, while the probability is low, we can not exclude the possibility that future users run EAR on other tasks and data sets to derive private information or profile vulnerable groups.

4.7.2 A Broader Outlook

Most of the discussion across the chapter focused on the task of hate speech detection and used EAR as a regularization to mitigate unintended bias while solving it.

However, EAR stands as a generic, computationally-lightweight regularization to force a stronger contextualization into token embeddings. This property might benefit LLMs beyond unintended bias mitigation, specifically in tasks where lexical overfitting prevents from learning transferable representations of words.

In future developments, we plan to test EAR in additional NLP downstream tasks: is a stronger contextualization effectively a *better* contextualization in different scenarios, e.g., in the GLUE (Wang et al., 2019b) or SuperGLUE (Wang et al., 2019a) suites?

Chapter 5

Benchmarking Post-Hoc Interpretability Approaches for Misogyny Detection

Transformer-based Natural Language Processing models have become the standard for hate speech detection. However, the unconscious use of these techniques for such a critical task comes with negative consequences. Various works have demonstrated that hate speech classifiers are biased. These findings have prompted efforts to explain classifiers, mainly using attribution methods. In this chapter, we provide the first benchmark study of interpretability approaches for hate speech detection. We cover four post-hoc token attribution approaches to explain the predictions of Transformer-based misogyny classifiers in English and Italian. Further, we compare generated attributions to attention analysis. We find that only two algorithms provide faithful explanations aligned with human expectations. Gradient-based methods and attention, however, show inconsistent outputs, making their value for explanations questionable for hate speech detection tasks.

5.1 Motivation

The advent of social media has proliferated hateful content online (Ypma et al., 2021) – with severe consequences for attacked users even in real life. *Women* are often attacked online. A study by Data & Society[1] of women between 15 to 29 years

[1] https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf

	You	**are**	**a**	**smart**	**woman**
ΔP (10^{-2})	-0.1	1.1	-0.0	0.8	-47.6
G	0.11	0.10	0.09	0.25	0.27
IG	-0.17	0.18	-0.09	-0.35	-0.20
SHAP	0.00	-0.14	-0.04	-0.03	0.78
SOC	0.07	-0.13	0.03	0.03	0.52

Table 5.1 Explanations generated by benchmarked methods. A fine-tuned BERT wrongly classifies the text as misogynous. Darker colors indicate higher importance.

showed that 41% self-censored to avoid online harassment. Of those, 21% stopped using social media, 13% stopped going online, and 4% stopped using their mobile phone altogether. These numbers demonstrate the need for automatic misogyny detection systems for moderation purposes.

Various Natural Language Processing (NLP) models have been proposed to detect and mitigate misogynous content (Attanasio and Pastor, 2020; Attanasio et al., 2022b; Basile et al., 2019; Fersini et al., 2020a; Guest et al., 2021; Indurthi et al., 2019; Lees et al., 2020; Safi Samghabadi et al., 2020). However, several papers have already demonstrated that hate speech detection models suffer from unintended bias, resulting in harmful predictions for protected categories (e.g., *women*). Table 5.1 (top row) reports a very simple sentence that a state-of-the-art NLP model misclassifies as misogynous content.

This issue shows the need to understand the rationale behind a given prediction. A mature literature on model interpretability with applications to NLP-specific approaches exists (Rajani et al., 2019; Ross et al., 2021; Sanyal and Ren, 2021, inter-alia).[2] As explanations become part of legal regulations (Goodman and Flaxman, 2017), a growing body of work has focused on the *evaluation* of explanation approaches (Hase and Bansal, 2020; Jacovi and Goldberg, 2020; Nguyen and Martínez, 2020; Nguyen, 2018, inter-alia). However, little guidance on which interpretability method suits best to the sensible context of misogyny identification has been given. For instance, some explanations in Table 5.1 hint to which token is wrongly driving the classification and even highlight a potential bias of the model. But not all of them.

[2] We refer the reader to Danilevsky et al. (2020) and Madsen et al. (2021) for a recent, thorough perspective on explanation methods for NLP models.

We bridge this gap. We benchmark interpretability approaches to explain state-of-the-art Transformer classifiers on the task of automatic misogyny identification. We cover two benchmark Twitter datasets for misogyny detection in English and Italian (Fersini et al., 2018, 2020b). We focus on single-instance, post-hoc input attribution methods to measure the importance of each token for predicting the instance label. Our benchmark suite comprises gradient-based methods (Gradients (Simonyan et al., 2014b) and Integrated Gradients (Sundararajan et al., 2017)), Shapley values-based methods (SHAP (Lundberg and Lee, 2017)), and input occlusion (Sampling-And-Occlusion (Jin et al., 2019)). We evaluate explanations in terms of plausibility and faithfulness (Jacovi and Goldberg, 2020). Table 5.1 reports an example of token-wise contribution computed with these methods. Furthermore, we study attention-based visualizations and compare them to token attribution methods searching for any correlation. To our knowledge, this is the first benchmarking study of feature attribution methods used to explain Transformer-based misogyny classifiers.

Our results show that SHAP and Sampling-And-Occlusion provide plausible and faithful explanations and are consequently recommended for explaining misogyny classifiers' outputs. We also find that, despite their popularity, gradient- and attention-based methods do *not* provide faithful explanations. Outputs of gradient-based explanation methods are inconsistent, while *attention does not provide any useful insights for the classification task*.

Contributions We benchmark four post-hoc explanation methods on two misogyny identification datasets across two languages, English and Italian. We evaluate explanations in terms of plausibility and faithfulness. We demonstrate that not every token attribution method provides reliable insights and that attention cannot serve as explanation. Code is available at https://github.com/MilaNLProc/benchmarking-xai-misogyny.

5.2 Benchmarking suite

In the following, we describe the scope (Section 5.2.1) of our benchmarking study, the included methods (Section 5.2.2), and the evaluation criteria (Section 5.2.2).

5.2.1 Scope

We consider *local* explanation methods (Guidotti et al., 2019; Lipton, 2018). Given a classification model, a data point, and a target class, these methods explain the probability assigned to the class by the model. *Global* explanations provide model- or class-wise explanations and are hence out of the scope of this thesis.

Among local explanation methods, we focus on *post-hoc* interpretability, i.e., we explain classification models that have already been trained. We leave out *inherently interpretable* models (Rudin, 2019) as they do not find widespread use in NLP-driven practical applications.

We restrict our study to input attribution methods. In Transformer-based language models, inputs typically correspond to the tokens' input embeddings (Madsen et al., 2021). We, therefore, refer to *token attribution* methods to generate a contribution score for each input token (or word, resulting from some aggregation of sub-word token contributions).

5.2.2 Methods

We benchmark three families of input token attribution methods. First, we derive token contribution using gradient attribution. These methods compute the gradient of the output with respect to each of the inputs. We compute simple gradient (G) (Simonyan et al., 2014b) and integrated gradients (IG) (Sundararajan et al., 2017). Then, we attribute inputs using approximated Shapley values (SHAP) (Lundberg and Lee, 2017). Finally, following the literature on input perturbation via occlusion, we compute contributions using Sampling-And-Occlusion (SOC) (Jin et al., 2019). See appendix B for all implementation details.

We refer the reader to Section 2.3.4 of this work for a principled introduction to Post-Hoc Local Explanation Methods.

Attention There is an open debate on whether attention can be used as an explanation or not (Bastings and Filippova, 2020; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Our benchmarking study provides a perfect test-bed to understand if attention aligns with attribution methods. We compare standard self-attention with effective attention (Brunner et al., 2020; Sun and Marasović, 2021). Further, we measure attribution between input tokens and hidden representations using Hidden Token Attribution (HTA) (Brunner et al., 2020).

5.2.3 Evaluation criteria

We use *plausibility* and *faithfulness* as evaluation criteria (Jacovi and Goldberg, 2020). A "plausible" explanation should align with human beliefs. In our context, the provided explanation artifacts should *convince* humans that highlighted words are responsible for either misogynous speech or not.[3] A "faithful" explanation is a proxy for the true "reasoning" of the model. Gradient attributions are commonly considered faithful explanations as gradients provide a direct, mathematical measure of how variations in the input influence output. For the remaining attribution approaches, we measure faithfulness under the *linearity assumption* (Jacovi and Goldberg, 2020), i.e., the impact of certain parts of the input is independent of the rest. In our case, independent units correspond to input tokens. Following related work (Feng et al., 2018; Jacovi et al., 2018; Serrano and Smith, 2019, inter-alia), we evaluate faithfulness by erasing input tokens and measuring the variation in the model prediction. Ideally, faithful interpretations highlight tokens that change the prediction the most.

5.2.4 Data

Automatic misogyny identification is the binary classification task to predict whether a text is misogynous or not.[4] We focus on two recently-released datasets for misogynous content identification in English and Italian, released as part of the Automatic Misogyny Identification (AMI) shared tasks (Fersini et al., 2018, 2020b). Both datasets have been collected via keyword-based search on Twitter. Table 5.2 reports the dataset statistics.

5.3 Experimental setup

Among the Transformer-based models, we focus on BERT (Devlin et al., 2019) due to its widespread usage. We fine-tuned pre-trained BERT-based models on the

[3] In this study, the human expectation corresponds to the authors'.

[4] Characterizing misogyny is a much harder task, possibly modeling complex factors such as shaming, objectification, or more. Here, we simplify the task to focus on benchmarking interpretability.

Dataset	# Train	# Test	Hate %	F1
AMI-EN	4,000	1,000	45%	68.78
AMI-IT	5,000	1,000	47%	79.79

Table 5.2 Summary of datasets in terms of the number of training, validation and test tweets, percentage of hateful records within the training split, and F1-score of BERT models on test sets.

AMI-EN and AMI-IT datasets. We report full details on the training in Section B. Table 5.2 reports the F1 macro performance of BERT models on the test splits.

We explain BERT outputs on both tweets from test sets[5] and manually-generated data. Please refer to Appendix B for model training details. On real data, we address two questions: 1) *Is it right for the right reason?*, i.e., we assess if the model relies on a plausible set of tokens; 2) *What is the source of error?*, i.e., we aim to identify tokens that wrongly drive the classification outcome. By explaining manually-defined texts, we can probe for model biases.

Tables 5.3-5.6 report token contributions computed with benchmarked approaches (5.2.2). We report contributions for individual tokens.[6]

We define table contents as follows. Separately by the explanation method, we first generate raw contributions, average them across the multi-dimensional input array, and then L1-normalize the contribution vector. Finally, we use a linear color scale between solid blue (assigned for contribution -1), white (contribution 0), and solid red (contribution 1). For all reported examples, we explain the `misogynous` class. Hence, positive contributions indicate tokens *pushing* towards the misogynous class, while negative contributions push towards the non-misogynous one. Lastly, the second top row reports the variation on the probability assigned by the model when the corresponding token is erased (ΔP).

5.4 Discussion

Error analysis Table 5.3 shows the explanations for a tweet incorrectly predicted as misogynous. IG, SHAP, and SOC assign a negative contribution to the word

[5] We rephrase and explain rephrased versions of tweets to protect privacy.

[6] While several work average sub-word contributions for out-of-vocabulary words, there is no general agreement on whether this brings meaningful results. Indeed, an average would assume a model that leverages tokens as a single unit, while there is no clear evidence of that.

	You	**pu**	**##ssy**	**boy**
ΔP (10^{-2})	-0.3	-0.2	-35.6	0.8
G	0.11	0.19	0.32	0.18
IG	0.26	0.00	0.14	-0.60
SHAP	-0.03	0.52	0.28	-0.17
SOC	-0.01	0.03	0.51	-0.14

Table 5.3 Example from AMI-EN test set, anonymized text on first row. Ground truth: `non misogynous`. Prediction: `misogynous` ($P = 0.78$).

	s*ck	**a**	**d*ck**	**and**	**choke**	**you**	**b*tch**
ΔP (10^{-2})	-0.02	0.2	0.8	0.3	-0.1	0.03	-13.4
G	0.10	0.08	0.14	0.07	0.08	0.10	0.25
IG	-0.14	-0.16	-0.08	-0.05	-0.20	-0.22	-0.16
SHAP	0.24	-0.03	0.07	-0.05	0.05	-0.06	0.50
SOC	0.20	-0.02	0.26	-0.02	0.07	0.00	0.29

Table 5.4 Example from AMI-EN test set, anonymized text on first row. Ground truth: `misogynous`. Prediction: `misogynous` ($P = 0.90$).

boy. This matches our expectations since the target of the hateful comment is the male gender. These explanations are thus plausible. Still, the tweet is classified as misogynous. The tokens *pu* and *##ssy* mainly drive the prediction to the misogynous class, as revealed by all explainers (SHAP and SOC in a clearer way). Explanations suggest the model is failing to assign the proper importance to the targeted gender of the hateful comment. These plausible explanations are also faithful. Removing the term *boy* increases the probability of the misogynous class while omitting tokens *pu* and *##ssy* decrease it.

We further analyze the term *p*ssy* and its role as a source of errors. Almost all tweets of the test set containing the term *p*ssy* are labeled by the model as misogynous. The false-positive rate on this set of tweets is 0.93 compared to the 0.49 of the overall test set. Similar considerations apply to English words typically associated with misogynous content as *b*tch* and *wh*re*.

Is it right for the right reason? Table 5.4 reports the explanation of a correctly predicted misogynous tweet. Gradient, SHAP, and SOC explanations assign a high positive contribution to slurs (*b*tch*, *s*ck*, and *d*ck*). These explanations align

	Ann	**is**	**in**	**the**	**kitchen**	**David**	**is**	**in**	**the**	**kitchen**
ΔP (10^{-2})	-40.4	15.4	12.7	-12.6	-24.3	-1.0	8.0	-1.3	-5.8	-6.7
G	0.25	0.16	0.08	0.10	0.21	0.19	0.18	0.09	0.09	0.28
IG	-0.15	0.18	0.12	-0.33	-0.22	-0.36	0.14	0.09	-0.25	-0.17
SHAP	0.27	-0.31	-0.15	-0.01	0.27	-0.29	-0.38	-0.19	-0.05	0.09
SOC	0.28	-0.19	-0.06	0.10	0.07	-0.25	-0.11	-0.03	0.04	0.05

Table 5.5 Manually-generated example. Text starts with a female (left) and male (right) name. Ground truth (both): `non-misogynous`. Prediction: `misogynous` ($P = 0.53$) (left), `non-misogynous` ($P = 0.14$) (right).

with human expectations. However, not all slurs impact the classification outcome. Explanations on *b*tch* are faithful but they are not for *s*ck* and *d*ck*. Differently, IG does not highlight any token with a positive contribution. This result goes against expectations as the predicted class is misogynous, so we cannot conclude anything specific.

Unintended bias We study explanations to search for errors caused by unintended bias, a known phenomenon affecting models for misogynous identification. A model suffering from unintended bias performs better (or worse) when texts mention specific identity terms (e.g., *woman*) (Dixon et al., 2018).

Table 5.1 reports the non-misogynous text "You are a smart woman" incorrectly labeled as misogynous. SHAP, SOC, and, to a lesser extent, Gradient explanations indicate the term *woman* as responsible for the prediction. This result matches with recent findings on the unintended bias of hateful detection models (Borkan et al., 2019; Dixon et al., 2018; Nozza et al., 2019) and therefore explanations are plausible. Removing the term *woman* causes a drop of 0.48 to the probability of the misogynous class. This validates the insight provided by the explanations. Similar to the previous examples, the explanation of IG is challenging to interpret.

Table 5.5 shows another example of unintended bias. The text “Ann is in the kitchen” is incorrectly labeled as misogynous. Gradients, SHAP, and SOC assign the highest positive contribution to the (commonly) female name *Ann*. Interestingly, the second most important word for Gradients and SHAP is *kitchen*, reflecting stereotypes learned by the classification model (Fersini et al., 2018). These explanations are faithful: the model prediction drops by a significant 0.40 and 0.24 when erasing the tokens *Ann* and *kitchen*, respectively. We substitute the name *Ann* with *David*, a common male name. We observe that the prediction and the explanations drastically

	p*rca	p*ttana	che	gran	pezzo	di	f*ga	p*ttana	che	gran	pezzo	di	f*ga
ΔP (10^{-2})	94.7	79.7	-0.8	-0.6	0.3	-0.7	-0.6	1.0	-2.3	-1.3	0.4	0.3	-22.9
G	0.17	0.15	0.06	0.07	0.11	0.07	0.13	0.20	0.08	0.10	0.14	0.08	0.21
IG	-0.25	-0.10	-0.09	-0.16	-0.04	0.21	0.13	-0.12	-0.03	-0.25	0.11	0.17	0.32
SHAP	-0.69	-0.01	0.01	0.05	0.05	0.05	0.14	0.15	0.10	0.13	0.10	0.10	0.43
SOC	-0.56	-0.07	0.00	0.04	0.05	-0.05	0.22	0.00	0.05	0.07	0.04	-0.12	0.57

Table 5.6 Manually-generated example. Complete text (left) and text without initial “p*rca” (right). Non-literal translation: “*holy sh*t what a nice piece of *ss*”. Ground truth (both): `misogynous`. Prediction: `non-misogynous` ($P = 0.03$) (left), `misogynous` ($P = 0.97$) (right).

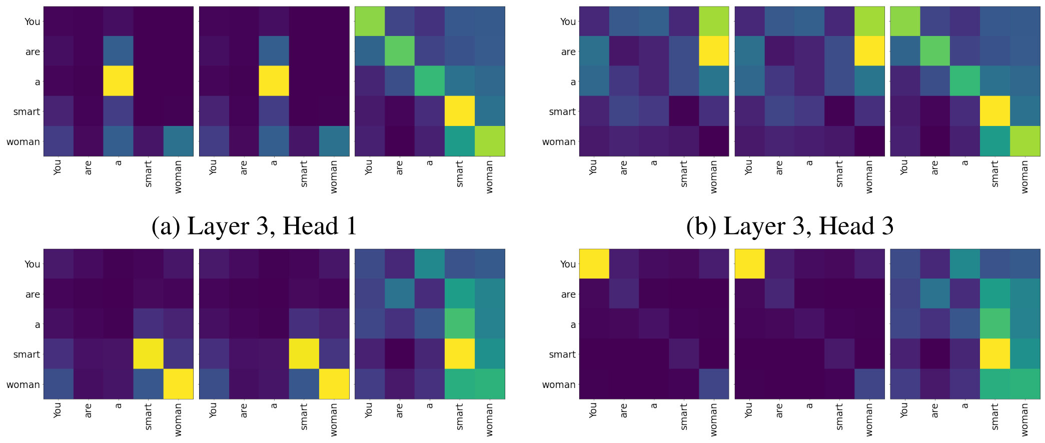
change. BERT correctly assigns it to the non-misogynous class, and IG, SHAP, and SOC give a high negative contribution to the word *David*. The all-positive contributions of Gradients do not provide valuable insights.

Bias due to language-specific expressions Table 5.6 (left) shows an example of incorrectly predicted misogynous text in Italian: "p*rca p*ttana che gran pezzo di f*ga" ("holy sh*t what a nice piece of *ss"). The expression "*p*rca p*ttana*" (literally *pig sl*t*) is a taboo interjection commonly used in the Italian language and does not imply misogynous speech.

The interpretation of the gradient explanation is hard since all contributions are positive and associated with the misogynous class. All explanation methods assign a positive contribution to the word *f*ga* (**ss*). SHAP, SOC, and, to a lesser extent IG, indicate that the main reason behind the non-misogynous prediction is the term *p*rca*. The bias of the model towards this expression was firstly exposed in (Nozza, 2021) and it thus validates IG, SHAP, and SOC explanations as plausible. When one of the two terms of the expression is removed, the probability increases significantly. This suggests that explanations by IG, SHAP, and SOC are faithful. Further, we inspect the behavior of explanation methods when we erase one of the terms. We omit the word *p*rca* and we report its explanations on Table 5.6 (right). The text is correctly assigned to the misogynous class and the word *f*ga* (**ss*) has the highest positive contribution for all the approaches.

5.4.1 Is attention explanation?

We follow up on the open debate on attention used as an explanation, providing examples on the misogyny identification task. Figure 5.1 shows self-attention maps in our fine-tuned BERT at different layers and heads for the already discussed



(a) Layer 3, Head 1

(b) Layer 3, Head 3

(c) Layer 10, Head 1

(d) Layer 10, Head 3

Fig. 5.1 Attention (left), Effective Attention (center), and Hidden Token Attribution (right) maps at different layers in fine-tuned BERT. Lighter colors indicate higher weights. Sentence: "You are a smart woman".

sentence "You are a smart woman". Based on our previous analysis (Section 5.4), we know that the model has an unintended bias towards the token "woman".

We cannot infer the same information from attention maps. Raw attention weights differ significantly for different layers and heads. In this example, there is a vertical pattern (Kovaleva et al., 2019) on the token "a" in layer 3 (Figure 5.1, a). However, the pattern disappears from heads in the same layer (Figure 5.1, b) and from the same head, on higher layers, where, instead, a block pattern characterizes "smart" and "woman" (Figure 5.1, c). This variability hinders interpretability as no unique behavior emerges.

Effective Attention (Brunner et al., 2020) is based on attention and shares the same issue.[7] These results further motivate the idea that attention gives only a *local* perspective on token contribution and contextualization (Bastings and Filippova, 2020). However, this does not provide any useful insight for the classification task. To further validate this limited scope, we use Hidden Token Attribution (Brunner et al., 2020) and measure the contribution of each input token (i.e., its first-layer token embedding) to hidden representations. There is a marked diagonal contribution on lower layers, meaning that tokens mainly contribute to their representation. Interestingly, on the upper layers, a solid contribution to "smart" and "woman"

[7] In most of our experiments, Effective Attention brings no perceptually different maps than simple Attention. The two methods are hence equivalent for local attention inspection.

appears for all the tokens in the sentence. Different patterns between HTA and attention suggest that attention weights do not measure token contribution even in the locality of a layer and a single head.

We observed similar issues in other examples and for Italian models (see Section B). Therefore, we cannot consider attention as a plausible or faithful explanation method and **discourage the use of attention to explain BERT-based misogyny classifiers**.

5.5 Related Work

Few works applied interpretability approaches to hate speech detection. Wang (2018) proposes an adaptation of explainability techniques for computer vision to visualize and understand the CNN-GRU classifier for hate speech (Zhang et al., 2018). Mosca et al. (2021) study both local and global explanations. They use Shapley values (Lundberg and Lee, 2017) to quantify feature importance on a *local* level and feature space exploration for a *global* explanation. Risch et al. (2020) analyze multiple attribution-based explanation methods for offensive language detection. The analysis includes an interpretable model (Naïve Bayes), model-agnostic methods based on surrogate models (LIME (Ribeiro et al., 2016), layer-wise relevance propagation (LRP) (Bach et al., 2015), and a self-explanatory model (LSTM with an attention mechanism). SHAP explainer is applied (Wich et al., 2020) to investigate the impact of political bias on hate speech classification. Sample-And-Occlusion (SOC) explanation algorithm has been used in its hierarchical version in different papers to show the results of hate speech detection (Kennedy et al., 2020; Nozza, 2021).

In this chapter, we specifically focus on hate speech against women. In this context, Godoy and Tommasel (2021) apply SHAP to derive global explanations to find any unintended bias in a misogyny classifier based on Random Forest.

While growing efforts are made for evaluating interpretability approaches for NLP models (Atanasova et al., 2020; DeYoung et al., 2020; Hase and Bansal, 2020; Jacovi and Goldberg, 2020; Nguyen and Martínez, 2020; Nguyen, 2018; Prasad et al., 2021), the evaluation is not domain-specific. Therefore, the benchmarking miss to consider specific sensitive problems and biases that are proper of the hate speech domain on which the explanation validation must focus. Here, we fill this gap by focusing on post-hoc feature attribution explanation methods on individual predictions for the task of hate speech against women.

5.6 Conclusion

In this chapter, we benchmarked different explainability approaches on Transformer-based models for the task of hate speech detection against women in English and Italian. We focus on post-hoc feature attribution methods applied to fine-tuned BERT models.

Our results address the second core research question of this work: can we infer lexical overfitting (and biased behaviors) using post-hoc explanations in hate speech classifiers based on Transformers? If so, which one is the more reliable? First, our qualitative study has shown that models over-rely on specific tokens to predict an outcome (see Tables 5.1, 5.3, 5.5, and 5.4). Second, quantitative results show that **SHAP and SOC provide plausible and faithful explanations** and are consequently recommended for explaining misogyny classifiers' outputs. In contrast, gradient- and attention-based approaches failed to provide reliable explanations.

In future work, we plan to add a systematic evaluation involving human annotators to the benchmarking suite. We also plan to include recently introduced token attribution methods (Sikdar et al., 2021) as well as new families of approaches, like natural language explanations (Narang et al., 2020; Rajani et al., 2019) and input editing (Ross et al., 2021). Finally, we will assess explanations of the most problematic data subgroups (Goel et al., 2021; Pastor et al., 2021; Wang et al., 2021).

5.7 Ethical Considerations

We explain BERT-based classifiers using a controlled subset of a large, fast-growing collection of explanation methods available in the literature. While replicating our experiments with different approaches, or on different data samples, from different datasets or explaining different models, we cannot exclude that some people may find the explanations offensive or stereotypical. Further, recent work has demonstrated that gradient-based explanations are manipulable (Wang et al., 2020), questioning the reliability of this widespread category of methods.

We, therefore, advocate for responsible use of this benchmarking suite (or any product derived from it) and suggest pairing it with human-aided evaluation. Moreover, we encourage users to consider this chapter as a starting point for model debugging (Nozza et al., 2022) and the included explanation methods as baselines for future developments.

Chapter 6

Conclusions

This thesis addresses a well-known issue in recent literature: predominant language tools are **far from perfect** both in their ability in language tasks and the social harm they perpetrate due to data-intensive training. Our research contributions span mainly three aspects.

First, we challenged modern Transformer-based sentence embedders for misogyny identification. Results on Twitter data show that they are *not* the expected top-tier solution. Indeed, we demonstrated how TF-IDF, curated lexicons, and semantic parsing enhance sentence embeddings. However, despite the improvement, we identified confounding factors in social media text that fool the model. The leading cause we found is poor generalization capabilities.

This evidence motivated our following study. Here, we discovered a tight interplay between self-attention, lexical overfitting, and unintended bias against minorities in Transformer-based language models. We contributed a novel regularization approach, dubbed **EAR** (Attanasio et al., 2022b), that mitigates lexical overfitting and bias in BERT. The method is domain, model, and language agnostic.

Our third main contribution is a benchmarking study of four post-hoc interpretability approaches applied to Transformer-based classifiers.

6.1 Contributions

The following is a summary of the main contributions of the thesis.

6.1.1 Improving Sentence Embeddings for Misogyny Identification in Italian

We proposed a novel multi-agent classification system to identify misogynous speech in Italian Tweets. We tested and submitted our system to the Automatic Misogyny Identification shared task (Fersini et al., 2020b) of the 2020 EVALITA campaign (Basile et al., 2020).

Our contributions can be summarized as follows.

- Although semantically rich, sentence embeddings extracted with modern language models are not enough to model misogyny. We have shown that a Support Vector Machine classifier trained on TF-IDF and a lexicon BoW provide a useful complement to sentence embedding when the latter has uncertain prediction scores.
- The resulting system is, however, **brittle in modeling the peculiar language used on social media**. Our analysis has highlighted chiefly four confounding factors that affect text classification models: words referring to parts of the body (e.g., “gola”, eng: *throat*), self-mocking references, reported misogynous speech, and hate expressed against the male gender.

6.1.2 Mitigating Unintended Bias in Pre-Trained Language Models

We proposed a novel mitigation technique for the unintended bias of hate speech classifiers towards demographic groups.

Our contributions can be summarized as follows.

- We **relate unintended bias** to lexical overfitting to specific words in the training corpus. These terms (not necessarily identity terms), if not regularized, drive the classification outcome regardless of the surrounding context.
- We demonstrate that overfitting tokens share a trait: the language models learn them using a **narrow self-attention**.
- We hence propose a novel regularization approach we call **Entropy-based Attention Regularization**. EAR maximizes self-attention entropy as part of

the training. **EAR works with any attention-based model, and it does not require any prior knowledge of identity terms**. Therefore, it generalizes better to different languages and contexts. We show the effectiveness of EAR as a bias mitigation technique in English and Italian hate speech datasets over several extrinsic bias metrics.

- Finally, we further explore our intuition on the relationship between narrow self-attention, lexical overfitting, and unintended bias. We automatically extract the overfitting terms probing for their entropy and show how the list reveals domain-specific words.

6.1.3 Benchmarking XAI for Misogyny Detection

We benchmarked post-hoc interpretability techniques on BERT-based misogyny detection models.

Our contributions can be summarized as follows.

- We benchmarked four techniques on two state-of-the-art misogyny detection datasets across two languages, English and Italian.
- Our results show that, among explainers, **only SHAP** (Lundberg and Lee, 2017) **and SOC** (Jin et al., 2019) **provide plausible and faithful explanations**.
- Gradient-based approaches failed to provide reliable explanations, while attention does not provide any useful insight for the classification task.
- Empirical analysis on synthetic samples shows that explanation methods can detect unintended bias in text classifiers.

6.2 Future Work

In the following section, we detail current research activities that stem from the work proposed in this thesis (the short-term) and several broader perspectives on language models and social biases (the long-term).

6.2.1 The Short-Term Direction

Current research activities expand the studies on Entropy-based Attention Regularization and, more broadly, generalization via auxiliary regularization.

Present evidence shows that EAR is effective in the task of hate speech detection to mitigate extrinsic bias. While promising, these results set clear research directions beyond the specific task and objective.

EAR is a generic, computationally-lightweight alternative to generalization approaches for Transformer models. Learning context-*richer* token representations can improve model performance in more NLP downstream tasks that require reasoning over the context. GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) are perfect candidate benchmarks to validate this hypothesis. Furthermore, reduced lexical overfitting can mitigate other forms of extrinsic bias. For example, Transformer-based Neural Machine Translation systems resolve poorly gender inflections when translating from English to a Grammatical Gender Language, reinforcing social stereotypes on job positions and gender (Stanovsky et al., 2019). It would be interesting to apply EAR to this case study.

On the technical side, further directions can study how attention-based regularization (i) behaves at different model and dataset scale, (ii) impacts pre-training, (iii) relates to emergent properties in language models (Teehan et al., 2022; Wei et al., 2022).

6.2.2 The Long-Term Direction

At the time of writing, the amount of research on language models is unprecedented. In the following, we review some long-term challenges ahead of the field, knowing they constitute a needle in a giant haystack. These are diachronic and demographic-aware adaptation, multi-modality, and multi-linguality.

From a time-aware perspective, all essential ingredients of our pipelines get outdated. We train and test models on data sampled in a given moment; even in later fine-tuning, we start from a checkpoint that had likely been trained months or even years before. In this *static* setup, the models we build are forced to a specific knowledge (ontological and about events) and language. For example, during the COVID-19 pandemic, sinophobia spread across social media: new ways, implicit and explicit, of targeting Asians, arose. Our best, already deployed hate speech detection models and APIs never occurred to meet that vocabulary (or, paradoxically, the word “COVID-19”). As **language is a reflex of time**, following people’s ideas and events, modeling or *updating* language models is a compelling area of research that comes across many fields and tasks.

To a more considerable extent, time is not the only latent factor we discard when training models with today’s procedures. Ingesting large corpora in an unsupervised pre-training and then fine-tuning them again on plain text leaves out individuals’ characteristics and nuances, such as their demographics and background. Especially in the area of hate speech detection, future language models should *situate* decisions in their context: for example, more informed decisions can arise if we explicitly bake the demographic factors of the speaker and the addressee in the training step (Dinan et al., 2020b). Other studies have developed similar ideas to factor uncertainty, and annotators’ agreement into the computational pipeline (Leonardelli et al., 2021; Rottger et al., 2022).

Beyond the textual dimension, even the medium **through which hatred spreads evolves**. New memes[1] combine images and text to reinforce stereotypical misconceptions. Recent work has had modest success in the automatic identification using only the bare content, i.e., an image and a caption: the hateful or stereotypical message is often hidden behind the surface content and requires common knowledge of the world (Attanasio et al., 2022a). Vision-language models do not seem to be the final solution in this active area of research.

Finally, we plan to extend hate speech detection and bias evaluation in **multilingual settings**. While efforts primarily target English, no established evidence confirms that both methods and results transfer to other languages. Along these lines, recent works advocate focusing on cultural differences between languages when modeling multilingual models (Talat et al., 2022) or developing new benchmarks for extrinsic bias evaluation in languages different from English (Ousidhoum et al., 2019; Röttger et al., 2022).

[1] https://en.wikipedia.org/wiki/Meme

6.3 Final Words

Writing this manuscript brought me a seesaw of emotions, and it has undoubtedly signed a crossroads for my academic career. But, in the long run, I think I have matured along the way.

Thank you, dear reader, if you have made it so far. If you have found weak spots, I think it is somewhat expected – a yet-to-become full-grown researcher carried out most of the work.

Ultimately, I hope our research has inspired you and made you come up with new exciting ideas.

Bibliography

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep Learning Models for Multilingual Hate Speech Detection. *arXiv:2004.06465 [cs]*, April 2020. URL http://arxiv.org/abs/2004.06465.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer, 2018.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.263. URL https://aclanthology.org/2020.emnlp-main.263.

Giuseppe Attanasio and Eliana Pastor. Politeam @ AMI: improving sentence embedding similarity with misogyny lexicons for automatic misogyny identification in italian tweets. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2765/paper142.pdf.

Giuseppe Attanasio, Luca Cagliero, and Elena Baralis. Leveraging the explainability of associative classifiers to support quantitative stock trading. In *Proceedings of the Sixth International Workshop on Data Science for Macro-Modeling*, DSMM '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380300. doi: 10.1145/3401832.3402679. URL https://doi.org/10.1145/3401832.3402679.

Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. Milanlp at semeval-2022 task 5: Using perceiver io for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, 2022a.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings* *of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland, May 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.88.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 100–112, Dublin, Ireland, May 2022c. Association for Computational Linguistics. URL https://aclanthology.org/2022.nlppower-1.11.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015. doi: 10.1371/journal.pone.0130140. URL https://doi.org/10.1371/journal.pone.0130140.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep Learning for Hate Speech Detection in Tweets. *WWW '17 Companion*, 2017. doi: 10.1145/3041021.3054223. URL http://arxiv.org/abs/1706.00188.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1023. URL https://aclanthology.org/P14-1023.

Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.1.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://aclanthology.org/S19-2007.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. EVALITA 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2765/overview.pdf.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3805. URL https://aclanthology.org/W19-3805.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.14. URL https://aclanthology.org/2020.blackboxnlp-1.14.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13, 2000.

Jayadev Bhaskaran and Isha Bhallamudi. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3809. URL https://aclanthology.org/W19-3809.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.96. URL https://aclanthology.org/2021.acl-short.96.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017a. doi: 10.1162/tacl_a_00051. URL https://aclanthology.org/Q17-1010.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017b.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819. URL https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 491–500. ACM, 2019. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

Cristina Bosco, Dell’Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. Overview of the EVALITA 2018 hate speech detection task. In *EVALITA 2018*, pages 1–9. CEUR, 2018.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/brunet19a.html.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BJg1f6EFDB.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186, 2017.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2029. URL https://aclanthology.org/D18-2029.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012.

Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1012. URL https://aclanthology.org/N16-1012.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. Multi-task learning in deep neural networks at evalita 2018. *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, pages 86–95, 2018.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1418. URL https://aclanthology.org/D19-1418.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019b. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL https://aclanthology.org/D17-1070.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 2021. doi: 10.1162/tacl_a_00425. URL https://aclanthology.org/2021.tacl-1.74.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.46.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of* *the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.190. URL https://aclanthology.org/2021.eacl-main.190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.656. URL https://aclanthology.org/2020.emnlp-main.656.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.23. URL https://aclanthology.org/2020.emnlp-main.23.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. *arXiv preprint arXiv:2112.06905*, 2021.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2006. URL https://aclanthology.org/P18-2006.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002. URL https://aclanthology.org/D18-1002.

Mary Ellsberg, Lori Heise, World Health Organization, et al. Researching violence against women: a practical guide for researchers and activists. 2005.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020. doi: 10.1162/tacl_a_00298. URL https://aclanthology.org/2020.tacl-1.3.

Samuel Fabrizi. fabsam@ ami: a convolutional neural network approach. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Bologna, Italy. CEUR. org*, 2020.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1407. URL https://aclanthology.org/D18-1407.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR-WS.org, 2018.

Elisabetta Fersini, Debora Nozza, and Giulia Boifava. Profiling Italian misogynist: An empirical study. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France, May 2020a. European Language Resources Association (ELRA). ISBN 979-10-95546-49-8. URL https://aclanthology.org/2020.restup-1.3.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. AMI @ EVALITA2020: Automatic Misogyny Identification. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, 2020b. CEUR.org.

Paula Fortuna and Sérgio Nunes. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30, July 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL https://doi.org/10.1145/3232676.

Simona Frenda, Bilal Ghanem, Estefanía Guzmán-Falcón, Manuel Montes-y Gómez, Luis Villasenor-Pineda, et al. Automatic expansion of lexicons for multilingual misogyny detection. In *EVALITA 2018*, pages 1–6. CEUR-WS, 2018.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752, 2019.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/content/115/16/E3635.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.265. URL https://aclanthology.org/2020.acl-main.265.

Hamidreza Ghader and Christof Monz. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1004.

Daniela Godoy and Antonela Tommasel. Is my model biased? exploring unintended bias in misogyny detection tasks. In *AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies*, pages 97–111, 2021. URL http://ceur-ws.org/Vol-2942/invited2.pdf.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-demos.6. URL https://aclanthology.org/2021.naacl-demos.6.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150. URL https://aclanthology.org/2021.acl-long.150.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL https://aclanthology.org/N19-1061.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Alex Graves. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013. URL http://arxiv.org/abs/1308.0850.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.114. URL https://aclanthology.org/2021.eacl-main.114.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820, 2018. URL http://arxiv.org/abs/1805.10820.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42, 2019. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online, July 2020. Association for Computational Linguistics. doi:

10.18653/v1/2020.acl-main.492. URL https://aclanthology.org/2020.acl-main.492.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.491. URL https://aclanthology.org/2020.acl-main.491.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL https://aclanthology.org/P16-2096.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2009. URL https://aclanthology.org/S19-2009.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5408. URL https://aclanthology.org/W18-5408.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

Emma Alice Jane. 'back to the kitchen, cunt': Speaking the unspeakable about online misogyny. *Continuum*, 28(4):558–570, 2014.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*, 2019.

Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.107. URL https://aclanthology.org/2021.eacl-main.107.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.483. URL https://aclanthology.org/2020.acl-main.483.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 267–280. Springer, 2019. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR, 2017. URL http://proceedings.mlr.press/v70/koh17a.html.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020. URL https://arxiv.org/abs/2009.07896.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1445. URL https://aclanthology.org/D19-1445.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.411. URL https://aclanthology.org/2021.findings-emnlp.411.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. Jigsaw@ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, 2020. CEUR.org.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.822. URL https://aclanthology.org/2021.emnlp-main.822.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220, 2016. URL http://arxiv.org/abs/1612.08220.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.488. URL https://aclanthology.org/2020.acl-main.488.

Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57, jun 2018. ISSN 1542-7730. doi: 10.1145/3236386.3241340. URL https://doi.org/10.1145/3236386.3241340.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc Interpretability for Neural NLP: A Survey. *arXiv preprint arXiv:2108.04840*, 2021. doi: 10.48550/ARXIV.2108.04840. URL https://arxiv.org/abs/2108.04840.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://aclanthology.org/N19-1063.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland, May

2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.132.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Edoardo Mosca, Maximilian Wich, and Georg Groh. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.socialnlp-1.8. URL https://aclanthology.org/2021.socialnlp-1.8.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. WT5?! Training Text-to-Text Models to Explain their Predictions. *arXiv preprint arXiv:2004.14546*, 2020. doi: 10.48550/ARXIV.2004.14546. URL https://arxiv.org/abs/2004.14546.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*, 2020. URL https://arxiv.org/abs/2007.07584.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1097. URL https://aclanthology.org/N18-1097.

Debora Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 2: Short Papers), pages 907–914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.114. URL https://aclanthology.org/2021.acl-short.114.

Debora Nozza, Elisabetta Fersini, and Enza Messina. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1026.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369343. doi: 10.1145/3350546.3352512. URL https://doi.org/10.1145/3350546.3352512.

Debora Nozza, Federico Bianchi, and Dirk Hovy. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.191. URL https://aclanthology.org/2021.naacl-main.191.

Debora Nozza, Federico Bianchi, and Dirk Hovy. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.bigscience-1.6.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1474. URL https://aclanthology.org/D19-1474.

Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2532–2542, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.199. URL https://aclanthology.org/2020.emnlp-main.199.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.329. URL https://aclanthology.org/2021.acl-long.329.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360, 2020.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1302. URL https://aclanthology.org/D18-1302.

Eliana Pastor and Elena Baralis. Explaining black box models by means of local rules. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 510–517, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359337. doi: 10.1145/3297280.3297328. URL https://doi.org/10.1145/3297280.3297328.

Eliana Pastor, Luca de Alfaro, and Elena Baralis. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 1400–1412, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383431. doi: 10.1145/3448016.3457284. URL https://doi.org/10.1145/3448016.3457284.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online, October 2020. Association for Computational Linguistics. doi: 10.18653 /v1/2020.emnlp-demos.7. URL https://aclanthology.org/2020.emnlp-demos.7.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. To what extent do human explanations of model behavior align with actual model behavior? In *Proceedings of the Fourth BlackboxNLP Workshop on*

Analyzing and Interpreting Neural Networks for NLP, pages 1–14, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.1. URL https://aclanthology.org/2021.blackboxnlp-1.1.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1146. URL https://aclanthology.org/N18-1146.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3550–3559, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1391. URL https://aclanthology.org/D18-1391.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487. URL https://aclanthology.org/P19-1487.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Julian Risch, Robin Ruff, and Ralf Krestel. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-56-6. URL https://aclanthology.org/2020.trac-1.22.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. What's in a name? Reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1424. URL https://aclanthology.org/N19-1424.

Alexis Ross, Ana Marasović, and Matthew Peters. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.336. URL https://aclanthology.org/2021.findings-acl.336.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *arXiv preprint arXiv:2206.09917*, 2022.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.13. URL https://aclanthology.org/2022.naacl-main.13.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019. URL https://doi.org/10.1038/s42256-019-0048-x.

Irene Russo, Francesca Frontini, and Valeria Quochi. OpeNER sentiment lexicon italian - LMF, 2016. URL http://hdl.handle.net/20.500.11752/ILC-73.

Shiva Omrani Sabbaghi and Aylin Caliskan. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. *CoRR*, abs/2206.01691, 2022. doi: 10.48550/arXiv.2206.01691. URL https://doi.org/10.48550/arXiv.2206.01691.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-56-6. URL https://aclanthology.org/2020.trac-1.20.

Svetlana Sagadeeva and Matthias Boehm. Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2290–2299, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383431. doi: 10.1145/3448016.3457323. URL https://doi.org/10.1145/3448016.3457323.

Soumya Sanyal and Xiang Ren. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.805. URL https://aclanthology.org/2021.emnlp-main.805.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2):336–359, 2020. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007/s11263-019-01228-7.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282.

Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997.

Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 358–364, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375865. URL https://doi.org/10.1145/3375627.3375865.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1339. URL https://aclanthology.org/D19-1339.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural NLP models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.71. URL https://aclanthology.org/2021.acl-long.71.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014a. URL http://arxiv.org/abs/1312.6034.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014*, 2014b. URL http://arxiv.org/abs/1312.6034.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1601. URL https://aclanthology.org/W14-1601.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL https://aclanthology.org/P19-1164.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.247.

Kaiser Sun and Ana Marasović. Effective attention sheds light on interpretability. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4126–4135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.361. URL https://aclanthology.org/2021.findings-acl.361.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017. URL http://proceedings.mlr.press/v70/sundararajan17a.html.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.bigscience-1.3.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. Emergent structures and training dynamics in large language models. 2022.

Ameya Vaidya, Feng Mai, and Yue Ning. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017b.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.132. URL https://aclanthology.org/2021.acl-long.132.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL https://aclanthology.org/D19-1221.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019a.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Cindy Wang. Interpreting neural network hate speech classifiers. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 86–92, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5111. URL https://aclanthology.org/W18-5111.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. Gradient-based analysis of NLP models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.24. URL https://aclanthology.org/2020.findings-emnlp.24.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.41. URL https://aclanthology.org/2021.acl-demo.41.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL https://aclanthology.org/N16-2013.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Maximilian Wich, Jan Bauer, and Georg Groh. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.7. URL https://aclanthology.org/2020.alw-1.7.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2020.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Patricia Ypma, Célia Drevon, Chloe Fulcher, Oriana Gascon, Kevin J Brown, Aleksandar Marsavelski, and Sylyie Giraudon. Study to support the preparation of the european commission's initiative to extend the list of eu crimes in article 83 of the treaty on the functioning of the eu to hate speech and hate crime. 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.380. URL https://aclanthology.org/2020.acl-main.380.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Ziqi Zhang, David Robinson, and Jonathan A. Tepper. Detecting hate speech on twitter using a convolution-GRU based deep neural network. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 745–760. Springer, 2018. doi: 10.1007/978-3-319-93417-4_48. URL https://doi.org/10.1007/978-3-319-93417-4_48.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1064. URL https://aclanthology.org/N19-1064.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.274. URL https://aclanthology.org/2021.eacl-main.274.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161.

Appendix A

Entropy-based Attention Regularization

Experimental Setup

Hyper-parameters

All our experiments use the Hugging Face transformers library (Wolf et al., 2020). We base our models and tokenizers on the `bert-base-uncased` checkpoint for English tasks and on the `dbmdz/bert-base-italian-uncased` checkpoint for Italian. We pre-process and tokenize our data using the standard pre-trained BERT tokenizer, with a maximum sequence length of 120 and right padding. We train all models with the following hyperparameters: batch size=64, learning rate=0.00002, weight decay=0.01, learning rate warmup steps=10%, full precision, maximum number of training epochs=30, and early stopping on non-improving validation loss after 5 epochs. Table 4.2 report results of BERT+EAR trained for 20 epochs with no early stopping, and regularization strength $\alpha = 0.01$. We chose the latter parameters with grid search on $\alpha \in [0.0001, 0.001, 0.01, 0.1, 1]$ and epochs $\in [10, 20, 30, 40, 50]$. When fine-tuning on MULTILINGUAL AND MULTI-ASPECT, we use a weighted cross-entropy classification loss ($\mathscr{L}_C$) to discount class unbalance. Specifically, we normalize the loss for data points belonging to class C by the prior probability of C, evaluated as its relative frequency in the training set.

For Kennedy et al. (2020), Nozza et al. (2019), Lees et al. (2020), and Ousidhoum et al. (2021), we kept all the parameters as specified by the respective authors. Please refer to our repository (https://github.com/g8a9/ear) for further details or the respective publications.

We trained all models with 10 different initialization seeds per parameter configuration and averaged over them to obtain stable results and meaningfully compute significance.

Statistical significance We compute the statistical significance of BERT+EAR over BERT and BERT with SOC mitigation via bootstrap sampling, following Søgaard et al. (2014), using $^{\circ}$ and $^{\triangle}$ (and their filled counterparts for a stronger significance) symbols, respectively. We use 1000 bootstrap samples and a sample size of %20. For Hate Speech, significance can only be computed on F1 scores since bias metrics require an assumption about the label distribution across identity terms that is not given.

Selection bias We computed the $\mathbf{B_2}$ metric following Ousidhoum et al. (2020). Specifically, we run the authors' code on each of our training dataset, using the query keywords used to sample each dataset. In case of queries composed of multiple words, we split and considered them separate keywords.

Dataset preprocessing The original MULTILINGUAL AND MULTI-ASPECT dataset comes in a multi-label, multiple class setting. Following Ousidhoum et al. (2021), we used the *Hostility* dimension of the dataset as target label and created a *Hate* binary from it as follows. We considered single-labeled "Normal" instances to be non-hate/non-toxic and all the other instances to be toxic.

Computation time We report NVIDIA Tesla V100 PCIE-16GB -equivalent computation time for the tested models. Averaging across the three presented data sets, training and evaluating 10 seeds of BERT+EAR (without early stop) requires 22 hours, compared to 72 hours for BERT+SOC and 7 hours for BERT. The regularization of attention entropy does not affect the computation time by a significant amount.

CO_2 emission Experiments were conducted using a private infrastructure, which has an estimated carbon efficiency of 0.432 kgCO_2eq/kWh. A cumulative of 319 hours of computation was performed on the hardware of type Tesla V100-PCIE-16GB (TDP of 300W). Total emissions are estimated to be 41.34 kgCO$_2$eq.

Estimations were conducted using the Machine Learning Impact calculator presented in (Lacoste et al., 2019).

List of identity terms

In the following, we report the list of identity terms used in the considered data sets and methods.

> (Kennedy et al., 2020): *muslim, jew, jews, white, islam, blacks, muslims, women, whites, gay, black, democrat, islamic, allah, jewish, lesbian, transgender, race, brown, woman, mexican, religion, homosexual, homosexuality, africans*
>
> (Nozza et al., 2019): *woman*, *women*, *daughter*, *girl*, *girls*, *mother*, *she*, *wife*, *lady*, *ladies*, *girlfriend*, *sister*
>
> (Fersini et al., 2020b): *nonne*, *matrone*, *mamme*, *casalinghe*, *compagne*, *morose*, *femmine*, *donne*, *fidanzate*, *nonna*, *matrona*, *casalinga*, *morosa*, *femmina*, *mamma*, *donna*, *fidanzata*, *compagna*
>
> (Dixon et al., 2018): *lesbian*, *gay*, *bisexual*, *transgender*, *trans*, *queer*, *lgbt*, *lgbtq*, *homosexual*, *straight*, *heterosexual*, *male*, *female*, *nonbinary*, *african*, *african american*, *black*, *white*, *european*, *hispanic*, *latino*, *latina*, *latinx*, *mexican*, *canadian*, *american*, *asian*, *indian*, *middle eastern*, *chinese*, *japanese*, *christian*, *muslim*, *jewish*, *buddhist*, *catholic*, *protestant*, *sikh*, *taoist*, *old*, *older*, *young*, *younger*, *teenage*, *millenial*, *middle aged*, *elderly*, *blind*, *deaf*, *paralyzed*

Appendix B

Benchmarking Post-Hoc Interpretability Approaches

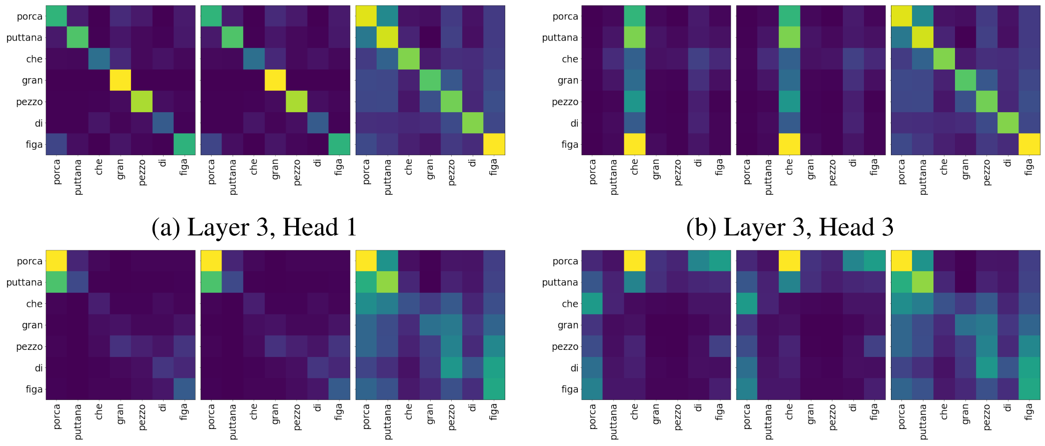
Experimental Setup

Training hyper-parameters

All our experiments use the Hugging Face transformers library (Wolf et al., 2020). We base our models and tokenizers on the `bert-base-cased` checkpoint for English tasks and on the `dbmdz/bert-base-italian-cased` checkpoint for Italian. We pre-process and tokenize our data using the standard pre-trained BERT tokenizer, with a maximum sequence length of 128 and right padding. We train all models for three epochs with a batch size of 64, a linearly decaying learning rate of $5 \cdot 10^{-5}$ and 10% of the total training step as a warmup, and full precision. We use 10% of training data for validation. We evaluate the model every 50 steps on the respective validation set. At the end of the training, we use the checkpoint with the best validation loss. We re-weight the standard cross-entropy loss using the inverse of class frequency for accounting class imbalance.

Explanation methods

We used the *Captum* library (Kokhlikyan et al., 2020) with default parameters to compute gradients (G) and integrated gradients (IG). Following (Han et al., 2020), for IG we multiply gradients by input word embeddings. For Shapley



(a) Layer 3, Head 1

(b) Layer 3, Head 3

(c) Layer 10, Head 1

(d) Layer 10, Head 3

Fig. B.1 Attention (left), Effective Attention (center), and Hidden Token Attribution (right) maps at different layers in fine-tuned BERT. Lighter colors indicate higher weights. Sentence: “p*rca p*ttana che gran pezzo di f*ga”, non-literal translation: “*holy sh*t what a nice piece of *ss*”.

values estimation (SHAP), we use the *shap* library[1] with PartitionSHAP as approximation method. For Sampling-And-Occlusion (SOC), we used the implementation associated with Kennedy et al. (2020).[2] Please refer to our repository (https://github.com/MilaNLProc/benchmarking-xai-misogyny) for further technical details.

Attention maps

We used attention weights provided by the transformers library for visualization. We implemented Effective Attention and Hidden Token Attribution following Brunner et al. (2020). We release the implementation on our repository.

Attention plots

Figure B.1 shows attention visualizations for the sentence “p*rca p*ttana che gran pezzo di f*ga” (Non-literal translation: “*holy sh*t what a nice piece of *ss*”). As discussed in Section 5.4 (**Bias due to language-specific expressions**), the text is misclassified as `non-misogynous` and most of explanation methods correctly highlight the Italian interjection “p*rca p*ttana”.

[1] https://github.com/slundberg/shap

[2] https://github.com/BrendanKennedy/contextualizing-hate-speech-models-with-explanations

Similar to results reported in Section 5.2.2, we cannot find useful insights in attention plots. Attention in layer 3 has a diagonal pattern in head 1, and a diagonal pattern in head 3 on the word *che* ("*what*"). However, these patterns disappear in layer 10, where attention focuses on *p*rca*. Moreover, at layer 10, HTA is more spread than attention, suggesting that the latter measures only a *local* token contribution.