The interpretability of the ReLU network to solve the problem of political correctness in the Black Myth of Wukong

Han Changhao

Abstract

This article addresses the influence of political correctness on the Chinese game Black Myth: Wukong, particularly through the lens of feminist criticism. Using Natural Language Processing (NLP) enhanced with ReLU networks, we explore how interpretability can be improved to analyze gender bias objectively within media content. Traditional NLP models often rely on complex, opaque neural networks, which lack transparency in analyzing sensitive topics like feminism. By transforming deep networks into more interpretable shallow networks with ReLU, this study seeks to provide insights into bias detection and offer a framework for evaluating feminist discourse. Through this model, we aim to illuminate how bias within game narratives can be detected and addressed, fostering a more inclusive gaming culture. The paper concludes by discussing potential future developments in NLP interpretability and ethical AI in gender bias analysis.

1 Introduction[1], [2], [3]

In recent years, there has been a growing global discourse surrounding feminism, expanding from well-known issues such as wage inequality to broader societal concerns. These discussions have also led to complex challenges, including political struggles and misuse of feminist rhetoric. A notable example is the Chinese single-player game "Black Myth: Wukong," which faced criticism from IGN France for perceived feminist issues and was reportedly subjected to extortion attempts by the organization Sweet Baby Inc. for a substantial sum of $7 million. This incident highlights the pervasive influence of political correctness and underscores the challenges in evaluating feminism, a concept often seen as subjective.Not only the "Black Myth：Wukong", but also the influence of feminist abuse exists in many fields, which has more or less constrained the development of many excellent creations. Moreover, the global problem of feminist abuse is causing huge economic losses, and various industries need to spend a lot of money to deal with the impact of feminist public opinion, which is still developing at a very fast pace。This matter is inevitable cause gender bias is a very subjective thing."In modern gender concepts, gender stereotypes often have a significant impact on our lives, and stereotypes can sometimes lead to bias.",[4]after such subjective problems become social behaviors, they will produce unpredictable results.

Traditional approaches to addressing these issues through Natural Language Processing (NLP) have shown potential, cause NLP appears to be able to objectively evaluate gender bias issues in media such as Black

Myth WuKong based on what is considered correct in sociology, thus accurately determining the legitimacy of Sweet Baby's extortion behavior and reducing the abuse of political correctness,yet People may think that NLP obtains data from biased media and is trained, which greatly reduces the credibility of NLP in analyzing gender bias issues and goes against our original beliefs.And the limitations of current algorithms have become apparent. When analyzing feminist topics, existing artificial intelligence models often rely on complex neural networks that are difficult to interpret, thereby limiting their utility as tools for evaluating such nuanced and subjective matters. This lack of interpretability presents a significant obstacle for researchers seeking to develop comprehensive and convincing NLP models capable of addressing feminist issues.

If a persuasive AI tool can be constructed to evaluate gender bias issues in works, it can greatly replace subjective judgments in feminism. Creators can use this network as evidence to support the political correctness of their works, reducing the economic and energy expenditure on feminist issues

2 Relu Network[5], [6]

A ReLU network is a type of feedforward neural network that uses ReLU activation functions to model complex functions. The network is structured with multiple hidden layers, where each layer applies a transformation to its input by multiplying it with a weight matrix, adding a bias vector, and then applying the ReLU activation function. This activation function compares each input value with zero and chooses the larger one, ensuring all outputs are non-negative.

A ReLU network N : Rn → Rm, is a composition of $L \in N$ hidden layers given by:

$\chi^{(l)} = \sigma(W^{(l)}\chi^{(l-1)} + b^{(l)})$

σ is a function that can compare the input number with zero and choose the larger one.In this Network $\chi^{(0)} = x$. N (x) is the output layer

$N(x) = W^{(L+1)}\chi^{(L)} + b^{(L+1)}$

The vector N = [n1, n2, . . . , nL] given the number of neurons in each layer, and all activations are in the positive reals, i.e. $\chi^{(l)} \in R^{nl} \geq 0$ for all $l \in \{1, . . . ,L\} = [L]$. We stress the dependence on x by writing $\chi^{(l)}(x)$.

The input to the network is processed through these layers, each producing an output that becomes the input for the next layer. The final layer produces the network's output. The architecture of the network is defined by specifying the number of neurons in each layer, and it is designed so that all activations (outputs of each neuron) are non-negative. This property is particularly useful in applications where the outputs need to be positive, such as certain types of regression tasks.

The ReLU network's structure allows it to model complex relationships in data, making it a powerful tool in various machine learning applications, including image recognition and natural language processing. The reliance on ReLU activation functions introduces non-linearity into the model, which is crucial for handling complex patterns and making the network more expressive.

By applying ReLU networks, it is possible to decompose complex deep neural networks into more interpretable three-layer shallow networks. This decomposition significantly enhances the interpretability of these models, making it easier for the public and researchers alike to understand the underlying criteria used in analyzing feminism through artificial intelligence,ReLU can better showcase the training data and ethical principles used in NLP to people,thus decline the worries of the data that Scientists used to train NLP.The ReLU network's ability to assign different weights to reference criteria and project them onto a shallow neural network allows for iterative optimization, improving the network's accuracy in analyzing feminist discourse. As we explore the latest developments in this field, it becomes clear that ReLU networks offer a promising avenue for creating more transparent and reliable NLP tools, particularly in contexts where subjective judgments are necessary.

3 Exploring Implicit Bias in Gaming Narratives: The Role of ReLU in Analyzing 'Black Myth: Wukong'[7]

The art design of the characters in the Black Mythical Sky follows traditional Chinese aesthetics, and there is a significant deviation between the requirements of mainstream feminism and the aesthetics of traditional Chinese aesthetics. This has led to a particularly serious problem of gender bias in the Black Mythical Sun Wukong, which has sounded the alarm for the spread and development of Chinese aesthetics to the world. It is not enough to judge whether there is bias solely based on the current mainstream feminism, especially for Chinese aesthetics, which is just emerging in countries other than China. Many inappropriate uses of feminism internationally will constrain the development of Chinese aesthetics.

The integration of ReLU activation functions into Natural Language Processing models has significantly enhanced the transparency and interpretability of these models, which is crucial for the detection and mitigation of biases. ReLU's nonlinearity plays a central role in how deep learning models process and understand complex language patterns. Specifically, ReLU generates sparse activations, meaning only a subset of neurons is activated in response to any given input. This sparsity facilitates a more transparent understanding of the model's decision-making process by allowing researchers to trace the influence of specific inputs, such as words or phrases, on the model's outputs.

Transparency is not just a technical benefit; it holds profound implications for ethical AI, particularly in fields like law, medicine, and media, where understanding the rationale behind a model's decision is critical. ReLU's ability to isolate and highlight specific neuron activations makes it possible to identify where biases may be introduced within a model. For instance, if certain neurons are consistently activated in response to particular demographic markers—such as gender or race—this could indicate that the model is associating these markers with specific, potentially biased, outcomes. This transparency is essential for identifying and addressing such biases, enabling developers to make informed decisions about how to adjust the model to reduce or eliminate these biases.

Moreover, the interpretability that ReLU brings to NLP models is vital in high-stakes applications. In contexts like legal reasoning or medical diagnostics, where model decisions can have significant consequences, being able to explain how and why a model arrived at a particular conclusion is indispensable. ReLU-enhanced transparency ensures that the decision-making process within the model can be scrutinized and understood, thereby building trust among users, regulators, and other stakeholders.

To illustrate how ReLU can be leveraged to analyze and mitigate biases in specific applications, consider the case of "Black Myth: Wukong," a video game that draws heavily on Chinese mythology. By employing advanced NLP techniques enhanced with ReLU activation functions, one can systematically analyze whether the game's narrative and character portrayals contain implicit biases. The process begins with the collection and preprocessing of diverse textual data related to the game, such as character dialogues, storyline descriptions, player reviews, and media coverage. This data is then analyzed using a ReLU-enhanced model, which introduces nonlinearity into the analysis, allowing the model to detect subtle and complex patterns that might not be immediately apparent.

For example, the model could reveal whether certain characters are consistently associated with positive or negative attributes based on their demographic characteristics, such as race or gender. By examining the activation patterns within the ReLU-based model, researchers can identify potential biases in how these characters are depicted. If the model shows that descriptions of female characters trigger different neuron activations compared to male characters, this could suggest a gender bias in the game's narrative. Similarly, if certain cultural backgrounds are linked to specific behaviors or attributes, this could indicate an ethnic bias.

The transparency afforded by ReLU is particularly valuable in this context because it allows researchers to trace the specific inputs that influence the model's outputs. This ability to "peer inside" the model's decision-making process is crucial for identifying biases and providing actionable insights. These insights can guide game developers in adjusting narratives, character designs, or marketing strategies to promote a more balanced and inclusive representation.

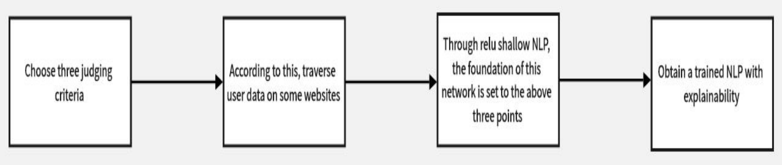
Furthermore, the findings from such an analysis could inform future storytelling approaches within game development, encouraging a more inclusive and culturally sensitive narrative structure. By leveraging ReLU-enhanced NLP models, this methodology provides a robust, transparent, and systematic approach to identifying and mitigating biases in complex media like "Black Myth: Wukong," ultimately contributing to the creation of a more equitable gaming experience.

4 Example

Web crawlers can traverse user data on some websites (such as Weibo and Steam), and can crawl some data according to the needs of users[8]

Now there are many ways to implement web crawlers, such as python[9] Once we have the review data, we can use the reviews as training data to train the relunetwork (NLP) and give weight to the reference for judging gender bias based on the emotional response of the player brought by different images。

After we have the review data, we can use the reviews as training data to train the relunetwork (NLP), and assign weight to the reference basis for judging gender bias according to the emotional reactions brought by different images, so as to train the NLP whose judgment results are truly in line with the psychology of the public.


Choose three judging criteria
According to this, traverse user data on some websites
Through relu shallow NLP, the foundation of this network is set to the above three points
Obtain a trained NLP with explainability

5 Conclusion and Outlook

ReLU network can transform from a deep network to a three-layer shallow network, which can be used to visualize the complex and cumbersome network structure of NLP, endowing the network with interpretability. With interpretable NLP, we can use NLP to make objective judgments in the subjective field of feminism, thereby preventing political correctness and the abuse of feminism. However, it is worth discussing that when transforming an extremely complex network into a shallow network, the shallow network is often very complex and difficult to establish moral standards for evaluation solely through these three layers. It is hoped that in the future, the structure of the relunenetwork shallow network can be further simplified to establish interpretability from a socially understandable level.

References

[1] H. Devinney, J. Björklund and H. Björklund, 《Theories of 〈Gender〉 in NLP Bias Research》, May 5, 2022, *arXiv*: arXiv:2205.02526. Seen on: August 31, 2024. [Online]. Published in: *http://arxiv.org/abs/2205.02526*

[2] J. Butler, *Who's Afraid of Gender?* Farrar, Straus and Giroux.

[3] 《 3_(Strong Ideas) Catherine D'Ignazio, Lauren F. Klein - Data Feminism-The MIT Press (2020)》.

[4] Tan Yi, 《 The Sociological Understanding about Prejudice》, *Adv. Psychol.*, Vol. 08, Issue 01, pp. 81-86, 2018,doi: 10.12677/AP.2018.81010.

[5] M. J. Villani and N. Schoots, 《Any Deep ReLU Network is Shallow》, June 20, 2023, arXiv: arXiv:2306.11827. Seen on: August 31, 2024. [Online]. Contained in: *http://arxiv.org/abs/2306.11827*

[6] P. Jin, "Shallow ReLU neural networks and finite elements," March 9, 2024, arXiv: arXiv:2403.05809. Seen on: November 21, 2024. [Online]. Published in: *http://arxiv.org/abs/2403.05809*

[7] Xian L. and Rui C., 《 Is Cross-Cultural Communication Influenced by Users' Cultural Identity? A Study on The Overseas Communication of 〈Wukong〉 Image:》, *MANDARINABLE J. Chin. Stud.*, Vol. 2, Issue 2, pp. 167-182, Oct 2023, doi: 10.20961/mandarinable.v2i2.887.

[8] Md. AbuKausar, V. S. Dhaka and S. Kumar Singh, 《 Web Crawler: A Review》, *Int. J. Comput. Appl.*,Vol. 63, Issue 2, pp. 31-36, February 2013, doi: 10.5120/10440-5125.

[9] Z. Chang, "A Survey of Modern Crawler Methods," in The 6th International Conference on Control Engineering and Artificial Intelligence, Virtual Event Japan: ACM, March 2022, pp. 21 - 28. doi: 10.1145/3522749.3523076.