Employing Neo-Psychometric Natural Language Processing in Classification of Anti-Trans Social Media Posts

Sofia Przyłuska[1] and Mikołaj Biesaga[2][0000−0001−5548−3546]

[1] College of Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Warsaw, Poland
s.przyluska@student.uw.edu.pl
[2] Robert Zajonc Institute for Social Studies, University of Warsaw, Warsaw, Poland
m.biesaga@uw.edu.pl

Abstract. While significant progress has been made in the area of prejudice detection, most contemporary attempts at hate speech classification lack theoretical backing, which is especially pronounced on the levels of feature extraction and classifier architecture. This limitation becomes problematic considering the limited size (or quality) of most datasets used, and is well illustrated in anti-transgender hate speech detection. In the present paper, we introduce the Neo-Psychometric Machine Learning approach. We define it as the use of Psychometric measures and domain knowledge for the broader purpose of addressing Machine Learning tasks. We propose a theory-driven classifier architecture using similarity metrics to questionnaire items from a psychometric operationalization of Transmisogyny theory. Our results show that the Neo-Psychometric NLP approach can be successfully applied to the task of classifying anti-trans social media posts from the TIDEs dataset, while reducing the risk of overfitting on construct-unrelated features and providing potential for theory-driven explanations. We show that anti-trans prejudice can be accurately classified even using similarity metrics to questionnaire items that do not include explicit statements about transgender people, but are rooted in the theoretical operationalization of the predictors of transphobia. While the Neo-Psychometric approach would, in our view, not be best suited in applications where the aim is to maximize accuracy, profit, or other performance metrics, we believe it can be applied in situations where theoretical backing, explainability and reducing construct-unrelated model bias is crucial — with good results.

Keywords: Neo-Psychometric Machine Learning · Psychometric Machine Learning · Natural Language Processing · Hate Speech · Explainable Machine Learning · Prejudice · Transgender · Transphobia · Transmisogyny · Sexism · Gender Essentialism

1 Introduction

Machine learning models can aid content moderation and make big data analysis possible. However, the quality of their performance depends on the datasets

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_41

they are trained on, which very often are prepared with the help of human annotators. In the context of prejudice detection, this poses ethical concerns related to exposing humans to large volumes of hate speech, which comes with psychological burden and mental health risks. To address this issue, the machine learning community has come up with two possible solutions. One approach is the manual labelling of the dataset by the researchers themselves and therefore mitigating the risk for unprepared annotators [e.g., 5, 10]. The other solution focuses on practices grounded in care for the mental wellbeing of annotators at every stage of the construction of a dataset [7].

One limitation of those approaches is the high cost and effort needed for the creation of datasets, which currently means that – in the context of hate speech classification – high quality and, especially, ethically produced datasets are not large, and large datasets are, in turn, often not high quality or not ethically sourced. This, consequently, increases the risk of classifier overfitting, inability to infer hate speech qualities adequately and, finally, misclassification.

To address this issue, we propose the Neo-Psychometric Machine Learning approach. Broadly speaking, this approach uses psychometric measures and domain knowledge for the broader purpose of addressing Machine Learning tasks. In the present study, we propose a theory-driven classifier architecture. We utilize Contextualized Construct Representations [1] – similarity metrics of the classified text and relevant psychometric scale item embeddings from a sentence transformer model – as intermediate features for traditional classification models. We demonstrate this approach on the task of classifying anti-trans social media posts from TIDEs dataset (Transgender and Nonbinary Community-Labeled Dataset for Transphobia Identification in Digital Environments).

Anti-Trans Hate Speech Classification. A number of strategies have been employed to classify anti-transgender hate-speech in social media posts.

In a study employing a Retrieval-Augmented Generation pipeline to classify sentiment towards transgender people Leitner et al. (2025) [10] examine the network structure of Anti-Trans actors on Tik-Tok. They analyze transcribed video content, descriptions, hashtags, usernames and tagged users from 59,860 videos. Their findings suggest that, while pro-trans users tend to either be very isolated or found in tight knit clusters, the anti-trans actors are deeply embedded with the neutral users and prevalent throughout the network. Notably, the classification pipeline they employ (Overall Accuracy = 0.67, F1 Anti-Trans = 0.48, F1 Pro-Trans = 0.83, F1 Neutral = 0.56), which uses RAG samples and taxonomy, can be described as theory driven. Nevertheless, the Llama3 model they utilize, even in the smallest variants, uses billions of parameters, which in some contexts (like CPU constrained computation or big data analysis) can be unfeasible.

Lameiro et al. (2025) [7] introduce the TIDEs dataset. It contains 3,509 annotated posts and comments (42.7% of them labelled as Transphobic) from YouTube, Tumblr, Truth Social, and Reddit. The authors compare a number of classification approaches, including metrics from Perspective API [9], a Logistic Regression model, and a state of the art DeBERTa classifier trained on their dataset. Because the TIDEs dataset was used in the present study, the perfor-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_41

mance metrics of the models can be found in Table 1 as a comparison for the model architectures proposed herein.

While these approaches are promising, they lack theoretical structure, which is especially pronounced on the levels of feature extraction and classifier architecture. This limitation becomes problematic considering the limited size (or quality) of most datasets used in the field.

Psychometric Machine Learning. Psychometric Machine Learning leverages classical ML and Deep Learning methods to indirectly study human behavior, characteristics and psychological states. This can include top-down (i.e., theory driven) and bottom up approaches, analyzing both multi-modal and semi-modal (e.g. textual) data [see. 4].

Notably, theory-driven Psychometric NLP can employ textual embedding based methods, which allow for operationalizing psychological constructs in vector spaces – akin to traditional Psychometric methods [see. 13]. One of those methods – Contextualized Construct Representations (CCR) [1] – focuses on utilizing similarity scores of sentence transformer embeddings between questionnaire scale items and the examined text. The method outperforms traditional text vectorization methods (i.e., Distributed Dictionary Representations and word-counting) in predicting psychological constructs and has the ability to capture nuanced context, while retaining the theoretically meaningful structure of the resulting vector space.

2 Neo-Psychometric Machine Learning

In the present paper, we introduce the Neo-Psychometric Machine Learning approach. We define it as the use of Psychometric measures and domain knowledge for the broader purpose of addressing Machine Learning tasks. In the case of NLP, we focus solely on analyzing text itself, which we treat as stimuli that can reflect a theoretical construct. We do this regardless of the authors intention. In contrast to Psychometric NLP, we are explicitly not interested in the psychological state, or characteristics, of the authors of the text. We posit that this framing is very well aligned with the goal of hate speech classification.

Herein, we utilize CCR – similarity metrics of the classified text and relevant psychometric scale item embeddings produced by a sentence transformer model – as intermediate features for such traditional classification models as logistic regression, Support Vector Machine and XGBoost. We test two theoretical approaches using our operationalization of the transmisogyny theory – Gender Essentialism, Social Determinism, and Ambivalent Sexism – one including the explicit transphobia scales (ATTM and ATTW) and one based solely on the constructs stemming from Transfeminist Theory.

3 Methods

The proposed Anti-Trans Hate Speech Classifier architecture assumes trans-related posts as inputs. We utilize the pre-trained all-MiniLM-L12-v2 model

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_41

(33.4M parameters, 384-dim output vector) [11], and the associated tokenizer to produce the text embeddings. We then calculate the CCR – dot product to the questionnaire items and use one of the three classifiers: Logistic Regression, XGBoost and Support Vector Machine.

Because all-MiniLM-L12-v2 normalizes the embedding vectors, in the present paper, we use dot product as the similarity metric, as, for normalized vectors, dot product is equivalent to cosine similarity. Higher scores in dot product reflect higher similarity of the post to a particular questionnaire item. Because, traditionally, in psychometric scales some items are inverted, we multiply the dot product by -1 for those items during feature calculation. This should not significantly impact model training, but it streamlines scale-wise interpretations of the similarity metrics – both in the case of dataset exploration and possibly in model-level and decision-level explainability.

Theory – Transmisogyny, used questionnaires. We used two operationalizations of the Transmisogyny theory [12]. We chose this theory, because of its potential to capture intersectional prejudice towards transgender women, comprehensive explanation of anti-trans prejudice and a number of empirical studies that align with it [e.g., 8, 2].

Serano [12] posits that prejudice towards trans people, especially trans women, is rooted in Oppositional Sexism (a construct well mirrored by psychological construct of Gender Essentialism), Traditional Sexism, Cissexism – the belief that the gender identity of cis people is more valid or real than the gender identity of trans people, and Transphobia – an irrational fear, aversion to, or discrimination against people, whose gender identities or expression deviates from societal norms. We operationalize those constructs using scales grounded in Psychological literature:

1. Gender Essentialism scale (18 items) [8] (used subscales: Inductive Potential, Biological Determinism, Immutability) to operationalize Oppositional Sexism.
2. Social Determinism scale (6 items) [8], as an additional dimension connected to Gender Essentialism, which is in line with contemporary feminist discourse.
3. Ambivalent Sexism Questionnaire (22 items) [6] – capturing Hostile and Benevolent aspects of sexist beliefs and narratives.
4. Attitudes Towards Transgender Men and Women Questionnaire (24 items) [3] as a direct measures of transphobia directed at binary trans people.

Dataset. In the present paper, we used the TIDEs dataset as it reflects the wide array of transphobic narratives, multi-platform sourcing and high quality of annotation. Therefore, we posit that the dataset is suitable for the examination of the proposed architecture, as well as for the training and, especially, evaluation of anti-trans hate speech models.

Figure 1 presents the distributions of total scale scores (CCR sum, not standardized). Note that the relationships seen in the TIDEs dataset plots are pronounced and align with the chosen theory.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_41

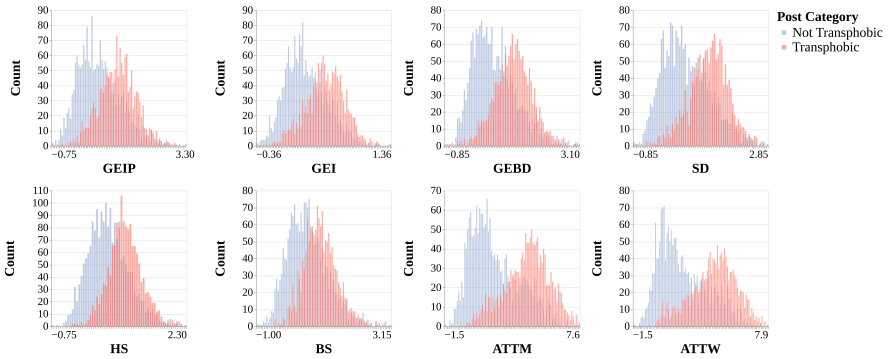

Fig. 1. Distributions of total scale scores by post category. GEIP : Gender Essentialism – Inductive Potential; GEI : Gender Essentialism – Immutability; GEBD : Gender Essentialism – Biological Determinism; SD : Social Determinism; HS : Hostile Sexism; BS : Benevolent Sexism; ATTM : Attitudes Towards Transgender Men; ATTW : Attitudes Towards Transgender Women.

4 Results

All classifier variants performed well in the task of classifying transphobic posts. SVM classifier achieved an F1 score of 0.79 (Accuracy = 0.81, Precision = 0.81, Recall = 0.76) over one run with a train-test split of 80:20. The XGBoost classifier achieved an F1 score of 0.78 (Accuracy = 0.81, Precision = 0.80, Recall = 0.76) over one run with a train-validation-test split with 80:10:10 ratio and early stopping. Logistic Regression achieved an F1 score of 0.75 (Accuracy = 0.79, Precision = 0.77, Recall = 0.74) over one run with a test-train split of 80:20. Table 1 shows comparisons to the models trained by Laimeiro et al. [7]. We argue that those results are comparable to the performance of a DeBERTa black box classifier [7] (Accuracy = 0.82, F1 = 0.81), while reducing the risk of overfitting on construct-unrelated features and providing potential for theory-driven explanations.

Table 1. Comparison of the Neo-Psychometric Model (NPM) to other classifiers.

Model	Accuracy	Precision	Recall	F1	Source
Perspective API (toxicity)	0.63	0.91	0.59	0.72	TIDEs
Perspective API (identity attack)	0.71	0.80	0.68	0.74	TIDEs
Logistic Regression	0.79	0.87	0.70	0.77	TIDEs
DeBERTa	**0.82**	0.89	0.74	**0.81**	TIDEs
NPM – Logistic Regression	0.79	0.77	0.74	0.75	Herein
NPM – XGBoost	0.81	0.80	0.76	0.78	Herein
NPM – SVM[1]	**0.81**	0.81	0.76	**0.79**	Herein

[1] Support Vector Machine.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_41

Alternative operationalization. An alternative operationalization stemming solely from the theoretical predictors of transmisogyny was tested in order to validate the approach and reduce potential construct-related bias. In the TIDEs dataset, the word "transgender" is more common in the Transphobic ($n = 895$), than in the Non-Transphobic ($n = 152$) category and has the highest mutual information with post category. Moreover, the discrepancy is also seen for words referring to men and women. While this can plausibly reflect real world differences in frequency distributions of these words, it can also introduce construct-related bias to any model trained on this dataset (i.e., classifying posts mentioning transgender people or gender as Transphobic). This is problematic in black box models and – in our approach – with the use of the ATTMW scales, because all items from both scales contain the phrases "transgender men" and "transgender women". This could potentially result in higher similarity of posts containing these words to all ATTMW scale items. Our approach allowed for reducing the risk of this bias by excluding the ATTM and ATTW scales. Even after this change, the models retained high performance metrics (see Table 2).

Table 2. Neo-Psychometric Model Performance over 100 runs with random 80:20 train-test splits.

Model	Accuracy	Precision	Recall	F1	ATTMW[1]
NPM – Logistic Regression	0.79	0.75	0.75	0.75	Yes
NPM – XGBoost[2]	0.79	0.75	0.76	0.75	Yes
NPM – SVM[3]	0.79	0.75	0.77	0.76	Yes
NPM – Logistic Regression	0.78	0.75	0.74	0.74	No
NPM – XGBoost[2]	0.78	0.74	0.75	0.75	No
NPM – SVM[3]	0.79	0.74	0.77	0.76	No

[1] ATTM and ATTW scales.
[2] No early stopping.
[3] Support Vector Machine.

5 Discussion

The present study introduced the notion of Neo-Psychometric Machine Learning. We showed that this approach can be successfully applied in an NLP task – classification of anti-trans hate speech. Utilizing CCR of psychometric scale items, we have represented unstructured textual data in a vector space with theoretically meaningful dimensions. Because the feature space becomes theoretically meaningful, we posit that this is more than simple dimensionality reduction.

Moreover, this approach does not require heavy compute power, as it does not involve fine-tuning of the sentence transformer model. All of the computations reported in the present paper were run on CPU (11th Gen Intel Core i5-1135G7 (8), 4.2GHz). Furthermore, because the scales used were already psychometrically validated, we could avoid the process of stimuli labelling.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_41

Explainability. A notable advantage of the proposed approach is its potential for improving explainability on every stage of a classifiers life cycle. As shown in Figure 1 the intermediate features can be used for dataset exploration in a theoretically structured manner. For instance, it is informative, that transphobic posts score higher on the Social Determinism scale than other trans-related posts in a particular dataset. Notably, this can be used to evaluate theoretical fit of a dataset prepared for a specific purpose. In our case, we would argue that the TIDEs dataset reflects the theoretical aspects of anti-trans prejudice very well.

Moreover, we argue that the Neo-Psychometric approach is very well suited for model-level and decision-level explainability. Using this approach, researchers can define the conceptual structure of the analyzed construct (e.g., anti-trans prejudice) themselves – instead of delegating the theoretical work to a machine learning model. This further allows for reducing the hard problem of explaining how transformers represent a complex construct to two simpler ones – sentence similarity, and explaining statistical relationships between the Neo-Psychometric dimensions and the predicted class label.

In other words, using this approach we take the conceptual work of defining what Transphobia is away from the model, giving the control over representing the construct in question back to the researcher and leaving the computation to the machine learning implementation.

Limitations. While the Neo-Psychometric Classification approach can, in our view, help reduce both construct-unrelated and construct-related bias, this requires robust dataset evaluation and theoretical examination. Moreover, it is still too early to tell how effective it is at reducing bias overall, and especially in real-world applications. A notable limitation of using psychometric scale items as references for theory-driven dimensionality reduction is the fact that similarity to questionnaire items might overall reflect not only the intended construct, but also the "questionnaire-ness" of the classified texts. This has been mentioned by [13] who propose the use of Correlational Anchored Vectors to ameliorate this problem in the context of Psychometric NLP.

Furthermore, as seen in Table 1, our approach reduced the classifiers performance metrics, however not drastically. We propose that future research should employ fine tuning alongside Neo-Psychometric feature extraction in order to maximize model performance and improve usability.

Conclusions. While the Neo-Psychometric approach would, in our view, not be best suited in applications where the aim is to maximize accuracy, profit, or other performance metrics, we believe it can be applied in situations where theoretical backing, explainability and reducing construct-unrelated model bias is crucial – with good results.

Acknowledgments. We thank Adam Zadrożny, Wojciech Przytuła, Joanna Rączaszek-Leonardi, Wiktor Rorot and Szymon Rynkun for making it possible for Sofia to write significant portions of the code used in the current project during their seminars, as well as for their valuable insights.

Disclosure of Interests. The authors declare no competing interests.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_41

References

[1] Atari, M., Omrani, A., Dehghani, M.: Contextualized Construct Representation: Leveraging Psychometric Scales to Advance Theory-Driven Text Analysis (2023)
[2] Atwood, S., Morgenroth, T., Olson, K.R.: Gender essentialism and benevolent sexism in anti-trans rhetoric. Social Issues and Policy Review **18**(1), 171–193 (2024)
[3] Billard, T.J.: Attitudes Toward Transgender Men and Women: Development and Validation of a New Measure. Frontiers in Psychology **9** (2018)
[4] Celli, F., Kartelj, A., Dordevič, M., Suhartono, D., Filipovič, V., Milutinovič, V., Spathoulas, G., Vinciarelli, A., Kosinski, M., Lepri, B.: Twenty Years of Personality Computing: Threats, Challenges and Future Directions. ACM Computing Surveys p. 3806009 (2026)
[5] Channon, L., Mathieson, N.: Automated Detection of Mainstreamed Transphobic Content on YouTube (2025). https://doi.org/10.57814/49JZ-0663
[6] Glick, P., Fiske, S.T.: Hostile and Benevolent Sexism: Measuring Ambivalent Sexist Attitudes Toward Women. Psychology of Women Quarterly **21**(1), 119–135 (1997)
[7] Lameiro, F., Dunagan, L., Card, D., Gilbert, E., Haimson, O.: TIDEs: A Transgender and Nonbinary Community-Labeled Dataset and Model for Transphobia Identification in Digital Environments. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. pp. 1411–1423 (2025)
[8] Lee, K.Y., Reis, H.T., Rogge, R.D.: Seeing the World in Pink and Blue: Developing and Exploring a New Measure of Essentialistic Thinking about Gender. Sex Roles **83**(11-12), 685–705 (2020)
[9] Lees, A., Tran, V.Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., Vasserman, L.: A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3197–3207 (2022)
[10] Leitner, M., Dorn, R., Morstatter, F., Lerman, K.: Characterizing Network Structure of Anti-Trans Actors on TikTok (2025)
[11] Reimers, N., Gurevych, I.: 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. pp. 3982–3992. Association for Computational Linguistics (2019)
[12] Serano, J.: Whipping girl: a transsexual woman on sexism and the scapegoating of femininity (2007)
[13] Teitelbaum, L., Simchon, A.: Neural text embeddings in psychological research: A guide with examples in R. Psychological Methods (2025)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_41