International Conference on Formal Methods
and Foundations of Artificial Intelligence
Eszterházy Károly Catholic University
Eger, Hungary, June 5–7, 2025

# Hungarian Case Study on Automated Detection of Body-Shaming Comments Using Machine Learning

**Franciska N. Gőz**[a], **Erika B. Varga**[b]

[a]Institute of Informatics, University of Miskolc
gozfanni@gmail.com

[b]Institute of Informatics, University of Miskolc
erika.b.varga@uni-miskolc.hu

## Abstract

Social media facilitates online interactions but also enables body-shaming comments which are often ambiguous. This paper presents a machine learning-based approach for detecting Hungarian body-shaming comments, an underrepresented area in NLP. A dataset of Facebook comments was collected and expanded with synthetic data. Using HuSpaCy and HuBERT, logistic regression and MLP classification models were trained with TF-IDF and SBERT embeddings. The best-performing model achieved 88% accuracy, demonstrating the potential of NLP techniques for moderating harmful online content in low-resource languages. The results highlight key challenges, including category overlap and class imbalance, emphasizing the need for context-aware classification methods in automated content moderation.

## 1. Introduction

Social networking sites offer remarkable opportunities for communication and connection, but at the same time they also serve as fertile ground for the spread of harmful behaviors. Damaging comments, encompassing various forms such as body shaming, hate speech, cyberbullying, and online harassment, have become increasingly widespread [7]. The widespread nature of these comments contributes to a toxic online environment, affecting not only individual well-being but also shaping

social attitudes and potentially fueling real-world discriminatory behaviors [8].

The exponential growth of user-generated content has facilitated the spread of such harmful language, posing serious problems in maintaining a respectful online environment [2]. The automatic detection and moderation of these harmful comments presents significant challenges due to the large amount of online content, the diversity of language, and the difficulty in distinguishing harmful intent from protected free speech [3].

Body shaming is defined in [10] as a type of negative social interaction which involves derogatory comments about an individual's physical appearance, often leading to diminished self-esteem, social withdrawal, and even mental health issues such as depression and eating disorders [6]. Social media platforms have attempted to address the spread of body-shaming content through community guidelines and moderation systems. However, these efforts are often insufficient due to the great volume of content and the ambiguous nature of body-shaming remarks, which can be disguised as humor [4]. The urgency to address this issue stems not only from its psychological impact but also from the legal and ethical obligations of these platforms to provide a safe environment for their users [1].

Effective interventions require a twofold approach. Social networking platforms should enhance their moderation systems with machine learning techniques to automatically detect and classify harmful content. These methods can help scale the identification and removal of body-shaming remarks, even when they are disguised [5]. On the other hand, comprehensive legal frameworks, such as the European Union's Digital Services Act (DSA) [9], should be established to hold platforms accountable for harmful content while balancing the principles of freedom of speech [11].

This study contributes to these efforts by exploring the development of a classification model for detecting body-shaming comments in Hungarian. By integrating machine learning techniques and considering the complexities of both negative and ambiguous remarks, the proposed solution aims to support the automated moderation systems of social media platforms.

# References

[1] G. Aitchison, S. Meckled-Garcia: *Against Online Public Shaming*, Social Theory and Practice (2020), doi: 10.5840/soctheorpract20201117109.

[2] A. Balayn, J. Yang, Z. Szlavik, A. Bozzon: *Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature*, Trans. Soc. Comput. 4.3 (2021), p. 11, doi: 10.1145/3479158.

[3] C. Gehweiler, O. Lobachev: *Classification of intent in moderating online discussions: An empirical evaluation*, Decision Analytics Journal 10 (2024), p. 100418, issn: 2772-6622, doi: 10.1016/j.dajour.2024.100418.

[4] V. Gongane, M. Munot, A. Anuse: *Detection and moderation of detrimental content on social media platforms: current status and future directions*, Soc. Netw. Anal. Min. 12 (2022), p. 129, doi: 10.1007/s13278-022-00951-3.

[5] H. Han, M. Asif, E. Awwad, N. Sarhan, Y. Y. Ghadi, B. Xu: *Innovative deep learning techniques for monitoring aggressive behavior in social media posts*, J Cloud Comp 13 (2024), p. 19, doi: 10.1186/s13677-023-00577-6.

[6] G. Holland, M. Tiggemann: *A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes*, Body Image 17 (2016), pp. 100–110, doi: 10.1016/j.bodyim.2016.02.008.

[7] E. A. Jane: *Online Abuse and Harassment*, in: The International Encyclopedia of Gender, Media, and Communication, Wiley & Sons, 2020, doi: 10.1002/9781119429128.iegmc080.

[8] M. Merino, J. F. Tornero-Aguilera, A. Rubio-Zarapuz, C. V. Villanueva-Tobaldo, A. Martín-Rodríguez, V. J. Clemente-Suárez: *Body Perceptions and Psychological Well-Being: A Review of the Impact of Social Media and Physical Measurements on Self-Esteem and Mental Health with a Focus on Body Image Satisfaction and Its Relationship with Cultural and Gender Factors*, Healthcare 12.14 (2024), doi: 10.3390/healthcare12141396.

[9] *Regulation (EU) 2022/2065 of the European Parliament and of the Council on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*, 2022.

[10] C. Schlüter, G. Kraag, J. Schmidt: *Body Shaming : an Exploratory Study on its Definition and Classification*, International Journal of Bullying Prevention 5 (2023), pp. 26–37, doi: 10.1007/s42380-021-00109-3.

[11] A. Turillazzi, M. Taddeo, L. Floridi, F. C. and: *The digital services act: an analysis of its ethical, legal, and social implications*, Law, Innovation and Technology 15.1 (2023), pp. 83–106, doi: 10.1080/17579961.2023.2184136.