Proceedings of the International Conference on
Formal Methods and Foundations of Artificial Intelligence
Eszterházy Károly Catholic University
Eger, Hungary, June 5–7, 2025
pp. 65–77
DOI: 10.17048/fmfai.2025.65

Hungarian case study on automated detection of body-shaming comments using machine learning

Franciska N. Gőz, Erika B. Varga

Institute of Information Technology,
Faculty of Mechanical Engineering and Informatics
University of Miskolc, H-3515 Miskolc, Hungary
gozfanni@gmail.com
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.

Keywords: Hungarian text analysis, toxic comment filtering, social media moderation, body-shaming detection, machine learning classification

AMS Subject Classification: 68T05, 68T07, 68T50

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 [15]. 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 [16].

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 [7].

Body shaming is defined in [25] 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 [11]. 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 [8]. 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 [10]. On the other hand, comprehensive legal frameworks, such as the European Union's Digital Services Act (DSA) [21], should be established to hold platforms accountable for harmful content while balancing the principles of freedom of speech [27].

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.

The novelty of this work lies primarily in the creation of the first annotated dataset of Hungarian body-shaming comments and in presenting a case study for using Hungarian text processing tools.

2. Related works

Body shaming has been widely recognized as a significant social issue, particularly in the context of social media. Schlüter et al. [24] conducted an exploratory study to provide a scientifically grounded definition and classification of body shaming. They define body shaming as an unrepeated act in which individuals express unsolicited, predominantly negative opinions about another person's body, often perceived negatively by the target. Importantly, their findings highlight the dimensional nature of body shaming, ranging from well-meant advice to overtly malicious

insults. The study emphasizes the distinction between body shaming and related concepts, such as appearance teasing, trolling, and cyberbullying, noting that body shaming is not necessarily repeated nor always intentional. Since body shaming is harmful to the victims' physical and mental health, numerous studies investigated its impacts, for example on women's body image concerns [19], on eating disorders [12, 23, 26], and on mental well-being [4].

Body shaming is increasingly prevalent on social media platforms [6]. While these platforms have their own community guidelines to regulate and remove harmful posts and comments, their policies are shaped by the legal framework of the countries in which they operate. In the United States, for example, the First Amendment of the Constitution guarantees broad freedom of speech, allowing individuals to express their opinions freely, even if they are offensive to others. However, specific cases like threats or defamation may still be restricted. Since major social media platforms like Facebook, Instagram, and Twitter are based in the U.S., they primarily adhere to these legal principles. While the platforms' community guidelines prohibit harassment, hate speech, and body-shaming content, their enforcement is not legally mandated but rather a voluntary application of their policies. According to the community standards of Facebook and Instagram, hate speech and harassment, including content targeting individuals based on their physical appearance are prohibited. Body shaming explicitly falls under this category, and such posts are subject to removal. Twitter's rules also ban behavior that harasses or intimidates others, including body shaming. The platform takes actions to suspend or ban offending users.

The U.S. prioritizes freedom of speech as a fundamental right, emphasizing minimal government intervention in online content moderation. This framework allows platforms significant authority in enforcing their policies without strict government oversight. In contrast, the European Union, through the DSA, imposes stricter obligations on social media platforms to tackle harmful content, including hate speech and harassment. The DSA emphasizes protecting users from harmful online behavior, requiring platforms to remove illegal content promptly, increase transparency in content moderation policies, and offer effective appeal mechanisms for users whose content is removed [3].

In this context, investigations on classifying comments on body shaming by means of machine learning algorithms seems to bring benefits for the users of social networking sites. What makes the task more complicated is to identify a writer's real intent. Body-shaming remarks are often intended as humor from the speaker's perspective or expressed through ambiguous language. This subjectivity makes it difficult, even for humans, to determine whether a particular comment falls into the category of body shaming. Consequently, the automatic detection of such remarks presents a significant challenge [28].

Some recent studies concentrate on detecting and analyzing body-shaming online comments. In [5] sentiment analysis is performed on Twitter comments using Naive Bayes Classifier. The dataset consists of 1,000 training tweets and 329 test tweets. The algorithm achieved an accuracy of 80.55 % and highlighted key words

like “overweight” and “thin”. The research underlines the potential of Naive Bayes for effective classification of negative and positive sentiments, though it acknowledges challenges in dealing with context and linguistic issues.

Jaman et al. [14] use the Naive Bayes Classifier to detect body-shaming sentiments in comments on YouTube beauty vlogs. The study involves 33,044 comments, of which 986 were labeled as body-shaming after manual validation. The model achieved a high accuracy of 98.48 %, demonstrating the effectiveness of tailored preprocessing and dataset splits.

Grasso et al. [9] examine the detection and classification of body shaming content in Italian social media applying contextual word embeddings and transformer-based architectures. The study highlights challenges in multilingual settings, such as data scarcity and linguistic diversity, and proposes advanced methods to overcome these.

Reddy et al. [20] explore body shaming online content detection in low-resource languages, using transfer learning and multilingual embeddings to bridge the gap caused by limited datasets. The authors show that pre-trained language models like `mBERT` can improve classification performance when fine-tuned on small, annotated corpora.

In our experiments, we collected comments from Facebook and supplemented them with synthetic texts to develop and compare machine learning models for classifying comments about an individual’s physical appearance into six categories. Five of these classes fall within the domain of body shaming, such as negative opinions on overweight, underweight, height, skin tone and body hair; while the sixth class is to contain sentiments on physical traits that are not considered as body shaming. Our experiments focus on Hungarian, an underrepresented language in social media. To the best of our knowledge, this is the first scientific paper that addresses the filtering of Hungarian-language texts on this specific topic.

3. Methods and data

3.1. Hungarian text processing tools

For preprocessing we used `HuSpaCy` [18], a Hungarian adaptation of the `spaCy` library providing tokenization, lemmatization, part-of-speech tagging, dependency parsing, and named entity recognition. `HuSpaCy` has been optimized for Hungarian linguistic resources and offers efficient pipelines suitable for machine learning applications.

For embeddings we employed `HuBERT` [17], a transformer-based language model following the `BERT` architecture [13] and trained on `Webkorpusz` 2.0, a large Hungarian corpus. To obtain sentence-level representations, we applied the `Sentence-BERT` (`SBERT`) approach [22] on top of `HuBERT` outputs. This combination yields rich contextual embeddings that have proven effective in Hungarian `NLP` tasks.

These tools ensured reliable preprocessing and rich representations for classification while keeping the methodology comparable with prior work in low-resource

language settings.

3.2. Data collection and synthetic data generation

The initial dataset was translated from comments collected from Facebook. The focus was on comments related to body-shaming, which were identified and annotated manually by one of the authors, based on her interpretation of the content. These comments represented diverse categories of harmful speech aimed at various physical traits, such as weight, height, body hair, and skin tone. Since annotation was carried out by a single annotator, we acknowledge that subjectivity may affect the labeling. In particular, ambiguous cases sometimes arise, for example when a comment such as *"Hűha, jó sokat fogytál!" – "Wow, you've lost a lot of weight!"* could be interpreted either as a positive remark (non-toxic) or as a negative body-related comment (skinny). The annotated comments were finally organized into five predefined categories:

- Body hair
  (e.g., *"Úgy néz ki a lábad, mintha évek óta nem találkozott volna borotvával." – "Your legs look like they haven't seen a razor in years."*)
- Height
  (e.g., *"Gondolom, nálad nem opció a legfelső polcra pakolás." – "I bet you can't even reach the top shelf."*)
- Fat (e.g., *"Toka az egész fejed." – "Your double chin is your entire face."*)
- Skinny (e.g., *"Lapos vagy, mint egy deszka." – "You're flat as a board."*)
- Skin tone
  (e.g., *"Annyira világos a bőröd, hogy egy nyaralás után is alig látszik változás." – "Your skin is so pale, even after a vacation, there's no difference."*)

To enrich the dataset and increase its diversity, synthetic comments were generated using ChatGPT-4. In the prompts, we specified the target category (e.g., body hair, skin tone), instructed the model to generate a short (1-2 sentences) negative comment in an informal style resembling Facebook posts, and included examples taken from our manually annotated data. Approximately 5 % of the final dataset consists of such synthetic comments. We did not evaluate model performance separately on real and synthetic comments due to the small dataset size, but the synthetic data was only used to increase diversity and balance across categories, not to replace authentic user comments.

The dataset was further augmented with synthetic statements that express positive sentiments regarding a person's physical appearance. These comments were labeled as the "non-toxic" category, e.g. *"Nagyon szép a bőröd ezen a fotón, természetes ragyogás!" – "Your skin looks beautiful in this photo, with a natural glow!"*. In this way, we can investigate whether positive sentiments can be clearly distinguished from negative ones through classification.

3.3. Dataset statistics

The dataset comprises 330 labeled comments with approximately equal representation across the six categories. On average, there are 55 comments per category, amounting to a balanced dataset for training purposes. The detailed statistics are presented in Tables 1 and 2.

Table 1. Dataset statistics before tokenization.

Label	**Number of comments**	**Num. of words**				**Num. of chars**			
		Avg	Std	Min	Max	Avg	Std	Min	Max
bodyhair	63	11.25	2.61	5	18	69.86	13.43	32	102
height	63	10.13	4.76	5	25	65.17	29.25	26	144
fat	55	8.16	3.41	2	16	48.87	23.02	11	102
skinny	64	8.75	2.36	2	16	52.28	15.8	12	110
skintone	35	10.14	1.83	7	14	61.97	10.52	42	85
non-toxic	50	8.92	1.12	6	11	56.16	6.22	47	70
Total	**330**								
Average	**55**	9.56	2.68	4.5	16.67	59.05	16.37	28.33	102.17

Table 2. Dataset statistics after tokenization.

Label	**Number of comments**	**Num. of words**				**Num. of chars**			
		Avg	Std	Min	Max	Avg	Std	Min	Max
bodyhair	63	5.43	1.35	2	9	37.01	8.48	17	62
height	63	5.05	2.05	2	11	33.48	14.81	14	86
fat	55	3.98	1.62	1	8	25.82	12.48	4	59
skinny	64	4.27	1.54	1	10	26.69	10.41	8	65
skintone	35	4.94	1.45	3	8	31	9.09	15	51
non-toxic	50	4.72	0.97	3	7	34.64	8.11	17	59
Total	**330**								
Average	**55**	4.73	1.5	2	8.83	31.44	10.57	12.5	63.67
Total	**330**								
Average	**55**	9.56	2.68	4.5	16.67	59.05	16.37	28.33	102.17

4. Results

The aim of this research is to build a model that can classify social media comments into predefined categories. The first step was to prepare the collected data before text processing. This involved tokenizing and lemmatizing words, and then tranforming sentences into numerical vectors using either `TF-IDF` vectorization or `SBERT` embeddings. Next, the data set was randomly split into training and test sets, and the training set was balanced by `SMOTE` technique (Synthetic Minority Oversampling Technique). Two models were considered for the special purpose: a simple logistic regression model and an MLP neural network. The neural network model was configured with 3 hidden layers containing 100, 50, and 25 neurons respectively; applies ReLU activation function and was optimized using the Adam

algorithm with a maximum of 300 iterations. Finally, four different setups were investigated:

1. Logistic regression model (sentences prepared with `TF-IDF` vectorization)
2. Logistic regression model using `SBERT`
3. `MLP` model (sentences prepared with `TF-IDF` vectorization)
4. `MLP` model using `SBERT`

We applied 5-fold cross-validation for all four model configurations in order to compensate for the randomness of a single train-test split and to ensure more generalizable performance estimates. In each configuration we report the best-performing results (highest accuracy) obtained across the folds.

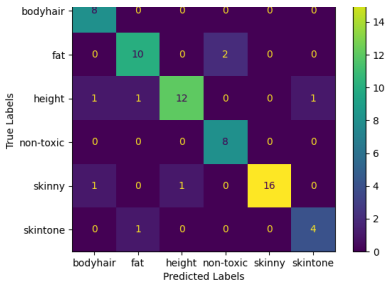
4.1. Logistic Regression model

Logistic regression is a widely used model for text classification tasks due to its simplicity and interpretability. It is particularly effective for categorizing text data into multiple predefined classes, even when only a limited amount of data is available. This makes it a useful algorithm for handling smaller datasets.

Logistic regression

	precision	recall	f1-score	support
bodyhair	0.80	1.00	0.89	8
fat	0.83	0.83	0.83	12
height	0.92	0.80	0.86	15
non-toxic	0.80	1.00	0.89	8
skinny	1.00	0.89	0.94	18
skintone	0.80	0.80	0.80	5
accuracy			0.88	66
macro avg	0.86	0.89	0.87	66
weighted avg	0.89	0.88	0.88	66

(a) Evaluation.



(b) Confusion matrix.

Figure 1. Results for the Logistic regression model with `TF-IDF`.

As depicted in Figure 1a and 1b the model achieved an overall accuracy of 88 %, correctly classifying 88 % of all examples. The macro average represents the average performance across all classes (precision: 0.86, recall: 0.89, F1-score: 0.87), balancing the influence of smaller and larger classes. The weighted performance average, which considers class support, yielded approximately 0.88 for all key metrics. Notably, despite their low support, the "bodyhair" and "non-toxic" categories achieved 100 % recall, meaning they were identified without errors. On the other hand, the "skinny" class with the highest support was predicted with 100 % precision, meaning that no other categories were misclassified here. The model produced the worst performance in the case of the "skintone" class, which has the lowest support in the test set and the lowest representation in the whole dataset.

4.2. Logistic Regression model using SBERT

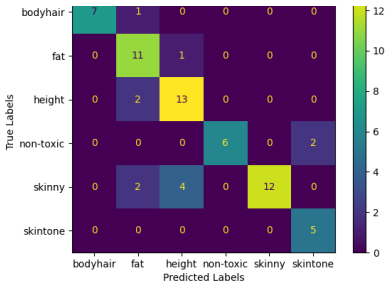
The goal of SBERT is to accurately measure the semantic similarity between sentences. It generates sentence-level embeddings, which provide a numerical representation of sentence semantics, thereby enabling the evaluation of multi-sentence sentiments. The results of the logistic regression model using SBERT embeddings are summarized in Figure 2a and 2b.

The model achieved an accuracy of 82 %, lower than that of traditional logistic regression. The reason for this is that the dataset contains mainly one-sentence long comments and the SBERT technique does not have significant effect here. Nevertheless, the model produced 100 % precision for the class with the highest support ("skinny") and also for the classes having low support ("bodyhair" and "non-toxic"). It is notable, that the model could produce 100 % recall in the case of the "skintone" category, which has the lowest support in the test set and the lowest representation in the whole dataset.

SBERT - Logistic regression

	precision	recall	f1-score	support
bodyhair	1.00	0.88	0.93	8
fat	0.69	0.92	0.79	12
height	0.72	0.87	0.79	15
non-toxic	1.00	0.75	0.86	8
skinny	1.00	0.67	0.80	18
skintone	0.71	1.00	0.83	5
accuracy			0.82	66
macro avg	0.85	0.85	0.83	66
weighted avg	0.86	0.82	0.82	66

(a) Evaluation.



(b) Confusion matrix.

Figure 2. Results for the Logistic Regression model with SBERT embeddings.

4.3. MLP model

The Multi-Layer Perceptron (MLP) is a neural network model adaptable to various problems through optimization of the number of hidden layers, neurons, and learning parameters. This flexibility allows it to perform effectively on text classification tasks. For our data set, the model achieved an accuracy of 85 %, which is a bit lower than the performance of logistic regression. The lower performance may be due to the small dataset size. The detailed evaluation is displayed in Figure 3a and 3b.

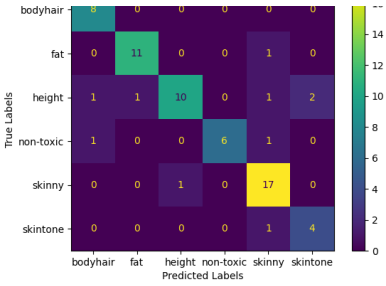
In comparison with logistic regression, this model could not predict the class with the highest support ("skinny") with 100 % precision or recall. On the other hand, it produced 100 % precision in the case of the "non-toxic" class and 100 % recall in the case of the "bodyhair" class, though these have low support in the test set. What is common with logistic regression, the MLP model produced the worst

performance in the case of the "skintone" class, which has the lowest support in the test set and the lowest representation in the whole dataset.

MLP

	precision	recall	f1-score	support
bodyhair	0.80	1.00	0.89	8
fat	0.92	0.92	0.92	12
height	0.91	0.67	0.77	15
non-toxic	1.00	0.75	0.86	8
skinny	0.81	0.94	0.87	18
skintone	0.67	0.80	0.73	5
accuracy			0.85	66
macro avg	0.85	0.85	0.84	66
weighted avg	0.86	0.85	0.85	66

(a) Evaluation.



(b) Confusion matrix.

Figure 3. Results for the `MLP` model with `TF-IDF`.

4.4. `MLP` model using `SBERT`

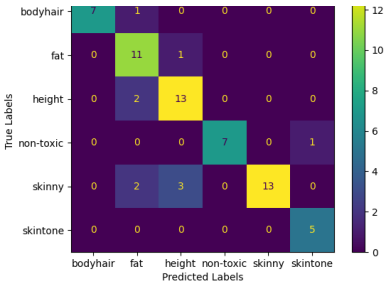
Figure 4a and 4b summarize the results of the `MLP` model using `SBERT` sentence embeddings. The model achieved an overall accuracy of 85 %, which is the same as in the case of the traditional `MLP` model.

The use of `SBERT` yields some similar results than in the case of the logistic regression model. Namely, the model produced 100 % precision for the class with the highest support ("skinny") and also for the classes having low support ("bodyhair" and "non-toxic"). Also, the model could produce 100 % recall in the case of the "skintone" category, which has the lowest support in the test set and the lowest representation in the whole dataset. `SBERT` performs more effectively with `MLP`, as its F1-scores match or exceed those of logistic regression across all categories.

SBERT - MLP

	precision	recall	f1-score	support
bodyhair	1.00	0.88	0.93	8
fat	0.69	0.92	0.79	12
height	0.76	0.87	0.81	15
non-toxic	1.00	0.88	0.93	8
skinny	1.00	0.72	0.84	18
skintone	0.83	1.00	0.91	5
accuracy			0.85	66
macro avg	0.88	0.88	0.87	66
weighted avg	0.88	0.85	0.85	66

(a) Evaluation.



(b) Confusion matrix.

Figure 4. Results for the `MLP` model with `SBERT` embeddings.

To provide a more transparent overview of the model performance across classes, Table 3 summarizes accuracy, macro-F1 scores and the support of each class in the

test set. This highlights the impact of small-sample categories on model performance.

Table 3. Summary of evaluation metrics across models.

Model	**Accuracy**	**Macro-F1**	**Support (per class)**
LogReg (`TF-IDF`)	0.88	0.87	[8, 12, 15, 8, 18, 5]
LogReg (`SBERT`)	0.82	0.83	[8, 12, 15, 8, 18, 5]
MLP (`TF-IDF`)	0.85	0.84	[8, 12, 15, 8, 18, 5]
MLP (`SBERT`)	0.85	0.87	[8, 12, 15, 8, 18, 5]
Ensemble model	**0.83**	**0.86**	**66** (incl. all classes)

5. Conclusions

Body shaming on social media platforms is a significant social and technological challenge. Despite existing moderation efforts, the high volume of content and the complexity of body-shaming remarks make automated detection difficult. This study demonstrates the potential of machine learning techniques in classifying body-shaming comments, particularly in the Hungarian language, which is underrepresented in such research.

The research findings demonstrate that `HuBERT` and `HuSpaCy` are highly effective tools for analyzing Hungarian texts. Their integration enabled the development of efficient classification models, even when working with a small corpus. Logistic regression models offered simplicity and interpretability, achieving an accuracy of up to 88 %. At the same time, `MLP` models utilizing `SBERT` embeddings provided enhanced flexibility in handling ambiguous or complex linguistic context. The models performed particularly well in distinguishing harmless content (the "non-toxic" class) and identifying body-shaming related to body hair, though challenges remained with underrepresented categories like "skin tone".

These findings emphasize the need for tailored machine learning techniques in social media moderation. Future research should focus on addressing class imbalance, expanding datasets with diverse linguistic and cultural contexts, and exploring advanced deep learning architectures to further improve classification accuracy.

6. Limitations and future work

The main limitations of this study are the small dataset size, the underrepresentation of certain categories, and the inherent overlap between them (e.g., skinny vs. fat). These issues are reflected in typical errors, such as misclassifying "Your skin is so pale, even after a vacation there's no difference." as non-toxic instead of skin tone, or labeling "You're flat as a board." as fat instead of skinny. Another limitation is that annotation was carried out by a single annotator, which may introduce

subjectivity, and that comments were analyzed in isolation without conversational context, potentially obscuring pragmatic signals (e.g., irony, humor). Due to limited dataset size and computational constraints, we did not include multilingual transformer baselines such as `mBERT` or `XLM-R` in the current study. Future work should focus on enlarging the dataset, improving class balance, and applying more context-aware models to better handle ambiguous or ironic expressions.

7. Ethical and legal compliance

All data were collected from publicly available Facebook comments. Personally identifiable information was removed during preprocessing to ensure full anonymity. The released dataset contains only anonymized text and complies with the ethical standards for research on social media content as well as the European Union's GDPR regulations.

8. Code and data availability

The code used for training and evaluation (requires `Python` 3.10.0, `spacy` 3.7.4, `huspacy` 0.12.1, and `sentence_transformers` 3.2.1) as well as the anonymized dataset of Hungarian body-shaming comments are openly available at `https://github.com/Fanni98/Diplomaterv`.

References

[1] G. Aitchison, S. Meckled-Garcia: *Against Online Public Shaming: Ethical Problems with Mass Social Media*, Social Theory and Practice 47.1 (2021), pp. 1–31, URL: https://www.jstor.org/stable/45378050.

[2] A. Balayn, J. Yang, Z. Szlávik, A. Bozzon: *Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature*, ACM Transactions on Social Computing 4.3 (2021), pp. 1–56, DOI: 10.1145/3479158.

[3] A. Candeub: *The Digital Services Act, the First Amendment, and deputised surveillance*, Journal of Media Law 16.1 (2024), pp. 65–73, DOI: 10.1080/17577632.2024.2362479.

[4] F. Deviantony, Y. Fitria, R. Rondhianto, N. Pramesuari: *An in depth review of body shaming phenomenon among adolescent: Trigger factors, psychological impact and prevention efforts*, South African Journal of Psychiatry 30 (2024), p. 2341, DOI: 10.4102/sajpsychiatry.v30i0.2341.

[5] K. Diantoro, Rinaldo, A. Sitorus, A. Rohman: *Analyzing the Impact of Body Shaming on Twitter: A Study Using Naive Bayes Classifier and Machine Learning*, Digitus: Journal of Computer Science Applications 1.1 (2023), pp. 11–25, DOI: 10.61978/digitus.v1i1.58.

[6] G. Fioravanti, S. B. Benucci, V. Vinciarelli, S. Casale: *Body shame and problematic social networking sites use: the mediating effect of perfectionistic self-presentation style and body image control in photos*, Curr Psychol 43 (2024), pp. 4073–4084, DOI: 10.1007/s12144-023-04644-8.

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

[8] V. Gongane, M. Munot, A. Anuse: *Detection and moderation of detrimental content on social media platforms: current status and future directions*, Social Network Analysis and Mining 12.1 (2022), p. 129, doi: 10.1007/s13278-022-00951-3.

[9] F. Grasso, A. Valese, M. Micheli: *Body-Shaming Detection and Classification in Italian Social Media*, in: In Natural Language Processing and Information Systems, 2024, doi: 10.1007/978-3-031-70239-6_18.

[10] H. Han, E. A. M. Asif, N. Sarhan, Y. Ghadi, B. Xu: *Innovative deep learning techniques for monitoring aggressive behavior in social media posts*, Journal of Cloud Computing 13.19 (2024), doi: 10.1186/s13677-023-00577-6.

[11] 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.

[12] A. Hummel, A. Smith: *Ask and you shall receive: Desire and receipt of feedback via Facebook predicts disordered eating concerns*, International Journal of Eating Disorders 48.4 (2015), pp. 436–442, doi: 10.1002/eat.22336.

[13] D. Jacob, M. Chang, K. Lee, K. Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, North American Chapter of the Association for Computational Linguistics, 2019.

[14] J. Jaman, H. Hannie, M. Simatupang: *Sentiment Analysis of the Body-Shaming Beauty Vlog Comments*, in: In Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, 2019, doi: 10.4108/eai.12-10-2019.2296530.

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

[16] M. Merino, J. Tornero-Aguilera, A. Rubio-Zarapuz, C. V. Tobaldo, A. Martín-Rodríguez, V. 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, doi: 10.3390/healthcare12141396.

[17] D. Nemeskey: *Natural Language Processing methods for Language Modeling*, PhD thesis, Eötvös Loránd University, 2020.

[18] G. Orosz, G. Szabó, P. Berkecz, Z. Szántó, R. Farkas: *Advancing Hungarian Text Processing with HuSpaCy Efficient and Accurate NLP Pipelines*, Speech, and Dialogue 14102 (2023), doi: 10.1007/978-3-031-40498-6_6.

[19] A. Raspovic, I. Prichard, A. Salim, Z. Yager, L. Hart: *Body image profiles combining body shame, body appreciation and body mass index differentiate dietary restraint and exercise amount in women*, Body Image 46 (2023), pp. 117–122, doi: 10.1016/j.bodyim.2023.05.007.

[20] V. Reddy, H. Abburi, N. Chhaya, T. Mitrovska, V. Varma: *You Are Big, S/he Is Small Detecting Body Shaming in Online User Content*, in: In: Social Informatics, 2022, doi: 10.1007/978-3-031-19097-1_25.

[21] *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)*, Official Journal of the European Union, 2022.

[22] N. Reimers, I. Gurevych: *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, in: Conference on Empirical Methods in Natural Language Processing, p. 2019.

[23] S. Santarossa, S. Woodruff: *SocialMedia: Exploring the Relationship of Social Networking Sites on Body Image, Self-Esteem, and Eating Disorders*, Social Media + Society 3 (2017), doi: 10.1177/2056305117704407.

[24] C. Schlüter, G. Kraag, J. Schmidt: *Body Shaming: an Exploratory Study on its Definition and Classification*, International Journal of Bullying Prevention 5 (2021), doi: 10.1007/s42380-021-00109-3.

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

[26] K. Suhag, S. Rauniyar: *Social Media Effects Regarding Eating Disorders and Body Image in Young Adolescents*, Cureus 16.4 (2024), doi: 10.7759/cureus.58674.

[27] A. Turillazzi, M. Taddeo, L. Floridi, F. Casolari: *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.

[28] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, H. Margetts: *Challenges and frontiers in abusive content detection*, in: In Proceedings of the Third Workshop on Abusive Language Online, pp. 88–93, url: https://aclanthology.org/W19-3509.pdf.