Emotion Intensity Assessment in Eating Disorders: A Polish BERT-Based Study

Anonymous ACL submission

Abstract

Expressing emotions directly can be a challenge, especially for those suffering from eating disorders. In such cases, identifying emotions can be very difficult for a psychologist, while emotion regulation is a key action supporting. There is a gap in the use of emotion recognition solutions in Polish. This article presents an innovative approach that is based not on the identification of emotions but on the estimation of their intensity in notes written in Polish. The solution uses Polbert - Polish language model BERT and NAWL dictionary – collection of words described in terms of emotion. The whole algorithm is based on dynamic coded word embeddings from Polbert and searching in this space the NAWL neighbors, which influence emotion intensities words in and out of the dictionary. The results demonstrate a difference between the emotional profiles of patients suffering from eating disorders and healthy ones despite the model used, cased or uncased. Detailed results of people suffering from eating disorders showed higher values in more difficult emotions such as sadness, anger, fear, and disgust than in healthy people. At the same time, happiness was higher among healthy people than among people with disorder.

1 Introduction

The development of Large Language Models (LLMs) has revolutionized natural language processing, enabling sophisticated analysis of textual data. These models, trained on extensive corpora, possess the ability to understand and generate human-like text, making them invaluable tools in various domains, including sentiment and emotion analysis. In the context of mental health, particularly eating disorders, analysis of the emotional content of patients' texts can offer profound insights into their psychological states. This study aims to leverage LLMs to assess ethe intensity of emotions in texts authored by individuals suffering from eating disorders, thereby contributing to more nuanced therapeutic evaluation.

2 Related works

The researchers Liu et al. (2024) implemented a series of designed emotional large language models (EmoLLMs) for comprehensive affective analysis. Their study demonstrated that EmoLLMs have an advantage over all open LLMs and other known tools in sentiment analysis. The authors Aroyehun et al. (2023) introduced the LEIA model in two variants: LEIA-base and LEIA-large, designed for emotion recognition in text. Trained on a dataset of more than six million posts with self-annotated emotion labels: happiness, affection, sadness, anger, and fear — LEIA offers a balanced approach to emotion detection. The model demonstrates strong generalization between different posts, users, and time periods. The authors Chu et al. (2024) used LLM to analyze Twitter data related to eating disorders (ED), showcasing the potential of LLM to uncover the implicit perspectives of online communities. Their study highlights the power of LLMs as a valuable tool for addressing eating disorders in the digital space, paving the way for more precise and effective public health strategies. The authors Chopra et al. (2024) utilized LLMs and topic modeling to analyze posts from the Reddit social media platform. Their study demonstrated that LLMs can significantly reduce the time and resources needed for large-scale analyses of social media data related to eating disorders. Furthermore, multiple studies have examined the use of LLMs in sentiment

analysis. One such study, conducted by Zhang et al. (2024), provides an in-depth assessment of LLMs' capabilities in this domain. The researchers compared the performance of LLMs with that of small language models (SLMs), finding that while LLMs handle simpler tasks effectively in a zero-shot setting, they face challenges with more complex ones. However, in a few-shot learning scenario, LLMs consistently outperform SLMs, highlighting their usefulness in situations where annotated data is limited.

3 Method

The purpose of the study was to estimate the intensity of 5 of the 6 basic emotions according to Paul Ekman: happiness, sadness, anger, fear, and disgust (Ekman, 1992). Surprise was not investigated when building the NAWL core dictionary (see 3.2) and was therefore not analyzed in this study. Figure 1 shows the algorithm schema.

3.1 Dataset

The dataset comprised 201 textual notes, including 115 from Polish-speaking patients diagnosed with anorexia and 86 from healthy controls. Eligibility criteria for the patient group included female sex, age between 12 and 19 years, a confirmed clinical diagnosis of anorexia with a disease duration not exceeding three years, and the absence of comorbid psychological disorders. The control group consisted of females within the same age range (12–19 years) with no diagnosed psychological conditions.

The average age in the group of patients was 15.7 ± 2. The average weight in this group was 35.1 ± 4.7 kg and the mean BMI (Body Mass Index) was 15.1 ± 2.8 kg/m^2.

In the healthy group, the mean age was 15.1 ± 1.9 years. The average weight in this group was 57.1 ± 10.1 kg and the mean BMI was 21.5 ± 3.4 kg/m^2.

Participants were asked the question: "How do you perceive your body?". Patients suffering from anorexia, who were undergoing therapy, provided their responses during consultations with a psychologist. The control group consisted of female adolescents and young adults enrolled in primary and secondary schools, and their responses were collected via a questionnaire. The study protocol, including the collection of research material, was reviewed and approved by the Bioethics Committee of the Medical University of Silesia.

3.2 NAWL dictionary

Nencki Affective Word List (NAWL) is a database of 2 902 Polish words including nouns (1676), verbs (614) and adjectives (612) with ratings of emotional valence, arousal and imageability Wierzba et al. (2015). It's polish counterpart of Berlin Affective Word List-Reloaded Võ et al. (2009). The word ratings were collected from 265 Polish participants. The words in the original study were scored using 7-point scales, this study uses min-max normalization, and the values are re-valued to 0-1.

3.3 PolBERT model

The Polish language model PolBERT, developed on the BERT architecture (12 layers, 786 hidden units, 12 attention heads, and approximately 110M parameters), was introduced in two variants: uncased and cased. The uncased version was trained on a corpus of 262 920 423 sentences; however, it presented certain limitations, including incorrect tokenization of some Polish characters and diacritical marks, the presence of duplicates in the training corpus, and the absence of Whole Word Masking. In contrast, the cased version, trained on 68 283 960 sentences, addressed these issues, thereby overcoming the limitations observed in the uncased model (Kłeczek, 2020).

3.4 Methodology

Methodology covering the steps to obtain the full intensity of emotions for the entire text notes. Tool was developed with Python programming language (Van Rossum and Drake, 2009). We used the SpaCy library (Honnibal and Montani, 2017) and the Hugging Face Transformers library, which includes the PolBERT model (Kłeczek, 2020), to support our implementation and text processing tasks.

3.4.1 Word embeddings generation

Two pretrained models, PolBERT cased and PolBERT uncased, were employed to generate dynamic word embeddings. Embeddings were obtained by averaging the hidden states from the last four layers of each model. This choice was motivated by previous studies indicating that higher layers capture more abstract se-

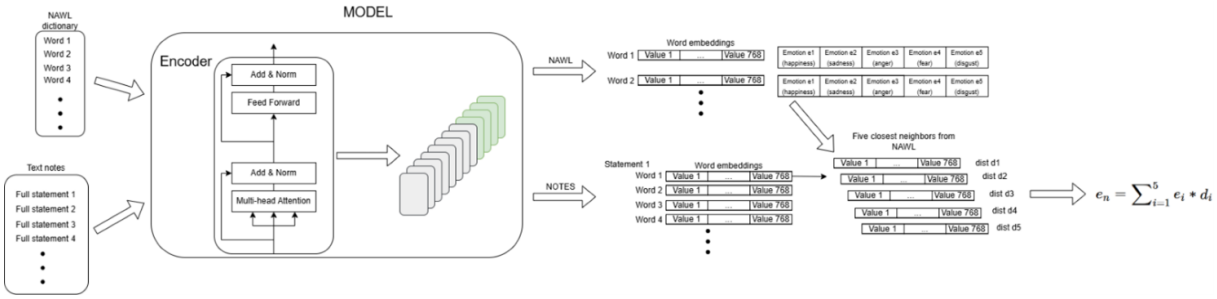

Figure 1: Schema of algorithm

mantic features (Tenney et al., 2019), with the best performance reported when concatenating representations from the top four hidden layers (Devlin et al., 2019). Word embeddings were generated for all words in the collected notes, including stopwords, as well as for all entries in the NAWL lexicon, which served as a context-free reference dictionary. Each resulting word embedding vector had a dimensionality of 768.

3.4.2 Interpolation

When each word is dynamically encoded as a vector, the interpolation process begins. For each word, including those present in the NAWL dictionary, five closest neighbors within the NAWL (context-free) dictionary are identified in the embedding space using Euclidean distance as the similarity metric. The number of closest neighbors was chosen based on article Barańska et al. (2022). The authors did not indicate a significant change between the 4- and 5-neighborhoods. In this study, we decided to choose 5-neighborhood due to the different approach to the problem and a larger number of neighbors in this case may be important, but this requires further research. Below is the formula for calculating Euclidean distance:

$$d(p,q) = \sqrt{\sum_{n=1}^{n}(p_i - q_i)^2} \tag{1}$$

The existing words in NAWL were also revaluated due to the different placement in the word embedding space depending on the statement's context. Then the Euclidean distances are normalized to 0-1, the sum of distances of 5 closest distances is 1. Based on how far the neighbor is placed, it determines the weight of the influence on the analyzed word. The following formula was used to calculate each emotion:

$$e_n = \sum_{i=1}^{5} e_i * d_i \tag{2}$$

where e_n - n-emotion of the analyzed word, e_i – n's emotion of i-neighbor of the analyzed word, d_i – distance i-neighbor from the analyzed word.

3.4.3 Emotion intensities for notes

To calculate the intensities of basic emotions, we used logarithmic normalization to eliminate outlier values. The formula was as follows:

$$e_{ns} = \frac{\sum_1^k log(1 + e_{nk})}{k} \tag{3}$$

where e_{ns}– n-emotion of sentence(s), k – count of words in sentence, e_{nk}- n-emotion of k word in sentence.

4 Results

Figure 2 shows results in the form of radar chart. This form was chosen to compare two models – Polbert uncased and Polbert cased. Additionally, the division between healthy people and patients suffering from anorexia was taken into account. Regardless of the model used, the happiness intensity was higher for healthy people (cased model – 0.32, uncased model – 0.33). Patients with eating disorders for both models obtained the same violence of happiness – 0.29. The rest of emotion in both models achieved higher values of more difficult emotions such as sadness, anger, fear and disgust for patients suffering from anorexia compared to healthy ones. The levels of emotion intensities between uncased model and cased model did not differ significantly. We also conducted an investigation on the neighbors of the NAWL word. Our main interest was focused on how frequently the words from NAWL have themselves from

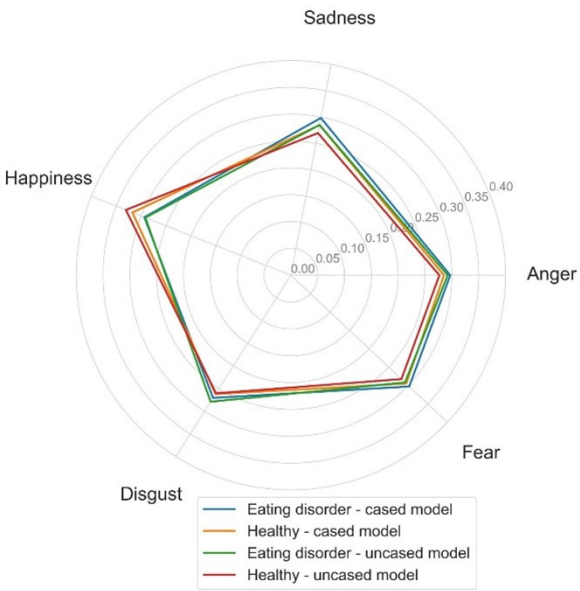

Figure 2: Radar profiles of patients suffering from anorexia and healthy people using uncased and cased Polbert models.

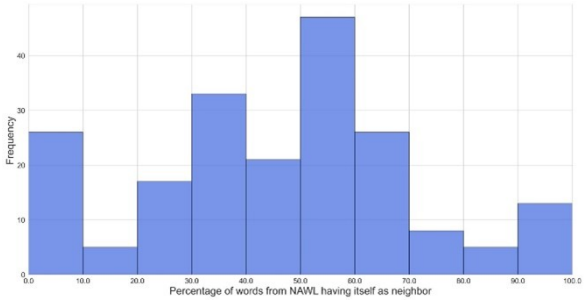

Figure 3: The frequency with which NAWL words co-occurred with other NAWL words in their context.

original placements. Figure 3 shows the histogram of frequency of NAWL word percentage coverage. The dominant peak occurs between 50% and 60%, indicating that in most analyzed notes, half to three-fifths of the NAWL words appeared adjacent to other NAWL words.

Only 25 notes did not have any coverage. The aim of this section was to check whether a word used in a complete sentence under the influence of other words still has the influence of itself from the original dictionary. More research is needed to examine the influence of specific words on others.

5 Conclusion

We proposed solution using LLM Polbert adjusted to Polish language to extract human intensities of emotions in text notes. Our results show no influence of original model training dataset (cased vs uncased) on emotion intensities in two analyzed groups – people suffering from eating disorders and healthy group.

6 Limitations

The tool we propose addresses a sensitive and significant topic. Limitations of our approach should be acknowledged. The data used to train the PolBERT model may include biases, potentially reflecting societal stereotypes related to gender, ethnicity, or other demographic characteristics, as noted by the author.

Another limitation concerns the interpolation process. Interpolation process could be improved by capturing the direction of surrounding neighbors.

Importantly, the results produced by our method should be regarded as supportive analytical indicators rather than conclusive diagnostic outcomes. Such biases could influence the interpretation of emotional content and lead to skewed or unfair conclusions. Future work will aim to refine this process to achieve more context-sensitive and ethically responsible emotion analysis.

References

Segun Taofeek Aroyehun, Lukas Malik, Hannah Metzler, Nikolas Haimerl, Anna Di Natale, and David Garcia. 2023. LEIA: Linguistic embeddings for the identification of affect. 12(1):1–21. Number: 1 Publisher: SpringerOpen.

Klaudia Barańska, Agnieszka Różańska, Stella Maćkowska, Katarzyna Rojewska, and Dominik Spinczyk. 2022. Determining the intensity of basic emotions among people suffering from anorexia nervosa based on free statements about their body. 11(1):138. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Medini Chopra, Anindita Chatterjee, Lipika Dey, and Partha Pratim Das. 2024. Deciphering psycho-social effects of eating disorder : Analysis of reddit posts using large language model(LLM)s and topic modeling. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 156–164. Association for Computational Linguistics.

Minh Duc Chu, Zihao He, Rebecca Dorn, and Kristina Lerman. 2024. Large language models help reveal unhealthy diet and body concerns in online eating disorders communities. *Preprint*, arxiv:2401.09647 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arxiv:1810.04805 [cs].

Paul Ekman. 1992. An argument for basic emotions. 6(3):169–200. Publisher: Routledge _eprint: https://doi.org/10.1080/02699939208411068.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Dariusz Kłeczek. 2020. Polbert: Attacking polish nlp tasks with transformers. In *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 5487–5496. Association for Computing Machinery.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. *Preprint*, arxiv:1905.05950 [cs].

Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Melissa L. H. Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The berlin affective word list reloaded (BAWL-r). 41(2):534–538.

Małgorzata Wierzba, Monika Riegel, Marek Wypych, Katarzyna Jednoróg, Paweł Turnau, Anna Grabowska, and Artur Marchewka. 2015. Basic emotions in the nencki affective word list (NAWL BE): New method of classifying emotional stimuli. 10(7):e0132305.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906. Association for Computational Linguistics.