Data Augmentation for Multi-Class Eating Disorders Text Classification

Ghofrane Merhbene and **Mascha Kurpicz-Briki**
Applied Machine Intelligence, Bern University of Applied Sciences
Biel/Bienne, Switzerland
ghofrane.merhbene@bfh.ch

Abstract

In this study, we tackle the challenge of detecting Eating Disorders (EDs) in German text, a relatively unexplored area in natural language processing (NLP) for mental health. In this project, we developed a manually annotated German dataset from YouTube comments. To address the class distribution imbalance, we employed back translation as a data augmentation technique. This process significantly enhanced the dataset's utility. Through a comprehensive grid search, we identified a Support Vector Machine (SVM) model as the most effective, achieving an average F1-score of 0.83. Our findings not only contribute to the research field of ED detection in German but also demonstrate the effectiveness of innovative data augmentation techniques in managing class imbalances in natural language processing.

1 Introduction

The application of Natural Language Processing (NLP) in mental health diagnosis represents great potential for the future of healthcare (see, e.g., (Rezaii et al., 2022)). Its use may help clinicians in their diagnostic processes, marking a pivotal shift in the treatment and understanding of mental health. Despite this potential, the field faces a significant limitation: the predominant focus on English in research, which hinders the applicability of findings across diverse linguistic contexts. Eating disorders (EDs) typically manifest as detrimental eating habits, disruptions in behaviors, thoughts, and emotions related to food, which can lead to significant weight loss or gain in some cases. These disorders affect not only mental well-being but also physical health. Classified under the F50 category in the ICD-10 (World Health Organization, 1992), EDs encompass various conditions such as anorexia, bulimia, and overeating[1].

[1]https://www.icd10data.com/ICD10CM/Codes/F01-F99/F50-F59/F50-

In response to this research gap, our study introduces an innovative method focused on the identification of various types of EDs in German text. This approach not only widens the linguistic scope of current research, but also contributes to a more global understanding of EDs in different cultures and languages.

Our paper presents several substantial advancements to the current state of the art. Primarily, we have trained and tested a classifier capable of detecting a range of Eating Disorders, thereby pushing the boundaries of automated diagnostic tools in NLP for mental health research. Additionally, we provide insights on the effectiveness of data augmentation with back translation for a task like ED detection in German with few training data being available.

The structure of this paper is as follows: Section 2 offers a comprehensive review of related work, laying the groundwork for our research. Section 3 presents our proposed methodology and the materials used. Section 4 is dedicated to the presentation and discussion of our experimental results, showcasing the efficacy and insights derived from our model. Section 5 concludes the paper with a summary of our findings and potential directions for future research. Lastly, Section 6 reflects on the limitations of our study, ensuring a balanced and critical understanding of our work.

2 Related Work

The current state-of-the-art on the application of NLP in detecting EDs in languages other than English remains relatively sparse. To date, very few efforts in this area have been primarily focused on specific languages, with only one study addressing Spanish (López Úbeda et al., 2019) and two dedicated to Polish (Spinczyk et al., 2020; Rojewska et al., 2022). Furthermore, the majority of the English focused studies have predominantly used data from social media platforms such as Red-

dit (Yan et al., 2019), Twitter (Benítez-Andrades et al., 2021; López Úbeda et al., 2019; Zhou et al., 2020; Benítez-Andrades et al., 2022; Wang et al., 2017; He and Luo, 2016), and Tumblr (He and Luo, 2016; De Choudhury, 2015). A significant number of these investigations have leveraged the CLEF eRisk dataset (Wang et al., 2018; Ragheb et al., 2018; Aguilera et al., 2021; Aragon et al., 2021; Mohammadi et al., 2019; Ramiandrisoa and Mothe, 2020; Paul et al., 2018; Trotzek et al., 2018; Ramiandrisoa et al., 2018; Ortega-Mendoza et al., 2018; Liu et al., 2018; Merhbene et al., 2023), a fundamental resource provided by the Conference and Labs of the Evaluation Forum (CLEF).

CLEF eRisk (Parapar et al., 2023), an annual event in the research community, is designed to evaluate and benchmark the capabilities of various NLP systems in identifying and analyzing high-risk and harmful content on social media, including language patterns indicative of EDs, self-harm, and suicidal tendencies. This initiative offers a repository of social media posts, primarily from Reddit, and facilitates a competitive environment for teams to develop and assess their models for early detection of such critical issues. The overarching goal is to propel advancements in NLP, specifically in the context of recognizing and interpreting high-risk content on social media platforms.

Using a common dataset like CLEF eRisk enables researchers to strive for optimal outcomes in a competitive setting. However, this approach is not without its limitations. A key concern is the potential for leading NLP technologies to become overly adapted to patterns unique to this dataset, possibly leading to reduced efficacy when applied to varied types of data. This underscores the necessity of integrating greater diversity in training and evaluation datasets for NLP models targeting ED detection.

Despite these challenges, the field has witnessed some promising developments. For instance, López Úbeda et al. (2019) reported an impressive F1 score of 0.91 using supervised machine learning models. Similarly, Wang et al. (2017) achieved an accuracy of 0.97 using a Support Vector Machine classifier, leveraging user-based metrics encompassing social status, behavior, and psychometry. These successes indicate the potential of NLP in the realm of automatic ED detection, pointing towards a promising direction for future research.

3 Materials and Methods

3.1 Dataset

German served as the principal language for this work. A high-quality dataset is pivotal in such research; therefore, we used the YouTube API[2] to compile a robust dataset of anonymized German YouTube comments. To methodically identify videos relevant to Eating Disorders (EDs), we developed search queries incorporating specific keywords indicative of various EDs. These keywords included Essstörungen (eating disorders), Anorexie (anorexia), Bulimie (bulimia), among others that are closely associated with eating behaviors and body image issues. We also formulated query phrases to capture a broad spectrum of personal and informative content, such as "Meine Erfahrung mit Essstörungen" (My experience with eating disorders), "Leben mit einer Essstörung" (Living with an eating disorder), and "Magersucht OR Bulimie: Wie ich es geschafft habe" (Anorexia OR Bulimia: How I overcame it). These queries were designed to ensure the inclusion of a diverse range of video content related to the spectrum of eating disorders.

After identifying relevant videos, we extracted all comments under each video, ensuring comprehensive coverage of public discourse on these topics. The collected comments underwent a meticulous annotation process. Three domain-specific annotators manually labeled the data, adhering to a detailed set of annotation guidelines we developed. We employed a majority agreement method to finalize the annotations. Given the multifaceted nature of EDs, we used a multi-label annotation framework with six labels based on the ICD-10 (World Health Organization, 1992), as detailed in Table 1.

To address potential inconsistencies in manual annotations and ensure the reliability of our dataset, we computed inter-annotator agreement using pairwise Cohen's kappa coefficient for each label. The kappa values ranged from fair to substantial agreement, highlighting a generally reliable annotation process despite the subjective complexities involved in interpreting comments related to EDs.

Table 2 gives some insights into the dataset, highlighting key statistics such as the total number of samples and the average and standard deviation of text lengths measured in characters. To further char-

[2]https://developers.google.com/youtube/v3

Label	Description
A (Anorexia)	For cases where individuals engage in extreme calorie restriction, excessive exercise, or purging to control their weight or body shape.
O (Overeating)	For cases where individuals engage in behaviors like eating large amounts of food rapidly, feeling a loss of control over their eating, or eating when not physically hungry.
B (Binge Eating)	For cases where individuals engage in binge eating followed by purging behaviors such as vomiting, using laxatives, or excessive exercise. Common features of bulimia include weight fluctuations, tooth decay, and dehydration.
N (No ED)	For individuals who do not show any eating disorder behaviors or symptoms. An example is someone with a healthy relationship with food and their body, displaying no signs of disordered eating.
P (Previous ED)	For individuals with a history of eating disorders but are currently in recovery, showing no symptoms. An example is someone who had anorexia but is now in remission, identified by their use of past tense when describing their experience with an eating disorder.
K (No label)	When none of the other labels applies.

Table 1: Labels Description.

No. of Samples	Mean Text Length (chars)	Std. Dev. of Text Length (chars)
743	561	726.7

Table 2: Dataset Statistics.

acterize the dataset, we analyzed the lexical diversity and the average number of words per comment. The results are summarized in Table 3.

Lexical Diversity	Avg. nb. of words per comment
0.18	92.14

Table 3: Dataset Characteristics.

The lexical diversity score is calculated by dividing the number of unique words by the total number of words (over all datapoints). The resulting score of this metric is 0.18. Additionally, the average word count of 92.14 per comment demonstrates that the comments are detailed enough to provide substantial textual content for analysis, allowing for the expression of personal experiences and insights crucial for understanding public perceptions and misconceptions related to eating disorders.

A significant challenge associated with working with data derived from social media is ensuring a balanced representation of the targeted classes within a study. Our dataset notably exhibited this issue of imbalance, as depicted in Figure 1.

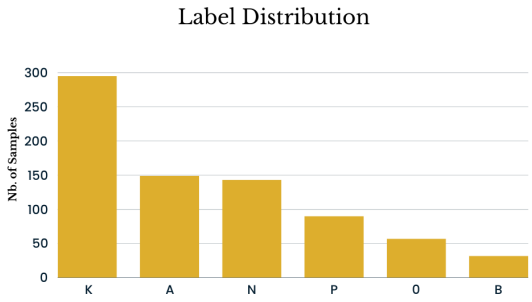

Figure 1: Label Distribution (A= Anorexia, O= Overeating, B= Binge Eating, N= No ED, P= Previous ED, K= No Label)

3.2 Textual Data Augmentation

In the realm of NLP, data augmentation is a crucial technique in machine learning, especially in the context of limited or imbalanced datasets. It involves artificially increasing the diversity of data, without actually collecting new data. This is achieved by creating modified versions of existing data points using techniques such as synonym replacement, sentence shuffling, and back translation (Pellicer et al., 2023).
Translation is a paraphrasing technique that has surged in popularity with advancements in machine translation technologies and the widespread availability of online translation APIs (Li et al., 2022).

Back translation involves translating text from the original language to a secondary language and then back to the original language again (see example in Table 4). This technique has demonstrated its effectiveness in a range of applications. For example, Beddiar et al. (2021) relied on back translation to augment their dataset and improve the performance of their model. This method effectively generates additional, syntactically accurate data points while preserving the original semantic content. This ensures the enrichment of datasets without the risk of introducing low-quality, noisy data, thereby upholding the overall data integrity.

In this work, we use *MarianMT*[3] (Tiedemann and Thottingal, 2020) a neural machine translation framework from Hugging Face. The back translation process was executed using three different languages: English, Dutch, and Luxembourgish.

However, it was not uniformly applied across all labels. The variation in label distribution, as shown in Figure 1, inspired this selective approach. For instance, label B had approximately 34 data points, whereas label O had about twice as many. Figure 3 illustrates the specific back translation augmentation applied to each label. Some labels, like A and N, underwent the process once, others like P twice, and labels O and P three times. It is important to note that label K, being the predominant class, was excluded from this augmentation process.

The implementation of back translation in these languages effectively contributed to creating a larger more balanced dataset with 1377 entries, thus enhancing the diversity of the data used in our analysis. Table 5 displays the lexical diversity and average word count per comment in the augmented dataset.

Figure 2 shows the new label distribution after augmenting the dataset using back translation. This technique improved the label distribution in the dataset, especially for the categories with previously moderate to high sample counts such as Anorexia (A) and No ED (N). This technique has effectively increased the representation of most labels, helping to reduce the initial imbalance.

3.3 Performance Metrics

To assess the performance of our proposal we rely mainly on F1-score and Balanced Accuracy.

- **F1-score**: The harmonic mean of precision and recall. Precision is the ratio of true positive predictions to the total number of positive predictions, while recall is the ratio of true positive predictions to the total number of actual positives.

$$F1\text{-}score = 2 \times \frac{\left(\frac{TP}{TP+FP}\right) \times \left(\frac{TP}{TP+FN}\right)}{\left(\frac{TP}{TP+FP}\right) + \left(\frac{TP}{TP+FN}\right)} \quad (1)$$

- **Balanced Accuracy**: It is particularly useful for evaluating classification performance on datasets with imbalanced class distributions. It is defined as the average of recall obtained on each class.

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right) \quad (2)$$

Where TP, FN, TN, and FP correspond to true positive, false negative, true negative and false positive respectively.

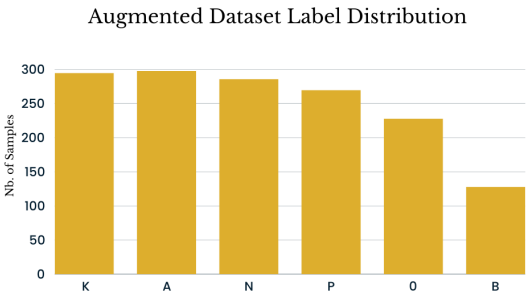

Figure 2: Augmented Dataset Label Distribution (A= Anorexia, O= Overeating, B= Binge Eating, N= No ED, P= Previous ED, K= No Label).

3.4 Classification

A critical aspect to consider, particularly when working with data derived from social media platforms, is text pre-processing. This step is instrumental in refining and standardizing the input text, thereby significantly enhancing its quality and consistency. Our dataset underwent a thorough preprocessing to prepare it for effective model training. This involved the following operations: URL removal, HTML tag removal, removal of special

[3]https://huggingface.co/docs/transformers/model_doc/marian

Process Step	Text
Original German Text	Die Essstörungen haben mein Leben kaputt gemacht.
Translated to Dutch	Eetstoornissen hebben mijn leven geruïneerd.
Back Translated to German	Essstörungen haben mein Leben ruiniert.

Table 4: Example of the back translation process used in data augmentation. The original German text translates to "Eating disorders have ruined my life." in English.

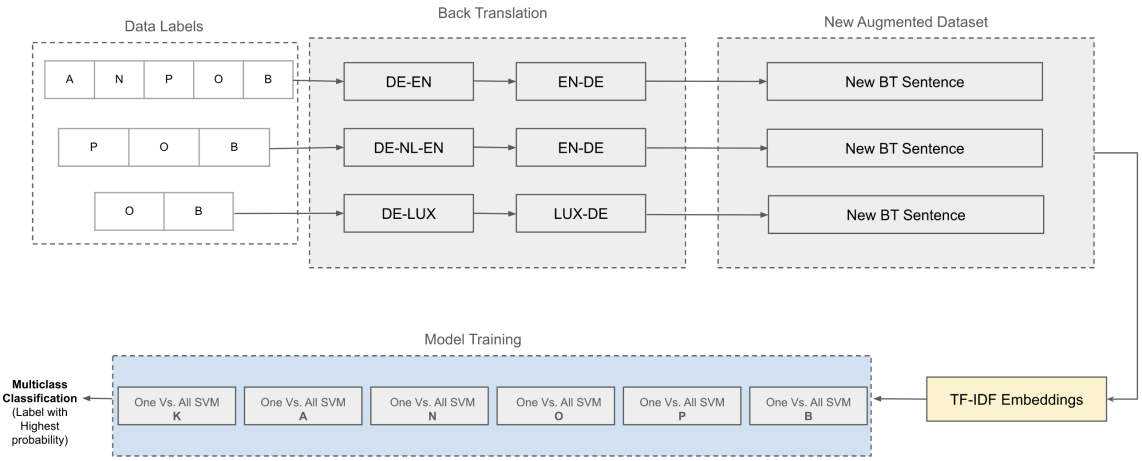

Figure 3: General scheme of the proposed architecture.

Lexical Diversity	Avg. nb. of words per comment
0.11	101.67

Table 5: Augmented Dataset Characteristics.

characters and numbers, case normalization, tokenization, and finally stop words removal. Following this and to ensure robust model evaluation, 30% of our original dataset was designated as the test set to assess model performance. The remaining 70% of the original data, supplemented with synthetic data, was allocated for training purposes. To compute the word embeddings, we rely on the Term Frequency-Inverse Document Frequency (TF-IDF) with a vector size of 5000. TF- IDF is a statistical measure that assesses the relevance of a word within a document set, or corpus. The TF component of TF-IDF increases proportionally with the number of times a word appears in a document, reflecting its importance. Conversely, the IDF component inversely scales the weight of the word based on its frequency across the entire corpus. This adjustment is crucial as it diminishes the influence of words that occur commonly across all documents, thereby helping to highlight more distinctive terms within each document.

To ensure that our classifier can accurately identify sentences with multiple labels, a *One-vs-All* model training approach is applied. Thus, a separate model is trained for each label. This method is especially relevant given the overlapping and multifaceted nature of eating disorders, allowing for a more nuanced and comprehensive classification, as demonstrated in Figure 3.

To identify the most effective model for each label, we conducted a grid search across various models, including Multinomial Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting, Logistic Regression, SVM, and Multilayer Perceptron (MLP) Classifier. For the SVM, we used a linear kernel to maintain computational efficiency while capturing linear relationships, set the regularization parameter C to 10 to allow for some misclassification but with a firm margin, and chose 'scale' for the gamma parameter to automatically adjust it according to the number of features, ensuring the model's adaptability. The MLP classifier, which is a type of neural network, was allowed 1000 iterations, giving the network ample opportunity to converge on a solution and learn from the data effectively. The other classifiers were used with their default parameters to establish a baseline performance.

The selection of the best-performing model for each label was based on the F1-score. This evaluation process was applied to both the original and the augmented datasets to ensure the most accu-

rate and effective model selection. Additionally, to avoid overfitting, we ensured that the synthetic data was used solely for training purposes, while deliberately excluding it from the testing phase.

4 Results and Discussion

4.1 Results

Tables 6 and 7 present the F1-score results from our experiments on the original and the augmented datasets, respectively. In these tables, we've accentuated the best results corresponding to each label. The Decision Tree algorithm achieved better results for the labels B, and O. Meanwhile, the Random Forest algorithm was the frontrunner for label K, the gradient-boosting classifier took the lead for label K, B, surpassing the performance of all other models. The MLP classifier was distinguished as the best for label A, while the SVM stood out for labels A and N, indicating its robustness across these particular categories.

The enhancement in performance after dataset augmentation is clear. Notably, we observed a substantial uptick in results, with an approximate 70% improvement for label B and a significant 50% boost for label A.

When looking at the average performance across all labels, it becomes clear that SVM classifier outperforms all others, achieving an F1-score of 0.41 before data augmentation and 0.83 after. Additionally, the implementation of back translation as a data augmentation technique significantly enhanced the average performance, yielding an approximate improvement of 40%.

Figure 4 shows the results in terms of balanced accuracy for each label. From the graph, it is observable that SVM when trained on the augmented dataset outperforms all other classifiers, while MLP follows close behind. We can also notice that most classifiers benefited from data augmentation, although the Random Forest and Gradient Boosting models did not show improvement on label K when comparing the augmented dataset to the original.

4.2 Discussion

Back translation for data augmentation has garnered increasing recognition for its potential to enhance datasets across various fields. In this study, applying this method to our dataset and integrating it with different machine learning models resulted in a notable improvement in both F1-score and Balanced Accuracy. Such performance enhancements are likely linked to the introduction of linguistic variations by back translation, which contribute to a more robust and varied dataset. This aspect is particularly important in the realm of Eating Disorders, where the subtle nuances of language and expression are key to accurately identifying and categorizing the different types of EDs.

It is also important to note that the efficiency of this method has been further emphasized by its successful application in other research contexts. For example, Corbeil and Ghadivel (2020) have demonstrated the efficiency of back-translation's paraphrasing capability and its ability to generate robust and diverse new data points, and Bédi et al. (2022) found that using back-translation to augment a dataset on hate speech was beneficial for their machine learning model. Furthermore, (Beddiar et al., 2021) reported a significant enhancement in their study, where the application of back translation on a novel cyberbullying detection dataset using a convolutional neural network (CNN) architecture led to a 42% improvement in the F1-score. This correlation between our results and those of other studies reinforces the broad applicability and effectiveness of this method.

It is important to note that the only label that showed little to no improvement of both F1-score and Balanced Accuracy metrics was label K. This was likely because the augmentation was omitted for this label as it is already the dominant class.

5 Conclusion

In this study, we addressed a notable research gap by focusing on the automatic detection of Eating Disorders (EDs) in German text, and thus contributing to the state-of-the-art of NLP for mental health. Our work led to developing a specialized, manually annotated dataset tailored for ED detection in German. Despite facing challenges with a significant class imbalance within the dataset, we successfully implemented back translation for data augmentation to tackle this challenge. This approach not only helped in balancing the dataset but also significantly enhanced the model performance. It resulted in a remarkable 40% overall improvement in F1-score and a notable increase in the Balanced Accuracy score when used with SVM for classification over our 6 data classes. Our findings underscore the potential of language-specific resources and targeted augmentation techniques in improving the accuracy of automatic ED detection systems.

Label	MNB	DT	RF	GB	LR	SVM	MLP
A	0.00	0.38	0.00	0.10	0.00	0.42	**0.50**
B	0.00	**0.33**	0.00	**0.33**	0.00	0.00	0.00
K	0.28	0.56	**0.73**	**0.73**	0.66	0.64	0.62
N	0.00	0.29	0.00	0.24	0.00	**0.46**	0.29
O	0.00	**0.55**	0.00	0.35	0.00	0.47	0.00
P	0.00	0.25	0.00	0.23	0.00	**0.50**	0.18
Average	0.05	0.39	0.12	0.33	0.11	**0.41**	0.27

Table 6: F1-scores for each label using different models on the Original dataset (MNB: MultinomialNB, DT: DecisionTree, RF: RandomForest, GB: GradientBoosting, LR: LogisticRegression, SVM: SVC, MLP: MLPClassifier)

Label	MNB	DT	RF	GB	LR	SVM	MLP
A	0.00	0.59	0.62	0.57	0.42	**0.86**	**0.86**
B	0.0	0.77	0.80	0.80	0.00	**0.96**	**0.96**
K	0.03	0.59	0.33	0.64	0.44	**0.73**	0.60
N	0.00	0.47	0.31	0.30	0.04	**0.72**	0.59
O	0.0	0.61	0.76	0.81	0.20	**0.84**	0.80
P	0.0	0.54	0.67	0.67	0.39	**0.88**	0.83
Average	0.005	0.60	0.58	0.63	0.25	**0.83**	0.77

Table 7: F1-scores for each label using different models on the Augmented dataset (MNB: MultinomialNB, DT: DecisionTree, RF: RandomForest, GB: GradientBoosting, LR: LogisticRegression, SVM: SVC, MLP: MLPClassifier)

Figure 4: Balanced Accuracy for each label using different models with the original and the augmented dataset (*df* corresponds to the orignal dataset and *df_a* to the augmented dataset).

6 Limitations

In considering the limitations of this study, it is important to acknowledge the constraints associated with the dataset. While we have made significant strides in data augmentation through back translation, the severe initial imbalance in class distribution may still have residual effects on the generalizability of our findings. Notably, some labels, such as label B with only 34 examples, required the use of the entire dataset for augmentation to ensure their inclusion in the testing process. This approach was essential for maintaining a balanced representation across different classes, albeit potentially limiting the variety of testing scenarios.

Moreover, the manual annotation process, despite being thorough, is subject to human error and interpretative variability, which could influence the reliability of the dataset. Additionally, the reliance on text from YouTube comments presents a limitation in terms of linguistic variety and depth, as it still remains a social media platform and it may not fully represent the broader spectrum of language use associated with eating disorders, or the general population. Another limitation is the diversity in the available training data. As the used comments were anonymous, no information about gender or other characteristics of the authors of the texts were available. This needs to be addressed in future work to ensure that the classifier works with the same efficiency for different groups of the population. Finally, the performance of the SVM model, while promising, was evaluated within the context of this specific dataset, and its applicability to other datasets or in a real-world scenario requires further validation.

Ethics Statement

The work presented in this paper is part of a research project investigating NLP for mental health. The data collection and processing followed an internal guideline that was established in collaboration with a legal advisor. All data was anonymized. The annotation was done by domain experts who were hired at adequate local conditions and who are familiar with the sensitivity of the texts provided. The targeted tools derived from this research aim to provide further insights to clinical professionals, not to replace them. Given the potential limitations of such methods and datasets, the authors consider it highly relevant to keep human experts in the loop.

Acknowledgements

The authors gratefully acknowledge the support of the Inventus Bern Foundation for our research in the field of augmented intelligence for the detection of eating disorders.

References

Juan Aguilera, Delia Irazú, Irazú Hernández Farías, María Ortega-Mendoza, and Manuel Montes-Y-G omez. 2021. Depression and anorexia detection in social media as a one-class classification problem.

Mario Ezra Aragon, Adrian Pastor Lopez-Monroy, Luis-Carlos Gonzalez-Gurrola, and Manuel Montes. 2021. Detecting mental disorders in social media through emotional patterns - the case of anorexia and depression. *IEEE Transactions on Affective Computing*, pages 1–1.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Branislav Bédi, Hakeem Beedar, Belinda Chiera, Nedelina Ivanova, Christèle Maizonniaux, Neasa Ní Chiaráin, Manny Rayner, John Sloan, and Ghil'ad Zuckermann. 2022. Using LARA to create image-based and phonetically annotated multimodal texts for endangered languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 68–77, Dublin, Ireland. Association for Computational Linguistics.

José Alberto Benítez-Andrades, José Manuel Alija-Pérez, Isaías García-Rodríguez, Carmen Benavides, Héctor Alaiz-Moretón, Rafael Pastor Vargas, and María Teresa García-Ordás. 2021. Bert model-based approach for detecting categories of tweets in the field of eating disorders (ed). In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 586–590.

José Alberto Benítez-Andrades, José-Manuel Alija-Pérez, Maria-Esther Vidal, Rafael Pastor-Vargas, and María Teresa García-Ordás. 2022. Traditional machine learning models and bidirectional encoder representations from transformer (bert)-based automatic classification of tweets about eating disorders: Algorithm development and validation study.

Jean-Philippe Corbeil and Hadi Abdi Ghadivel. 2020. Bet: A backtranslation approach for easy data augmentation in transformer-based paraphrase identification context.

Munmun De Choudhury. 2015. Anorexia on tumblr: A characterization study. *Digital Health*.

Ling He and Jiebo Luo. 2016. What makes a pro eating disorder hashtag: Using hashtags to identify pro eating disorder tumblr posts and twitter users.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Ning Liu, Zheng Zhou, Kang Xin, and Fuji Ren. 2018. Tua1 at erisk 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*.

Pilar López Úbeda, Flor Miriam Plaza del Arco, Manuel Carlos Díaz Galiano, L. Alfonso Urena Lopez, and Maite Martin. 2019. Detecting anorexia in spanish tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 655–663. INCOMA Ltd.

Ghofrane Merhbene, Alexandre R. Puttick, and Mascha Kurpicz-Briki. 2023. BFH-AMI at erisk@clef 2023. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 727–735. CEUR-WS.org.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. Quick and (maybe not so) easy detection of anorexia in social media posts.

Rosa M Ortega-Mendoza, A Pastor López-Monroy, Anilu Franco-Arcega, and Manuel Montes-Y-Gómez. 2018. Peimex at erisk2018: Emphasizing personal information for depression and anorexia detection.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

Sayanta Paul, Jandhyala Sree Kalyani, and Tanmay Basu. 2018. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803.

Waleed Ragheb, Bilel Moulahi, Jérôme Azé, Sandra Bringay, and Maximilien Servajean. 2018. Temporal mood variation: at the clef erisk-2018 tasks for early risk detection on the internet.

Faneva Ramiandrisoa and Josiane Mothe. 2020. Early detection of depression and anorexia from social media: A machine learning approach. In *CEUR-WS*, volume 2621 of *Proceedings of the Conference CIRCLE 2020*, Samatan, France.

Faneva Ramiandrisoa, Josiane Mothe, Farah Benamara, and Véronique Moriceau. 2018. Irit at e-risk 2018. In *9th Conference and Labs of the Evaluation Forum, Living Labs (CLEF 2018)*, pages 1–12, Avignon, France.

Neguine Rezaii, Phillip Wolff, and Bruce H Price. 2022. Natural language processing in psychiatry: the promises and perils of a transformative approach. *The British Journal of Psychiatry*, 220(5):251–253.

Katarzyna Rojewska, Stella Maćkowska, Michał Maćkowski, Agnieszka Różańska, Klaudia Barańska, Mariusz Dzieciątko, and Dominik Spinczyk. 2022. Natural language processing and machine learning supporting the work of a psychologist and its evaluation on the example of support for psychological diagnosis of anorexia. *Applied Sciences*, 12(9).

Dominik Spinczyk, Maciej Bas, Marcin Dzieciatko, Mateusz Ma ckowski, Katarzyna , and Sylwia Ma ckowska. 2020. Computer-aided therapeutic diagnosis for anorexia. *BMC Medical Informatics and Decision Making*, 20(1):251.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia fhdo biomedical computer science group (bcsg).

Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media.

Yu-Tseng Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. A neural network approach to early risk detection of depression and anorexia on social media text. In *CLEF*.

World Health Organization. 1992. *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organization.

Hao Yan, Ellen Fitzsimmons-Craft, Micah Goodman, Melissa Krauss, Sanmay Das, and Patty Cavazos-Rehg. 2019. Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention. *International Journal of Eating Disorders*, 52.

Sicheng Zhou, Yunpeng Zhao, Jiang Bian, Ann F Haynos, Rui Zhang, and Rui Zhang. 2020. Exploring eating disorder topics on twitter: Machine learning approach.