Healthcare of the Future 2022
T. Bürkle et al. (Eds.)

doi:10.3233/SHTI220318

Burnout and Depression Detection Using Affective Word List Ratings

Sophie HAUG[a] and Mascha KURPICZ-BRIKI[a,1]
[a] *Applied Machine Intelligence, Bern University of Applied Sciences, Biel, Switzerland*

Abstract. Burnout syndrome and depression are prevalent mental health problems in many societies today. Most existing methods used in clinical intervention and research are based on inventories. Natural Language Processing (NLP) enables new possibilities to automatically evaluate text in the context of clinical Psychology. In this paper, we show how affective word list ratings can be used to differentiate between texts indicating depression or burnout, and a control group. In particular, we show that depression and burnout show statistically significantly higher arousal than the control group.

Keywords. psychology, natural language processing, affective word lists, augmented intelligence

1. Introduction

Burnout syndrome and depression are prevalent mental health problems in many societies today. Whereas currently inventories including multiple-choice questions are used in the clinical intervention and research (e.g., [1] [2]) it has been shown that these have limitations and that novel methods using text data such as interview transcripts or free-text answers are promising. However, they generate a large overhead when evaluated manually [3]. Furthermore, it can be difficult to demarcate different mental health conditions, such as for example depression and burnout that might have overlapping symptoms [4]. Most of the related work in detecting indication for depression on text is done for the English language. In this paper, we provide insights into how affective word list ratings can be used to differentiate between texts indicating depression or burnout, and a control group, for the German language. In particular, we show that depression and burnout show statistically significantly higher arousal than the control group. The promising results of our work enable the development of new technologies to support clinical practitioners to automatically find indication for the presence of depression or burnout in a text written by a patient or an interview transcript. Previously, NLP methods have been applied to different clinical use cases, such as e.g., information extraction for cancer-related electronic health record (EHR) notes [15] or computational phenotyping [16]. Such new technologies can in the future provide new smart tools to support clinical practitioners in their daily work (so called *Augmented Intelligence*).

[1] Correspondence: Mascha Kurpicz-Briki, Applied Machine Intelligence, Bern University of Applied Sciences, Höheweg 80, CH-2502 Biel/Bienne, Switzerland; E-mail: mascha.kurpicz@bfh.ch.

2. Background: Affective Word Lists

Affective word ratings are normative ratings of words in a given language relating to the emotional content of a word. For example, say we have the category *valence*, which describes how strongly a word is associated with positive emotions. Then the word “happy” will have a very high score in the *valence* category and the word “sad” a very low one. Affective Word Lists (AWL) are dictionaries with word ratings for different emotional categories. They have been previously used both for studying emotional and physical response to language as well as for studying emotional meaning within language itself, e.g. [5]. Other works, while not having directly used affective word lists, have used features such as “percentage of positive words” [6] or “percentage of words associated with body image” [7] as specified features in their classifiers.

The first corpus of affective ratings for the English language was presented by Bradley and Lang in 1999 [13] and since then, various such corpora have been compiled in different languages. For our analysis we used two existing affective word lists for the German language: The first is the Affective Norms corpus, a dictionary of 350’000 lemmatised German words, that was introduced by Köper and Schulte im Walde [8] in 2016. For this list, all ratings were obtained from a supervised machine learning algorithm, which is why the list is substantially larger than other affective word lists which were labelled by hand.

The second corpus, the Berlin Affective Word List Reloaded (BAWL-R), introduced by Võ et al. [14], consists of 2900 German words and has been rated by 200 human annotators. For the final rating the average of the individual ratings was taken. Unlike the Affective Norms corpus, the BAWL-R has a part-of-speech column and thus allows, e.g., to only consider verbs or nouns in the analysis.

The affective categories used in both Affective Norms and BAWL-R are the following:

- *Imageability,* which describes the degree of visual imageability.
- *Valence*, which determines the positiveness or pleasantness that a word is associated with.
- *Arousal,* which relates to the intensity of the emotional activation level, ranging from excitement to relaxation. High arousal is “perceived as a sensation of being reactive to stimuli and mentally awake” [5].

Moreover, the Affective Norms corpus has an additional category, *Abstractness/ Concreteness*, which is somewhat related to *imageability* and describes the degree of sensual perceivability of a word (the higher the value the more concrete a word is). The work of Mäntylä et al. [5] uses a slightly different categorization, the *Valence-Arousal-Dominance (VAD)* model, which is frequently used in psychology for emotion categorization. The categories used are *valence*, *arousal* and *dominance*. *Dominance* describes the degree to which we have control on a stimulus, ranging from submission to feeling in control. The authors use this model to categorize emotions in Jira[2] issues and ultimately assess productivity and burnout in software engineering, with burnout being associated with low valence, low dominance and high arousal [5]. Unfortunately, there is no available German lexicon which uses the *dominance* category, which would have been especially enlightening, to compare the findings directly with [5].

[2] Jira is a software used for project management and collaboration, often used for IT projects.

3. Experimental Setup

Data set: An extended version of the data set from [9] consisting of texts of the following three classes was used: (i) Burnout (ii) Depression (online testimonials, transcripts of documentaries, online forums) and (iii) Control. Each category contains anonymized texts based on publicly available data originating from or having a strong relation to individuals suffering from burnout, depression or none of them (control group). We extended the data set, and in particular added the category (ii) for depression. For this work, the data set had to be lemmatized, so that when looking up words in the affective word lists, we would not miss a match due to a different case, affix or grammatical form.

Analysis of the Affective Word Lists (AWL): From analyzing the AWLs, the following could be inferred

- Both dictionaries use the infinitive form for verbs
- Both dictionaries use the singular form for nouns; the Affective Norms corpus stores nouns in capitalized form, whereas in BAWL-R all words are stored as both uppercase and lowercase but not capitalized (however this is superfluous because BAWL-R has a part-of-speech column which allows us to isolate nouns)
- In the Affective Norms corpus, participle forms (e.g., *laufend*) are kept as is and not reduced to their base verb form (unlike existing NLP libraries for German). BAWL-R, being significantly smaller, does not contain any participle forms.
- The Affective Norms corpus contains a small number of punctuation tokens; however, all punctuation was ignored in further analysis. BAWL-R does not contain any punctuation tokens.
- The Affective Norms corpus includes stop words, on the other hand, BAWL-R contains only nouns, verbs and adjectives.

In a first step, we extended the AWLs by an additional column *p_word* (pre-processed word) which stores the lowercase lemma of that word. Additionally, the Affective Norms corpus was extended by a column which stores the *part-of-speech*. This would allow a more refined analysis of e.g., only verbs used in the text. However, we found that computing a word's part of speech using SpaCy[3] out of context, i.e., processing the standalone word and not a sentence where that word occurs, only works for nouns and verbs, as SpaCy seems to consistently tag all adjectives as adverbs. Thus, using this method, it is only possible to tag nouns and verbs correctly.

Categorization: When using affective word lists for categorizing text one must decide on how to attribute a score to a piece of text based on the individual word scores from the affective word list. To do so, we used the approach presented by Mäntylä et al. [5] to define a VAD (Valence-Arousal-Dominance) score based on the so-called SentiStrength algorithm [10]. When looking up words in the AWL we always converted a word to its lowercase lemma and looked it up in the *p_word* column to ensure we do not miss a word due to cased spelling or grammatical inflection. For each category, *valence*, *imageability*, *arousal* and, in the case of the Affective Norms dictionary, *abstractness-concreteness*, we extended the data set by an additional column which stores the score for the respective category.

[3] SpaCy is a Python library for Natural Language Processing

Statistical Tests: Using the Shapiro Wilk Test we established non-Gaussianity for all affective word list categories. For verifying the statistical significance of the differences found in the scores across the categories Control, Burnout and Depression, we used the Kruskal-Wallis Test. For example, if the p-value is significantly smaller than $\alpha = 0.05$ for *valence*, we can assume the differences found in *valence* across the classes Burnout, Depression and Control are statistically significant and not just by chance.

4. Results

4.1. BAWL-R

The results on the smaller corpus BAWL-R are shown in Table 1 (for all categories the average score was used, scores range from -5 to 5).

Table 1. Mean and standard deviation (σ) for all affective word categories in the BAWL-R corpus (Valence, Arousal, Imageability) for the data set, with the classes Control Group, Burnout and Depression.

Class	Val. mean	Val. σ	Arou. mean	Arou. σ	Img. mean	Img. σ
Control	4.234	0.525	0.811	0.324	2.282	0.723
Burnout	4.247	0.610	0.936	0.389	2.573	0.825
Depress.	4.246	0.421	0.900	0.310	2.352	0.576

Table 2 presents the results for on the Affective Norms corpus (for all categories the average score was used, scores range from -5 to 5).

Table 2. Mean and standard deviation (σ) for all affective word categories in the Affective Norms corpus (Valence, Arousal, Imageability and Abstract-/Concreteness) for the data set, with the classes Control Group, Burnout and Depression.

Class	Val. mean	Val. σ	Arou. mean	Arou. σ	Img. mean	Img. σ	Abstr. mean	Abstr. σ
Contr.	3.303	0.682	2.971	0.519	3.370	0.689	3.144	0.624
Burn.	3.617	0.818	3.046	0.570	3.555	0.811	3.215	0.641
Depr.	3.518	0.653	2.928	0.520	3.331	0.570	3.046	0.504

For BAWL-R, we found that only the differences found for *arousal* and *imageability* are statistically significant, whereas the differences for *valence* yielded a high p-value of 0.856 in the Kruskal-Wallis test. Therefore, the differences in the *valence* category have to be considered coincidental. For the Affective Norms corpus, we found that all of the differences were statistically significant, i.e., each yielded p-value < 0.05 in the Kruskal-Wallis test.

Table 3. p-values for BAWL-R corpus and for the Affective Norms corpus using the Kruskal-Wallis Test. The p-value is <0.05 in all cases except BAWL-R *valence*, which means that the differences in this case have to be considered coincidental.

Corpus	Valence	Arousal	Imageability	Abstract-/Concreteness
Affective Norms	$4.29 \times 10{-}9$	$4.05\times 10{-}3$	$1.57 \times 10{-}4$	$2.53 \times 10{-}4$
BAWL-R	**0.856**	$3.62\times 10{-}5$	$2.13\times 10{-}6$	-

5. Discussion

BAWL-R: The results show that the Burnout texts scored highest in both the *imageability* and the *arousal* category. The latter is in accordance with [5], where Burnout is associated with high arousal. It has to be noted that the differences are on the small scale, being smaller than the average standard deviation for both cases. The higher value for *arousal* compared to depression would be in accordance with the notion of the "upward" cycle of burnout and the "downward" cycle of depression by [11]. There is little to be found in literature on the interpretation of the *imageability* score. A study by Raghunath et al. [12] has found an association between high imageability in certain word classes and an elevated anxiety level. However, the words whose usage was found to be correlated to anxiety were not just high in imageability but also low in valence, therefore we cannot transfer these findings directly to our scenario. Also, and maybe more importantly, the authors seem to have a different understanding of imageability than e.g., the creators of the Affective Norms corpus [8], for whom high imageability relates to "things we can actually see" [8], whereas Raghunath et. al. [12] define imageability as "the degree to which words evoke mental images" [12] - which is clearly not the same.

Affective Norms Corpus: We recall that this corpus differs from the BAWL-R corpus, in that it is much larger and contains most words used in daily conversation. Thus, in this case virtually every word in a sentence has a score, of which again we took the average score. Again, the Burnout class scored highest in all categories. Notably, both Burnout and Depression scored higher in the *valence* category than Control group, which is counterintuitive and contradicts the findings of [6], which found that essays written by depressed college students contained a lower ratio of positively valanced words than those written by students who do not suffer from depression or have recovered from it (note that [6] used a different dictionary for the English language and also a different score). The high valence score for Burnout also contradicts the findings of [5], who used the same score as we did. The higher *arousal* score of the Burnout class compared to Depression is again clearly visible and confirms the Burnout-Depression-dichotomy that is advocated by [11].

Limitations: One problem that stands out in the analysis of *valence* scores, is the AWL approach's inability to capture negation or negative modifiers. For example, the following high scoring text for the Valence category illustrates this well:

«Ungenügende Kontrolle und Wertschätzung führten bei mir zu einer hohen Selbstüberforderung.»

The word "Selbstüberforderung" (a state where one has overburdened oneself and is struggling to meet self-imposed or external demands) has the lowest valence score (1.435) and the word "Wertschätzung" (appreciation) the highest one (8.493). The negative modifier "ungenügend" (insufficient), which basically negates the presence of the word it modifies, is present twice. However, this is completely missed by this model ("ungenügend" is not even factored in because there is another word with a lower valence score) and this sentence's valence score is dominated by the outlier "Wertschätzung", which is in fact semantically negated. It stands to discussion why the works by [6] and [5] seem not be affected by this problem. We do have to take into account that these works analyzed texts in a different language, and usage of negation of a positive word vs. usage of a negative word could also depend on language and culture.

6. Conclusion

By analyzing the affective scores of texts from three different groups (Burnout, Depression and Control group), it was found that both Burnout and Depression show significantly higher *arousal* than the Control group. This gives first indication that this research direction is promising to enable new methods for clinical intervention in the future using text-based data. However, limitations have been identified that will need to be addressed in future work. Further research is required in the field of NLP for mental health in order to extend the technology and – in collaboration with clinical partners - to define the clinical requirements for tools based on such technologies.

The authors gratefully thank the Swiss National Science Foundation (SNSF), since the Bachelor thesis leading to this paper was conducted in the context of a SNSF Spark project.

References

[1] Maslach, C., Jackson, S. E., & Leiter, M. P. (1997). Maslach burnout inventory. Scarecrow Education.
[2] Demerouti, E., & Bakker, A. B. (2008). The Oldenburg Burnout Inventory: A good alternative to measure burnout and engagement. Handbook of Stress and Burnout in Health Care, 65, 78.
[3] Jaggi, F. (2019). Burnout praxisnah. Lehmanns Media.
[4] Schonfeld, I. S., & Bianchi, R. (2016). Burnout and depression: two entities or one?. Journal of Clinical Psychology, 72(1), 22-37.
[5] Mäntylä, M., Adams, B., Destefanis, G., Graziotin, D., & Ortu, M. (2016, May). Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity?. In Proceedings of the 13th International Conference on Mining Software Repositories (pp. 247-258).
[6] Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. Cognition & Emotion, 18(8), 1121-1133.
[7] Spinczyk, D., Bas, M., Dzieciątko, M., Maćkowski, M., Rojewska, K., & Maćkowska, S. (2020). Computer-aided therapeutic diagnosis for anorexia. Biomedical Engineering Online, 19(1), 1-23.
[8] Köper, M., & Im Walde, S. S. (2016, May). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 german lemmas. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 2595-2598).
[9] Nath, S., and Kurpicz-Briki, M. (2021). BurnoutWords - detecting burnout for a clinical setting. In Proceedings of the 10th International Conference on Soft Computing, Artificial Intelligence and Applications (SCAI 2021), CS & IT Conference Proceedings, Zurich, Switzerland.
[10] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2011). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 62(2), 419.
[11] Brühlmann, T. O. N. I. (2010, February). Burnout und Depression Überschneidung und Abgrenzung. In Swiss Medical Forum (Vol. 10, No. 08, pp. 148-151). EMH Media.
[12] Raghunath, B. L., Mulatti, C., Neoh, M. J. Y., Bornstein, M. H., & Esposito, G. (2021). The associations between imageability of positive and negative valence words and fear reactivity. *Psychiatry International*, *2*(1), 32-47.
[13] Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical report C-1, center for research in psychophysiology, University of Florida.
[14] Võ, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534-538.
[15] Datta, S., Bernstam, E. V., & Roberts, K. (2019). A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *Journal of biomedical informatics*, *100*, 103301.
[16] Zeng, Z., Deng, Y., Li, X., Naumann, T., & Luo, Y. (2018). Natural language processing for EHR-based computational phenotyping. *IEEE/ACM transactions on computational biology and bioinformatics*, *16*(1), 139-153.