BURNOUTWORDS - DETECTING BURNOUT FOR A CLINICAL SETTING

Sukanya Nath and Mascha Kurpicz-Briki

Institute for Data Applications and Security IDAS,
Bern University of Applied Sciences, Biel/Bienne, Switzerland

ABSTRACT

Burnout, a syndrome conceptualized as resulting from major workplace stress that has not been successfully managed, is a major problem of today's society, in particular in crisis times such as a global pandemic situation. Burnout detection is hard, because the symptoms often overlap with other diseases and syndromes. Typical clinical approaches are using inventories to assess burnout for their patients, even though free-text approaches are considered promising. In research of natural language processing (NLP) applied to mental health, often data from social media is used and not real patient data, which leads to some limitations for the application in clinical use cases.

In this paper, we fill the gap and provide a dataset using extracts from interviews with burnout patients containing 216 records. We train a support vector machine (SVM) classifier to detect burnout in text snippets with an accuracy of around 80%, which is clearly higher than the random baseline of our setup. This provides the foundation for a next generation of clinical methods based on NLP.

KEYWORDS

Natural Language Processing, Psychology, Burnout, Machine Learning.

1. INTRODUCTION

Stress causes several syndromes or diseases and is a major problem of today's society. The *Stress in America Report 2019* from the American Psychological Association has shown that Americans believe that a healthy stress level is on average 3.8 (scale ranging from 1 to 10, where 10 is "a great deal of stress" and 1 is "little or no stress"), however, they report to have experienced in average a stress level of 4.9 [1]. Nearly 8 in 10 Americans say that the coronavirus pandemic is a significant source of stress in their lives [2].

Sometimes, this stress can lead to a burnout syndrome. Last year, the WHO has included burnout in the 11th Revision of the International Classification of Diseases (ICD-11) as a syndrome[1]. In ICD-11, burnout is defined as follows:

"Burnout is a syndrome conceptualized as resulting from chronic workplace stress that has not been successfully managed. It is characterised by three dimensions: 1) feelings of energy

[1] https://icd.who.int/browse11/l-m/en\#/http://id.who.int/icd/entity/129180281

David C. Wyld et al. (Eds): MLDS, NECO, SEMIT, IBCOM, SPPR, SCAI, CSIA, ICCSEA - 2021
pp. 177-191, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111815

depletion or exhaustion; 2) increased mental distance from one's job, or feelings of negativism or cynicism related to one's job; and 3) a sense of ineffectiveness and lack of accomplishment." [2]

In the global pandemic crisis around COVID-19, research has shown an increase of burnout, in particular on frontline personnel in the health sector [3] [4].

Burnout identification is complex, because it overlaps with other syndromes [5] and multiple definitions exist [6]. For example, fatigue is a common symptom for both depression and burnout [7]. To identify burnout in clinical intervention, inventories are used. Inventories are psychological tests, where the person concerned fills out a questionnaire. The currently used metric, in both practice and most studies, measures burnout with self-test inventories, and has been criticized in the literature [8] [9].

Inventories with scaling questions, even though often used in practice, have major limitations. In personality inventories, people tend to fake their results [10]. This risk might be increased with a delicate topic such as burnout. Furthermore, extreme response bias (ERB), i.e., the tendency of some respondents to choose the highest or lowest option, is a well-known issue [11]. Other research also reports defensiveness (denying symptoms) and social desirability (to show an exaggeratedly positive image) in inventories [12].

Literature agrees that the Maslach Burnout Inventory (MBI) [13] [14] and the Tedium Measure [15], later called Burnout Measure, are most commonly used in practice and research [16] [9]. The MBI is an introspective psychological inventory and consists of 22 items in three dimensions: emotional exhaustion, depersonalization and personal accomplishment. The Tedium Measure [15] is often used as an alternative to MBI. This inventory consists of 21 items, and each item has to be classified by frequency.

Apart the methods using inventories, burnout components can also be identified by independent judges in interview extracts [17] [9]. The drawback of interviews or free-text questions is that they result in large overhead (interviews, transcription and interpretation) and therefore promising approaches are often not further explored [9].

In this paper, we present BurnoutWords, a dataset based on extracts from conversations with burnout patients, a control group and experts. We provide insights into the wording of burnout patients, extract features from the dataset and allow further research to develop new approaches to enable new clinical methods with text-based burnout identification. We train a burnout classifier using Support Vector Machine (SVM) models and reach an accuracy with a clear improvement over the random baseline.

Whereas most related work in the field investigates social media content, our dataset includes extracts from interviews with burnout patients, aggregated from different previous work and pre-processed for automated evaluation. Using interview extracts from real burnout patients reduces the noise as compared to social media data. This allows to fill the gap between natural language processing (NLP) and the application of the new technologies to develop new methods for application in clinical psychology.

Our paper contributed to the current state-of-the-art by providing a new type of burnout detection dataset, which is based on interview extracts instead of social media data. We furthermore demonstrate how such a dataset can be used to develop new technologies for clinical psychology.

[2] https://icd.who.int/browse11/l-m/en\#/http://id.who.int/icd/entity/129180281

Therefore, our work creates the foundation for future work in the field and new methods for clinical burnout detection.

In section 2, we described the related work with regard to the application of NLP methods and existing datasets for burnout and depression detection. Then, in section 3, we describe the BurnoutWords dataset and how it has been assembled. In section 4 we describe our experimental setup. We then describe (section 5) and discuss (section 6) the results before concluding the article in section 7.

2. RELATED WORK

2.1. Natural Language Processing for Burnout/Depression Detection

Few works exist for burnout detection in general, and as far as of our knowledge, no comparable dataset exists. Burnout is not a disease, but considered as a syndrome, making its detection often hard, because its symptoms overlap with other syndromes or diseases [5] in particular depression. Therefore, we consider additionally the related work for depression detection with the help of NLP.

In particular, we focus on text-based data and therefore do not further discuss approaches based on biomarkers, vocal data (e.g., [18]) or image-based approaches (e.g., [19]). It has been shown that in clinical psychology, written language often plays a central role in diagnosis [20]. The authors summarize the linguistic and social indicators that have been applied to automatic depression detection in different contexts, e.g., narratives written by college students with or without depression [21].

The work focusing on data from social media is the majority of available research in the field, but also work about online forums exist, e.g., [22]. In general, in the related work, the concerned users have been identified by using a screening survey, their public sharing of a diagnosis, or their membership in an online forum [23].

Whereas depression is often diagnosed as being present or absent (e.g., presence of depression symptoms on facebook posts for college students [24]), a model based on survey answers and the language used in facebook posts assesses the severity of depression [25] using the depression facet scores of the Big 5 item pool [26]. Changes in depression across seasons are observed which confirms results from clinical psychology. It has been shown that explicit depression references are rare, but when they appear, they are strong indicators for a real-life depression [27]. Other research targets particular types of depression: a study analyzing Twitter data provides an approach to identify mothers at risk for postpartum depression [28] (complemented later with shared facebook data [29]).

Different technologies are used for the detection of depression on social media, e.g., based on the linguistic inquiry word count (LIWC) lexicon [30] containing multiple psychological constructs. Some more recent approaches use deep learning architectures to achieve better results, which requires a large amount of data. Research has shown how a neural network can be designed to detect depression with limited data and without any exhaustive feature engineering [31], presenting a neural network architecture that optimizes word embeddings. SenseMood [32] is applying a CNN-based classifier and Google's Bert model [33] on posted images and tweets from users with or without depression, combining visual and textual features. They are using a dataset previously presented in research [34], which contains a set of users with anchor tweets matching the strict pattern that they have been diagnosed depression.

In an enterprise context, studies have used the Valence-Arousal-Dominance (VAD) model [35] to study productivity or risk for burnout in data such as issues and comments from a software development management tool [36]. Burnout risk is measured as low valence and dominance, and high arousal.

Whereas most work focusses on social media data only, it has been shown that the language from Facebook posts can be used to predict depression for consenting individuals recorded in medical records [37]. Another study has examined how language patterns on social media change prior to emergency department visits [38].

When mental health and in particular depression is studied from social media data only, the user's mental state is reflected from published posts and thoughts. This can lead to limitations in the data sets. In our work, we train our model based on text data transcribed from interviews with confirmed burnout patients. This provides the basis for new approaches to be applied not only on new data from social media, but also as a tool support for professionals in clinical intervention.

2.2. Existing Datasets for Burnout/Depression Detection

Social Media: The dataset from ReachOut triage shared task [39] consists of 65'024 forum posts that were manually labelled by expert judges by their severity. This dataset addresses different types of mental health crisis, not focusing particularly on a specific diagnosis or syndrome like depression or burnout. The dataset from the CLPsych 2015 shared task on depression detection was constructed using tweets from users that have stated explicitly that they have been diagnosed with depression or PTSD [40] [41]. For each user up to 3000 recent public tweets were added to the dataset. Other work has collected data from Twitter, creating three datasets: Depression, Non-Depression, and Candidate-Depression [34]. Based on the user's information and current tweet, an anchor tweet [41] to determine the mental health was selected. To simulate observation over time as in a clinical setup, tweets following in the next month after the anchor tweet were also considered for the selected users.

Clinical Interviews/Crisis Counselor: The Crisis Text Line dataset [42] is based on the data collected by a 24/7 text-based crisis support hotline. It can be made available for researchers upon application and contains labeled data concerning different topics including depression. The Distress Analysis Interview Corpus (DAIC) [43] [44] contains clinical interviews conducted by humans and agents in English language. The semi-structured clinical interviews provide a contribution to detect psychological distress conditions such as anxiety, depression or post-traumatic stress disorder. Each interview does also include a depression score from the PHQ-8 inventory [45]. The data has been transcribed and annotated for a variety of features[3]. The AVEC 2014 challenge about depression detection [46] included also interviews in German language which are available as video and audio traces. The original challenge did not provide transcripts. However, other work has used automatic speech recognition technology to transcribe interviews from this corpus and make it available upon request [47].

Disease Information from Clinical Notes: The SemEval-2014 Task 7 dataset [48] includes clinical notes which are annotated with disease/disorder mentions. The dataset is based on the Shared Annotated Resources (ShARe) project[4]. Recent research has provided a large public dataset for clinical motivated symptom extraction from clinical notes [49]. Such approaches are helpful to automatically process and aggregate clinical data. In the dataset presented in this paper,

[3] https://dcapswoz.ict.usc.edu/
[4] http://share.healthnlp.org

we go one step further: in addition to the clinical symptoms, thoughts and sentiments expressed by the patients are included in the dataset and thus in the classifier.

Based on our literature research, we conclude that there is currently no publicly available dataset that provides labeled patient interview data in text form for burnout detection, neither for the English nor the German language. In this paper, we therefore introduce a first version of the BurnoutWords dataset for the German language to the research community to enable future research in the field.

3. THE BURNOUTWORDS DATASET

3.1. Dataset Content and Origin

This paper presents the BurnoutWords dataset, containing extracts from interviews in the German language. The interviews have been conducted in the context of different previous work where extracts have been published, and have been collected and aggregated in our research. The dataset contains texts and corresponding labels, whether the texts describe the current burnout situation (label *burnout*), or the view of a person that has no (more) burnout, or potential measures and prevention of burnout (label *noburnout*). Table 1 shows an extract from the dataset to illustrate the format.

Table 1. Example: one record of the BurnoutWords dataset. Label 1 indicates burnout (label 0 would indicate the class noburnout). Translation to English: (The) doctor made the diagnosis. (The) doctor is a good colleague of mine and takes a lot of time for me. I only realized it ca. 2 months later. It needs time until one can accept it. I informed myself about the disease. The information on the internet did not help me. The doctor could give me advice.

Text	Label
Arzt stellte Diagnose. Arzt ist guter Kollege von mir und nimmt sich viel Zeit für mich. Ich realisierte es aber erst ca. zwei Monate später. Braucht Zeit bis man es akzeptieren kann. Informierte mich über die Krankheit. Die Informationen im Internet haben mir nicht geholfen. Der Arzt konnte mir Auskunft geben.	1

A part of the extracts has been published in a thesis [50] and, in agreement with the author of the thesis, the extracts have been re-used in this research. The thesis studies burnout in medical doctors. The author conducted interviews with confirmed burnout patients (i.e., the medical doctors), and with a control group of medical doctors not having burnout. We collected and pre-processed the relevant data from the thesis and labeled it into the following classes: burnout patients (*burnout*), and control group (*noburnout*).

More interview extracts have been collected from a book investigating on burnout in an enterprise context [51]. The author conducted interviews with seven different patients that previously had a burnout, working in different domains (i.e., IT, marketing, engineering, health care, coaching and journalism). We identified the questions concerning descriptions of their burnout, including thoughts and symptoms of that time, and included their answers to the class *burnout* of our dataset. Furthermore, we included answers to questions discussing their current life after burnout, and hints to avoid or handle burnout in the data labeled as *noburnout*.

The author [51] also conducted interviews with experts, including medical doctors, psychologists and coaches. We selected answers describing the symptoms and emotions of burnout patients and labeled them as *burnout*. We identified questions concerning the handling and prevention of burnout and included the answers in the class *noburnout*.

The original data provides some demographic information about the participants in the interviews; however, it is partially modified for privacy reasons. Since only extracts from the interviews are available (the entire interviews cannot be used for further research due to data protection law limitations), the text content is not equally distributed between the participants. Due to this and the partial anonymization of demographic data, it is not considered in the dataset. The dataset contains an overall number of 293 records, 216 in the class *burnout* (73.72%) and 77 in the class *noburnout* (26.28%). Due to the fact that only extracts of the interviews were published, a majority of those extracts comes from the burnout patients and not the control group.

3.2. Data Ethics

The local ethics commission has approved to use this publicly available data for our research. Technology itself is neither good nor bad. However, the ethical aspects of the application of such technologies must be discussed. In our research, we focus on providing the required technology to allow future clinical methods to assess mental health. Such methods should be based on the voluntariness of the involved individuals and provide benefits for individuals and the society. When such technologies are misused as an instrument of power, for example using them in a company, to automatically assess written communication of collaborators without their explicit consent or with negative consequences in case of not giving their consent, this would be a major ethical issue. We therefore provide access to our dataset only upon request for future research in an academic or clinical context.

4. EXPERIMENTAL SETUP

4.1. The Words of Burnout Patients

Preprocessing: First of all, characters such as brackets or quotation marks have been removed. Then, double whitespaces have been removed. Finally, the character *ß* is replaced with *ss*, since this character is not common in all the German speaking areas. The pre-processed data is stored in an additional column of the data table. The original data is also available for further research. Responses with less than 200 characters were excluded, since we assume that they do not contain enough information.

In a first step, we examine the words that are being used by the burnout patients and the control group respectively. We therefore consider the ten most used nouns (lemmatized) for each group, excluding the words that appear in both groups. We assume that those words are conversational words that do not add value for the resolution of the question whether there is a burnout or not.

4.2. Model Training

We split the data into a training and a test set. We are using around 70% of the data for feature extraction and training of the model, and around 30% of the data to evaluate the model against completely new data that it was not trained with. Figure 1 shows the experimental setup as explained in detail in the following subsections.

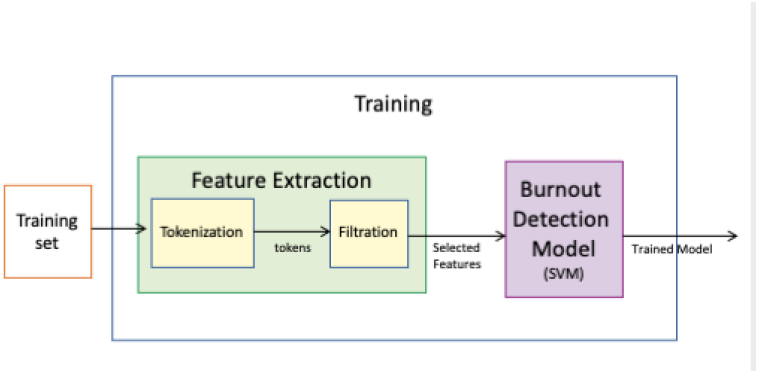

Figure 1. The experimental setup of our training including feature extraction and the burnout detection model.

4.2.1. Feature Extraction

The first part of the training process is the feature extraction, which can be further divided into two phases. The first is the tokenization phase. At the end of this phase, we produced a list of all the tokens which could be potential features. The second phase is the feature filtration stage. In this phase, evaluated the goodness of the features. The features which passed the filtration phase are considered as selected features.

Tokenization: In order to generate features, we divided the training set into the two classes - *burnout* and *noburnout* - and aggregated the contents of each class. At the end of this first step, we had one large file containing the text from all the *burnout* cases, and another large file of text from all the *noburnout* cases. The *burnout* text file was larger than the *noburnout* text file, as we had a higher number of burnout cases. We fed both text files to a function which created a list of tokens (using SpaCy[5]). We filtered the stop-words as they contain little information about the topic. Similarly, we filtered out the punctuations as they also contain little information about the topic and in particular because the interview excerpts are a form of notes and not directly written by the interviewee. We produced a list of tokens for the *burnout* class and another list of tokens for the *noburnout* class.

Feature Filtration: In this phase, our aim is to select those features which can help us achieve a clear boundary between the two classes during the model training. As such, we are interested in those features which occur with a much higher frequency in one class but not in the other class. At first, the lists of tokens of both classes were counted for frequency of occurrence, and a feature dictionary was created such that the key is the token/feature and the value is the frequency of this token/feature. As mentioned before, we have more *burnout* cases than *noburnout* cases. To reduce the advantage of length for the *burnout* class, we further normalized each of the values of this feature dictionary with the total number of tokens in the respective class text file. We then computed the difference between the frequencies of both classes: we took all the features present in the *burnout* and *noburnout* feature dictionaries and note the difference of their normalized frequencies:

$$\text{diff} = \text{burnout}_{\text{norm.freq}} - \text{noburnout}_{\text{norm.freq}}$$

We created a new dataset (dataframe) containing the following columns: feature, burnout frequency normalized, noburnout frequency normalized, difference. We sorted using the

[5] https://spacy.io

frequency difference. Note that if we have a positive difference, the *burnout normalized frequency* is greater than the one for *noburnout*. In other words, the feature has occurred relatively more frequently in the *burnout* text rather than *noburnout* text. In the same way, if we have a negative difference, the feature has occurred more frequently in the *noburnout* text than in the *burnout* text.

For a given feature set size, say N (where N is an even number), we selected N/2 features which had the greatest positive difference and another N/2 features with the greatest negative difference. In this way, we selected those features which had the greatest (positive or negative) difference between the two classes in the training set. We did not consider features which occur relatively commonly in both classes as they are not very discriminatory.

4.2.2. Burnout Detection Model

This section describes the second part of the training phase. In this phase, we transform the training and test sets into their corresponding feature vector matrix according to the previously selected features. For the purpose of training our burnout detection model, only the training set is used, while the test set is used for model evaluation.

During training, we loop over the different features identified in the previous experiment, using different feature set sizes (N=10,20,...,100). For each iteration we created Support Vector Machine (SVM) models using different configurations. Support Vector Machine models [52] [53] are supervised learning models that attempt to find a hyperplane separating the classes in N dimensions where N is the number of features. A set of data points, called the support vectors, is located close to the hyperplane and helps to orient the hyperplane, such that the maximum separation of the classes, i.e., larger margins, may be achieved. Support Vector Machines use mathematical functions called kernels to transform the feature space such that nonlinear boundaries are transformed to linear boundaries separable by a hyperplane. The linear kernel is the simplest and works well with linearly separable data. Some other Kernels which are commonly used [54] are RBF (Gaussian radial basis function), polynomial and sigmoid (sometimes being referred to as neural network model) kernels. Therefore, we applied the linear, RBF, polynomial and sigmoid kernels in our experiments. The parameter C is the regularization parameter controlling the trade-off between misclassifications and the width of the margin. A large value of C leads to an overfit wiggly boundary whereas a low value of C causes a smoothened boundary [54]. In our proposed model, the value of C used was 1.

In Figure 2 and 3, we show ten selected features with large positive difference and another ten features with large negative difference.

Word	BurnoutFreqNorm	NoBurnoutFreqNorm	Difference
Und	0.018621	0.010431	0.008190
Sachen	0.004325	0.000348	0.003977
mal	0.016338	0.012865	0.003473
Du	0.003604	0.000695	0.002909
Sie	0.005646	0.002782	0.002865
teilweise	0.003003	0.000348	0.002656
sagen	0.009851	0.007302	0.002549
hab	0.004085	0.001739	0.002346
Praxis	0.002643	0.000348	0.002295
Da	0.002883	0.000695	0.002188

Figure 2. Words with large positive difference. Translation: *Und: and, Sachen: things/stuff, mal: one time, Du: you, Sie: she/her/they/you(formal), teilweise: partially, sagen: to say, hab: (I) have, Praxis: (medical) practice, Da: there*

Word	BurnoutFreqNorm	NoBurnoutFreqNorm	Difference
Also	0.005887	0.009388	-0.003501
Arbeit	0.003484	0.006606	-0.003122
Patienten	0.004205	0.007302	-0.003097
Menschen	0.001081	0.004172	-0.003091
Unternehmen	0.000000	0.002782	-0.002782
Nein	0.000000	0.002782	-0.002782
irgendwas	0.000721	0.003477	-0.002756
okay	0.001321	0.003825	-0.002503
Ich	0.012134	0.014604	-0.002470
sage	0.004445	0.006606	-0.002161

Figure 3. Words with large negative difference. Translation: *Also: so/thus, Arbeit: work, Patienten: patients, Menschen: people, Unternehmen: company, Nein: no, irgendwas: something/anything, okay: okay, Ich: I, sage: (I) say*

5. RESULTS

5.1. The Words of Burnout Patients

Figure 4 and 5 show the word clouds for the ten most used nouns for the text labeled as *burnout* and *noburnout* respectively. The size of the font reflects the quantity (number of times) of the words appearing in the dataset.

Figure 4. Top ten nouns from the class *burnout*; translation to English: *Beispiel: example, Prinzip: principle, Monat: month, Chef: boss, Termin: appointment/deadline, Dienst: shift/service, Recht: right/law, bisschen: a bit, Ja: yes, Oberärztin: senior physician (female)*

Figure 5. Top ten nouns from the class *noburnout*; translation to English: *Fähigkeit: capability/competence, Besseres: something better, Radiologie: radiology, Auto: car, Bürokauffrau: office clerk (female), Lob: praise, Feedback: feedback, Abstrich: tradeoff, Berufsbild: job profile, Medizin: medicine*

We observe in the results that for burnout patients, stress-related topics are of high importance (e.g. senior physician, shift/service, boss), as well as factors concerning the time (e.g. month,

appointment/deadline). On the other side, for the texts from the class *noburnout*, positive work-related topics are more common (e.g. competence, praise, feedback).

5.2. Burnout Detection Model

Figure 6 depicts the results of our experiments, relating the number of features considered to the accuracy obtained for models using linear, RBF, polynomial and sigmoid kernel SVM. The random baseline is 73.4% (percentage of burnout cases), since the test and training set contain a majority of records of the class *burnout*.

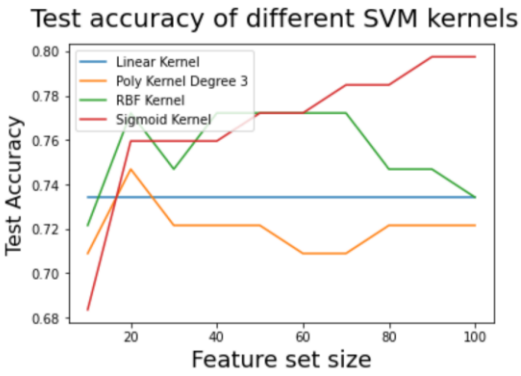

Figure 6. Test accuracy for different feature set sizes of the SVM classifiers using linear, RBF, polynomial and sigmoid kernels

The sigmoid kernel performs the best (79.7%) at around feature set size 90. It can be observed that increasing features continues to improve the accuracy in case of sigmoid kernel. We have shown the features set size up to 100, after this point increasing the feature set size caused a fall in test accuracy for the sigmoid kernel. The linear kernel likely predicts only one class and therefore lies on the random baseline. This indicates that our data is likely not linearly separable. The polynomial kernel of degree 3 goes up to an accuracy of 74.7% at around feature set size 20. The RBF kernel goes up to 77.2% at feature set size 40 and then stabilizes before falling down.

We are thus able to train a classifier to detect burnout in text snippets with an accuracy clearly above the random baseline. Figure 7 shows the confusion matrix for the SVM with sigmoid kernel for a feature set size of 90.

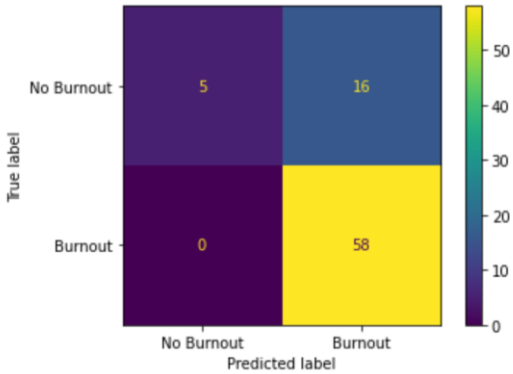

Figure 7. Confusion matrix of the SVM classifier with sigmoid kernel for a feature set size of 90.

6. DISCUSSION

In the feature selection we identified the most typical words for the classes *burnout* and *noburnout*. Interestingly, the word *Du (engl: you)* is a feature for the class *burnout*, whereas the word *Ich (engl: I)* is a feature of the class *noburnout*. This is counterintuitive since previous work has shown that the common use of first-person singular can be interpreted as a sign of depression [20] [21].

Our burnout classifier reaches an accuracy of 79.7%, which is clearly over the baseline of 73.4%, even though the dataset contains a limited number of records. We therefore see a large potential for text-based approaches to identify burnout.

The confusion matrix in Figure 7 indicates that our approach is more likely to present false positives (i.e. patients **not** having burnout being classified as having burnout) and less false negatives (i.e. patients having burnout, being classified as **not** having burnout). This can be partially explained by the fact that the dataset contains more records for the class *burnout*. In a clinical setting, this is preferable, since symptoms of burnout often overlap with other diseases or syndromes. We prefer that a doctor checks closely a patient because the tool indicated a potential burnout, and finds out he/she does have another syndrome, rather than following the prediction of the tool leading to the release of a person that might need further help.

The interaction between the tool and the clinical professional is very important. The machine can provide tools for clinical professionals, but cannot replace them. Machines are not able to take ethical decisions, and cannot carry the consequences of their decisions. Therefore, a good tool provides decision-support in a clinical context, in a similar way nowadays inventories are doing. Such inventories, as for example the Maslach Burnout Inventory, have known limitations (e.g. users faking their results [10], extreme response bias [11], defensiveness and social desirability bias [12]). Free-text questions or interview transcripts can provide interesting new possibilities in the detection of burnout. The work presented in this paper confirms that such approaches are feasible and we assume that with more data, even better results than the ones presented in this paper can be achieved.

Based on the results presented in this paper, the following challenges will need to be addressed in future work. Currently, the dataset does not differentiate between men and women, since the data is completely anonymized. However, in future work more detailed profiles of patients will need to be considered and carefully examined. It has been shown that the way men and women communicate is different [55], and that gendered wording can have an important impact. Due to the available data, we currently only include German extracts from interviews in the dataset. We expect to connect to clinical partners and extend our dataset with extracts in English and French. This will allow us to target a larger community and we plan to explore whether our findings for German are also applicable for the other languages.

7. CONCLUSION

Most work using NLP to detect depression uses data from social media, whereas the specific case of burnout is rarely addressed. In this paper, we presented BurnoutWords, a dataset based on interview extracts from burnout patients in the German language. The dataset was assembled with data from previous research and not from social media, as opposed to most of the existing research in the field. It allows first insights into the wording of burnout patients, and will be extended in the future with additional data. We also plan to investigate on the wording of burnout patients in other languages.

Since the existing work in the field of burnout detection is very limited, a comparative evaluation to similar datasets is currently not possible. Therefore, in future work, we want to address also the area of burnout detection in social media, to provide comparable validation measures of our approaches.

We showed that upon such data, a classifier using Support Vector Machines with an accuracy clearly higher than the random baseline can be trained. With the sigmoid kernel, an accuracy of almost 80% was achieved, as compared to the random baseline of 73.4% (percentage of burnout cases in the dataset). Given that this first version of the dataset is very limited, the result is very promising. This work creates the foundation for future work in the field and new methods for clinical burnout detection.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the funding for this work by the Swiss National Science Foundation in the context of the Spark call.

REFERENCES

[1] American Psychological Association. 2019. Stress in america 2019.
[2] American Psychological Association. 2020. Stress in america 2020: a national mental health crisis.
[3] Elie Azoulay, Jan De Waele, Ricard Ferrer, Thomas Staudinger, Marta Borkowska, Pedro Povoa, Katerina Iliopoulou, Antonio Artigas, Stefan J Schaller, Manu Shankar Hari, et al. 2020. Symptoms of burnout in intensive care unit specialists facing the covid-19 outbreak. 10(1):1–8.
[4] Takahiro Matsuo, Daiki Kobayashi, Fumika Taki, Fumie Sakamoto, Yuki Uehara, Nobuyoshi Mori, and Tsuguya Fukui. 2020. Prevalence of health care worker burnout during the coronavirus disease 2019 (covid-19) pandemic in japan. *JAMA network open*, 3(8):e2017271–e2017271.
[5] Ferdinand Jaggi. 2019. *Burnout praxisnah*. Lehmanns Media.
[6] Wilmar B Schaufeli, Christina Maslach, and Tadeusz Marek. 2017. The future of burnout. In Professional burnout, pages 253–259. Routledge.
[7] Irvin Sam Schonfeld and Renzo Bianchi. 2016. Burnout and depression: two entities or one? *Journal of clinical psychology*, 72(1):22–37.
[8] Sybille Hautle. 2012. Das burnout-syndrom: Erhebung typischer merkmale zur herleitung diagnostischer fragen für ein selbstbeurteilungsinstrument.
[9] M Burisch. 2010. Das burnout-syndrom (4. aktual. aufl.).
[10] Christine Elizabeth Lambert. 2013. *Identifying faking on self-report personality inventories: Relative merits of traditional lie scales, new lie scales, response patterns, and response times*. Ph.D. thesis.
[11] Gaël Brulé and Ruut Veenhoven. 2017. The '10 excess' phenomenon in responses to survey questions on happiness. *Social Indicators Research*, 131(2):853–870.
[12] Margot M Williams, Richard Rogers, Allyson J Sharf, and Colin A Ross. 2019. Faking good: An investigation of social desirability and defensiveness in an inpatient sample with personality disorder traits. *Journal of personality assessment*, 101(3):253–263.
[13] Christina Maslach, Susan E Jackson, Michael P Leiter, WB Schaufeli, and RL Schwab. 1996. Maslach burnout inventory manual. menlo park. *CA: Mind Garden*, pages 191–218.
[14] C Maslach, SE Jackson, MP Leiter, WB Schaufeli, and RL Schwab. 1986. Maslach burnout inventory instruments and scoring guides forms: General, human services, & educators. *Health and Quality of life Outcomes*, 7:31.
[15] Elliot Aronson, Ayala M Pines, and Ditsa Kafry. 1983. Ausgebrannt. *Vom überdruß zur selbstentfaltung*.
[16] Viviana Abati. 2007. *Burnout: Erkennen - vorbeugen - verhindern*. SPEKTRAmedia.
[17] Cary Cherniss. 1980. *Professional burnout in human service organizations*. Praeger Publishers.
[18] Nadee Seneviratne and Carol Espy-Wilson. 2020. Deep learning based generalized models for depression classification. *arXiv preprint arXiv:2011.06739*.
[19] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6:1–12.

[20] Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A cross-modal review of indicators for depression detection systems. In *Proceedings of the fourth workshop on computational linguistics and clinical psychology—From linguistic signal to clinical reality*, pages 1–12.

[21] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

[22] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

[23] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

[24] Megan A Moreno, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon, and Tara Becker. 2011. Feeling bad on facebook: Depression disclosures by college students on a social networking site. *Depression and anxiety*, 28(6):447– 455.

[25] H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125.

[26] Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

[27] Yaakov Ophir, Christa SC Asterhan, and Baruch B Schwarz. 2019. The digital footprints of adolescent depression, social rejection and victimization of bullying on facebook. *Computers in Human Behavior*, 91:62–71.

[28] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3267–3276.

[29] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638.

[30] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

[31] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

[32] Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. Sensemood: Depression detection on social media. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 407–411.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understand- ing. *arXiv preprint arXiv:1810.04805*.

[34] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

[35] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. 47. University of Illinois press.

[36] Mika Mäntylä, Bram Adams, Giuseppe Destefanis, Daniel Graziotin, and Marco Ortu. 2016. Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity? In *Proceedings of the 13th international conference on mining software repositories*, pages 247–258.

[37] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preotiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

[38] Sharath Chandra Guntuku, H Andrew Schwartz, Adarsh Kashyap, Jessica S Gaulton, Daniel C Stokes, David A Asch, Lyle H Ungar, and Raina M Merchant. 2020. Variability in language used on social media prior to hospital visits. *Scientific reports*, 10(1):1–9.

[39] David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 118–127.

[40] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31– 39.

[41] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

[42] Ge Ge Jackie Chen. 2014. *Visualizations for mental health topic models*. Ph.D. thesis, Massachusetts Institute of Technology.

[43] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128.

[44] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10.

[45] K Kroenke, RL Spitzer, and JBW Williams. 2001. The patient health questionnaire (phq-9)–overview. *J9 Gen Intern Med*, 16:606–16.

[46] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3– 10.

[47] Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE spoken language technology workshop (SLT)*, pages 136–143. IEEE.

[48] Sameer Pradhan, Wendy Chapman, Suresh Man, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Citeseer.

[49] Jackson M Steinkamp, Wasif Bala, Abhinav Sharma, and Jacob J Kantrowitz. 2020. Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. *Journal of biomedical informatics*, 102:103354.

[50] Edith Rahner. 2011. *Das Burnout-Syndrom bei Ärzten: eine qualitative Studie zur Selbstwahrnehmung von Ursachen und Lösungsansätzen*. Ph.D. thesis.

[51] Simone Albrecht. 2008. Burnout–der weg danach. *Burnout im Lichte von Theorie und Praxis. VDM, Saarbrücken*.

[52] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

[53] Corinna Cortes and Vladimir Vapnik. 1995. Support- vector networks. *Machine learning*, 20(3):273–297.

[54] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. 2001. *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[55] Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.

AUTHORS

Sukanya Nath

Sukanya Nath is currently a PhD student at the Computational Linguistics group at the University of Neuchâtel. Her area of research is multiple authorship attribution, authorship verification and profiling. She is also a research assistant under Dr. Mascha Kurpicz-Briki at the Bern University of Applied Sciences for the BurnoutWords project. She has a master in Computer Science (Data Science Specialization) from the University of Bern. Contact: sukanya.nath@bfh.ch

Mascha Kurpicz-Briki

Dr. Mascha Kurpicz-Briki obtained her PhD in the area of energy-efficient cloud computing at the University of Neuchâtel. After her PhD, she worked a few years in industry, in the area of open-source engineering, cloud computing and analytics. She is now professor for data engineering at the Bern University of Applied Sciences, investigating how to apply digital methods and in particular natural language processing to social and community challenges. Contact: mascha.kurpicz@bfh.ch