IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 13, 2022, accepted March 7, 2022, date of publication March 28, 2022, date of current version April 5, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3162685

Detection of Anorexic Girls-In Blog Posts Written in Hebrew Using a Combined Heuristic AI and NLP Method

YAAKOV HACOHEN-KERNER[1], NATAN MANOR[1], MICHAEL GOLDMEIER[1], AND EYTAN BACHAR[2]
[1]Department of Computer Science, Jerusalem College of Technology, Jerusalem 9116001, Israel
[2]Department of Psychology, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 9190501, Israel
Corresponding author: Yaakov HaCohen-Kerner (kerner@jct.ac.il)

ABSTRACT In this study, we aim to detect in social media texts written in Hebrew girls who are suspected of being anorexic. We constructed a dataset containing 100 blog posts written by females who are probably anorexic, and 100 blog posts written by females who are likely to be non-anorexic. The construction of this dataset was supervised and approved by an international expert on anorexia. We tested several text classification (TC) methods, using various feature sets (content-based and style-based), five machine learning (ML) methods, three RNN models, four BERT models, three basic preprocessing methods, three feature filtering methods, and parameter tuning. Several insights were found as follows. A set of 50-word n-grams (mostly word unigrams) given by an expert was found as a good basic detector. A heuristic process based on the random forest ML method has overcome a combinatorial explosion and led to significant improvement over a baseline result at a level of $P = .01$. Application of an iterative process that tests combinations of "k out of n'" where $n' < n$ (n is the number of feature sets) lead to a result of 90.63%, using a combination of 300 features from ten feature sets.

INDEX TERMS Mental disorders, natural language processing, supervised machine learning, text analysis, text classification, text processing.

I. INTRODUCTION

A mental disorder is a behavioral or mental pattern that causes significant distress or impairment of personal functionality. Szmukler [1] and James *et al.* [2] estimated that there are approximately 971 million people worldwide who suffer from various mental disorders, e.g., ∼ 284 million suffer from anxiety, ∼ 264 million suffer from depressive disorders, and ∼ 50 million suffer from dementia. Furthermore, according to Wang *et al.* [3], between 76% and 85% of people with mental disorders receive no treatment.

A mental disorder can be diagnosed only by a mental health professional. In our opinion, plausible suspicion of many mental disorders can be generated by an intelligent supervised machine learning (ML) system based on social texts, such as Twitter messages and blog posts. Such a system will be able to detect various suspected mental disorders among people participating in social networks. The output of such a system can be presented to professional experts, to whom the individuals can be referred.

Online social networks, such as Facebook, Twitter, and blog forums, are extremely popular, with hundreds of millions of users, and many people express their mental state and feelings on social media. Therefore, social media enable the construction of datasets that can serve as excellent testbeds for the detection of various mental disorders.

In recent years, an increasing amount of research on mental health has focused on social media, e.g., Twitter [4], Facebook [5], and Reddit Reddit[1] (news and social website dedicated to thousands of communities) [6]. Coppersmith *et al.* [7] showed that several mental disorders (e.g., anxiety, eating disorders, and schizophrenia) can be detected in Twitter messages by using character n-gram language models (CLMs). Previous mental health studies have

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Martalo.

[1] https://www.redditinc.com/

largely focused on several specific mental disorders, such as depression [8], post-traumatic stress disorder (PTSD) [4], suicide [9], anxiety [10], schizophrenia [11], and bipolar disorder [12].

Anorexia nervosa (AN) is an eating disorder (ED) that is defined by the Diagnostic and Statistical Manual of Mental Disorders [Edition, 2013, pages 338-339] as follows:

"Anorexia Nervosa. Diagnostic Criteria A. Restriction of energy intake relative to requirements leading to significantly low body weight in the context of age, sex, developmental trajectory, and physical health. Significantly low weight is defined as a weight that is less than minimally normal or, for children and adolescents, less than that minimally expected. B. Intense fear of gaining weight or becoming fat, or persistent behavior that interferes with weight gain, even though at a significantly low weight. C. Disturbance in the way in which one's body weight or shape is experienced, undue influence of body weight or shape on self-evaluation, or persistent lack of recognition of the seriousness of the current low body weight."

Briefly, AN is an ED that is characterized as the maintenance of extremely low body weight, an intense fear of putting on weight despite being severely underweight and having amenorrhea and a distorted body image [13]. Adolescents and young adults are at particularly high risk. The peak onset for AN occurs during the teenage years and early twenties, with a preponderance in girls and women [6], [14], [15]. Therefore, the target population of the current study is adolescent and young adult females. According to James *et al.* [2], there are approximately 3.36 million people worldwide who suffer from anorexia.

AN is a disabling, deadly, and costly mental disorder that considerably impairs physical health and disrupts the psychosocial functioning of the individual. AN has one of the highest death rates of any psychiatric disorder [15] and is considered the most dangerous among the cluster of eating disorders [16]. The physical complications of AN, which are the results of extreme starvation and subsequent thinness, are enormous. Various possible complications of AN are amenorrhea, i.e., the cessation of menstruation in menstruation-aged women [17], constipation, severe abdominal pain, slowness, low body temperature, and a failure to thrive to the body's full potential. If normal eating restoration is not gradual, severe edema may appear. Bone density may also significantly decrease [18], [19].

Various interesting ethical issues are involved in anorexia-related studies. For example, forced feeding of people with anorexia [20], participation of people with anorexia in fitness classes Giordano [21], and the use of deep brain stimulation to treat patients with anorexia [22].

In our study as well, there are some relevant ethical issues, e.g. (1) Whether and how can we present posts authored by girls suspected as being anorexic? (2) How can we ensure that people will not take advantage of vulnerable people?; (3) If there is any intention to contact the girls identified as being anorexic, how can we help them?; and (4) Whether and how can we persuade them to be involved in the research?

We hypothesize that the application of ML method(s) using heuristic combinations of suitable content-based and style-based feature sets on supervised social datasets related to anorexia can successfully identify, from their text data, females that are likely to be anorexic.

The key contributions and innovations of this research are as follows: A labeled anorexia-related dataset of social media posts written in Hebrew was constructed, approved by an international expert in the domain of anorexia, and made available to the public for reproducibility or benchmarking. A set of 50-word n-grams (mostly word unigrams) provided by an expert was found as a good basic detector. In addition, two heuristic methods lead to significant improvements over a baseline result at a level of $P = .01$: (a) a hill-climbing process that leads to a result of 89.07% using a combination of 372 features from nine feature sets and (b) an iterative process that tests combinations of "k out of n'" where $n' < n$ (the total number of feature sets) for different values of k and n' lead to the best result of 90.63%, using a combination of 300 features from ten feature sets.

The rest of this paper is organized as follows: Section II introduces previous anorexia detection systems related to eRisk tasks. Section III introduces previous mental disorder detection systems. Section IV presents previous datasets along with the dataset constructed for this research. Section V introduces the preprocessing methods, ML methods, and the experimental setup. Section VI presents the experimental results and an analysis of the main results. In Section VII, we discuss the implications that this study has and suggest future research. Finally, Section VIII summarizes and concludes this study.

II. PREVIOUS ANOREXIA DETECTION SYSTEMS RELATED TO ERISK TASKS

Since 2017, the eRisk (Early risk prediction on the Internet) lab[2] (under the platform of CLEF[3] – Conference and Labs of the Evaluation Forum) organized tasks related to early detection of various mental disorders such as depression and self-harm or anorexia. In 2018 and 2019, eRisk organized tasks related to the early detection of anorexia. The datasets are collections of hundreds of Reddit users labeled as anorexic or non-anorexic along with hundreds of posts and comments (for each user on average) written in English that were recorded chronologically. The positive set is composed of users who explicitly mentioned that they were diagnosed with anorexia, while the negative set is mainly composed of random users from the same social media platform (including users who have close relatives suffering from anorexia).

These datasets and the evaluation methodology were constructed using the same methodology and sources as the datasets described in Losada and Crestani [23].

[2] https://early.irlab.org/

[3] http://www.clef-initiative.eu/

The evaluation of the tasks involves standard measures, such as F1, Recall, and Precision. These measures are time-unaware and do not penalize late decisions. Therefore, Losada *et al.* [23] defined ERDE (Early Risk Detection Error), an error measure for early risk detection for which the fewer posts and alerts required to raise an alert, the better. Otherwise, there is a penalty for late decisions.

An overview of the early risk detection tasks (anorexia and depression) in eRisk 2018 is given in Losada *et al.* [24]. The dataset of the anorexia task is imbalanced. The organizers received 35 submissions from 9 different teams (each team could submit up to 5 variants or runs). The highest F1 score 0.85 was achieved by the FHDO-BCSGE model [25], which consists of a simple late fusion ensemble approach that has been calculated as the unweighted mean of the outputs obtained from three bags-of-words, including metadata models and two CNN models. The highest precision, recall, and the lowest ERDE scores (0.91, 0.88, 11.4%, respectively) were achieved by various runs of the UNSL models of Funez *et al.* [26], which were based on a model that uses a semantic representation of documents and a model that carries out an incremental estimation of the association of each user to each class.

About 80%–100% of the non-anorexia users were correctly identified by most of the systems (nearly all non-anorexia users fall in the range meaning that at least 80% of the systems labeled them as non-anorexic). In contrast, the distribution of anorexia users is flatter and, in many cases, they are only identified by less than half of the systems. An interesting result is that all anorexia users were identified by at least 10% of the systems. Most of the teams ignored penalties for late decisions and mostly focused on classification accuracy.

Aragon *et al.* [27] showed that early detection of signs of depression and anorexia of social media authors could be based on the presence of the emotions expressed by the authors. Their study was based on the eRisk-2018 anorexia dataset, which contains posts and comments written by 472 users where only 12.9% of them were categorized as positive. Aragon *et al.* [27] created groups of sub-emotions using the EmoLEX lexicon [28] and some word-embedding algorithms. The highest baseline result (F1 = 0.82) was obtained by a support vector machine (SVM) using BoSE unigrams (Bag of Sub-Emotions). Using the late fusion method, they improved to F1 = 0.84. The performance of deep learning (DL) models (word2vec and Glove) was also tested and was very low.

Losada *et al.* [29], in their overview paper, provided an overview of eRisk 2019, related to early risk detection of three tasks related to health and safety: anorexia, depression, and self-harm. The anorexia dataset of eRisk 2019 is also highly imbalanced. The organizers received submissions from 13 teams. Nine teams processed the entire thread of messages (around 2,000 iterations).

The highest F1 score (0.71) in eRisk 2019 was achieved by an ensemble approach developed by the ClaC team [30]. Their method employed several attention-based neural sub-models that extracted features and predicted class probabilities. These features served as input features to an SVM model. The highest precision score (0.77) and a relatively high F1 score (0.68) were achieved by the LIRMM team [31], which applied a deep mood module that activates several attention-based DL models.

The datasets provided in 2018 and 2019 by eRisk, which are highly imbalanced are composed of posts and comments written in English by Reddit users labeled as anorexic or non-anorexic, along with hundreds of posts and comments (for each user in average) that were recorded chronologically. In 2020 and 2021, eRisk did not suggest any task or dataset relevant to Anorexia.

Uban *et al.* (2021) proposed a DL model to detect signs of people suffering from anorexia in social media. They also tried to explain the behavior of their model. They trained a hierarchical attention model and used its internal encodings to discover different clusters of anorexia symptoms. They interpreted the identified patterns from emotional expressions, personality traits, and psycho-linguistic features. They found patterns of word usage in some users with anorexia, which show that they feel less as being part of a group compared to control cases, as well as that they have abandoned explanatory activity as a result of a greater feeling of helplessness and fear.

In contrast, in our study, we worked with a balanced dataset we constructed, which is composed of posts written in Hebrew in Israeli blog forums (or sub-forums) that are located in various public-domain Israeli websites. This dataset was supervised and approved by an international expert in the domain of anorexia (more details in Section IV part B).

III. OTHER PREVIOUS MENTAL DISORDER DETECTION SYSTEMS

Tsugawa *et al.* [33] presented a model based on the results of a web-based questionnaire answered by 209 Twitter users regarding their social media activities, to measure their degree of depression. Their best result (accuracy of 0.66 and F-measure of 0.46) was obtained using an SVM based on 17 features: 10 topics extracted using the latent Dirichlet allocation (LDA) model [34], a ratio of positive-affect words contained in tweets, a ratio of negative-affect words contained in tweets, number of tweets per day, overall retweet rate, a ratio of tweets containing a URL, number of followers, and number of users followed.

Wang *et al.* [35] developed a method that automatically gathers individuals who self-identify as ED in their profile descriptions, as well as in their social network connections with others on Twitter. They also built predictive models to classify users with and without an ED. They explored three different ML methods: naive Bayes (NB), an SVM with various kernels, and k-nearest neighbors. The best accuracy result (above 97%) was obtained using a linear SVM with default settings. In their best model, each user is represented as a vector of 97 features composed of 6 social-status features, 11 behavioral features, and 80 psychometric features that match each of the 80 psychologically-relevant categories in the Linguistic Inquiry and Word Count (LIWC) lexicon [36].

Shen and Rudzicz [10] presented models that detect anxiety in posts submitted in Reddit. They collected 22,808 posts over three months, 9,971 of them were anxiety-related posts (''anxiety'') and 12,837 were general posts (''control''). They applied n-gram language modeling, vector embeddings, topic analysis, and emotional norms to generate features that accurately classify posts related to binary levels of anxiety. They obtained an accuracy of 98% in two models: word2vec embeddings combined with LIWC features, and n-gram probabilities combined with LIWC.

Birnbaum *et al.* [11] introduced models that distinguish between Twitter messages written by 146 users with self-disclosed diagnoses of schizophrenia and 146 users from a control group. The performance was evaluated using a 10-fold cross-validation method (70% training and 30% validation). Their models used the TF-IDF values of the top-500 n-grams and 50 LIWC categories. Then, feature filtering using the ANOVA F-test reduced the feature space from 550 to 350 features. They applied four supervised ML methods: Gaussian naïve Bayes, random forest (RF), logistic regression (LR), and SVM. The best results (accuracy of 0.81 and F-measure of 0.80) were obtained by RF.

Sekulić *et al.* [12] presented a study on the prediction of bipolar disorder from user comments. on Reddit posts written by 3,488 users with self-disclosed diagnoses of bipolar disorder and 3,931 users that were sampled from the general Reddit community. For each user, they extracted three types of features: (1) psycholinguistic features composed of syntactic features (e.g., pronouns and articles), topical features (e.g., work and friends), and psychological features (e.g., emotions and social context) based on LIWC categories, and words using similarities based on neural embeddings found through Empath [37]; (2) lexical features composed of TF-IDF weighted bag-of-words, stemmed using the Porter stemmer from NLTK [38]; (3) and several Reddit user features that attempt to model the user's interaction patterns. They applied three classifiers: SVM, LR, and RF. The best results (an accuracy of 0.869 and an F1-score of 0.863) were obtained by RF.

Ramírez-Cifuentes *et al.* [39] proposed several models for the early detection of anorexia on a collection composed of writings (posts or comments) from a set of Reddit users. Their model used 5,093 features composed of 64 frequencies of words belonging to the categories of the LIWC dictionary [40], 9 anorexia-related categories (anorexia, body image, food and meals, eating, caloric restriction, binging, compensatory behavior, and exercise), 4,303 word unigrams, 665 word bigrams, and 50 topics using LDA. The best results (0.85 for all three measures: F1, precision, and recall) were obtained using an SVM with 50 LDA topics, TFIDF values, 64 LIWC features, and the text length threshold (TLT) feature.

Zhou *et al.* [41] proposed a mental disorder aided diagnosis model that detects people with high probabilities of suffering from five common adult mental disorders: anxiety disorder, bipolar disorder, depressive disorder, obsessive-compulsive disorder, and panic disorder. The tested documents were tweets collected using relevant mental disorder-related hashtags and timestamp information. The supervised dataset contained 396 users with 5,323 tweets who were considered to have one of the five mental disorders and 400 users with 6,683 tweets who were considered to have no mental disorder. Using the stochastic gradient descent method they obtained precision, recall, and F1-measure scores of 0.77, 0.92, and 0.84, respectively.

Tadesse *et al.* [42] proposed several models that distinguish between depressed and non-depressed users in Reddit. The experiments were conducted on a dataset built by Pirina and Çöltekin [6]. The dataset contains 1,293 depression-indicative posts and 548 standard posts. The depression-indicative posts were collected from Reddit forums devoted to depression, in which the depressed users asked for support. Standard posts written by non-depressed users were collected from Reddit forums related to family or friends. The authors applied five supervised ML methods (SVM, LR, RF, AdaBoost, and MLP). The best results (accuracy of 0.9 and F1-score of 0.93) were obtained by MLP using word bigrams, LIWC, and LDA features.

Aragón *et al.* [43] proposed a method called a Bag of Sub-Emotions (BoSE) that represents social media documents. This set of fine-grained emotions is automatically generated using a lexical resource of emotions and subword embeddings from Fast-Text. Using this representation capture topics and emotions that are used for depression detection. The usage of their simple and interpretable method improved the results compared to proposed baselines and a representation based on the core emotions and obtained competitive results in comparison to the state of the art approaches (i.e., related eRisk task winners) that are much more complex and difficult to interpret (most of the participants used plenty of different features and a vast range of models, including deep).

Alhuzali *et al.* [44] described a method that detects a sign of depression from users' posts. Their method applied pre-trained models that extract features for all user's posts and then feed them into an RF classifier, achieving an average hit rate of 32.86% in sub-task 3 of the CLEF 2021 e-risk shared task. Their method achieved reasonable performance. The evaluation showed that different SpanEmo-encoder layers produced different results. The choice of which layer to choose depends on the metric of interest. They also reported some negative results, and hope that it will inspire the community to investigate the correlations/associations between different aspects.

IV. CONSTRUCTED SUPERVISED DATASETS

In this section, we describe, on the one hand, the construction of previous datasets based on social media and on the other hand, the construction of our supervised dataset.

A. CONSTRUCTION OF PREVIOUS SUPERVISED DATASETS BASED ON ONLINE SOCIAL MEDIA FROM STUDIES

Several studies have been conducted on the detection of users considered to have a mental illness based on online forums, without being identified as such through a clinical diagnosis. Guntuku *et al.* [45] introduced 12 studies on automatically detecting mental disorders without relying on diagnoses made by clinicians. The datasets from five studies were based on posts from Twitter, Facebook, and other web forums, written by users who have been self-declared as having a certain mental illness, as well as posts, are written by control users. Most of the described datasets are balanced or relatively balanced.

Wongkoblap *et al.* [46] presented a review of 48 studies dealing with, among other things, the prediction of various mental health disorders based on data from social media. The datasets in these studies were obtained using two main approaches: (1) directly collecting data from the participants with their consent, using surveys and electronic data collection instruments, and (2) indirectly collecting public posts from social network platforms, based on regular expressions used to search for relevant posts, e.g., "I was diagnosed with [condition]." The authors did not provide details about the balance degree of the studies' datasets they reviewed.

The construction of the eRisk datasets was described in Section II. As mentioned at the end of Section II, in our study, we constructed a dataset that is composed of posts written in Hebrew by Israeli web-users in blog forums (or subforums) that are located in various public-domain Israeli websites. Our dataset is composed of balanced positive and negative sets and was supervised and approved by an international expert on anorexia (see next sub-section). This is in contrast to the previous usually imbalanced datasets, whose positive cases are of users who explicitly mentioned that they were diagnosed with anorexia.

B. CONSTRUCTION OF OUR SUPERVISED DATASET

In this study, we constructed a dataset containing 100 blog posts written in Hebrew that are likely to have been written by girls with anorexia, and 100 blog posts that are likely to have been written by girls without anorexia. The blog posts written by girls probably with anorexia were collected from blog forums (or sub-forums) dedicated to anorexic girls that are located in the following public-domain Israeli websites: http://israblog.org/, https://www.tapuz.co.il/, https://saloona.co.il/, and https://www.fxp.co.il/. In these forums, only posts that were most likely written by girls with anorexia were labeled as "positive posts." No other posts (e.g., posts written by family members or medical doctors) were selected. "Negative posts" were collected from forums, which are not connected to mental disorders. The construction of this dataset was supervised and approved by Professor Eytan Bachar, an international expert in the field of anorexia [16], [19], [47]–[49]. Professor Bachar approved every post that was labeled as "positive." We did not approve even posts that are likely to have been written by girls with bulimia, which is a related ED, but less dangerous than anorexia in respect to the chances of dying. The constructed dataset will be made available to the public for reproducibility or benchmarking.

TABLE 1. General details about the dataset.

Detail	SUB-DATASET OF ANOREXIC GIRLS	SUB-DATASET OF NON-ANOREXIC GIRLS
Number of posts	100	100
Number of words	31,376	36,887
Number of characters	165,790	198,921
Average words per post	313.76	368.87
Minimum words per post	29	40
Maximum words per post	1,516	2,136
Average characters per post	1,657.9	1,989.21
Minimum characters per post	142	220
Maximum characters per post	8,084	12,094
Average author age	20	21
Minimum author age	13	13
Maximum author age	25	25

Table 1 provides general details about the dataset. Most of the posts in the dataset (both positive and negative) were written by teenage girls or young women in their twenties. This is because almost all anorexic females are within these age groups.

In addition, 25% of the posts that were labeled as "negative" are likely to have been authored by athletic girls or girls who want to diet. These posts were chosen according to the following simple heuristic rule: posts containing at least one of 50 keywords (e.g., body, sports, athlete, calories, breast, binge, weight loss, starvation, menu, dietitian, food, diet, lean, slim, flat, bones, excess, and weight) that are used by anorexic girls according to an international expert. This was applied to achieve a more challenging dataset in terms of classification because many anorexic girls write about sports and dieting.

V. PREPROCESSING METHODS, ML METHODS, MODEL, AND EXPERIMENTAL SETUP

In this section, we introduce the preprocessing methods, ML methods, and the experimental setup of our study.

A. PREPROCESSING METHODS

In many cases, preprocessing the datasets can "clean" the datasets and improve their quality. There are basic types of preprocessing methods e.g., conversion of uppercase letters into lowercase letters, HTML tag removal, punctuation mark removal, stop-word removal, and word stemming, as well as advanced preprocessing methods such as correction of misspelled words, expansion of abbreviations, and word lemmatization.

Jianqiang and Xiaolin [50] tested six types of preprocessing methods (expanding acronyms, removing numbers, removing stop words, removing URL links, replacing negative mentions, and reverting words that contain repeated letters into their original English form) on five sentiment datasets. The best preprocessing method in their experiments was the replacement of negative mentions in the n-grams model. This method leads to a significant improvement in almost all classifiers on all datasets.

HaCohen-Kerner *et al.* [51] investigated the impact of all possible combinations of six preprocessing methods (spelling correction, HTML tag removal, converting uppercase letters into lowercase letters, punctuation mark removal, reduction of repeated characters, and stopword removal) on TC in three benchmark mental disorder datasets. In one dataset, the best result showed a significant improvement of approximately 28% over the baseline result using all six preprocessing methods. In the other two datasets, several combinations of preprocessing methods showed minimal improvements over the baseline results.

In another study, HaCohen-Kerner *et al.* [52] explored the influence of various combinations of the same six basic preprocessing methods mentioned in the previous paragraph on TC in four benchmark text corpora using a bag-of-words representation. The general conclusion was that it is always advisable to perform an extensive and systematic variety of preprocessing methods, combined with TC experiments because this contributes to improving TC accuracy.

B. ML METHODS

A wide variety of supervised ML methods are applied in TC tasks. Various classical supervised ML methods were implemented, such as support vector classifier (SVC), RF, and LR. During the last decade, DL methods (e.g., RNN and CNN) and then word embeddings (e.g., Word2vec, GloVe, ELMo, and BERT) become popular in TC.

In this research, at the first stage, we applied five classical supervised ML methods: SVC, RF, MLP, LR, and multinomial naïve Bayes (MNB).

An SVC is a variant of an SVM [53] implemented in Scikit-Learn. An SVC uses LibSVM [54], which is a rapid implementation of the SVM method. An SVM is a supervised ML method that classifies vectors in a feature space into one of two sets, given the training data. It operates by constructing an optimal hyperplane that divides the two sets, either in the original feature space or in higher dimensional kernel space.

An RF is an ensemble learning method for classification and regression [55], which constructs a multitude of decision trees. Each tree in the ensemble is generated by randomly selecting the attributes to split at each node, and these features on the training set are used to estimate the best split.

An MLP is an artificial neural network [56] based on a network of computational units (perceptrons) interconnected in a feed-forward manner. Typically, perceptrons apply a sigmoid function to the input they obtain and feed the next layer with the output of the function. This model is useful, particularly when the data are not linearly separable.

An LR [57], [58] is a linear classification model in which the output value is represented as a linear combination of the input values. A sigmoid function is used to model the probability of "success."

An MNB [59], a version of naive Bayes, is a probabilistic generative ML method. MNB is based on Bayes' theorem with the "naive" assumption of conditional independence between every pair of features, given the value of the class variable. In MNB, each document is viewed as a collection of words whose order is considered irrelevant.

C. EXPERIMENTAL SETUP

We used the accuracy measure to assess the usefulness of the various models. Accuracy is a suitable measure because our dataset is balanced (100 posts of anorexic girls and 100 posts of non-anorexic girls). To indicate which results are statistically significant compared to the baseline results, we ran 20 times 5-fold cross-validation experiments on the dataset to generate 100 performance estimates for Scheme A (any baseline experiment) and 100 estimates for Scheme B (any other experiment). These estimates can be paired because they are generated on the same splits of the dataset. Because the 100 estimates for each of the schemes are not statistically independent, having been generated from different subsets of the same dataset, many researchers (e.g., [60]–[62]) have applied the corrected resampled paired t-test developed by Nadeau and Bengio [63], [64], which has been found to be reliable (providing a false positive rate at the significance level when evaluated on synthetic data).

VI. EXPERIMENTAL RESULTS

In this section, we present the experimental results and an analysis of the main results

A. BASELINE WORD N-GRAM RESULTS

To select the word unigrams for use by the baseline models, we decided to work only with words that appear in at least three blog posts in the training set. An examination of the five pairings of the training and test sets showed that 2,245 is the minimal number of different words that appear. To achieve a reasonable accuracy baseline, based on the number of different words mentioned above, we decided to perform classification experiments on 100, 500, 1,000, 1,500, and 2,000 word unigrams according to both their TF and TF-IDF values, using five different common ML methods. Additionally, we examined the classification according to a list of 50 key expressions (46 word unigrams and 4 word bigrams), provided by an expert on anorexia, that characterize anorexic girls. The rationale behind the experiment was to define reasonable baselines for the discussed task. Table 2 lists the baseline accuracy results when using the above-mentioned features and ML methods.

Analysis of the results in Table 2 shows the following: The best baseline result (79.75%) was achieved by RF when

TABLE 2. Baseline accuracy results for the detection of anorexic girls.

#	# of word unigrams	TF/ TF-IDF	SVC	RF	MLP	LR	MNB
1	100	TF	64.85	74.85	63.70	65.88	69.93
2	100	TF-IDF	66.10	74.25	65.05	68.88	70.10
3	500	TF	71.57	79.42	69.15	69.80	72.97
4	500	TF-IDF	76.38	79.75	70.90	76.00	73.95
5	1000	TF	70.90	79.00	70.28	69.68	70.95
6	1000	TF-IDF	76.05	79.18	74.22	75.32	74.67
7	1500	TF	70.53	78.90	70.50	69.22	69.17
8	1500	TF-IDF	75.25	78.95	72.05	75.10	72.85
9	2000	TF	70.62	78.80	70.58	69.30	68.37
10	2000	TF-IDF	75.53	79.02	72.40	75.62	72.18
11	Expert's 50 terms	-	75.12	74.03	75.70	75.53	62.50

using only the top-500 word unigrams (according to their TF-IDF values). In addition, the TF-IDF results are higher than the TF results for all tested ML methods for almost all tested numbers of word unigrams (except for one out of the 20 cases). Therefore, from now on, in principle, the following experiments will use only the TF-IDF values instead of the TF values. Another noteworthy finding was the relatively high result of 75.70% obtained by MLP using 50 keywords of the expert, which was better than all results achieved using 100 word unigrams by all tested ML methods. That is, the 50 keywords of the expert are better for a basic classification than 100 words with the highest TF-IDF values. A plausible explanation is that expertise is an important asset when we want to apply classification using a relatively low number of word unigrams; it can reduce the number of word unigrams and yet improve the results.

In comparison with English, in Hebrew, there are fewer available NLP tools in general, and fewer available preprocessing methods in particular. In this research, we applied only three preprocessing methods: L - conversion of uppercase letters into Lowercase letters only for words in English; A - removal of '.' from Acronyms, e.g., I.B.M. into IBM; and H - removal of stop words using a basic list of 47 stop words in Hebrew [65]–[67].

The tools, libraries, and lists that we used in this study include Python (https://www.python.org/); Scikit-learn (https://scikit-learn.org/stable), a library for ML methods in Python; and NLTK (https://www.nltk.org/), a library that produces various n-gram features and a corpus of synonyms.

We conducted classification experiments with all possible combinations of preprocessing methods using the TF-IDF values of the top-100, -500, -1000, -1500, and -2000 frequent words. The best result (80.22%) was obtained by RF with 500 words using the H, A, and L preprocessing methods. This result shows a small and insignificant improvement of 0.47% in comparison with the best baseline result. Both the A and H preprocessing methods substantially change some of the texts in the dataset and prevent the ability to extract various features, e.g., spelling and function words. In contrast, the L preprocessing method, which converts uppercase letters in English into lowercase letters (for all the English letters in our dataset, which is written mainly in Hebrew), is a relatively small change in the texts in which there are no deletions or insertions of any letters. The application of L leads to a result of 80.0% (a small and insignificant improvement of 0.25% relative to the baseline). Although this improvement is insignificant, we implemented the L preprocessing method in the following experiments because this method is simple to implement easily and quickly and it leads to an improvement.

B. FEATURE SETS AND A HILL-CLIMBING MODEL

28 feature sets were defined semi-automatically. After reading many of the post blogs, we manually defined 28 feature sets where each set contains a basic list of features. Most of the feature sets are content-based, e.g., food and drink, hunger, vomiting and fasting, ana, calories and weight, anorexia, fat, sickness/illness, weakness and pain, sleep, and sports. Some of the feature sets are style-based, e.g., quantitative and average values, orthographic, limiters, intensifiers, repetition of words and letters, and language richness. Some of the feature sets are sentiment-based, e.g., positive and negative words. Typically, a set contains 5–112 word unigrams and their declensions that are relevant to the set. For instance, the "ana" set contains a few symbolic words describing anorexic girls, e.g., אנה ("ana"), לאנה ("to ana"), האנה ("the ana"), and פרו-אנה ("pro-ana"). Table 3 presents the general details of these feature sets. The Hebrew declensions were generated using regular expressions. The resulting words were checked and the illegal ones were filtered out.

To determine the best combination of feature sets for a TC task, all possible combinations of the feature sets should be attempted. However, for $n = 28$ (the number of feature sets), there are 2^{28} (134,217,728) possibilities. To overcome this combinatorial explosion for non-small values of n, several variants of hill-climbing have been proposed, (e.g., [68]). An application of TC to hill-climbing using feature sets was successfully demonstrated in HaCohen-Kerner *et al.* [69].

In this research, we apply the following hill-climbing process. In the first step, TC is applied to each feature set alone. The best feature set is selected from among n feature sets. In the second step, all possible combinations of two feature sets (where one of them is the set chosen in the first stage) are tested, that is, $(n-1)$ possible pairs of feature sets are verified in the second step. If the best combination of two sets achieves

TABLE 3. Single feature sets.

#	Set Code	Feature set (in English)	Feature set (in Hebrew)	# of word n-grams
1	ACF	quantitative and average values	מאפיינים כמותיים וממוצעים	7
2	AGF	anger, fear, despair, laziness, sadness, and depression	כעס, פחד, ייאוש, עצלות, עצב ודיכאון	101
3	ANF	ana	אנה	28
4	AOF	anorexia	אנורקסיה	26
5	CAF	calories and weight	קלוריות ומשקל	51
6	DEF	limiters	מצמצמים	41
7	E50TH	50 original keywords of an expert on anorexia e.g., vomiting, body image, pro-ana, thinness, diet, clavicles, food, underweight, starvation	50 ביטויים מקוריים של פרופ' לפסיכולוגיה	50
8	E50TTH	50 keywords of the expert translated from English e.g., hip bones, pro-ana, target weight, meal plan, dietitian, hunger, fasting, disgusting, pain, jealous	50 ביטויים של פרופ' לפסיכולוגיה המתורגמים מאנגלית	50
9	FDF	food and drink	אוכל ושתיה	112
10	FRC	orthographic	אורתוגרפית	15
11	FTF	fat	שמנה	37
12	HUF	hunger	רעב	31
13	INF	intensifiers	מגבירים	50
14	LOF	love, warmth, affection	אהבה, חיבה וחום	21
15	MEF	first person pronouns (I)	אני	32
16	NW	negative words	מילים שליליות	82
17	PNF	weakness and pain	חולשה וכאב	101
18	PW	positive words	מילים חיוביות	74
19	REF	repetition of words and letters	חזרות על מילים וחזרות על תווים	5
20	SIF	sick/illness	חולה/מחלה	41
21	SLF	sleep	שינה	25
22	SPF	sports	ספורט	35
23	TE	time	זמן	17
24	THF	thinness	רזון	73
25	VOF	vomiting and fasting	הקאה וצום	48
26	VUF	sex, cursing, insulting phrases, negative phrases, alcohol	מין, גסויות, ביטויים מזלזלים, ביטויים שליליים, ואלכוהול	83
27	WEF	language richness	עושר שפה	61
28	XTE	extended time	זמן מורחב	321

a better accuracy result than the best single feature set, then the process continues. This process proceeds step by step until no further improvement occurs. Such a hill-climbing model tests a maximal number of $n+(n-1)+\ldots+1$ combinations of feature sets. That is, the complexity of this heuristic model is $O(n^2)$ instead of $O(2^n)$. The rationale behind this experiment was to heuristically find an ML method and a combination of feature sets that achieves an improved classification result using a polynomial run time algorithm.

It should be noted that although the method should be discontinued when there is no improvement from one stage to the next, we ran the process to the end even though the result was not always improved because the order of magnitude remained the same and in some cases, the application of the ''extended'' version improved the result.

Table 4 presents the accuracy results for the hill-climbing process for feature sets using a 5-fold cross-validation process 20 times. Some of the results are marked with a ''V'' or ''$*$,'' which indicate that a specific result is statistically better or worse than the best baseline result, respectively. To compare the different results, we conducted statistical tests using a corrected paired two-sided t-test with a confidence level of 95%. In cases where the result is statistically better than the best baseline result with a confidence level of 99%, the result

TABLE 4. Accuracy results of the hill-climbing process for feature sets.

# of feature sets	Best ML method	Best set of feature sets	Accuracy result
1	LR	{9}	76.47
2	RF	{9, 15}	81.4
3	RF	{9, 15, 7}	83.83
4	RF	{9, 15, 7, 10}	86.17V
5	RF	{9, 15, 7, 10, 25}	86.7V
6	RF	{9, 15, 7, 10, 25, 1}	87.5W
7	RF	{9, 15, 7, 10, 25, 1, 4}	87.78W
8	RF	{9, 15, 7, 10, 25, 1, 4, 12}	88.08W
9	RF	{9, 15, 7, 10, 25, 1, 4, 12, 5}	89.07W

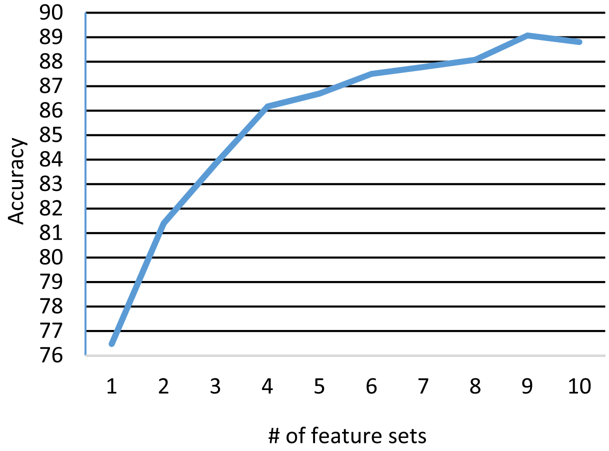


FIGURE 1. Best results during the hill-climbing process.

will be marked with "W". The highest accuracy for a single set is 76.47%, which was obtained using LR applied to the FDF (food and drink) set that contains 112 words. This result is lower (but not significantly lower, at a significance level of P = .05) than the baseline result of 79.75%, which was obtained using RF applied to the top-500 words according to their TF-IDF values.

The highest accuracy that was obtained during the hill-climbing process was 89.07%W (an improvement of 9.32% over the baseline result, which is statistically significant at a level of P = .01) with the RF using the following nine sets: ACF, AOF, CAF, E50TH, FDF, FRC, HUF, MEF, and VOF, containing 372 features. These sets are composed of seven content-based sets (four of them from the five sets that enabled the best results as single sets) and two style-based sets (ACF (quantitative and average values) and FRC (orthographic features)). The addition of a 10th set reduced the result to 88.8%.

Figure I shows the best results for each of the 10 steps of the hill-climbing process described in Table 4. We see a high growth rate until the end of the fourth step. After the fourth step, we see the first result (86.17%V), which was significantly better than the best baseline result a level of P = .05. In the next five steps (5–9), we see a slight growth rate. After the sixth step, we see the first result (87.5%W), which was significantly better than the best baseline result at a level of P = .01. The best result (89.07%W) was obtained after the ninth step. In the tenth step, the accuracy decreased and the hill-climbing process stopped.

1) CLASSIFICATION USING RNN MODELS

In this stage, we focused on the application of recurrent neural networks (RNNs), which are a class of neural networks that is suitable for modeling sequence data e.g., natural language processing or time series. We applied three common types of RNN models that have tools for sequence analysis, which are implemented by Keras application programming interface (API): SimpleRNN,[4] Long Short Term Memory (LSTM) [70], and Gated Recurrent Unit (GRU) [71].

TABLE 5. Results of the applied RNN models on the dataset.

# of Model	Preprocessing method	RNN layer type	Accuracy Result
1	Doc2vec	LSTM	50.0*
2	Text vectorization	LSTM	67.48*
3	Text vectorization	GRU	55.83*
4	Text vectorization	SimpleRNN	57.03*

Each RNN model contained five hidden layers (with 64 nodes in each such layer): encoder layer, embedding layer, a bidirectional layer that contains the RNN layer type (SimpleRNN/LSTM/GRU), and two dense layers. Default parameter values have been used. We ran 20 times 5-fold cross-validation experiments on the dataset and their results are presented in Table 5. The run time on our server of each model was around 2 days. The rationale behind this experiment was to determine whether the use of an RNN method can improve the best classification result.

All the results were significantly lower than the baseline result. The best RNN result (67.48%) was obtained using the LSTM model with the text vectorization embedding mode.

2) CLASSIFICATION USING BERT MODELS

We also applied classification using different BERT models based on the Hugging Face's BERT mode[5] (BertTokenizer for tokenizing and BertForSequenceClassification for classification) with several pre-trained models. The BertTokenize[6] is a class that builds an instance based on some pre-trained model. This class has a function named encode_plus that receives a sequence of words (a string) and returns the corresponding tokens of the sequence and the attention mask. The rationale behind this experiment was to determine whether the use of various models of BERT (the state-of-the-art DL method) can improve the best classification result.

3) THE APPLIED BERT MODELS ON THE DATASET

For our dataset, we used 4 different pre-trained models. Two of them were trained especially for Hebrew (AlephBERT and HeBERT) and two of them were trained for several languages, including Hebrew (WikiBERT and BERT multilingual base). Details about these BERT models are presented below.

1. **AlephBERT:** A large pre-trained model for Modern Hebrew introduced by Seker *et al.* [72] at Bar-Ilan University. This model was trained on 98.7M sentences from 3 different Hebrew text sources: the Hebrew portion of the OSCAR database, tweets in Hebrew collected from Twitter between 2014 - 2018, and the texts of the Hebrew Wikipedia.

[4] https://www.tensorflow.org/api_docs/python/tf/keras/layers/SimpleRNN
[5] https://huggingface.co/
[6] https://huggingface.co/transformers/model_doc/bert.html#berttokenizer

TABLE 6. Results of the applied BERT models on the dataset.

Pre-trained model	Vocab size	Accuracy result
AlephBERT	52,000	74.5
HeBERT	30,522	77.5
WikiBERT	20,101	61.0
BERT multilingual base (cased)	119,547	56.5

2. **HeBERT:** A Hebrew pre-trained language model introduced by Chriqui and Yahav [73]. This model was trained on over 24.6M sentences from three different Hebrew text sources: the Hebrew portion of the OSCAR, Hebrew dump of Wikipedia, and comments collected between January 2020 to August 2020 from Israeli news websites (Ynet, Israel Hayom, and Be-Hadre Haredim).

3. **WikiBERT:** A collection of BERT models for several languages built from Wikipedia texts that was introduced by Pyysalo *et al.* [74]. We applied the Hebrew version, which was trained on 166M tokens from Hebrew Wikipedia.

4. **BERT multilingual base (cased):** a pre-trained model on the top 104 languages with the most extensive Wikipedia that was introduced by Devlin *et al.* [75]. During the training, the entire Wikipedia was dumped into the model for each one of those 104 languages.

The results of these BERT models (all of them with 12 hidden layers and a hidden size of 768) on the dataset are presented in Table 6. It is important to note that the run time of each model was around two hours and that the results are relatively low. Therefore, we decided to run these models only one time of 5-fold cross-validation, instead of 20 times 5-fold cross-validation.

As expected, the two pre-trained models for Hebrew (HeBERT and AlephBERT) obtained significantly better results than the results of the two multilingual BERT models. Even though the "Vocab size" of the AlephBERT model is higher than the "Vocab size" of the HeBERT model, the result of the HeBERT model was better. A similar phenomenon was found for the two multilingual BERT models. Even though the "Vocab size" of the BERT multilingual base (cased) model is much higher than the "Vocab size" of the WikiBERT model, the result of the WikiBERT model was better.

C. THE HEURISTIC METHOD

Before presenting the heuristic experiments, we will mention that the best result that has been achieved so far (89.07%W), was obtained using the hill-climbing method. This result using the hill-climbing method was by applying RF using a combination of 9 sets. One of the disadvantages of the hill-climbing method is the risk of falling into the local optimum and not finding the global optimum. On the other hand, as mentioned above, the brute force method that tests all possible combinations of feature sets; the number of such combinations is 2^n where n is the number of features sets (28 in our case), and this is unpractical.

Therefore, we decided to apply a heuristic algorithm that will test only combinations of "k out of n'" items, where $n' < n$ (n is the number of feature sets) and $k <= n'$. In addition, we must remember that there is a non-negligible run time for each combination (depending on the number of feature sets in the combination, the number of features in each feature set, the applied ML method(s), the time needed to generate the model(s) in the training subset, and the time to activate the constructed model(s) in the test subset). A set composed of hundreds or thousands of combinations might take from a few hours to a few days on our available server (a virtual machine with the following specifications: Intel Xeon Platinum 8168 processor, 8 virtual cores (and later 16 cores), RAM of 32GB, and SSD of 127GB). For instance, the run time of a set of 8,008 combinations while applying only one ML method (RF) was 3 full days, one hour, and one minute.

The rationale behind the various experiment of the heuristic method was to apply an iterative heuristic process that tests much more combinations than the $O(n^2)$ combinations that were tested by the hill-climbing process to achieve a better classification result. Table 7 presents details about various combinations of "k out of n'" ($n' < n$; n is the number of feature sets) best feature sets and their results. After the application of various experiments, we analyzed all combinations that obtained an accuracy result of at least 88% using the RF ML method (including combinations from the hill-climbing process). We saw that the best combinations contained sets that were not always the best feature sets. We concluded that we should perform experiments of heuristic sets composed of 15 sets, but not the top 15 sets that achieved the best results on their own, but rather the 15 sets with the highest number of occurrences in combinations that achieved 88% and above. These are the new 15 selected feature sets: acf, fdf, frc, vof, aof, e50th, mef, caf, huf, anf, pw, nw, pnf, agf, and wef. It is important to point out that, as expected, many of the selected feature sets, are anorexia (directly- or indirectly-) related sets, e.g., FDF (food and drinks), AOF (anorexia phrases), CAF (calories and weight), ANF (phrases with inflections of "Anna"), HUF (hunger), AGF (anger), and E50TH (expert's 50 terms in Hebrew). Another important finding is that among these 15 selected sets, three sets do not appear in the top 15 feature sets that achieved the best results on their own as follows: (1) PNF obtained 53.98* alone using LR, 21st place; (2) AGF obtained 53.08* alone using RF, 22nd place; and (3) WEF obtained 52.23* alone using MNB, 25th place.

We run RF on all possible combinations of 9 and 10 sets out of these 15 sets (5,005 and 3,003 combinations, respectively). Table 7 presents the accuracy results ($\geq 89.35\%$) of set combinations in descending order using the RF ML method.

The first combination {vof, huf, aof, pnf, anf, agf, frc, mef, acf, fdf} obtained the best accuracy result (89.62W).

At this point, we decided to try two additional well-known directions for further improvements: feature filtering and parameter tuning. The rationale behind these experiments

TABLE 7. Best accuracy results (≥ 89.3%) of set combinations in descending order.

#	# of sets	# of features	Combination	Best Accuracy
1	10	501	vof, huf, aof, pnf, anf, agf, frc, mef, acf, fdf	89.62W
2	10	451	vof, huf, aof, anf, caf, agf, frc, mef, acf, fdf	89.43W
3	9	400	acf, fdf, frc, vof, aof, mef, huf, anf, pnf	89.35W

was to test common improvement methods to improve the best classification result.

1) FEATURE FILTERING

In this stage, using three types of feature filtering methods (Chi^2, ANOVA, and Mutual Information), we performed various experiments to improve the accuracy results achieved by the four combinations that obtained results ≥ 89.3%. The total run time of these experiments was 47 minutes.

The use of the Mutual Information feature filtering method on the 1st, 3rd, and 4th combinations and the application of RF on the resulting features led to higher results compared to the results without the filtering. The highest accuracy result (89.8%) was obtained using RF and 300 features after applying the "Mutual Information" feature filtering method on the 1st combination {vof, huf, aof, pnf, anf, agf, frc, mef, acf, fdf}. That is, a tiny improvement of 0.18% was obtained compared to the accuracy result (89.62%) achieved by the 1st combination, which consists of 10 feature sets containing 501 features without any feature filtering and/or parameter tuning.

2) PARAMETER TUNING

We applied parameter tuning on the combination {vof, huf, aof, pnf, anf, agf, frc, mef, acf, fdf} that achieved the highest result without any feature filtering as follows. Using the RandomizedSearchCV class of Sikict-Leran, we tried 150 random combinations of parameters. The best result (90.38W) was obtained by the following combination of parameters: n_estimators = 1900, max_features = sqrt, max_depth = 105, min_samples_split = 5, min_samples_leaf = 3, and bootstrap = True. Using the same feature set combination, we performed an extended experiment for various parameter combinations (a slightly larger range than the previous range). Although in this stage we tried 625 combinations of various parameters, we did not obtain any improvement.

3) FEATURE FILTERING AND PARAMETER TUNING

In this stage, first we applied feature filtering using Mutual Information on the discussed set combination {vof, huf, aof, pnf, anf, agf, frc, mef, acf, fdf} resulting in 300 features. Then we applied parameter tuning on these 300 features using 150 random combinations of parameters. The total run time for these 150 parameter combinations was 12 minutes. The best result 90.63W was obtained by the following combination of parameters: n_estimators =1200, max_features = sqrt, max_depth = 71, min_samples_split = 3, min_samples_leaf = 3, and bootstrap = True.

TABLE 8. Best accuracy results using feature filtering and/or parameter tuning.

	Without parameter tuning	With parameter tuning
Without feature filtering	89.62W	90.38W
With feature filtering	89.8W	90.63W

4) CONCLUSION OF FEATURE FILTERING AND/OR PARAMETER TUNING EXPERIMENTS

Using the best combination {vof, huf, aof, pnf, anf, agf, frc, mef, acf, fdf} that obtained an accuracy result of 89.62W, we applied thousands of feature filtering and/or parameter tuning experiments. The best results are presented in Table 8.

The contribution of parameter tuning was higher than the contribution of feature filtering. The application of only feature filtering led to 89.8W (a slight improvement of 0.18) while the application of only parameter tuning led to 90.38W (an improvement of 0.76). The best result 90.63W was achieved using both parameter tuning and feature filtering.

Our best result (an accuracy of 0.9063) is statistically better than the best baseline result with a confidence level of 99%. This result is competitive in comparison to the state-of –the-art results achieved in previous early detection of anorexia tasks (eRisk 2018 and eRisk 2019). The highest F1 score (0.85) in eRisk 2018 was achieved by the FHDO-BCSGE model [25], which consists of a simple late fusion ensemble approach and CNN models. The highest F1 score (0.71) in eRisk 2019 was achieved by an ensemble approach developed by the ClaC team [30].

The eRisk datasets and our dataset are composed of blog posts. However, there are many differences, e.g., (1) Our posts are written in Hebrew while the posts in the eRisk are written in English; (2) The eRisk datasets are also composed of comments to the posts while our dataset contains only posts; (3) Each eRisk dataset contains hundreds of Reddit users with hundreds of posts and comments (for each user on average), while our dataset contains only 200 posts; (4) Our dataset is balanced (therefore, the selected measure is accuracy) while the eRisk datasets are imbalanced (therefore, the selected measure is F1); and (5) In the eRisk datasets, the positive posts are of users who explicitly mentioned that they were diagnosed with anorexia, while in our dataset, the positive

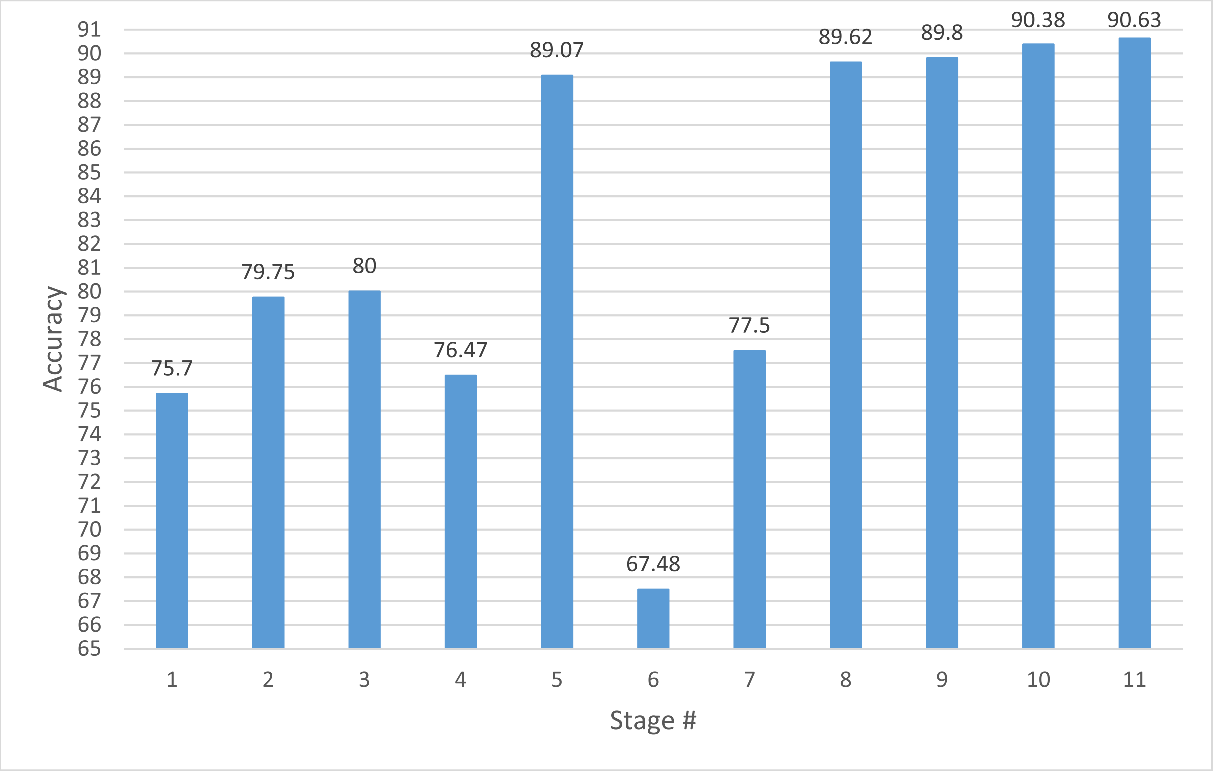


FIGURE 2. Accuracy results that were obtained along the main stages of this study.

TABLE 9. The best accuracy results obtained along the main stages of this study.

Stage #	TC model	ML	# of feature sets	# of features	Parameter tuning	Feature filtering	Accuracy result
1	Baseline using expert's 50 keywords	MLP	-	50	-	-	75.70
2	Best baseline using a bag of words	RF	-	500	-	-	79.75
3	The best baseline & conversion of uppercase letters into lowercase letters	RF	-	500	-	-	80.00
4	Best single feature set	LR	1	112	-	-	76.47
5	Best result in the hill-climbing process	RF	9	372	-	-	89.07W
6	Best RNN model	LSTM with text vectorization embedding mode					67.48*
7	Best BERT model	HeBERT 12 Hidden layers Vocab size: 30,522					77.5
8	Best heuristic combination of feature sets	RF	10	501	-	-	89.62W
9				300	-	∨	89.8W
10				501	∨	-	90.38W
11				300	∨	∨	90.63W

posts were probably written by anorexic girls and approved by our international expert.

VII. DISCUSSION AND FUTURE RESEARCH

There have been a few systems that dealt with the detection of anorexic girls. Current studies have various limitations (some of them are detailed below).

Lack of datasets: There is a great lack of datasets on anorexics in general and in the Hebrew language in particular. Current studies mainly apply supervised ML methods that require manual annotation. However, there are not enough (both in number and size) annotated datasets, especially in social datasets. There is also a lack of standards for dataset construction.

Annotated constructed datasets are not clinical ground truth: The annotated constructed datasets are not clinical ground truth. That is to say, the blog posts that were labeled as "positive" were probably (and not certainly) written by anorexic girls. These posts were collected from blog forums (or sub forums) dedicated to anorexic girls. Professor Eytan Bachar guided us to collect them as positive posts. He said that the cases of cheaters are negligible and we do not need to worry about them.

Balanced data vs. imbalanced data: Posts written by anorexic girls are a tiny proportion of social posts (even relative to other mental disorders such as depression and anxiety). However, many datasets related to mental disorders are either balanced or relatively balanced datasets rather than ill-balanced datasets as it is in reality.

There is no complete successful detection of anorexic girls: The current ML methods failed to completely detect posts written by anorexic girls. The success of these methods is partial. Nonetheless, there are several learned various statistical clues.

There are various challenges for future work. Some of them are presented below.

Application of non-classical ML methods: DL methods in general and word embedding vectors such as BERT and other models have boosted research in many domains. Extensive research is expected to also use these methods to detect various mental disorders including anorexia.

Extend the number and the size of relevant datasets using automatic generation of labeled data: New labeled data can be automatically generated without manual annotation in various methods, e.g., by collecting ground reference data, by using unsupervised or semi-supervised learning, and by using advanced simulators or generative models.

For example, Ko and Seo [76] suggest a TC method based on unsupervised or semi-supervised learning. Their method launches TC tasks with unlabeled documents and the title word of each category for learning, and then it automatically learns text classifier by using bootstrapping and feature projection techniques. Labeled data can be generated by generative models from a small amount of labeled data, which can be used for training the classifiers. Examples of such generative models are latent Dirichlet distribution (Blei *et al.* [77]), restricted Boltzmann Hinton [78], generative adversarial networks (Goodfellow *et al.* [79]), and a combination of reinforcement learning, generative adversarial networks, and recurrent neural networks (Li *et al.* [80]).

Extend the social text dataset(s) with clinical notes: The addition of clinical notes about the discussed users (in cases where it is possible) can strengthen the predictive ability and reliability of the classification models.

Another interesting future direction is to identify and analyze changes over the temporal information of users participating in social networks. Modeling the temporal track of users' posts can effectively monitor the change of mental status and it is essential for predicting early signs of anorexia that can be presented to professional experts, to whom the individuals can be referred.

A better understanding of anorexia: Many factors are relevant to anorexia. A better understanding of this mental disorder can lead to guidelines for better detection ability such as definition and use of more suitable feature sets.

VIII. SUMMARY AND CONCLUSION

In this study, we conducted an extensive and systematic set of experiments for several TC models (e.g., the half-interval search method and the hill-climbing method) on a dataset of post blogs written in Hebrew that we constructed for this study. We defined 28 feature sets and used them with five ML methods, preprocessing methods, three feature filtering methods (Chi^2, ANOVA, and Mutual Information), and parameter tuning.

Table 9 presents the main accuracy results that were obtained along the stages of this study, as explained above. Figure II presents these results graphically.

An interesting finding that repeats itself in most systems that are described in Sections II and III and also in our system (Table 9) is that classical ML methods (e.g.; SVM and RF) are currently better than DL methods.

Possible explanations for this finding are: (1) the classical ML methods have been explored and tried much more than the DL methods over the years and on the various datasets; (2) DL methods require large amounts of data to train (Gupta and Gupta [81]; Brauwers and Frasincar [82]), which is not the case in most of the above-mentioned datasets; (3) DL methods are computationally intensive and therefore need advanced computational resources (Montesinos-López *et al.* [83]; Ayoub *et al.* [84]); and (4) Optimization of DL methods requires extensive experiments using different combinations of number of layers, number of units, and number of epochs as well as other parameters [83].

The main contributions of this study are a labeled anorexia-related dataset of social media posts written in Hebrew, which was constructed and approved by an international expert in the domain of anorexia. This dataset was made available to the public for reproducibility or benchmarking. In addition, three insights and novelties were found: A set of 50-word n-grams supplied by an expert was found as a good basic detector. A heuristic process based on the RF ML method, feature filtering, and parameter tuning overcame a combinatorial explosion and lead to significant improvements over a baseline result at a level of P = .01. This process iteratively tested various combinations of "k out of n′" where $n' < n$ (the total number of feature sets) for different values of k and n' lead to the best result of 90.63% using a combination of 300 features from ten feature sets.

REFERENCES

[1] G. Szmukler and D. Bolton, *What is Mental Disorder? An Essay in Philosophy, Science, and Values*. London, U.K.: Oxford Univ. Press, 2008, p. 321.

[2] S. L. James *et al.*, ''Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the global burden of disease study 2017,'' *Lancet*, vol. 392, pp. 1789–1858, Nov. 2018.

[3] P. S. Wang, S. Aguilar-Gaxiola, J. Alonso, M. C. Angermeyer, G. Borges, E. J. Bromet, R. Bruffaerts, G. De Girolamo, R. De Graaf, O. Gureje, and J. M. Haro, ''Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys,'' *Lancet*, vol. 370, no. 9590, pp. 841–850, Sep. 2007.

[4] G. Coppersmith, C. Harman, and M. Dredze, ''Measuring post traumatic stress disorder in Twitter,'' in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 1–4.

[5] R. L. Frost and D. J. Rickwood, ''A systematic review of the mental health outcomes associated with Facebook use,'' *Comput. Hum. Behav.*, vol. 76, pp. 576–600, Nov. 2017.

[6] I. Pirina and Ç. Çöltekin, ''Identifying depression on reddit: The effect of training data,'' in *Proc. EMNLP Workshop SMM4H, 3rd Social Media Mining Health Appl. Workshop Shared Task*, 2018, pp. 9–12.

[7] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, ''From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses,'' in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*, 2015, pp. 1–10.

[8] A. Trifan, R. Antunes, S. Matos, and J. L. Oliveira, ''Understanding depression from psycholinguistic patterns in social media texts,'' in *Advances in Information Retrieval*, vol. 12036. Cham, Switzerland: Springer, 2020, pp. 402–409.

[9] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, ''Detection of suicide ideation in social media forums using deep learning,'' *Algorithms*, vol. 13, no. 1, p. 7, Dec. 2019.

[10] J. H. Shen and F. Rudzicz, ''Detecting anxiety through reddit,'' in *Proc. 4th Workshop Comput. Linguistics Clin. Psychol. From Linguistic Signal Clin. Reality*, 2017, pp. 58–65.

[11] M. L. Birnbaum, S. K. Ernala, A. F. Rizvi, M. De Choudhury, and J. M. Kane, ''A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals,'' *J. Med. Internet Res.*, vol. 19, no. 8, p. e289, Aug. 2017.

[12] I. Sekulić, M. Gjurković, and J. Šnajder, ''Not just depressed: Bipolar disorder prediction on reddit,'' 2018, *arXiv:1811.04655*.

[13] A. Honey and C. Halse, ''The specifics of coping: Parents of daughters with anorexia nervosa,'' *Qualitative Health Res.*, vol. 16, no. 5, pp. 611–629, May 2006.

[14] D. E. Pawluck and K. M. Gorey, ''Secular trends in the incidence of anorexia nervosa: Integrative review of population-based studies,'' *Int. J. Eating Disorders*, vol. 23, no. 4, pp. 347–352, May 1998.

[15] R. L. Palmer, ''Death in anorexia nervosa,'' *Lancet*, vol. 361, no. 9368, p. 1490, May 2003.

[16] E. Bachar, Y. Latzer, S. Kreitler, and E. M. Berry, ''Empirical comparison of two psychological therapies: Self psychology and cognitive orientation in the treatment of anorexia and bulimia,'' *J. Psychotherapy Pract. Res.*, vol. 8, no. 2, p. 115, 1999.

[17] *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., Amer. Psychiatric Assoc., Washington, DC, USA, 2013, vol. 21.

[18] H. Bruch, ''Four decades of eating disorders,'' in *Handbook of Psychotherapy for Anorexia Nervosa and Bulimia*. New York, NY, USA: The Guilford Press, 1985, pp. 7–18.

[19] E. Bachar and A. Verbin, *Psychodynamic Self Psychology in the Treatment of Anorexia and Bulimia*. London, U.K.: Routledge, 2020.

[20] P. C. Hébert and M. A. Weingarten, ''The ethics of forced feeding in anorexia nervosa,'' *Can. Med. Assoc. J.*, vol. 144, p. 141, Jan. 1991.

[21] S. Giordano, ''Risk and supervised exercise: The example of anorexia to illustrate a new ethical issue in the traditional debates of medical ethics,'' *J. Med. Ethics*, vol. 31, no. 1, pp. 15–20, Jan. 2005.

[22] H. Maslen, J. Pugh, and J. Savulescu, ''The ethics of deep brain stimulation for the treatment of anorexia nervosa,'' *Neuroethics*, vol. 8, no. 3, pp. 215–230, Dec. 2015.

[23] D. E. Losada and F. Crestani, ''A test collection for research on depression and language use,'' in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.*, 2016, pp. 28–39.

[24] D. E. Losada, F. Crestani, and J. Parapar, ''Overview of eRisk: Early risk prediction on the internet,'' in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.*, 2018, pp. 343–361.

[25] M. Trotzek, S. Koitka, and C. M. Friedrich, ''Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia,'' in *Proc. CLEF, Working Notes*, 2018, pp. 1–15.

[26] D. G. Funez, M. J. G. Ucelay, M. P. Villegas, S. Burdisso, L. C. Cagnina, M. Montes-y-Gómez, and M. Errecalde, ''UNSL's participation at eRisk 2018 lab,'' in *Proc. CLEF, Working Notes*, 2018, pp. 1–11.

[27] M. E. Aragon, A. P. Lopez-Monroy, L.-C.-G. Gonzalez-Gurrola, and M. Montes, ''Detecting mental disorders in social media through emotional patterns—The case of anorexia and depression,'' *IEEE Trans. Affect. Comput.*, early access, Apr. 27, 2021, doi: 10.1109/TAFFC.2021.3075638.

[28] S. M. Mohammad and P. D. Turney, ''Crowdsourcing a word–emotion association lexicon,'' *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, Aug. 2013.

[29] D. E. Losada, F. Crestani, and J. Parapar, ''Overview of eRisk 2019 early risk prediction on the internet,'' in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.*, 2019, pp. 340–357.

[30] E. Mohammadi, H. Amini, and L. Kosseim, ''Quick and (maybe not so) easy detection of anorexia in social media posts,'' in *Proc. CLEF (Working Notes)*, 2019, pp. 1–14.

[31] W. Ragheb, J. Azé, S. Bringay, and M. Servajean, ''Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media,'' in *Proc. CLEF (Working Notes)*, 2019, pp. 1–15.

[32] A. S. Uban, B. Chulvi, and P. Rosso, ''Understanding patterns of anorexia manifestations in social media data with deep learning,'' in *Proc. 7th Workshop Comput. Linguistics Clin. Psychol., Improving Access*, 2021, pp. 224–236.

[33] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, ''Recognizing depression from Twitter activity,'' in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 3187–3196.

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan, ''Latent Dirichlet allocation,'' *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[35] T. Wang, M. Brede, A. Ianni, and E. Mentzakis, ''Detecting and characterizing eating-disorder communities on social media,'' in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, Feb. 2017, pp. 91–100.

[36] Y. R. Tausczik and J. W. Pennebaker, ''The psychological meaning of words: LIWC and computerized text analysis methods,'' *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.

[37] E. Fast, B. Chen, and M. S. Bernstein, ''Empath: Understanding topic signals in large-scale text,'' in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 4647–4657.

[38] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Sebastopol, CA, USA: O'Reilly Media, 2009.

[39] D. Ramírez-Cifuentes, M. Mayans, and A. Freire, ''Early risk detection of anorexia on social media,'' in *Proc. Int. Conf. Internet Sci.*, 2018, pp. 3–14.

[40] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, ''The development and psychometric properties of LIWC2015,'' Dept. Psychol., Univ. Texas Austin, Austin, TX, USA, Tech. Rep., 2015.

[41] T. Zhou, G. Hu, and L. Wang, ''Psychological disorder identifying method based on emotion perception over social networks,'' *Int. J. Environ. Res. Public Health*, vol. 16, no. 6, p. 953, Mar. 2019.

[42] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, ''Detection of depression-related posts in reddit social media forum,'' *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

[43] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes-y-Gómez, ''Detecting depression in social media using fine-grained emotions,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 1481–1486.

[44] H. Alhuzali, T. Zhang, and S. Ananiadou, ''Predicting sign of depression via using frozen pre-trained models and random forest classifier,'' in *Proc. Working Notes CLEF*, 2021, pp. 21–24.

[45] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, ''Detecting depression and mental illness on social media: An integrative review,'' *Current Opinion Behav. Sci.*, vol. 18, pp. 43–49, Dec. 2017.

[46] A. Wongkoblap, M. A. Vadillo, and V. Curcin, ''Researching mental health disorders in the era of social media: Systematic review,'' *J. Med. Internet Res.*, vol. 19, no. 6, p. e228, Jun. 2017.

[47] L. Canetti, E. Bachar, and E. M. Berry, "Food and emotion," *Behav. Process.*, vol. 60, no. 2, pp. 157–164, 2002.

[48] Y. Latzer, Z. Hochdorf, E. Bachar, and L. Canetti, "Attachment style and family functioning as discriminating factors in eating disorders," *Contemp. Family Therapy*, vol. 24, pp. 581–599, Dec. 2002.

[49] U. Pinus, L. Canetti, O. Bonne, and E. Bachar, "Selflessness as a predictor of remission from an eating disorder: 1–4 year outcomes from an adolescent day-care unit," *Eating Weight Disorders Stud. Anorexia, Bulimia Obesity*, vol. 24, no. 4, pp. 777–786, Aug. 2019.

[50] Z. Jianqiang and G. Xiaolin, "Comparison research on text preprocessing methods on Twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017.

[51] Y. HaCohen-Kerner, Y. Yigal, and D. Miller, "The impact of preprocessing on classification of mental disorders," in *Proc. 19th Ind. Conf. Data Mining (ICDM)*, New York, NY, USA, 2019, pp. 52–66.

[52] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS ONE*, vol. 15, no. 5, May 2020, Art. no. e0232525.

[53] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995.

[54] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[55] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[56] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, nos. 5–6, pp. 183–197, Jul. 1991.

[57] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc. B, Methodol.*, vol. 20, no. 2, pp. 215–232, 1958.

[58] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.

[59] M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for text categorization revisited," in *Proc. Australas. Joint Conf. Artif. Intell.*, Berlin, Germany, 2004, pp. 488–499.

[60] E. Frank and R. R. Bouckaert, "Naive Bayes for text classification with unbalanced classes," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*, 2006, pp. 503–510.

[61] T. Kanamori, "Deformation of log-likelihood loss function for multiclass boosting," *Neural Netw.*, vol. 23, no. 7, pp. 843–864, Sep. 2010.

[62] M. Sabzevari, G. Martínez-Muñoz, and A. Suárez, "Vote-boosting ensembles," *Pattern Recognit.*, vol. 83, pp. 119–133, Dec. 2018.

[63] C. Nadeau and Y. Bengio, "Inference for the generalization error," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 2000, pp. 307–313.

[64] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003.

[65] Y. HaCohen-Kerner, E. Malin, and I. Chasson, "Summarization of Jewish law articles in Hebrew," in *Proc. CAINE*, 2003, pp. 172–177.

[66] Y. HaCohen-Kerner and S. Y. Blitz, "Initial experiments with extraction of stopwords in Hebrew," in *Proc. KDIR*, 2010, pp. 449–453.

[67] Y. HaCohen-Kerner, R. Dilmon, M. Hone, and M. A. Ben-Basan, "Automatic classification of complaint letters according to service provider categories," *Inf. Process. Manage.*, vol. 56, no. 6, Nov. 2019, Art. no. 102102.

[68] S. Baluja, "An empirical comparison of seven iterative and evolutionary function optimization heuristics," Dept. Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. NASA-CR-201901, 1995.

[69] Y. HaCohen-Kerner, H. Beck, E. Yehudai, M. Rosenstein, and D. Mughaz, "Cuisine: Classification using stylistic feature sets and/or name-based feature sets," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, pp. 1644–1657, 2010.

[70] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 473–479.

[71] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[72] A. Seker, E. Bandel, D. Bareket, I. Brusilovsky, R. Shaked Greenfeld, and R. Tsarfaty, "AlephBERT: A Hebrew large pre-trained language model to start-off your Hebrew NLP application with," 2021, *arXiv:2104.04052*.

[73] A. Chriqui and I. Yahav, "HeBERT & HebEMO: A Hebrew BERT model and a tool for polarity analysis and emotion recognition," 2021, *arXiv:2102.01909*.

[74] S. Pyysalo, J. Kanerva, A. Virtanen, and F. Ginter, "WikiBERT models: Deep transfer learning for many languages," 2020, *arXiv:2006.01538*.

[75] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[76] Y. Ko and J. Seo, "Text classification from unlabeled documents with bootstrapping and feature projection techniques," *Inf. Process. Manage.*, vol. 45, no. 1, pp. 70–83, Jan. 2009.

[77] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[78] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2012, pp. 599–619.

[79] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[80] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria, "A generative model for category text generation," *Inf. Sci.*, vol. 450, pp. 301–315, Jun. 2018.

[81] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Syst. Appl.*, vol. 121, pp. 49–65, May 2019.

[82] G. Brauwers and F. Frasincar, "A survey on aspect-based sentiment classification," *ACM Comput. Surv.*, Dec. 2021, p. 35.

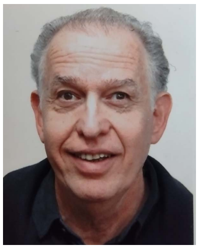
[83] A. Montesinos-López, O. A. Montesinos-López, D. Gianola, J. Crossa, and C. M. Hernández-Suárez, "Multi-environment genomic prediction of plant traits using deep learners with dense architecture," *G3, Genes, Genomes, Genetics*, vol. 8, no. 12, pp. 3813–3828, Dec. 2018.

[84] J. Ayoub, X. J. Yang, and F. Zhou, "Combat COVID-19 infodemic using explainable natural language processing models," *Inf. Process. Manage.*, vol. 58, no. 4, Jul. 2021, Art. no. 102569.

YAAKOV HACOHEN-KERNER received the Ph.D. degree in computer science from Bar-Ilan University, Ramat-Gan, Israel. He is the Dean of the Faculty of Engineering and Computer Science, Jerusalem College of Technology (JCT). He is also with the CS Department, JCT. He has authored and coauthored 104 papers. His current research interests include text classification, sentiment analysis, author profiling, word completion and prediction, key-phrase extraction, citation extraction and analysis, plagiarism detection, and composition of chess problems.

NATAN MANOR received the bachelor's degree in software engineering from the Jerusalem College of Technology (JCT). Together with his partner Michael Goldmeier, they finished their final graduation project under the guidance of Yaakov Hacohen-Kerner. He is a coauthor of two previous articles in the text classification domain together with Yaakov Hacohen-Kerner.

MICHAEL GOLDMEIER received the bachelor's degree in software engineering from the Jerusalem College of Technology (JCT). Together with his partner Natan Manor, they finished their final graduation project under the guidance of Yaakov Hacohen-Kerner. He is a coauthor of a previous article in the text classification domain together with his partner and Yaakov Hacohen-Kerner.



EYTAN BACHAR received the Ph.D. degree in psychology from The Hebrew University of Jerusalem, Jerusalem, Israel. He is the Head Psychologist of the Hadassah University Medical Center, Jerusalem, an Associate Professor with the Hebrew University of Jerusalem, and a Former Chair of the Israeli Association of Eating Disorders. He is a highly experienced Expert in eating disorders and post-traumatic stress disorders (PTSD). His research interests include psychological factors contributing to the onset, maintenance and treatment of two mental disorders: eating disorders and post-traumatic stress disorders.