Bio, psycho, or social: supervised machine learning to classify discursive framing of depression in online health communities

Renáta Németh¹ · Fanni Máté¹ · Eszter Katona¹ · Márton Rakovics¹ · Domonkos Sik¹

Accepted: 8 December 2021 / Published online: 8 January 2022
© The Author(s) 2022

Abstract

Supervised machine learning on textual data has successful industrial/business applications, but it is an open question whether it can be utilized in social knowledge building outside the scope of hermeneutically more trivial cases. Combining sociology and data science raises several methodological and epistemological questions. In our study the discursive framing of depression is explored in online health communities. Three discursive frameworks are introduced: the bio-medical, psychological, and social framings of depression. ~80 000 posts were collected, and a sample of them was manually classified. Conventional bag-of-words models, Gradient Boosting Machine, word-embedding-based models and a state-of-the-art Transformer-based model with transfer learning, called DistilBERT were applied to expand this classification on the whole database. According to our experience ‘discursive framing’ proves to be a complex and hermeneutically difficult concept, which affects the degree of both inter-annotator agreement and predictive performance. Our finding confirms that the level of inter-annotator disagreement provides a good estimate for the objective difficulty of the classification. By identifying the most important terms, we also interpreted the classification algorithms, which is of great importance in social sciences. We are convinced that machine learning techniques can extend the horizon of qualitative text analysis. Our paper supports a smooth fit of the new techniques into the traditional toolbox of social sciences.

Keywords Supervised machine learning · Online forums · Natural language processing · Framing of depression · Gradient boosting machine · Transformer-based model

✉ Renáta Németh
nemethr@tatk.elte.hu

¹ Faculty of Social Sciences, Research Center for Computational Social Science, Eötvös Loránd University of Budapest, Budapest, Hungary

1 Big textual data in psychological research

Our paper was written from a sociological perspective, with the purpose of identifying the challenges of combining social sciences with machine learning. Such a methodological attempt raises questions concerning the epistemological core of social research. Some of these questions expand the horizon of older ones, others are entirely new. Our article discusses (1) whether an automated language processing method can be applied to a sociological problem outside the scope of hermeneutically more trivial business applications. Our paper also highlights (2) the need for a deeper understanding of the annotation process and the use of annotated data. Furthermore, it (3) reveals the connection between inter-annotator agreement and predictive performance. Finally, the question concerning (4) the relationship between prediction and explanation, and the need to interpret classification models is analysed.

We present initial experiments on understanding individual framing of depression in online health communities. As such a huge corpus is impossible to read and process by human capacity, automated text analytical methods originating from the field of Natural Language Processing (NLP) are required. Another specific empirical characteristic of digital data is that they are produced independently from the researcher as opposed to researcher-controlled surveys or randomized trials. At first sight, this gives an advantage of objectivity, as we can avoid all the validity issues that arise in the case of self-reported studies. However, digital traces are not directly interpretable, therefore, their objectivity is also limited, especially because of their observational nature. We must interpret the observed digital behaviour (here: text) without the opportunity to ask questions from the observed persons, or at least to conduct a detailed individual-level investigation (here: to read the text). One solution for this challenge is to apply supervised machine learning (SML) (Hastie et al. 2009), which makes it possible to teach the result of a traditional qualitative text analysis of a small sub-corpus to an algorithm that expands it to the whole corpus. In our study we employed coders (or ‘annotators,’ a term coming from the programming community) to read a small sub-corpus, to interpret its content, after which we expanded this knowledge to the whole corpus. The method is a clear example of mixed methods: the result of a human annotation (a qualitative approach) is given as an input to the quantitative algorithm. SML techniques for automatic text classification are widely used in industrial environments, where commercial success is paramount (see e.g. Bernardi et al. 2019). Although promising social science applications are also known, it is still an open question whether these methods can be used to solve social science problems outside the scope of hermeneutically more trivial cases (for some of the challenges, see Chen et al. 2018).

2 Motivation of the topic of depression

Depression is a disease of modernity. From a sociological perspective, it can be viewed as a form of social suffering. It refers to social relations dominated by uncertainty, where collective representations impose increased responsibility on the individual, while the individual cannot change the circumstances. Depression, in this sense, is embedded in the more general problem of distorted social integration and is characterized by blocked individual ability to act (Sik 2018).

The discursive framing of depression is a social construction, which constitutes a set of concepts and perspectives on how patients perceive and interpret their illness. The framing defines the meaning of depression, it provides causal explanation, and affects treatment preferences. The current clinical explanations of depression point to bio-medical, psychological, and social discourses (e.g., Comer 2015). However, the sociology of mental disorders also raises the question: how is depression framed by the patients themselves? According to related research, psychotherapy tends to transform ‘social suffering’ into ‘individual suffering’, which has a medicalizing and individualizing consequence (see, e.g., Ehrenberg 2009; Flick 2016).

Online space, as an empirical base, is of great significance in this context, since (1) participants of discussions try to utilize their peers’ experiences in their everyday praxis, as lessons learned provide practical help for them; and, (2) while texts do not contain elaborated views of their authors, they are part of a dynamic discourse happening in real time.

Our approach has been shared by previous studies aiming at describing the dominant narrative panels identified in online depression forums (see e.g. Lachmar 2017, Pan et al. 2018, Sik 2020). Our aim was to continue the course set by these studies in understanding depression narratives. In our earlier papers (Sik et al. 2021; Németh et al. 2021) the inductively elaborated topical structure of the corpus was introduced. The present study approaches the corpus from the opposite direction: instead of an exploratory analysis, an attempt is made to classify our data according to a predefined conceptual framework based on the literature. Initial results of this study were published in Németh et al. (2020), where we introduced our learning path from the difficulties of human annotation to the limitations of algorithmic language processing and concluded that meaningful interpretation of massive data sets requires the integration of both qualitative and quantitative methodological and theoretical insights.

In the present paper, we discuss our latest results with a complete discussion of the quantitative methodological challenges arising when we apply SML in social research. We also highlight some points in which the logic of this data-driven approach is highly different from traditional, theory-based statistical modelling used in social sciences. We discuss the measurement issue of intercoder agreement, identify hard-to-annotate cases, and prove that their presence misleads the machine learner. We also complete our initial, already published results with more advanced models. Because these models often look like ‘black boxes’ in the sense that they do not explain their predictions, we also try to make their prediction interpretable, which is of great importance in the field of social sciences.

3 Data collection

We used SentiOne, a web-based social listening and text analytics platform, as the source of our data, and chose the most popular English-speaking online health forums, which were selected via Google search for “depression forum” and “depression online.”

We collected only depression-related posts, where ‘depression-related’ was defined as follows: (1) we selected threads which contained the word “depression” or “depressed” in their title or at least in one of their posts, then (2) we selected posts the link, topic, or content of which contained a depression-related term, for instance: “unipolar depression,” “mood disorder,” or “depressant.” The data set, collected by SentiOne in compliance with GDPR regulations, contained 79 889 articles posted between February 15, 2016 and

Fig. 1 Distribution of posts by domains

February 15, 2019 covering only publicly available posts, which were shared willingly by their authors.

After removing duplicate and too short (< 20 words) posts, our final corpus contained 67 857 posts. The frequency of posts varied between the domains(Fig. 1)—the most dominant forums being healthunlocked.com and depressionforums.org.

4 Log of supervised machine learning

Our initial research question was whether our large corpora can be classified into the three framing categories in an automated way. We planned to employ human annotators to classify a smaller sample of the posts and apply different supervised machine learning algorithms to expand this knowledge.

A machine learning algorithm makes the computer ‘learn’ from data without being explicitly programmed. SML on textual data helps researchers to perform analysis on data sets too large for human processing. For example, in case of supervised sentiment analysis, getting several movie reviews labelled (=training data) as positive, negative, or neutral, the algorithm will learn by example and be able to predict the polarity of an unlabelled review using its textual patterns. It is supervised, because there is a supervisor, who attaches the polarity labels to a sample of the data. In case of unsupervised machine learning, there are no pre-labelled data, and the algorithm will, for example, try to cluster the reviews into groups based solely on their textual similarity (for more details on unsupervised learning see e.g. Eisenstein 2019).

Pre-labelled training data in some cases are readily available—see, the study of Hasan et al. (2014), who aimed at identifying emotions expressed by Twitter messages by using machine learning algorithms. Pre-labelled data were based on hashtags as emotion labels (like “#sad,” “#excited,” or “#happy”), and the algorithm expanded this knowledge to other non-hashtagged messages using only the textual patterns of the hashtagged ones.

When such pre-labelled data are not at hand, human annotators are employed to read and label (or ‘annotate’) a sample of the data manually. In order to avoid overfitting, predictive performance is assessed on an “out-of-sample” dataset by dividing the annotated set into a training set and a validation set. Using the training set, an algorithm is constructed to predict labels by identifying textual patterns, which explain the annotators’ labels. Using these patterns, the algorithm predicts labels for texts in the validation set as if we did not know their labels given by the annotators. The algorithm’s predictive performance can be evaluated by comparing these labels with the ones given by the annotators (for a good summary on practical and methodological challenges of the annotation task, see e.g. Eisenstein 2019).

Typically, several different algorithms are tried, which may differ in their mathematical approach, their language model, the actual values of their parameters, and the textual features they consider. Based on the evaluation, we can choose the best one among them. If its performance is acceptable, it can be applied to label the whole unannotated data set.

In this paper, we illustrate the specific nature of a new modelling approach, and we also try to integrate it into the traditional social science toolbox. The data-driven modelling approach presented here is very different from traditional, theory-based statistical modelling used in social sciences. Social scientists for explanatory purposes typically build only one data-generating model based on their theoretical assumptions, then estimate and test the model parameters. Their models are rather simple and easily interpretable; however, they mostly ignore out-of-sample performance. On the contrary, data scientists build predictive models, that is, try to predict responses for future inputs. While treating the underlying data-generating process as unknown, they build complex models, train them, and evaluate their predictive performance to discover the best one. These models perform well on out-of-sample data, but often produce black-box results which are hard to interpret as the link between the outcome and the input cannot be easily described. Data scientists in the last years have realized that interpretability is important to them as well, as it can help them to know more about the reason why a model might fail. New methods for the interpretation of black-box machine learning models are being developed and published (for a summary see Molnar 2019). These are indirect methods in the sense that do not (cannot) give the direct interpretable link between the outcome and the input but, for example, may explain how individual predictions were made or give the average behaviour of the model on a dataset.

5 Producing training data: the process of annotation

Everyday machine-learning tasks are basic and easy for humans—see, e.g., image annotations to mark shape, pattern, or size of objects in images. To perform these tasks, it is common to recruit annotators through internet services (e.g., Amazon Mechanical Turk, a marketplace for recruiting online workers). This solution outsources annotation tasks in bulk, with low costs and fast completion rates (Sorokin and Forsyth 2008). In our case, the annotation task was not as simple, requiring basic social science knowledge. Framing is partly an unconscious process: to identify the framing at work, a simple referential reading is not enough. One has to take into consideration: (1) ontological models (depression as bodily, mental, or social phenomenon), (2) explanatory models (depression as dysfunction or illness, result of maladaptation/trauma, or consequence of social distortions caused by peripheral role experiences in the social network), and (3)

therapeutic models (medical interventions, psychotherapy or social work / social political intervention). As these factors are present in many cases only in a latent way, the annotation requires complex hermeneutic interpretation, which also needs to be coordinated in order to reach a unified approach (Gius and Jacke 2017).

We recruited our annotators carefully from social science university students having an interest in the topic and the method. The training set of ~4 500 posts was selected randomly. For the classification task, with many real examples, a detailed classification guideline was prepared, which was recursively updated in an inductive way based on regular discussions with the annotators. The detailed definition of each category continuously evolved as more of the training set was annotated. This process is substantially different from quantitative research, which applies a pre-established set of explicit and unambiguous categories. Our case was rather a kind of qualitative coding, as our categories were derived from a prior abstract theory and inductively developed during the research. The categories represented the researchers' concepts, and not those of the annotators' own, which is a type of qualitative categorization called "theoretical" by Maxwell (2005).

The conceptual clarification was complemented with a list of examples:

- Bio-medical: *In the past I would get severely depressed by the long winter nights. I feel like medication and proper treatment has given me a second chance at living it more fully.* (Long nights as a natural phenomenon to explain depression. Concepts of "medication" and "treatment" to describe improvement.)
- Psychological: *You have to sacrifice the story that you tell about yourself. You have to be willing to start writing a new story. You have to be brave enough to consider what life would be like if you were ok. You have to stick your neck out just a tiny bit and risk having hope. You have to get clear about what being OK means. You have to be willing to picture yourself bright and happy.* (Speaks about herself/himself from a psychological self-analytic perspective. Revises herself/himself in the framework of cognitive therapy.)
- Sociological: *Every piece of climate news increasingly comes with a sense of dread: is it too late to turn around? The notion that our individual grief and emotional loss can actually be a reaction to the decline of our air, water, and ecology rarely appears in conversation or the media. It may crop up as fears about what kind of world our sons or daughters will face. But where do we bring it? Some bring it privately to a therapist. It is as if this topic is not supposed to be publicly discussed.* (Explicit links to the society with a critique of the everyday praxis, which transforms social issues into suffering related to the self.)

Five labels were used, as 'unclassifiable' (it is about depression, but the framing cannot be identified) and 'irrelevant' (it is not about depression) were also added to the three meaningful framing types. This non-trivial annotation task implied that (1) annotators were instructed to assign two labels to the texts if needed, a primary and an (optional) secondary one. 34% of the posts got a second label from at least one of the annotators. Additionally, (2) we had two independent annotators for each text, which gave a maximum of four labels together. The final, integrated label was based on majority voting, therefore, the most frequent label among the four was chosen. We asked a third annotator (a senior researcher) to resolve the (very few, 12.3%) ambiguous cases. A secondary integrated label was also assigned, if there was an unambiguous second winner of the voting, e.g., if both of the annotators used the same primary and the same secondary label.

The majority of the first labels pertained to psychological (33%) or bio-medical (29%) framing, and only a smaller proportion of them (9%) was labelled as sociological. 30% of the texts were irrelevant or unclassifiable. Sociological framing was present more frequently among the second labels. We also investigated label distribution among those cases where two annotators agreed on the primary label. We call these cases ‘easy’, as they are less controversial in human annotation. When determining annotator agreement, we used only 4 categories, not making a difference between ‘irrelevant’ and ‘unclassifiable.’ 57% of all cases were easy. The type of framing has a strong relation with easiness: the proportion of easy cases was 72%, 55%, 37% among the bio-medical / psychological / sociological labels, respectively.

To better understand why a post is easy or difficult to annotate, we also investigated their linguistic characteristics, by comparing the distribution of the features in the easy and hard groups. After normalizing the distribution by the standard deviation for each feature, the difference between the medians was calculated (see Fig. 3 in the 14). Although the differences are not very large (they differ by no more than 0.15 standard deviations), they make sense. Lexically more diverse posts and posts with fewer misspelled words tend to be easier to annotate. Easy posts tend to have a lower proportion of verbs and higher proportion of nouns. As nouns are more frequent in formal styles, and verbs are more frequent in informal styles (e.g. Heylighen and Dewaele 1999), our results suggest that easy posts are presumably more accurate, more formal, while hard posts are more flexible, more informal, closer to the oral style.

6 Inter-annotator agreement

Inter-annotator agreement is at the heart of automated text analysis: if the codes are not reliable, the learner cannot effectively learn from them, and its prediction cannot be trusted. The question arises how to assess agreement.

Our research design was specific from two points of view: we used more than two categories for labelling, and annotators were allowed to assign secondary labels to the posts. There are many different agreement measures defined for different contexts (for a good review—see, Lombard et al. 2002). Simple percent agreement, despite being very widely used, does not take into account the agreement occurring by chance. Note that for uniform categorical distributions if two annotators select between two labels randomly, the agreement between them is expected to be 50%. If there are three labels, the expected chance agreement is only 33%. Therefore, the same observed agreement shows a different picture depending on whether it was calculated in a 2-label or 3-label context. We used Cohen’s kappa instead (Cohen 1960), which tells us how much better the classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class:

$$\kappa = \frac{p - p_e}{1 - p_e}$$

where p is the proportion of units where there is agreement, and p_e is the proportion of units which would be expected to agree by chance. The nominator tells us how much better the observed agreement is than the expected agreement by chance. This difference is compared to the denominator, which gives the theoretical maximum of the difference ($1 - p_e$, if

$p = 1$). In the case of perfect agreement ($p = 1$), kappa equals 1. On the contrary, if annotators select labels randomly, $p = p_e$, so kappa equals 0.

Kappa has been widely used in bio-medical or psychological sciences for a couple of decades. A typical application is when two or more judges categorize diagnostic images or personality traits in parallel. However, in the field of machine learning, kappa has not received much attention as a measure for accuracy (Ben-David 2008). Commonly used evaluation measures in machine learning context are precision, recall, and F1-score. Each of these measures are defined for two labels only. For more than two labels, they are usually simply averaged over all the possible binary label pairs (sociological vs. others, psychological vs. others, etc.), which is ill-advised if we have an unbalanced label distribution like we have here. Finally, they do not account for agreement by chance.

Another specificity of our annotation was that annotators were allowed to assign secondary labels to the posts. If we define agreement as the match of primary labels, simply discarding the optional secondary label, we get an overly conservative measure. Rosenberg and Binkowski (2004) proposed a solution for multiple labels, but it does not account for the rank order that our primary and secondary labels have. Hence, we considered a new, ‘liberal’ version of kappa, when agreement means that one of the primary labels agreed with either the primary or the secondary label given by the other annotator (a similar solution is used by Flor et al. 2016). The advantage of liberal kappa over its original, conservative version is that it also takes into account the secondary meaning of the texts. While liberal kappa may show an overly optimistic picture, its conservative counterpart leads to a too strict evaluation. The “truth” is somewhere between, hence we present both of them when comparing different classification algorithms.

Measured in the conservative way, our observed agreement was 58.3%, and the liberal measure was 69.7%. For observed agreement, the criterion of 70% is often used for exploratory research. Turning to kappa: according to Landis and Koch (1977), a kappa between 0.4 and 0.6 shows a moderate strength of agreement, while kappa greater than 0.6 shows a substantial agreement. Our conservative kappa was 0.42, while the liberal version was 0.58, which shows a nearly substantial agreement.

7 Preprocessing and feature extraction

The first phase of all automated text analysis is preprocessing to transform texts into numerical data (see Fig. 2 below). During our analysis, we used Python’s NLTK package (Bird et al. 2009). Most text mining approaches rely on a word-based representation of texts (e.g., Aggarwal and Zhai 2012), assuming that texts are ‘bags of words,’ without taking word order into account. As Fig. 2 shows, we also tested bag-of-words classifiers in the first modelling phase. Distribution of words within texts forms the numeric input data for the analysis. Instead of simple frequency distribution, ‘term frequency–inverse document frequency’ (Tf-idf) weights are often used, which take normalized frequency into account where each word frequency is divided by the number of texts this word appears in. We tested both approaches.

We followed another frequently used preprocessing step by treating two-word collocations as single terms (‘significant bigrams’ in NLP terminology) during the analysis. For example, significant bigrams in our corpus were “prefrontal cortex,” “family member,” and “side effect.” We also treated the name of most common mental disorders as single terms (for example, “bipolar depression” or “obsessive-compulsive disorder”).

Fig. 2 The analytic process

We also tried to find three-word collocations (significant trigrams) and proper nouns (‘named entity recognition’ in NLP terms), but as they had low frequencies, we finally skipped these steps.

We employed lemmatization to standardize different forms of the same word (running, runs, ran → run). We also discarded punctuation and capitalization; deleted URLs, e-mail addresses, and repost part of the texts; as well as removed duplicate and too short (<20 words) posts. We tested whether the removal of very common words (‘stop words’) improves the learner; examples of stop words are “the,” “in,” and “when.” The idea behind stop word removal is to put more emphasis on those words which define the meaning of the given post.

We also tested whether replacing the word-based representation for a lower-level representation gives more efficient learning algorithms. For this purpose, we used character n-grams ($n = 2$ to 5), a sequence of n characters, with the hope to make the model more robust to spelling errors which often occur in online communication. These steps may seem to result in a serious reduction of information. The simplifying language model itself makes the validation phase highly important.

As a result, we got a numeric dataset with texts in rows, terms (lemmatized words or character n-grams) in columns, and occurrence (weighted) frequencies in cells. Terms are playing the role of independent variables in our classification models. In order to build some background knowledge into the model, we tried to define further variables (‘document features’ in NLP terms) which might help the learner when classifying the posts. Such features were the number of times she/he mentions mental health medications or dosage of drugs (“8 mg”). Our reason for adding these features was that it may tend to relate to bio-medical framing. We also used the sentiment score of the post as a potential predictor. Further features characterizing language usage of the authors were: number of words / sentences / abbreviations, proportion of nouns / verbs / adjectives / stop words / emojis / numbers / misspelled words / punctuation marks / commas, average length of sentences / words, diversity of emojis and lexical diversity of the text. For details on their calculation see the 14.

8 Automated classification experiment I: multinomial logistic regression

Figure 2 presents the whole analytic process of our research. In the modelling phase, first we experimented with bag of words classifiers, that treat words as independent units, ignoring their order.

All the source code developed for the classification experiments have been released at <https://github.com/RC2S2/QUQU-depression>. In our first experiments, Python's Scikit-learn library (Pedregosa et al. 2011) was used with different algorithms to classify the posts into the framing categories. First, we trained a multinomial logistic regression classifier. The outcome variable was the primary label—two variants were tested: with five labels (bio-medical, psychological, sociological, irrelevant, and unclassifiable) or with four labels (by merging the latter two labels). We used all available tokens in several configurations: (1) as terms, ignoring those that have a document frequency lower than 5, (2) as character n-grams ($n = 2$ to 5), and (3) as word n-grams ($n = 1$ to 3). We tested different values of the regularization parameter: a C value¹ of 1 turned out to be the best.

As we have already discussed earlier, when evaluating the model, we split the labelled data into a training set and a validation set. In order to prevent potential bias caused by the splitting itself, we used a more developed variant of the splitting procedure: the five-fold cross-validation. The procedure splits the dataset into five parts (called folds) randomly, executes a validation for each group, by taking four groups jointly as a training set and taking the remaining one group as a validation set, and retains the evaluation measure. The five evaluation measures were averaged to give a single summary measure. When defining folds, we used stratified sampling in order to assure identical distribution of labels across the five folds (on importance of stratification in cross-validation—see Kohavi's widely cited paper, 1995).

When evaluating our classifier models, we had to compare predicted primary labels with the ones given by the annotators, which defines the same measurement problem we had in case of measuring inter-annotator agreement: it is a multi-class classification problem with unbalanced classes. Therefore, we used Cohen's kappa instead of the standard performance measures (recall, F1-score, etc.). Again, we also considered a liberal version of the kappa accepting a predictive label if it agrees with either the primary or the secondary label given by the annotators.

The best model (Table 1) gives a liberal kappa of 0.53 (a moderate agreement) and uses Tf-idf weights. It outperformed the baseline by a moderate margin (0.53 vs. 0.4). When interpreting it, we have to consider the inter-annotator agreement as well: it may provide an estimate of the objective difficulty of the classification task (an assumption we will investigate below). The liberal version of kappa was 0.58, that is, our learner nearly reaches the human inter-annotator agreement.

Adding features and word n-grams marginally improved performance, while character n-grams had a negligible effect. The table shows the results of models with stop word removal and with a four-category outcome; the presence of stop words and the five-category outcome did not change the results.

¹ Regularization is used to reduce overfitting in predictive models. The C parameter is defined, as smaller values specify stronger regularization.

Table 1 Logistic regression performance results (averaged over folds)

Document-term matrix vectorization	Features used?	Conservative kappa	Liberal kappa
Term frequency	No	0.40	0.48
Tf-idf	Yes	0.41	0.49
Term frequency	No	0.44	0.51
Tf-idf weights	Yes	0.44	0.52
Tf-idf using word n-grams (n=1 to 3)	Yes	0.44	0.53
Tf-idf using character n-grams (n=2 to 5)	Yes	0.45	0.53

Standard performance measures, precision, and recall give another insight into the performance of the best model. They can be defined in a very simple way. If our aim is to detect posts with sociological framing of depression, precision is the proportion of correctly classified sociological posts within those predicting the label ‘sociological.’ Recall is the proportion of correctly classified posts within those truly having the label ‘sociological,’ known as sensitivity in biostatistics. Note that only the denominators differ in the two definitions (for a good summary on these and other measures—see, Powers 2011).

We calculated these measures for each label separately and calculated a weighted average for them (weighted by label percentage). As Table 4 in the 14 shows, averaged recall and precision were both 0.68, showing an acceptable achievement. These results are discussed in more detail from a hermeneutic point of view in our previous paper (Németh et al. 2020).

9 Automated classification experiment II: other bag-of-words classifiers

We also experimented with other classifiers incorporating features that proved to give the best performance in case of the logistic regression: Tf-idf weights, word n-grams ($n = 1$ to 3), features, stop word removal, and four categories in the outcome. Some often used classifiers were chosen: Bernoulli Naïve-Bayes, Support Vector Machine (SVM), Random Forest, and Gradient Boosting Machine (GBM), all from Python’s Scikit-learn library (Pedregosa et al. 2011). Bernoulli Naïve-Bayes was used with an alpha parameter equalling 1. In the case of the other models, we used a grid search for finding the best combination of hyper-parameters. For SVM, we tuned C , gamma, and kernel parameters. For Random Forest, the number of trees and maximum number of features, and for GBM, the number and maximum depth of trees, and the learning rate.

Liberal version of kappa (averaged over folds) was calculated as a performance measure. Kappa was 0.41 for Bernoulli Naïve-Bayes, 0.50 for SVM, 0.48 for Random Forest, and 0.52 for GBM. Tables 5, 6, 7, 8, 9, 10 in the 14 show all the other performance measures for the models. Logistic regression remarkably outperforms the other classifiers (kappa: 0.53), even GBM which is widely considered state-of-the-art among out-of-the-box models.

10 Automated classification experiment III: language modelling-based classifications

Since the best models using features derived from the bag-of-words approach achieved similar performance regardless of model structure, we decided to turn to methods that implicitly or explicitly deal with the sequential nature of textual data.

First, we experimented with word embeddings. In these models (Eisenstein 2019), words are embedded in a multidimensional vector space, where the distance between two words corresponds to the similarity of their contexts in the corpus. Context here is typically a symmetric window around the word. Word embedding models are semantic models based on the distributional hypothesis that words occurring in a similar context tend to have similar meanings.

Using the gensim Python library (Řehůřek and Sojka 2021), we created a word2vec word embedding (Mikolov et al. 2013) for the words in our full corpus (choosing a 100-dimensional embedding size, and a window size of 10), then averaged these word vectors for the posts to get document-level features that give a (semantic) representation of the posts.

We added these new features to the previous ones for use in the best logistic regression and GBM models (see the word-embedding-based classifiers in Fig. 2). Using the average position of the words contained by the document as a set of classification features is a well-known procedure, see e.g. Krishnan and Kamath (2020). As a result, even after rerunning the grid search for the best hyper-parameters, neither model produced better kappa values than previously (for detailed results see Table 9 and Table 10 in the 14).

Since the simple addition of word-embedding features can only be considered as implicit sequence modelling—through the sequentially moving window in the embedding algorithm—we turned to a second, explicit sequence modelling approach, employing the DistilBERT deep-learning (DL) language model (Sanh et al. 2020, see language based modelling step in Fig. 2). We used the transformers library (PyTorch implementation) from Hugging Face (huggingface.co) (Wolf et al. 2020), which provides an open source library of DL language models. DistilBERT is a large-scale pre-trained transformer with ~66 million parameters, which we were able to use with transfer learning. Transfer learning is a technique where a deep learning model pre-trained on a very large dataset is used as a starting point for a related task on another dataset. Being a very efficient machine learning technique, it made a breakthrough in computer vision in around 2012 and 13, while in NLP, it has become a viable option since the ground-breaking work of Vaswani et al. (2017).

DistilBERT was pre-trained on the Toronto Book Corpus and English Wikipedia. We have fine-tuned it for the classification task at hand, training the added fully-connected layer's weights with our own data. This means that the DistilBERT language-model worked as a sophisticated feature-extractor in our case. With a batch size of 32, and a learning rate of $2e-5$ for the AdamW optimizer, three training epochs were needed to reach the best test set performance. Other parameters were used as default. The fine-tuned DistilBERT classifier reached a conservative kappa of 0.53, and a liberal kappa of 0.60 on the test set for the four-category task. Kappa and all the other performance measures (see Table 11) show that the DistilBERT classifier had a better performance than the so far best logistic regression model.

11 Interpreting the classifiers

From a machine learning perspective, classifiers are usually chosen based on performance measures. However, even developers of these algorithms shouldn't just trust the model and ignore why it made a certain decision (Molnar 2019). How the given decision was made is also important from the traditional explanatory point of view social sciences have. Not just because interpretation helps us to determine whether the predictions are sensible, but also to put findings into the context of domain knowledge. We can uncover which factors have the greatest impact on the model's decisions, which reduce or increase the probability of a classification or what is the effect of the factors we considered a priori to be important.

There are challenges many machine learning algorithms face when it turns to interpretability. The best performance for large datasets is often achieved by complex models that look like black boxes and whose interpretation is challenging even for experts. Our best performing DistilBERT model is also one such model. We have used a SHAP explainer auxiliary model (Lundberg and Lee 2017) to extract interpretable feature information, based on the logits of the output layer.

As a comparison and for a better understanding, we also performed the interpretation for logistic regression. Interpretation in this case is easier: we may consider the model itself and look for those factors that have the highest effect on the prediction, that is, the ones with the largest standardized coefficients. For the multinomial logistic regression in the Scikit-learn implementation, the exponentiated coefficients for the features can be interpreted as the odds ratios for the given label compared to the others. This is the factor by which the probability of a given label changes for a one unit increase in the value of the feature.

The best logistic regression model uses character n-grams which are not easily interpretable. However, we do not sacrifice too much interpretability if we choose the second-best model, which has only slightly worse predictive performance (in the penultimate row of Table 1). This model uses word n-grams which are well interpretable.

Feature importance values have both a magnitude (how strongly the feature affects the prediction) and a direction (positive or negative - whether the feature increases or reduces the prediction probability). We listed the terms with largest positive values by labels in Table 2. Some terms are common abbreviations, see *ect* = electroconvulsive therapy, *re* = about, concerning (regarding, referencing). The two classifiers show a similar picture. Most bio-medical features are related to medical treatment and the human body, while most psychological features are related to psychological therapy, emotions, and the human mind. Sociological features relate to social relationships (child, family), social institutions (school, university, job, government), social exclusion (gay, abuse, lonely) and material deprivation (afford, rent, pay). In case of the DistilBERT classifier there are also word fragments among the most important terms, see *##* that denotes any string of characters. One of these terms is *##mg*, that refers to drug dosage, hence it is sensible that the term is related to the bio-medical label.

In case of logistic regression, only one of the features ('NOUN') pertains to those which are not terms but were defined by us to characterize language usage of the author. NOUN is the number of nouns used in the post, and it characterizes bio-medical framing. As we already mentioned earlier, nouns are more frequent in formal styles (e.g. Heylighen and Dewaele 1999). Although it requires further analysis, the fact that NOUN

Table 2 Interpretation of the DistilBERT and the logistic regression classifier. 20 terms with largest positive feature importance values. ##: any string of characters

Predicted label	DistilBERT	Logistic Regression
Bio-medical	medication drug psychiatrist med ##mg doctor 300 hospital medicine medicinal ##de drugs alcohol depression anti dose ##press hormone addict could	medication med doctor antidepressant drug take effect depression dose psychiatrist sleep ect body brain NOUN medicine pill treatment well week
Psychological	depression anxiety psychologist therapist feel emotion psychiatrist depressed thought get think help like mood people go make mental therapy	therapist depression thought write therapy psychologist step help negative emotion break interest person anger counsel anxiety self happiness cop mind
Sociological	job military school industry related university male government charity gay social steel worker society funding abuse paid rent public wealth	job people money friend child group lonely family pay school men support social afford shit re mental_health support_group everyone society

Table 3 Liberal kappa of the DistilBERT model when trained and validated with/without hard cases (with train/validation size in brackets)

		Validation		
		Easy	Hard	Easy + hard
Training	Easy	0.74 (1896/621)	0.34 (1896/493)	0.56 1896/1114)
	Hard	0.59 (1444/621)	0.33 (1444/493)	0.46 (1444/1114)
	Easy + hard	0.79 (3340/621)	0.37 (3340/493)	0.60 (3340/1114)

increases the probability of bio-medical framing as opposed to psychological or sociological framings may be explained by the more factual nature of this approach.

12 Automated classification experiment IV: Relationship between inter-annotator disagreement and the learning algorithm

It can help to better understand the nature of our learning task if we try to find out what information can be inferred from inter-annotator disagreement regarding the learning algorithm. In the first step we examined whether the information if the post was easy or hard to annotate helps the learner. We experimented with adding this new feature to our best bag-of-words classifier, a logistic regression model. The new model gave only slightly better performance measures than the original one (see Table 12 in the 14). The model coefficients pertaining to the new feature were 0.5092, 0.2257, -0.4200 for the bio-medical, psychological, sociological label, respectively. The picture confirms our earlier descriptive results: easiness increases the probability of having a bio-medical (and to a lesser extent, of having a psychological) label, even after controlling for every other feature and term. If we allow *easy* to interact with all the other features, performance measures (Table 13) do not change. That is, the knowledge of whether a post is easy or hard does not increase the efficiency of the predictive use of the other features.

In the second step, we examined whether the moderate performance of our learners is affected by the inter-annotator disagreement we observed. The relation is not self-evident. Interpretative process of the human annotators may work completely differently than the algorithms of the learner. Although negative impact of annotation inconsistencies on learning performance was proved for certain specific tasks (Schwartz et al. 2011; Rehbein and Ruppenhofer 2011), in principle, it is possible for our annotators and our models to find completely different cases difficult to classify.

In this experiment we aimed to answer the question whether the level of inter-annotator disagreement provides a good estimate for the objective difficulty of the machine learning task; whether the cases hard for annotators are also hard for the learner.

Our aim was to measure the effect of ‘hard cases’ (where two annotators did not agree on the primary label, 43% of the posts) which, if they are also hard for the learner, can lead to unfair performance results found in either the test or the training data. We trained the best model (DistilBERT classifier), filtering out from the training (or validation) set all hard (or easy) cases that resulted in a reduction of the sample (Table 3). It is important to note that the size of the training set obviously affects the success of classification (smaller training size—worse performance), and the size of the validation set

is related to the reliability of our performance measure (smaller validation size—greater variance in performance measure). Hence, we also presented the size of the two sets in the table. We ran nine different models by varying validation and training sets as below (Table 3).

As Table 3 shows, the system that is both trained and validated on easy cases attains the best performance. Cases hard for humans are hard for learners as well: we found a massive benefit from filtering the validation set (column ‘easy’ vs. column ‘hard’ or ‘easy + hard’). Validation sets with easy cases are easier to predict even if we train on hard cases (0.59 vs. 0.33).

The presence of hard cases in training data significantly misled the learning of easy cases (first column, 0.74 vs. 0.59)—a phenomenon called “hard case bias” by Klebanov and Beigman (2014). Negative effect of the presence of hard cases in training data can be eliminated if we mix them with easy cases (first column, 0.74 vs. 0.79).

At the same time, training (partly or totally) on easy cases did not make the prediction of hard cases worse (second column)—even though training size was substantially smaller after filtering. This finding suggests that some of the hard cases are similar to the easy ones, but our learner failed to recognize this similarity when it was trained on hard cases alone.

We repeated the same analysis for the second best (logistic regression) model, and got very similar results in tendency, which shows the robustness of our results (see Table 14 in the 14).

13 Discussion

Combining sociology and machine learning raises several methodological and epistemological questions. While our paper is written from a sociological perspective, we aim to reflect on these issues as well. Our paper illustrates what logic these methods work along, why they might be advantageous, and what their limitations are.

We trained machine learning classifiers to predict framing of depression of posts written by users of online health communities. A state-of-the-art Transformer-based model with transfer learning, DistilBERT was compared to Gradient Boosting Machine, word-embedding-based models and more conventional bag-of-words learning algorithms. Our results indicate that moderate accuracy (precision 60%, kappa 0.53) can be achieved by using a simple logistic regression classifier, but performance can be improved by using the DistilBERT classifier (precision 73%, kappa 0.60). The latter surpasses human inter-annotator agreement (precision 70%, kappa 0.58). Our finding illustrates that transfer learning models can be well utilized in social sciences. These disciplines have only limited capacities to annotate training data but transfer learning models can potentially achieve better performance even with even relatively smaller corpora.

We showed that, instead of treating labels as unambiguous inputs, it can be worthwhile to investigate more closely the annotated data and identify hard cases that show substantial disagreement between annotators. We examined what information can be inferred from the hard cases regarding the nature of the classification task. Considering machine learning performances with and without the hard cases, easy cases were less controversial not only in human annotations but also for machine learners. This finding confirms that the level of inter-annotator disagreement provides a good estimate for the objective difficulty of the classification.

According to our results the presence of hard training cases misled the learners. At the same time, training (partly or totally) on easy cases did not make the prediction of hard cases worse. We do not think our results suggest that performance may be generally improved if models are trained on easy data only, although there are studies where hard cases are removed from both training and validation data in the first step of the analysis (see Dagan et al., 2006).

According to our results, easy posts tend to follow a bio-medical framing, while hard posts are rather sociological. At the same time, easy posts show language features that suggest these posts are more punctual, more formal, while hard posts are more flexible, more informal, closer to the oral style. The latter finding may be explained on the one hand by the more factual nature of the bio-medical approach and the more subjective nature of the sociological approach. On the other hand, it may be explained by the difference of the elaboration of the bio-medical vocabulary (exact) and the everyday descriptions of social hardships (vague).

One of the main methodological contributions of this paper is in giving additional evidence of the need to pay attention to problematic cases in annotated data, an argument already put forward in the literature (e.g.: Klebanov and Beigman 2014). Most NLP annotation tasks aggregate labels into an integrated label. However, as our results show, aggregation eliminates potentially useful information from ambiguous cases. A potential future direction for research would be a more complex exploitation of annotated data, e.g. by incorporating annotation disagreement and annotator's reliability into the model.

When interpreting the models, we identified the most important terms in our classifier algorithms, which gave us insight into the meaning of the framing types. Based on the most important terms, predictive rules behind our models seem to be sensible, e.g. sociological features mostly relate to social relationships, social institutions, social exclusion and material deprivation. One direction for further investigation would be to predict labels for the whole corpus and to take a closer look at the sociological frame of interpretation. Another exciting future direction is to take into account temporality and investigate whether online depression forums have a homogenizing effect: it may be analysed if users converge to a collective framing of depression which potentially integrates newcomers as well.

Machine learning techniques can extend the opportunities of qualitative text analysis, giving a promising approach in social research. As social science concepts are generally complex in nature, our results may contribute to the discussion on the potential of NLP techniques in knowledge-driven textual analyses. We hope that our paper supports a smooth fit of the new techniques into the traditional toolbox of social sciences.

Appendix

Extracted document features

- The number of occasions any mental health medication occurs in the post. Outer databases of list of mental health drugs:

en.wikipedia.org/wiki/List_of_psychiatric_medications (retrieved: 15-04-2019) and www.mind.org.uk/information-support/drugs-and-treatments/antidepressants-a-z/overview/?o=60248#.XOPLU8gzbiU (retrieved: 15-04-2019).

- - The number of occasions some dosage ('8 mg') occurs in the post.

- - Number of sentences / words.
- - Proportion of nouns / verbs / adjectives / emojis / misspelled words / stop words / numbers / punctuation marks / commas in the post.
- - Diversity of emojis.
- - Average length of sentences / words.
- - Lexical diversity of the post: root-TTR (type-token ratio), that is, number of different tokens divided by the root square of the total number of tokens.
- - Sentiment score of the post. We used Textblob sentiment analyzer for the polarity rating (positive/negative) calculation of the posts. Loria et al. (2018).
- - Number of abbreviations in the post that are commonly used in chat forums (e.g., 2nite=tonight). Outer database used:

<https://harshad.wordpress.com/list-of-acronyms-text-messaging-shorthand>.

(retrieved: 15-04-2019)

Feature distribution within easy and hard categories

See Fig. 3

Fig. 3 Difference between feature medians for easy and hard cases. Feature distributions are normalized by standard deviations.

Performance measures for the classification models

(train/validation size: 3562/891)

Bag of Words classifiers

See Tables 4,5,6,7,8.

Table 4 Performance evaluation of the best logistic regression models. The values are averaged over folds

Document-term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
Count vectorizer	Yes	0.58	0.64	0.58	0.64	0.58	0.64	0.41	0.49
	No	0.58	0.64	0.58	0.63	0.58	0.63	0.40	0.48
TF-idf vectorizer	Yes	0.61	0.68	0.61	0.67	0.60	0.66	0.44	0.52
	No	0.60	0.67	0.61	0.67	0.59	0.65	0.44	0.51
TF-idf vect., word n-grams	Yes	0.61	0.68	0.61	0.67	0.59	0.66	0.44	0.53
TF-idf vect., char. n-grams	Yes	0.61	0.68	0.62	0.68	0.60	0.67	0.45	0.53

Table 5 Performance evaluation of the best SVM model. The values are averaged over folds

Document- term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
TF-idf vect., word n-grams	Yes	0.60	0.66	0.60	0.66	0.58	0.65	0.42	0.50

Table 6 Performance evaluation of the best Naïve-Bayes model. The values are averaged over folds

Document- term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
TF-idf vect., word n-grams	Yes	0.58	0.65	0.52	0.58	0.52	0.58	0.34	0.41

Table 7 Performance evaluation of the best Random Forest model. The values are averaged over folds

Document- term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
TF-idf vect., char. n-grams	Yes	0.61	0.69	0.62	0.69	0.60	0.67	0.46	0.54

Table 8 Performance evaluation of the best GBM model. The values are averaged over folds

Document- term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
TF-idf vect., word n-grams	Yes	0.60	0.68	0.60	0.68	0.59	0.67	0.43	0.52

Word embedding based classifiers

See Tables 9,10.

Table 9 Performance evaluation of the logistic regression model with word-embedding based features added. The values are averaged over folds

Document- term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
TF-idf vect., word n-grams	Yes	0.61	0.68	0.62	0.68	0.60	0.67	0.46	0.54

Table 10 Performance evaluation of the GBM model with word-embedding based features added. The values are averaged over folds

Document- term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
TF-idf vect., word n-grams	Yes	0.60	0.67	0.61	0.67	0.59	0.66	0.44	0.52

DistilBERT language model classifier

See Table 11.

Table 11 Performance evaluation of the best DistilBERT model

Precision		Recall		F1-score		Kappa	
Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
0.68	0.73	0.68	0.73	0.66	0.71	0.53	0.60

Logistic regression and easy/hard posts

See Tables 12,13,14.

Table 12 Performance evaluation of the best logistic regression model expanded with a feature indicating easy/hard posts. The values are averaged over folds

Document- term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
TF-idf vect., char. n-grams	Yes	0.61	0.69	0.62	0.69	0.60	0.67	0.46	0.54

Table 13 Performance evaluation of the best logistic regression model expanded with feature *easy* and its interactions with the other features. The values are averaged over folds

Document- term matrix vec.	Features used?	Precision		Recall		F1-score		Kappa	
		Cons.	Lib.	Cons.	Lib.	Cons.	Lib.	Cons.	Lib.
TF-idf vect., word n-grams	Yes	0.61	0.69	0.62	0.69	0.61	0.68	0.46	0.54

Table 14 Liberal kappa of logistic regression models when trained and validated with/ without hard cases (with train/ validation size in brackets), *: cross-validated, averaged over folds

		Validation		
		Easy	Hard	Easy+hard
Training	Easy	0.71* (2013/505)	0.29 (2517/1937)	0.53* (2013/891)
	Hard	0.45 (1937/2517)	0.30* (1549/387)	0.38* (1549/891)
	Easy + hard	0.69* (3562/505)	0.31* (3562/387)	0.53* (3562/891)

Acknowledgements This research has been supported by the Higher Education Excellence Program of the Ministry of Human Capacities, Hungary at Eötvös Loránd University (ELTE–FKIP). The authors thank Gergő Morvay for suggestions regarding the organization of the annotation process. We would like to thank also to the students who participated in the annotation.

Funding Open access funding provided by Eötvös Loránd University. This research has been supported by the Higher Education Excellence Program of the Ministry of Human Capacities, Hungary at Eötvös Loránd University (ELTE–FKIP).

Data Availability We used SentiOne, a web-based social listening and text analytics platform. We chose forums which are public and accessible without registration, in order to follow data protection and ethical regulations. The data set was collected by SentiOne in compliance with GDPR regulations. We obtained only publicly available posts, which are shared willingly by their authors.

Code Availability During the preprocessing, we used Python's NLTK package (Bird et al. 2009). During the analysis, we used the transformers library (PyTorch implementation) from Hugging Face (huggingface.co) (Wolf et al. 2019), Python's Scikit-learn library (Pedregosa et al. 2011) and gensim library (Řehůřek and Sojka 2021). All the source code developed for the classification experiments have been released at <https://github.com/anonymuspixel/depression>.

Conflicts of interest The authors declare that they have no conflict of interest. The authors declare that they have no competing financial interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aggarwal, C.C., Zhai, C.: Mining Text Data. Springer, Boston (2012)
- Ben-David, A.: Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert. Syst. Appl.* **34**(2), 825–832 (2008)
- Bernardi, L., Mavridis, T., Estevez, P.: 150 Successful machine learning models: 6 lessons learned at booking.com. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1743–1751. Association for Computing Machinery, New York: (2019)
- Bird, S., Edward, L., Ewan, K.: Natural language processing with Python. O'Reilly Media Inc., Sebastopol (2009)
- Chen, N.C., Drouhard, M., Kocielnik, R., Suh, J., Aragon, C.R.: Using machine learning to support qualitative coding in social science: shifting the focus to ambiguity. *ACM Trans. Interact. Intell. Syst.* **9**(4), 39 (2018)
- Cohen, J.: A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960)
- Comer, R.J.: Abnormal psychology. Worth Publishers, New York (2015)
- Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognizing textual entailment challenge. In: Quiñero-Candela, J., Dagan, I., Magnini, B., d'Alch-Buc, F. (eds.) *MLCW 2005*, LNAI 3944, 17–190. Springer-Verlag, Berlin, Heidelberg (2006)
- Ehrenberg, A.: The weariness of the self: diagnosing the history of depression in the contemporary age. McGill-Queen's University Press, Montreal (2009)
- Eisenstein, J.: Natural language processing. MIT Press, Georgia Tech (2019)
- Flick, S.: Treating social suffering? Work-related suffering and its psychotherapeutic re/interpretation. *Distinktion: J. Soc. Theory.* **17**(2), 149–173 (2016)
- Flor, M., Yoon, S.Y., Hao, J., Liu, L., von Davier, A.: Automated classification of collaborative problem solving interactions in simulated science tasks. In: Tetreault, J., Burstein, J., Leacock, C., Yannakoudakis, H. (eds.) Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 31–41. Association for Computational Linguistics, San Diego: (2016)
- Gius, E., Jacke, J.: The hermeneutic profit of annotation: on preventing and fostering disagreement in literary analysis. *Int. J. Humanit. Arts Comput.* **11**(2), 233–254 (2017)
- Hasan, M., Agu, E., Rundensteiner, E.: Using hashtags as labels for supervised learning of emotions in twitter messages. *ACM SIGKDD Workshop on Health Informatics*, New York. <http://web.cs.wpi.edu/~emmanuel/publications/PDFs/C25.pdf> (2014). Accessed March 30 2020
- Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Second Edition. Springer Science & Business Media: (2009)
- Heylighen, F., Dewaele, J.-M.: Formality of language: definition, measurement and behavioral determinants. Internal Report. Center Leo Apostel, Free University of Brussels (1999)
- Klebanov, B.B., Beigman, E.: Difficult cases: from data to learning, and back. In: proceedings of the 52nd annual meeting of the association for computational linguistics (Short Papers). pp. 390–396. Curran. Red Hook, Baltimore: (2014)
- Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: proceedings of the 14th international joint conference on artificial intelligence (II) (IJCAI '95). pp. 1137–1143. Morgan Kaufmann Publishers Inc., San Francisco: (1995)
- Krishnan, G.S., Kamath, S.S.: Hybrid text feature modeling for disease group prediction using unstructured physician notes. In: Krzhizhanovskaya, V., et al. (eds.) *Computational science – ICCS 2020*, vol. 12140. Springer, Cham (2020)
- Lachmar, E.M., Wittenborn, A.K., Bogen, K.W., McCauley, H.L.: #My depression looks like: examining public discourse about depression on twitter. *JMIR Ment. Health.*, **4**(4), e43 (2017)
- Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174 (1977)
- Lombard, M., Snyder-Duch, J., Bracken, C.C.: Content analysis in mass communication: assessment and reporting of intercoder reliability. *Hum. Commun. Res.* **28**, 587–604 (2002)
- Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., Dempsey, E.: Textblob: simplified text processing. <https://textblob.readthedocs.org/en/dev/> Accessed 26 March 2018
- Lundberg, S., and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30: (NIPS 2017). arXiv:1705.07874 (2017)
- Maxwell, J.A.: Qualitative research design: an interactive approach, 2nd edn. Sage Publications, Thousand Oaks (2005)
- Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

- Molnar, C.: Interpretable machine learning: a guide for making black box models explainable (eBook). Lulu, Morrisville (2019)
- Németh, R., Sik, D., Máté, F.: Machine learning of concepts hard even for humans: the case of online depression forums. *International J. Qualitative Methods* **19**, 1–8 (2020)
- Németh, R., Sik, D., Katona, E.: The asymmetries of the biopsychosocial model of depression in lay discourses - Topic modelling online depression forums. *SSM Popul. Health* 14 Paper: 100785: (2021)
- Pan, J., Liu, B., Kreps, G.L.: A content analysis of depression-related discourses on Sina Weibo: attribution, efficacy, and information sources. *BMC Public Health*. **18**, 772 (2018)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
- Powers, D.M.W.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2**(1), 37–63 (2011)
- Rehbein, I., Ruppenhofer, J.: Evaluating the impact of coder errors on active learning. In: proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies - Volume 1, HLT '11, pp 43–51. Association for Computational Linguistics, Stroudsburg, PA, USA. (2011)
- Řehůřek, R., Sojka, P.: Gensim—statistical semantics in python. pypi.org/project/gensim (2021). Accessed 20 February 2021
- Rosenberg, A., Binkowski, E.: Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In: proceedings of HLT-NAACL 2004: short papers. pp 77–80. Association for Computational Linguistics, Boston: (2004)
- Sanh, V., Debut, L., Chaumond, J., Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*: (2020)
- Scholz, B., Crabb, S., Wittert, G.A.: “Males don’t wanna bring anything up to their doctor”: men’s discourses of depression. *Qual. Health Res.* **27**(5), 727–737 (2017)
- Schwartz, R., Abend, O., Reichart, R., Rappoport, A.: Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In: proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies - Volume 1, HLT '11, pp 663–672. Association for Computational Linguistics, Stroudsburg, PA, USA. (2011)
- Sik, D.: From mental disorders to social suffering. *Europ. J. Soc. Theory.* **19**(4), 556–573 (2018)
- Sik, D., Németh, R., Katona, E.: Topic modelling online depression forums: beyond narratives of self-objectification and self-blaming. *J. Ment. Health* 30 p. (2021)
- Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. 2008 IEEE computer society conference on computer vision and pattern recognition workshops. pp 1–8. Anchorage, AK, USA: (2008)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *arXiv:1706.03762* (2017)
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., ... Rush, A.M.: HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*: (2020)