Bio, psycho or social - Discursive framing of depression in online health communities

Renáta Németh, Domonkos Sik, Fanni Máté, Eszter Katona
Eötvös Loránd University of Budapest, Faculty of Social Sciences
Research Center for Computational Social Science (rc2s2.eu)

We present initial experiments on understanding individual framing of depression in online health communities. Three discursive frameworks were introduced: the bio-medical, psychological, and social framing of depression. We applied different supervised learning algorithms for classification. The main challenge was that framing is a hermeneutically difficult concept, which affects both inter-annotator agreement and predictive performance. As we had a complex annotation design, we had to introduce a new augmented measure of inter-annotator agreement and prediction performance. Our results show that social framing cannot be effectively predicted, therefore, we presume that social discourse — being inferior to the others — is present only in an implicit way. We identified hard-to-annotate cases and proved that their presence misleads the learner, while training on easy data only might improve performance.

Motivation

Depression is a disease of modernity. As a form of social suffering, it refers to social relations dominated by uncertainty. **As a social construct,** framing defines the meaning of depression, gives its causal explanation, and can even determine treatment preferences. A current question in sociology is how mental disorders are framed by the patient. **Previous research in this field has been primarily qualitative.**

Data collection

We collected depression-related posts from the most popular English-speaking health forums. We filtered the corpus in two rounds: (1) we selected threads that contained the word "*depression" or "depressed"* in the title or in at least one post, then (2) we selected posts whose link, topic, or content contained a depression-related term, for instance: *"unipolar depression", "mood disorder", or "depressant".* The dataset, collected by SentiOne, contained 79,889 posts from February 2016 to February 2019. They are publicly available posts, which were shared willingly by their authors.

A difficult annotation task

The training set of 4,500 posts was selected randomly. The main challenge of the annotation was the hermeneutic interpretation of the posts. Framing is an unconscious process; a simple referential reading is not enough. **A detailed classification guideline** was prepared, which was recursively updated. The interpretation was practiced in several collective and individual turns, until the annotators developed a unified approach.

This non-trivial annotation task implied that (1) annotators were instructed to assign **a secondary label to the texts, if needed** (20% of the posts got a second label). Additionally, (2) we had **two independent annotators** for each text **with a gold standard annotator** (a senior researcher) in case of disagreement between the annotators (12.3% of the posts were concerned). The final, integrated label was based on majority voting (with a secondary label, if needed).

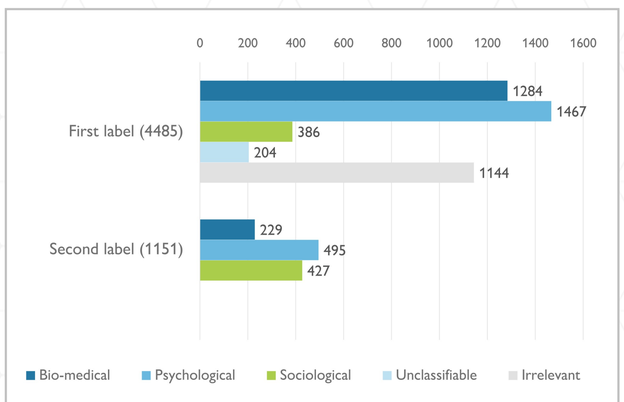


Figure 1. Distribution of the labels in the annotated corpus. Irrelevant = it is not about depression, unclassifiable = framing cannot be identified.

I feel like medication and proper treatment has given me a second chance at living it more fully.

You have to sacrifice your ego. You have to stop needing to be right about yourself. You have to stop needing to be the victim of depression.

We aren't; and this world isn't perfect. We may be led to believe that by the culture we live in, the TV ads, Hollywood, beauty magazines, etc., but what that does is allow our minds to make comparisons to convince us that we don't add up.

Augmenting measure of inter-annotator agreement to multiple labels

If labeling is not reliable, the analysis cannot be trusted. Simple percent agreement does not take those agreements into account which occur by chance. We used Cohen's kappa, which tells us how far the observed agreement is better than the agreement expected by chance.

As we had **secondary labels, an augmentation to kappa** was necessary. If we define agreement as the match of the two primary labels, simply discarding the optional secondary ones, we got an overly conservative measure. We introduced a liberal criterion as well, when the match of either of the secondary and either of the primary labels defines agreement, too. As Table (1) shows, we had a nearly substantial agreement.

	Conservative measure	Liberal measure
observed agreement	58.3%	69.7%
agreement expected by chance	27.9%	28.3%
kappa	0.42	0.58

Table 1. Inter-annotator agreement. For observed agreement, the criterion of 0.7 is often used for exploratory research. A kappa of 0.4–0.6 indicates moderate agreement, 0.6–0.8 substantial agreement.

Preprocessing

- deletion of repost part of the text
- removal of duplicate and too short (<20 words) posts
- deletion of URLs, e-mail addresses
- identification of name of mental disorders (n-grams)
- lemmatization (WordNet lemmatizer from Python NLTK)
- significant bigram detection (BigramCollocationFinder, NLTK, with PMI measure and human evaluation)
- stop-word removal (NLTK Stopwords Corpus)
- significant trigrams and named entity recognition for persons' name turned out to be non-relevant.

Our final, preprocessed corpus contains **67,857** posts.

Some words with the largest TF-IDF values from the corpus:
`smile, chronic, grief, kratom`[1]`, pessimistic, kai`[2]`, zopliclone`[3]`, schizoaffective_disorder, eft`[4]

1: leaves of a tropical tree, used to self-treat depression; 2: South Korean singer, opened up about his depression.; 3: sleeping pill; 4: emotional freedom therapy

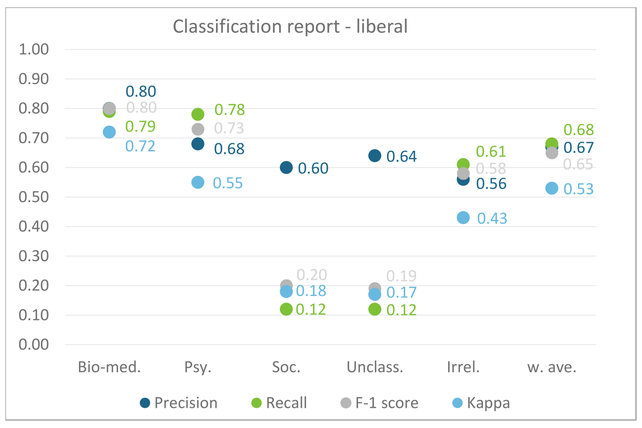
Experiment I: multinomial logistic regression

We trained a multinomial logistic regression classifier on the (final) primary labels. We used several configurations: (1) unigrams, bigrams, and trigrams, ignoring terms that have a document frequency lower than 5, as well as (2) character n-grams (n = 2 to 5), which should make the model more robust to spelling errors.

As we face a **(1) multi-class classification problem with (2) highly unbalanced classes, standard performance measures like Recall or F-score are not adequate,** while **Cohen's kappa can handle** both problems. Like in the case of inter-annotator agreement, we considered a liberal version of the kappa accepting a predictive label if it agrees with either the primary or the secondary label. Performance was assessed by stratified 5-fold cross-validation.

Document-term matrix vectorization	Features	Conservative kappa	Liberal kappa
Count-vectorized	Without feat.	0.38	0.46
	With features	0.39	0.47
TF-IDF vectorized	Without feat.	0.43	0.51
	With features	0.44	0.52
TF-IDF using word n-grams (n=1 to 3)	Without feat.	0.41	0.50
	With features	0.44	0.53

The best result (Table 2) gives a liberal kappa of 0.53 (a moderate agreement) and uses TF-IDF-vectorization. It outperforms the baseline by a moderate margin. Features and word n-grams do not improve performance, character n-grams have only a slight effect. Stop word retention and aggregation of irrelevant and unclassifiable labels (4 classes in the outcome) neither changes the results. The regularization parameter, C was also tested, a value of 1 turned out to be the best.



A detailed classification report gives a deeper insight into the performance of the best model (Figure 2). Bio-medical framing is the most predictable, while sociological framing is the least predictable one — especially its recall is very low.

Figure 2. Standard performance measures by labels (averaged over folds)

Some of the most important words from the best model using word n-grams and 4 classes in the outcome:
Bio-medical: `medication, antidepressant, drug, take, effect, brain, week`
Psychological: `therapist, depression, write, help, person, anxiety, self, mind`
Sociological: `job, people, money, child, group, family, pay, men, school, shit, society`

Experiment II: other classifiers

We also experimented other Scikit Learn classifiers, with TF-IDF-vectorized document-term matrix using word n-grams (n=1 to 3) and features, with stop word removal and four categories in the outcome. Liberal version of kappa (averaged over folds) was
0.41 for Bernoulli Naïve-Bayes (alpha=1)
0.50 for SVM (grid search for C, gamma and kernel) and
0.48 for Random Forest (grid search for number of trees and maximum number of features).
That is, **logistic regression remarkably outperforms the other classifiers**.

Experiment III: Do hard cases mislead the machine learner?

Our aim was to measure the effect of "hard cases" (items with contradictory primary labels, 43% of the posts), which, if they are also hard for the learner, can lead to unfair performance results either found in test or training data. We trained the best logistic regression model by filtering the training/test data (Table 3).
We found a massive benefit from test filtering, so cases hard for humans are also hard for learners. The presence of hard cases in training data misleads the learning of easy cases. However, training (partly or totally) on easy cases does not make prediction of hard cases worse.

Training	Test	Kappa (Liberal)	Train/test size
Easy	Easy	0.71*	2013/505
Easy	Hard	0.29	2517/1937
Easy	Easy+Hard	0.53*	2013/891
Hard	Easy	0.45	1937/2517
Hard	Hard	0.30*	1549/387
Hard	Easy+Hard	0.38*	1549/891
Easy+Hard	Easy	0.69*	3562/505
Easy+Hard	Hard	0.31*	3562/387
Easy+Hard	Easy+Hard	0.53*	3562/891

Table 3. Liberal kappa when trained and tested with/without hard cases, *: crossvalidated, averaged over folds

Conclusions

Our results indicate that **considerable accuracy** (precision 60%, kappa 0.53) can be achieved by using a simple logistic regression classifier. This result approaches human interrater agreement (70%, 0.58).
Bio-medical and psychological framing could be effectively predicted, while social framing is less treatable. We suppose that being inferior to other discourses, social discourse is present in a more latent way, without recognized semantic entities.
We empirically identified hard cases and showed that the **presence of hard training cases misleads the learner,** which suggests that performance might be **improved if** model is **trained on only easy data.**
As most sociological concepts, similarly to framing, are complex in nature, our results contribute to the discussion on the potential of NLP techniques in knowledge-driven textual analysis.

This research was supported by the Higher Education Excellence Program of the Ministry of Human Capacities at Eötvös Loránd University (ELTE-FKIP).