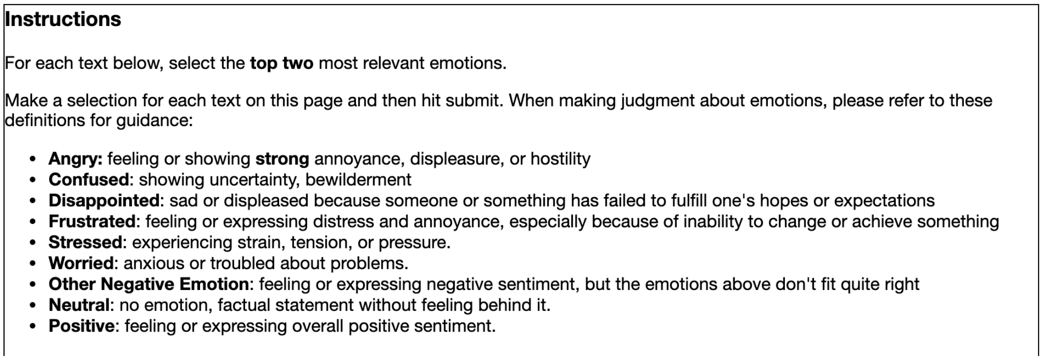
Supplementary Materials

A Dataset Details

Karenina is available at `https://github.com/sudhof/karenina` in JSONL format, and is released under a Creative Commons 4.0 International license.[3] Appendix F provides a Datasheet [10] for it. The authors bear all responsibility in case of violation of rights.

B Labeling Task



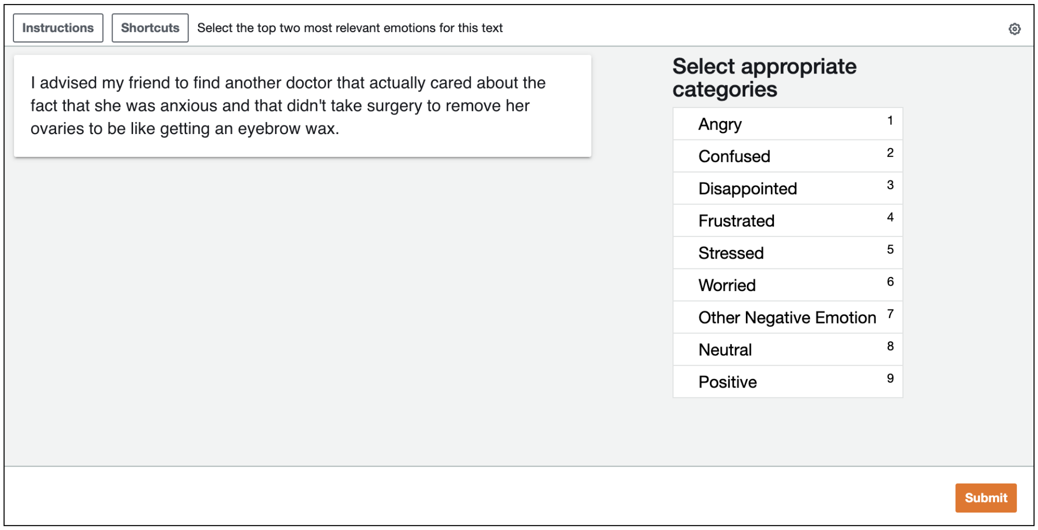
Instructions

For each text below, select the **top two** most relevant emotions.

Make a selection for each text on this page and then hit submit. When making judgment about emotions, please refer to these definitions for guidance:

- **Angry:** feeling or showing **strong** annoyance, displeasure, or hostility
- **Confused**: showing uncertainty, bewilderment
- **Disappointed**: sad or displeased because someone or something has failed to fulfill one's hopes or expectations
- **Frustrated**: feeling or expressing distress and annoyance, especially because of inability to change or achieve something
- **Stressed**: experiencing strain, tension, or pressure.
- **Worried**: anxious or troubled about problems.
- **Other Negative Emotion**: feeling or expressing negative sentiment, but the emotions above don't fit quite right
- **Neutral**: no emotion, factual statement without feeling behind it.
- **Positive**: feeling or expressing overall positive sentiment.

(a) Instructions provided to raters.



Select the top two most relevant emotions for this text

I advised my friend to find another doctor that actually cared about the fact that she was anxious and that didn't take surgery to remove her ovaries to be like getting an eyebrow wax.

(b) Sample annotation interface for a single example.

Figure 4: Mechanical Turk annotation interface.

B.1 Task Set-up

Figure 4 shows the interface for the annotation task used. Each Human Interface Task (HIT) included instructions that defined each emotion as well as ten sentences to be annotated. Workers were paid US$0.50 per HIT, and all workers were paid for all their work, regardless of whether we retained their labels. Examples were uploaded to Amazon's Mechanical Turk in batches of around 5K examples.

---

[3] `https://creativecommons.org/licenses/by/4.0/`

Table 4: Model learning rates selected by hyperparameter search.

Model	Learning Rate
Binary(Angry)	$9e{-}6$
Binary(Confused)	$9e{-}6$
Binary(Disappointed)	$9e{-}6$
Binary(Frustrated)	$1e{-}5$
Binary(Neutral)	$1e{-}5$
Binary(Positive)	$9e{-}6$
Binary(Stressed)	$1e{-}5$
Binary(Worried)	$2e{-}5$
Multi-class classifier	$1e{-}5$

B.2 Exclusion Criteria

We filtered the annotations according to a number of criteria. First, all annotators that uniformly answered the same thing for each example were filtered out. Second, any annotator that spent less than 100 seconds on a batch of examples was filtered out. Third, we labeled 1000 example texts in-house and filtered out any annotator that got at least two of the gold labeled positive examples wrong. Where this filtering led to an example having too few annotations, we had it relabeled by a pool of annotators that had previously done high quality work on our task.

To remove workers from our pool, we used a method of ‘unqualifying’, as described in [23]. This method does no reputational damage to workers but allows us to disqualify them from participating in future rounds. This iterative approach allowed us to gather annotations from a broad set of workers while continuously monitoring and guiding annotation quality. We encouraged a broad range of interpretations and participation in the annotation task by having at least 9 unique workers annotate each sample text, and, even after quality filtering, the vast majority of our examples have at least 6 annotators. While we think our method mainly increased label quality, we recognize that it can introduce unwanted biases, and we provide further commentary on the dataset’s potential biases in 6. The estimated hourly wage paid to participants was US$5.25 and the total amount spent on participant compensation was $26,020.20.

C Model Details

Our Random models are implemented using scikit-learn’s DummyClassifier with the ‘stratified’ guessing strategy [18].

All of our non-Random models begin with `bert-base-uncased` parameters from the HuggingFace library [27] with a dense layer added on top for the purposes of fine-tuning for classification. We explored learning rates in the range $1e{-}4$ to $1e{-}5$. Throughout, we use the Adam optimizer, a batch size of 16, and a maximum sequence length of 32. We used the early stopping function from keras on the dev set loss on each epoch with a patience of 4 to decide when to stop training in the hyper parameter tuning process. We report the models that had the best macro F1 average scores.

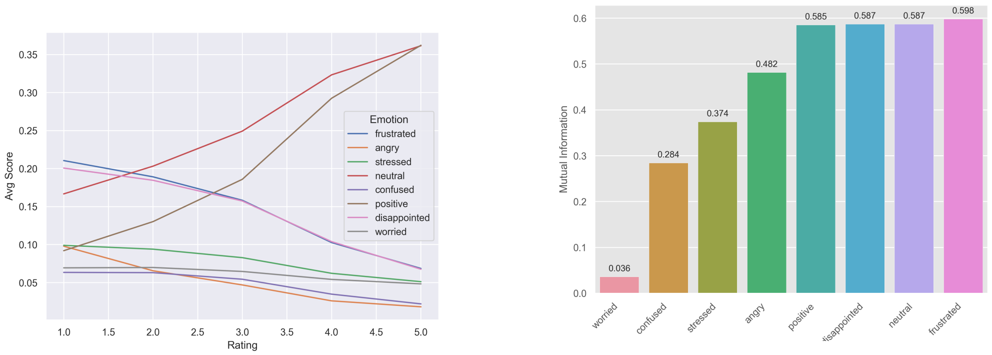
For our binary classifiers, we conducted hyperparameter tuning on a restricted hyperparameter space based on our results for the multi-class classifier. For the binary models, we only considered learning rates in the range $5e{-}5$ to $1e{-}5$.

The final learning rates used for the reported models can be found in Table 4.

All models were trained in Google Cloud’s AI notebooks environment with access to 4 CPUs, a total of 15 GB of RAM, and 1 NVIDIA Tesla T4 GPU.

D Additional Connections Between Emotions and Review Ratings

Section 5.1 briefly studied the relationship between review-level ratings and emotions. Figure 5b shows the full set of relationships, and figure 5b provides all of the mutual information values.



(a) Average emotion scores by review rating.

(b) Mutual information between each emotion and review ratings.

Figure 5: Relationship between emotions and review ratings.

Table 5: Randomly selected “Worried” sentences from 5-star reviews.

Sentence	Rating	Predicted Emotion
I was obviously nervous about getting eye surgery.	5	Worried
All these one star reviews had me worried, but not as worried as I was about getting my husband some help ASAP.	5	Worried
My PSA showed abnormalities.	5	Worried
I had huge fears regarding how bad I had let my teeth go over the span of 10 years.	5	Worried
Years after a very high risk first pregnancy (I had severe pre-eclampsia and diabetes) I wasn’t even sure I wanted to get pregnant again.	5	Worried

Worried is the only emotion that is only weakly related to the ratings, which we trace to the complexity of this emotion in healthcare contexts. Table 5 provides randomly sampled examples to show that the Worried emotion is extremely diverse in terms of its causes, which helps explain why it has relatively little correlation with overall star ratings as compared to the other categories.

We also conducted an experiment where we fit a linear regression model to predict ratings based on emotion scores. We train on 80% of the overall reviews and test on the remaining 20%. This yields R-squared values of 0.809 on the training set and 0.810 for the test set. Our ability to predict ratings based on predicted emotion scores demonstrates that our model is capable of capturing distinctions in experience feedback that meaningfully predict a customer’s overall evaluation of the experience.

E Keyword-Analysis of Care-Related Terms

Section 5.2 reported on a keyword-based analysis focusing on non-care-related terms. Figure 6 provides a comparable analysis for terms that relate more directly to care. We see that customers are generally favorable of the frontline service they receive (see results for “staff” and “dr”), but mentions of the “receptionist” contain comparatively high levels of anger and confusion, suggesting that additional training or staffing of the front desk would improve overall experience.

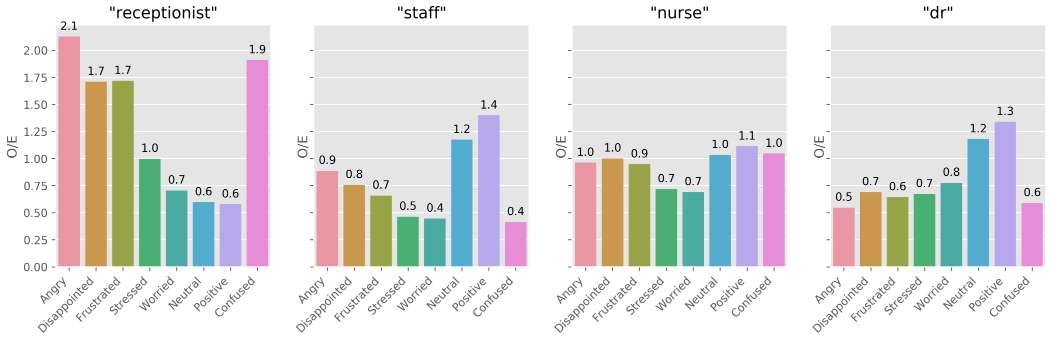


Figure 6: Observed over expected profiles for sample care-related key terms.

F Datasheet

F.1 Motivation

F.1.1 For what purpose was the dataset created?

Karenina was created to facilitate the development of emotion analysis systems that can accurately capture a range of negative emotions that are indicative of particular customer experiences.

F.1.2 Who created the dataset and on behalf of which entity?

The dataset was created by Moritz Sudhof (Motive Software), Liam Croteau (Motive Software), and Christopher Potts (Stanford University). All members of the team were functioning as independent researchers within their organizations.

F.1.3 Who funded the creation of the dataset?

The effort was funded by Motive Software.

F.2 Composition

F.2.1 What do the instances that comprise the dataset represent?

The instances are English-language sentences with labels and additional metadata. These sentences are records of acts of linguistic communication involving product and service evaluations in the space of consumer healthcare.

F.2.2 How many instances are there in total?

Version 1 (the current version) has 25,817 instances.

F.2.3 Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset contains a sample of sentences from the Yelp Academic Dataset, as described in Section 3.1 of the paper.

F.2.4 What data does each instance consist of?

The dataset is released in the JSON lines (JSONL) format. The README.md file in our project Github repository documents the dataset format.

F.2.5 Is there a label or target associated with each instance?

Yes, examples in Karenina have 1–4 labels drawn from the set ‘Confused’, ‘Disappointed’, ‘Frustrated’, ‘Angry’, ‘Stressed’, ‘Worried’, ‘Neutral’, ‘Positive’, and ‘Other Negative Emotion’. In addition, each example includes the full response distribution from our crowdsourcing effort, with anonymized worker ids. Each example in the dataset also has a number of metadata fields that could be used as labels as well. See our project README.md for a complete description.

F.2.6 Is any information missing from individual instances?

We have included all relevant information from our crowdsourcing effort, and we provide links into the Yelp Academic Corpus, which contains additional metadata.

F.2.7 Are relationships between individual instances made explicit?

Yes.

F.2.8 Are there recommended data splits (e.g., training, development/validation, testing)?

Yes, we have defined a train/dev/test split, using a procedure described in Section 3.7 of our paper.

F.2.9 Are there any errors, sources of noise, or redundancies in the dataset?

There are likely to be errors stemming from the fact that the corpus is a naturalistic one that was processed in heuristic ways and labeled using a large-scale crowdsourcing effort. As we discover errors, we will update the dataset and associated documentation.

F.2.10 Is the dataset self-contained, or does it link to or otherwise rely on external resources?

The dataset is self-contained, but it can be linked with the separate Yelp Academic Dataset for additional context. Yelp controls the Yelp Academic Dataset and could stop distributing it at any time.

F.3 Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

We are distributing the dataset with a Creative Commons 4.0 International license.[4]

There are no fees associated with using our dataset. We impose no restrictions on its usage ourselves, but we are also not in a position to adjudicate the question of whether it inherits the terms of the Yelp Academic Dataset[5] in virtue of using text snippets from that resources.

F.3.1 Does the dataset contain data that might be considered confidential?

No. All the data in Karenina is already public data taken from the Yelp site and included by Yelp in its Yelp Academic Dataset.

F.3.2 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

It is possible that some texts included in our dataset would be regarded as offensive. The texts are derived from online reviews, which can be very negative along many dimensions. We have not tried to identify or remove sentences that might cause offense.

F.3.3 Does the dataset relate to people?

Yes, it contains records of communicative acts by people, and often about people.

---

[4] `https://creativecommons.org/licenses/by/4.0/`

[5] `https://s3-media3.fl.yelpcdn.com/assets/srv0/engineering_pages/bea5c1e92bf3/assets/vendor/yelp-dataset-agreement.pdf`

F.3.4 Does the dataset identify any subpopulations (e.g., by age, gender)?

It does not do this in any way that we are aware of, and it is not out intention to identify specific populations.

F.3.5 Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

Yes, there is a path to such identification, since the Yelp Academic Dataset includes identifying information about its business and users, and our dataset links into that dataset.

F.3.6 Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

We believe the answer to be “No”, except insofar as the Yelp Academic Dataset might itself contain such information.

F.4 Collection Process

F.4.1 How was the data associated with each instance acquired?

The example texts were extracted from the Yelp Academic Dataset according to the procedure described in Section 3.1, and they were labeled according to the procedure described in Section 3.3 and Appendix B.

F.4.2 What mechanisms or procedures were used to collect the data?

All the information in the corpus was collected using Web applications.

F.4.3 If the dataset is a sample from a larger set, what was the sampling strategy?

Our sampling strategy is described in Section 3.1 of our paper.

F.4.4 Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection was done through Amazon Mechanical Turk. The methods, including payment information, are documented in Section 3.3 and Appendix B of our paper.

F.4.5 Over what timeframe was the data collected?

March 2021 to June 2021. The texts in the Yelp Academic Dataset are from the period October 2004 to December 2019.

F.4.6 Were any ethical review processes conducted?

No. The annotation process was done using tools and techniques developed in-house by Motive Software.

F.4.7 Did you collect the data from individuals directly, or obtain it via third parties or other sources (e.g., websites)?

The texts were extracted from the Yelp Academic Dataset, and the labels were assigned by individual crowdworkers.

F.4.8 Were the individuals in question notified about the data collection?

Yes.

F.4.9 Did the individuals in question consent to the collection and use of their data?

We assume that participation in our tasks constitutes consent.

F.4.10 If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

N/A, though workers were free to contact us to request that we not use their input. We have not received such requests. All our labels are assigned anonymously, and no information about crowdworkers' identities is included in our dataset.

F.4.11 Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

We do not consider our dataset to belong to the sort of high-risk category that would require such an analysis.

F.5 Preprocessing/cleaning/labeling

F.5.1 Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Our preprocessing steps are described in Section 3.1 of our paper.

F.5.2 Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw data in our case is the publicly available Yelp Academic Dataset.

F.5.3 Is the software used to preprocess/clean/label the instances available?

We use the NLTK sentence tokenizer [4]. We also use pretrained classifier models heuristically to help with sampling. These models are not publicly available, since they belong to Motive Software.

F.6 Uses

F.6.1 Has the dataset been used for any tasks already?

As of this writing, it has been used only for the experiments in our paper.

F.6.2 Is there a repository that links to any or all papers or systems that use the dataset?

No.

F.6.3 What (other) tasks could the dataset be used for?

We are not aware of tasks outside of emotion analysis that the dataset could be used for, though it's possible that it will find uses in tasks that involve emotion analysis as a component or that would benefit from emotion labels.

F.6.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Yes. The dataset is focused on consumer healthcare experiences as reported publicly on the Yelp website. This will delimit its range of useful and responsible applications.

F.6.5 Are there tasks for which the dataset should not be used?

We feel that the dataset should be used only for emotion analysis of publicly reported consumer healthcare experiences. Applications beyond that are speculative and should be explored cautiously.

F.7 Distribution

F.7.1 Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes, the dataset is publicly available at `https://github.com/sudhof/karenina`.

F.7.2 How will the dataset will be distributed?

Via `https://github.com/sudhof/karenina`.

F.7.3 When will the dataset be distributed?

It is presently available.

F.7.4 Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

We have released the dataset under a Creative Commons Attribution 4.0 International License.[6]

F.7.5 Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

F.7.6 Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

F.8 Maintenance

F.8.1 Who is supporting/hosting/maintaining the dataset?

The creators of the dataset are supporting and maintaining it.

F.8.2 How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Sudhof can be contacted on Github or by email.

F.8.3 Is there an erratum?

Not at present.

F.8.4 Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes.

F.8.5 If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

We are not aware of such limits.

F.8.6 Will older versions of the dataset continue to be supported/hosted/maintained?

Yes, they will be version-controlled. The only exception is that any instances we are required to remove will be removed from archived versions as well.

---

[6] `https://creativecommons.org/licenses/by/4.0/`

F.8.7 If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Yes. People can contact Sudhof via Github or by email.