ReAct: A *Re*view Comment Dataset for *Act*ionability (and more)

Gautam Choudhary[1], Natwar Modani[1], and Nitish Maurya[2]

[1] Adobe Research
[2] Adobe System
{gautamc, nmodani, nmaurya}@adobe.com

Abstract. Review comments play an important role in the evolution of documents. Before a document is agreed upon by the various stakeholders, there are usually multiple rounds of reviews, wherein the stakeholders provide their feedback through review comments. For a large document, the number of review comments may become large, making it difficult for both the authors and the other stakeholders to quickly grasp what the comments are about. It is important to identify the nature of the comments to identify which comments require some action on the part of document authors, along with identifying the types of these comments. In this paper, we introduce an annotated review comment dataset. The review comments are sourced from OpenReview site. After some preprocessing and data preparation, we conducted a crowd-sourcing task for getting these comments annotated. We analyze the properties of the dataset and validate the quality of annotations. The full dataset is publicly available at `https://github.com/gtmdotme/ReAct`. We also benchmark our data with standard baselines for classification tasks and analyze their performance.

Keywords: review dataset, actionability, taxonomy, text classification

1 Introduction

Review comments play an important role in the evolution of documents. Academic publications routinely go through a peer-review process, where the reviewers provide both their opinion about the suitability of the articles for the publication venue and also feedback to the authors for potentially improving the contributed article. Further, several publication venues are providing the authors a chance to respond to the review comments (for example, ACL, NAACL, EMNLP, etc., in addition to most journals). Therefore, it is important for the authors to be able to quickly digest the review comments so that they can address the concerns of the reviewers and clarify certain points which may not have been communicated adequately by the article itself.

Similarly, in business environments, documents play a key role in developing shared understanding between stakeholders, as well as plans for execution. Before a document is agreed upon by the various stakeholders, there are usually

multiple rounds of review of such documents, and the stakeholders provide their feedback through review comments. The review comments may be communicated independent of the document (say, in an email), or maybe sent as part of the document itself (e.g., as a sticky note). Often, the number of stakeholders is large, and for a large document, the number of review comments may become large. Therefore, it is useful not only for the authors of the document but also for other stakeholders, to be able to quickly grasp what the comments are about. While the dataset we are proposing is from an academic publication review domain, from our own experience and small internal informal survey within our organization, the categorization of business document review comments into similar categories is also appropriate.

In this work, we focus on two aspects of *understanding* the review comments. First, determining if a review comment requires some action on the part of document authors. This motivates the need for the classification of review comments based on 'Actionability'. Second, what type of review comment it is. Here, we take the types of review comments as *Agreement, Disagreement, Question, Suggestion, Shortcoming, Statement of Fact*, and *Others*, similar to (but not exactly the same as) [19]. We provide the reason for our choice of these specific types and their justification in Section 3.2. While there are some prior works on user comment classification, there are two gaps in those prior works that we address in this paper. First, as far as we know, there are no classification systems proposed for *document review comments*. Second, there is no publicly available dataset for the task.

The review comments are sourced from the OpenReview [12] site. After some preprocessing and data preparation, we conducted a crowd-sourcing task for getting these comments annotated. We analyze the properties of the dataset and open source it to the research community. Also, we present some baseline systems for the task and analyze their performance.

Our key contributions in this paper are the following:

- A review comment dataset consisting of $1,250$ labeled comments for identifying actionability and their types. We also have $\sim 52k+$ unlabelled (but otherwise processed) comments in this dataset for future extensions and/or use of semi-supervised approaches. [3]
- A taxonomy for types of review comments.
- Establishing strong baselines for the proposed dataset.

The rest of the paper is organized as follows: relevant background and prior works are discussed in Section 2. Section 3 describes the proposed dataset, starting from raw data collection to survey design to final preprocessed data. Benchmarking results are discussed in Section 4 along with experimental setup, followed by comparisons of our dataset with other related datasets. Finally, we conclude our work with some future directions that this dataset opens for the research community in Section 5.

[3] Full dataset available at: `https://github.com/gtmdotme/ReAct`

2 Related Work

Text classification has long been an active area of research, as the classification can help the users to be able to efficiently process a large amount of content. Finding actionable comments on social media (tweets) was addressed in [22] using new lexicon features. A *specificity* score was explored in [5] for an employee satisfaction survey and product review settings to understand actionable suggestions and grievances (complaints) for improvements. A shift from human crafted features to automatic feature extraction using LSTMs leveraging word embeddings was observed in [17] for political text classification. All of these works address only the actionability aspect of our problem, and the datasets used in these papers are not publicly available except in [13], where the actionability of review comments for code review is investigated using lexical features, and the dataset (*Chromium Conversations*) is made available publicly. However, the characteristics of comments are very different compared to document review comments as discussed later in the paper (refer Section 4.4).

Other binary classifications in prior work include Question classification [28], agreement/disagreement classification [1, 14, 24, 27] and suggestions/advice mining [6, 25, 26]. However, such binary classifications only provide information on a single dimension in isolation and fall short in providing a more extensive set of categorization as done in [15], where the authors investigated comments on product reviews in an e-commerce setting. They classify the comments on reviews into thumbs-up, thumbs-down, agreement, disagreement, question and answer acknowledgment categories. Feedback comments regarding library refurbishment were analyzed in [21] both for their actionability and sentiment (positive, negative, and neutral). Another work that takes a somewhat similar set of categories (question, suggestion, agreement, disagreement) is reported in [19]. Nevertheless, the datasets are again not publicly available, and the categories proposed are inadequate in providing a comprehensive set. Thus, supporting our taxonomy proposed in Section 3.2. A recent work that analyzes fine-grained emotions is [3], which creates a taxonomy of 27 emotions (or neutral) over comments obtain from a popular online forum. While this work makes the dataset public, the categorization of the emotions is not suitable for a document review setting.

OpenReview [12] is a popular online forum for reviewing research papers bearing similarities with our problem setting. In fact, the choice of gathering data from OpenReview is motivated by a comprehensive study for analyzing the review process [23] followed by studies centered around rating prediction [4, 10]. [8] also present *PeerRead* dataset consolidating reviews from a lot of conferences. Our dataset provides finer-grained annotation by providing two labels per review comment sentence and thereby opens up a new research direction.

3 Dataset: ReAct

While the prior art focuses on feature engineering and model architecture, we note a lack of publicly available datasets in this problem set. This section describes how we arrive at the proposed annotated dataset, *ReAct*.

In this paper, We use Fleiss' kappa κ [7] as the measure of inter-annotator agreement. It is used to determine the level of agreement between two or more annotators when the response variable is measured on a categorical scale. The measure calculates the degree of agreement in annotations over that, which would be expected by chance. It takes the ratio of the observed agreement above the expected agreement by chance, and the maximum possible agreement above the expected agreement by chance.

3.1 Raw Data Collection and Preprocessing

The proposed dataset is gathered from an online public forum OpenReview [12] where research papers are reviewed and discussed. Multiple anonymous reviewers review the papers and write free-form comments (along with people other than reviewers also writing comments) related to the paper. We extract 911 papers submitted to ICLR (The International Conference on Learning Representations) 2018 from OpenReview and filter out the comments written by people other than the reviewers and get a dataset where each record is a paper associated with its reviews and other metadata (final decision, rating, link to the paper, timestamps, abstract, etc.). Each paper is reviewed by at least 3 reviewers who provide their comments in free-form text. An average review spans about 19 sentences. On manual inspection, we find that giving a specific label to the whole review is not appropriate, since a review often contains some facts about the paper along with some merits/demerits and some questions. A natural choice is to chunk the review into smaller units. The paragraph structure in reviews is not always consistent and may still contain multiple types of comments. Therefore, we choose a sentence as the atomic unit and refer to it as a *comment*. A consequence of this choice is that some review comments become slightly less interpretable or less clear by themselves if a single argument of the reviewer is spread across multiple sentences. Also, some sentences still contain more than one type of comment. However, based on our inspection, we find the trade-off by choosing the sentence level granularity to be acceptable (only a small fraction of sentences needed additional context not captured by the sentence itself, and only a small number of sentences have more than one dominant comment type). We use a python tool pySBD [18] for sentence splitting and disambiguation of long paragraphs of reviews into logical sentences.

3.2 Classification Taxonomy

Given the motivating scenario of helping the user quickly be able to respond to review comments, say during rebuttal period, the choice of binary classification (as *actionable* or not) is fairly straightforward and has been used in prior literature as already discussed. However, the choice of *types* for the finer-grained classification is non-obvious.

To arrive at the appropriate class labels, we randomly selected 50 review comments and three volunteers started categorizing them independently, with an initial *types* seed list of *Suggestion*, *Agreement*, *Disagreement*, and *Question*

Table 1. Fine-grained classification taxonomy for document review comments (on right) based on initial labels (on left).

Initial Label(s)	Category Label
appreciation/agreement	*agreement*
conflict/disagreement	*disagreement*
inquiry/question	*question*
demand/ask/advice/suggestion	*suggestion*
problem/issue/shortcoming	*shortcoming*
opinion/statement of fact	*fact*
miscellaneous/others	*other*

inspired by [19]. Whenever a volunteer felt that the comment didn't fall in these types, the volunteer added a new type. After completing the independent categorization, a pool of all labels for types across the three volunteers was created. Now, a set of labels was consolidated if the volunteers agreed that the individual labels in that set had the same semantics. Table 1 shows the initial labels in the left column and the final proposed taxonomy of type labels in the right column.

3.3 Designing Survey

We selected 125 reviews for the main survey such that they were sufficiently long for having (at least) 10 comments as part of each review, and retained 10 randomly selected sufficiently long comments (having at least 10 words) corresponding to each of these reviews. These 1250 review comments are then annotated by a popular crowdsourcing platform, Amazon Mechanical Turk (AMT). Each comment is annotated by 5 different human annotators, also known as turkers. The annotators are given a set of instructions for annotating the comments. While the *actionable* (or *non-actionable*) label is fairly easy to understand (seen in Figure 1 as an explanation for 'Task 1'), we provide appropriate explanations using examples for the proposed finer-grained taxonomy. The examples can be seen in Figure 1, and the explanations were seen upon hovering over the help icon ('?'), and are listed in Table 2. We also asked annotators to provide feedback for the survey and find that most of the annotators found the survey self-sufficient and easy to understand, as also described in Section 3.4. The task of an annotator is to read the review comments and assign two labels to each comment. The first label $Label_1$ is to be assigned based on the actionability of the comment, i.e., among {*yes*, *no*} and constitutes $Task_1$. Similarly, another label $Label_2$ is to be assigned from the proposed taxonomy, one of {*agreement*, *disagreement*, *suggestion*, *question*, *fact*, *shortcoming*, *other*} based on the type of comments and constitutes $Task_2$. We also included a note indicating the presence of validation questions in the survey. While we didn't formally introduce any validation question, we randomly selected a few responses from a few HITs and manually validated them for obvious incorrect categorization. We found the quality of re-

Table 2. Explanations given in survey for $Label_2$ categories.

$Label_2$	**Explanation**
Question	Reviewer is asking a direct question to the authors.
Agreement	Reviewer is expressing agreement or highlighting some points from paper in positive manner.
Disagreement	Reviewer is disagreeing with some statements in the paper, or the overall hypothesis/conclusion of the paper itself.
Suggestion	Reviewer has clearly indicated the need for a task to be done.
Shortcoming	Reviewer is pointing out some type of shortcoming or problem, but not suggesting what the authors should do to fix those problems.
Fact	Reviewer just mentions a fact, or an opinion, which does not seem like a positive or negative statement about the paper.
Other	Comments that cannot be put in one of the above categories.

sponses to be reasonably high in terms of consensus among annotators compared to labels perceived by us in these random checks

The survey was available to annotators based on certain filters that AMT provides. We restricted the survey to *Mechanical Turk Masters* who had acceptance scores $\geq 95\%$ to get high-quality annotations. The reward of one complete survey (comprising two types of labels for 10 comments) was set to \$ 0.75 based on the feedback received on the pilot surveys floated initially, described next. A time limit of 30 minutes was set before the survey expired.

3.4 Analyzing Responses

Pilot Survey Instead of rolling out the survey fully in one go, we followed an iterative approach. A pilot survey was conducted to check if the tasks and instructions were clear and to get an estimate of the quality of responses. We handpicked 5 reviews (different from the ones used in the main survey) having a total of 50 comments using the above survey design. Post completion, we analyzed each of the responses one by one and noted a 'moderate' inter-annotator agreement score (Fleiss kappa), $\kappa \approx 0.48$ among the annotators [9]. Further, upon our manual inspection on our inspection, the responses still seemed to have a reasonable basis for those annotations A few other analyses were done as described in the next section to ensure the quality of responses. A noteworthy thing was the feedback received from annotators which substantially supported our comprehensive, yet simple survey design. The annotators expressed no lack of clarity and also found the survey task to be appropriate and complete. The time to answer was also analyzed and seemed to match with the time taken by the volunteers (close to 10 minutes per survey). Hence, the feedback received from this survey was sufficiently positive to go ahead with the main survey.

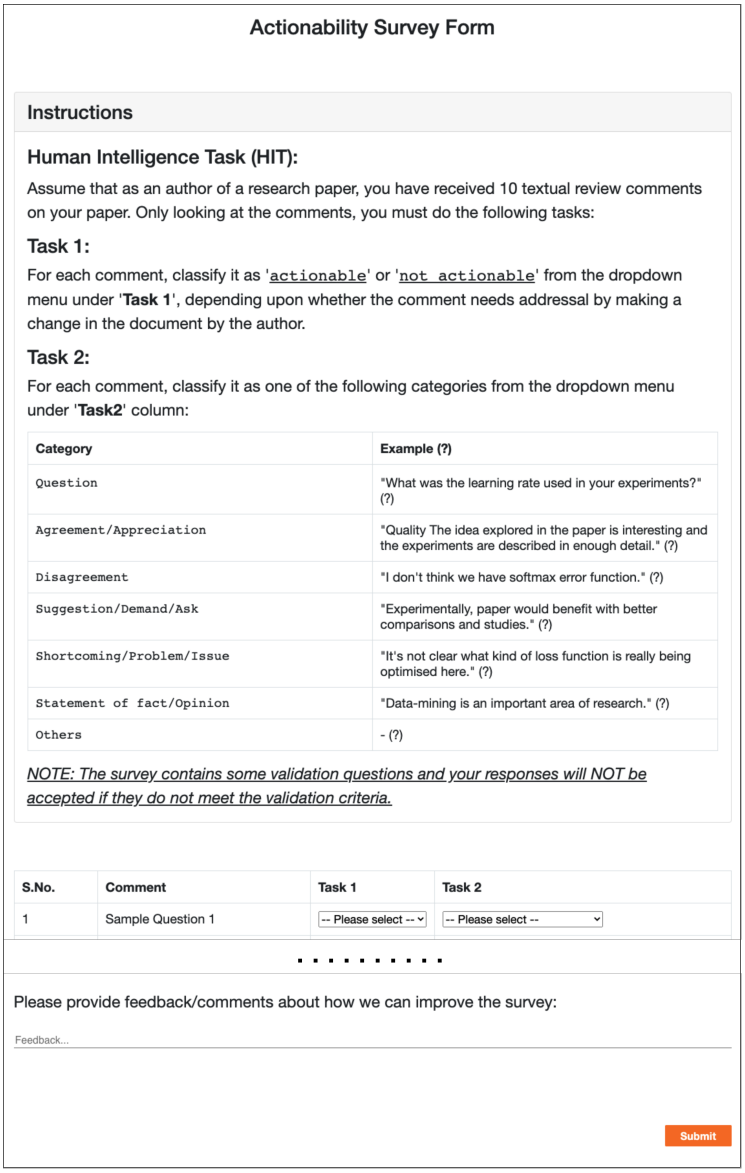

Fig. 1. AMT Survey Design for collecting responses.

Main Survey Post successful completion of the survey, we obtain a set of 6, 250 annotations for our dataset comprising of 125 reviews each containing 10 comments labeled by 5 different annotators. Each annotation consists of labels for the comment along with other metadata such as characteristics of the annotator (IDs, timestamps, etc.) and that of the survey such as (IDs, duration, times-

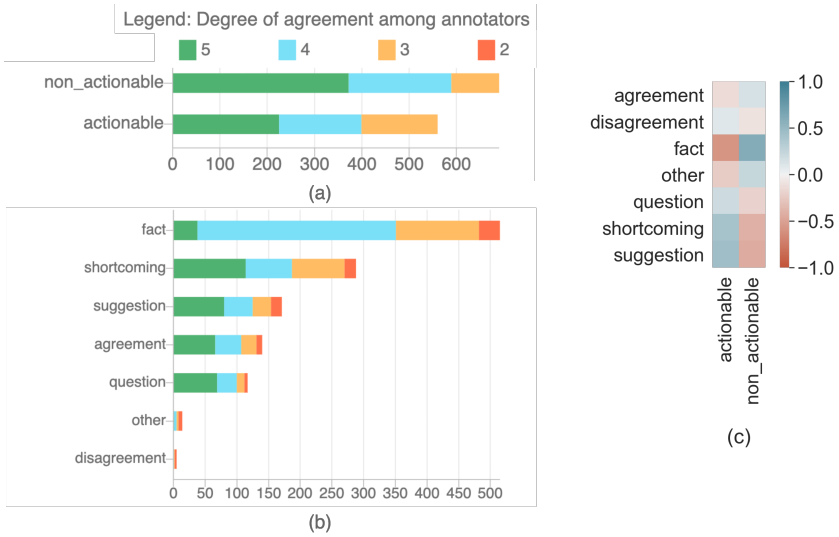

Fig. 2. (a,b) Distribution of review comments based on their count, hued by the fraction of agreement in annotators annotating the same comment for $Label_1$ and $Label_2$ respectively, (c) correlation observed in the two kinds of category labels annotations

tamps, etc.). We found that a total of 33 unique annotators participated in the survey with an average completion time of $\sim$ 10 minutes.

First, we analyze the Fleiss kappa scores on individual labeling tasks, i.e., for $Label_1$ and $Label_2$. For the $Label_1$ denoting actionability, we observe a 'moderate' inter-annotator score of 0.49 and a slightly higher score of 0.53 for $Label_2$ based on the proposed taxonomy [9].

Next, we analyze at a deeper level by looking at proportions of responses ranging from having a clear agreement to strong ambiguity as shown by proportions of stacked bars in Figure 2(a) and (b). At an aggregate level, for $Label_1$, almost 50% of annotations have a clear consensus where all 5 annotators vote for the same category label, while 30% of annotations have 4 out of 5 votes and the rest 20% have 3 out of 5 votes as shown in Figure 2(a) at per category label basis. Similarly, for $Label_2$, more than 70% of annotations have at least 4 out of 5 annotators agree on a specific category label as shown in Figure 2(b). We observe that *disagreement* is a rare class (with a low agreement between reviewers), suggesting this label may not be essential.

The correlation analysis using the Pearson Correlation Coefficient between the two sets of category labels (Figure 2(c)) strengthens the hypothesis that suggestions, shortcomings, and questions are more of an actionable item than the other categories. We found some noisy responses where annotators labeled shortcomings as non-actionable. Another example of noise found in the data is when annotators annotate an agreement as actionable. The proportion of such noisy responses is very less ($\sim$ 6.5%) in the whole dataset.

3.5 Moderation

To improve the quality of data further, we selectively moderated the labeling. In particular, we reviewed $Label_2$ for review comments in the case of maximum disagreement, i.e., when the maximum number of annotators agreeing on that label was 2. In addition, if the maximum number of annotators agreeing on $Label_1$ was 3 for any of *these* comments, then this label was also reviewed. In the cases where maximum number of annotators agreeing $Label_2$ was 3 (or more), indicating a good level of agreement (3 out of 5), neither of the labels were reviewed (even if for $Label_1$, the maximum number of agreeing annotators were 3, the borderline case). The rationale is that given the correlation between labels of the two types, we feel that often the finer-grained label, $Label_2$, is like a cause for the coarser-grained label, $Label_1$, which is an effect, and therefore, if there is a high agreement for the fine-grained label (to the extent of 3 out of 5 annotators agreeing on one label out of possible 7), we don't need to review the coarser-grained label even if the agreement is marginal for it (3 annotators agreeing out of 5 for a choice out of two) since probably there is a good agreement for the root-cause. Therefore, we didn't want to override their annotations.

There were a total of 91 cases where we reviewed $Label_2$, out of which 49 cases where we also reviewed $Label_1$. Ignoring these cases, the inter-annotator agreement for $Label_1$ increased to 0.52, and for $Label_2$, it increased to 0.57. Finally, in review, we ended up changing $Label_1$ for 19 cases. For $Label_2$, there were 34 cases where there was a tie and we picked one of the tied labels. Further, we changed labels for 18 other cases, where we assigned a label as ground truth for $Label_2$, which was not voted as (one of) majority label(s) by annotators. Given the small number of cases where we had to change the labels (about 1.5% cases), we believe the annotations' quality is very good.

3.6 Processed Dataset

Based on the above analysis, we assign the ground truth labels based on majority, i.e., out of 5 votes for a given comment, we chose the majority vote as the ground truth label. While a tie is not possible in $Task_1$, it may take place in $Task_2$ (consider the case where all the votes are for different labels or two votes each for two labels). Total number of tied cases for $label_2$ was 34. As mentioned before, such ties were resolved through the process of moderation. The final prepared dataset consists of $1,250$ comments with two sets of labels, $Label_1$ and $Label_2$, a sample of which is shown in Table 3. A summary of the descriptive statistics is shown in Table 4.

4 Benchmarking Experiments

Given a review comment from the proposed dataset, two classification scenarios arise:

- Binary classification ($Task_1$) to identify whether the comment is actionable to the author or not, and

Table 3. Sample comments and annotations from our proposed dataset.

Comment	$Label_1$	$Label_2$
It would enhance readability of the paper if the results were more self-contained.	actionable	suggestion
It lacks a few references and important technical aspects are not discussed.	actionable	shortcoming
Could you explain how classes are predicted given a test problem?	actionable	question
Indeed, the authors have succeed in showing that this is not necessarily the case	non-actionable	agreement
If the text is from the user, a named entity recognizer is used.	non-actionable	fact

Table 4. Summary statistics of the processed dataset.

No. of examples	1250
No. of labels for actionability	2
No. of labels for proposed taxonomy	7
No. of unique raters	33
No. of raters per comment	5
Average words per comment	∼23

– Multiclass classification ($Task_2$) to identify the nature of comment from the proposed taxonomy.

To the best of our knowledge, this is the first of its kind dataset in the domain of document review comments. To establish benchmark results, we model each of the text classification tasks.

4.1 Feature Extraction

Most of the recent works on producing contextual embeddings have shown to improve results over the human crafted features, although at the cost of interpretability of features themselves. We experiment with the following state-of-the-art sentence embeddings:

– Universal Sentence Encoder [2] (**USE**): The model is trained and optimized for text, such as sentences, phrases, or short paragraphs and encodes it into a 512 dimensional space.
– **DistilBERT** Embeddings [20] are a distilled version of BERT with faster performance and fewer parameters (768 dimensional vectors).
– **RoBERTa** Embeddings [11] are built out of tweaking the BERT model hyperparameters to produce robust embeddings that are shown to perform best for STS tasks. These are 1024 dimensional vectors.

Table 5. Test accuracy and F1 scores of the classification models for each task on our proposed dataset. Here, DistB is DistilBERT and RoB is RoBERTa.

Models	$Task_1$						$Task_2$					
	Accuracy			**F1-Score**			**Accuracy**			**F1-Score**		
	USE	**DistB**	**RoB**	**USE**	**DistB**	**RoB**	**USE**	**DistB**	**RoB**	**USE**	**DistB**	**RoB**
Baseline-Random	0.504	0.528	0.500	0.551	0.487	0.480	0.132	0.116	0.136	0.164	0.180	0.159
Baseline-Majority	0.588	0.588	0.588	0.435	0.435	0.435	0.408	0.408	0.408	0.236	0.236	0.236
LR	0.788	0.812	0.812	0.788	0.813	0.813	0.616	0.688	0.688	0.598	0.683	0.683
SVM	**0.796**	0.832	0.832	**0.796**	0.832	0.832	**0.636**	**0.72**	**0.72**	**0.621**	**0.708**	**0.708**
XGBoost	0.784	0.788	0.788	0.785	0.788	0.788	0.604	0.684	0.684	0.591	0.673	0.673
FNN (128, 32)	0.764	**0.848**	**0.832**	0.765	**0.849**	**0.833**	0.6	0.692	0.696	0.594	0.687	0.689

4.2 Text Classification Models

We experiment with the following text classifiers:

- *Baseline-Random*: This model predicts a class uniformly at random (out of 2 for $Task_1$ and out of 7 for $Task_2$).
- *Baseline-Majority*: This model predicts the most frequently occurring class in the training dataset.
- ML-Classifiers: We use standard classification models like Logistic Regression (*LR*), Support Vector Machine (*SVM*), *XGBoost*, and Feedforward Neural Network (*FNN*).

4.3 Experimental Setup

Data All the experiments involve using the proposed dataset by keeping 80% of the data for training and the rest 20% for evaluating the model. The (random) split is such that the proportions of the $Label_2$ are preserved in the two sets, i.e., stratified split.

Models We use a standard implementation of these models from the scikit-learn python library [16] keeping the default parameters fixed for a fair comparison across variations in models and embeddings.

Evaluation We evaluate the model performances using the standard metric of accuracy (fraction of correct predictions out of total) on the test set in Table 5. Given the imbalance in our dataset, we also report the f1-scores (weighted average of class-wise F1 scores). We note that the models are able to achieve a fair degree of accuracy (significantly higher than baselines), and also that using larger embeddings (RoBERTa and DistilBERT) results in better accuracy than smaller USE embeddings, although between RoBERTa and DistilBERT, there is no significant difference in performance. We fine-tune the hidden layer parameter for the FNN model and find the combination of using two layers of sizes 128 and

Table 6. Accuracy for Test data (with number of such cases in parenthesis) for each task grouped by agreement in annotators for one of the performant model, SVM using DistilBERT embeddings.

# Agreeing Annotators	$Task_1$	$Task_2$
5	0.88 (118)	0.85 (80)
4	0.85 (78)	0.74 (102)
3	0.70 (54)	0.63 (50)
2	—	0.33 (18)

Table 7. Performance comparison of Logistic Regression model on our proposed dataset *ReAct* and *Chromium Conversations* (CC) dataset.

Train Dataset	Test Dataset	Features	AUC	Accuracy	Precision	Recall	F1
CC	CC	Lexical	0.648	0.779	0.70	0.78	0.71
CC	CC	RoBERTa	0.665	0.736	0.73	0.74	0.73
CC	ReAct	RoBERTa	0.507	0.480	0.53	0.48	0.43
ReAct	ReAct	RoBERTa	0.877	0.804	0.81	0.80	0.80
ReAct	CC	RoBERTa	0.409	0.504	0.63	0.50	0.55

32 as the best configuration. Interestingly, a simple model (SVM) performs at par with a sophisticated model (FNN) as noted from the accuracies in Table 5.

We also attempt to correlate the accuracy achieved by the models with the degree of consensus among the annotators in Table 6. We see that as the number of annotators annotating these comments similarly decreases, the model accuracy also decreases. This is as expected, since some comments are more difficult to classify clearly in one category versus another, and therefore, both the human annotators and models find them harder.

4.4 Comparison with Related Datasets

Chromium Conversations To the best of our knowledge, the only public dataset having labels for actionability class is the *Chromium Conversations* dataset [13]. It is based on code reviews and not surprisingly, in general, the comments are *short* (e.g., 'merge the two curlies') and *cryptic* (e.g., 'sort') to a normal user. Also, this dataset is in a code review setting where the characteristics of review comments are very different from than review of academic and business documents. The average length in terms of tokens in this dataset is 7 (in our dataset it is 23). We do a cross-domain training, i.e., train on Chromium Conversations dataset and evaluate on our proposed test set and vice-versa and present results in Table 7. Note that, the authors of this dataset also share their feature set. But when used with BERT features, there's is a slight improvement (the first two rows) suggesting the suitability of using this set of features for future experiments. The subsequent rows capture those results and the decline in performance suggests that the nature of their dataset is different from ours.

Table 8. Distribution of top-5 emotions when our proposed dataset is tested on pre-trained GoEmotions' classifier (Please see Figure 3 to find the ground truth labels distribution corresponding to each predicted label).

Predicted Emotion	# Comments (out of 1250)
neutral	879
curiosity	78
approval	77
admiration	62
confusion	43

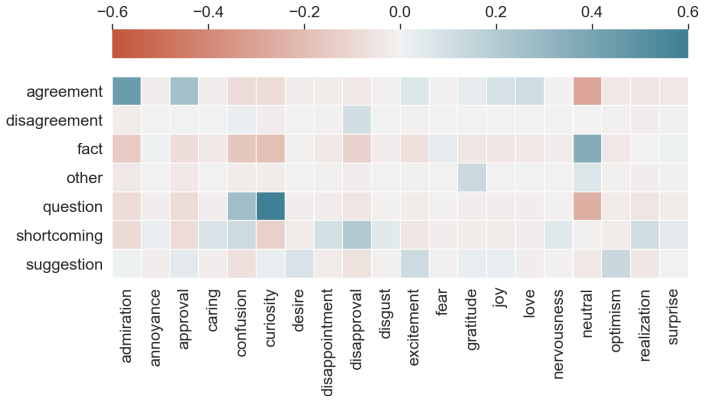

Fig. 3. Heatmap showing correlation between $Task_2$ labels, i.e., $Label_2$ and emotions, the latter predicted on proposed dataset using GoEmotions' pre-trained classifier.

GoEmotions Another recent dataset having sentiment labels is GoEmotions dataset [3]. It has 28 emotions and is manually annotated using crowdsourcing methods. We compare this emotion space with our 5 dimensional ($Label_2$) categorization by noting the predicted labels for the same set of texts. For this, we take their pre-trained model (F1 Score $\sim$ .51 when trained/tested on their dataset) and predict labels for our entire dataset (train and test). The top 5 predicted emotions are shown in Table 8. Majority of the reviews fall under the neutral category, implying less utility of this dataset in the domain of review comments. But we also observe some interesting connections between the two spaces as highlighted in Figure 3, such as *admiration*, *love* and *joy* emotions show a strong correlation with *agreement* category. Similarly, *question* tops *curiosity* emotion and *disagreement* tops in *sadness*, *disapproval*, and *disgust* emotion. The experiment suggests the non-redundancy of our $Label_2$ in our dataset.

5 Conclusion and Future Work

Review comments play an important role in the evolution of many types of documents. For some documents, the number of review comments may become large

requiring quick redressal. One important aspect of understanding the review comments is being able to determine which comments require some action on the part of document authors, and which ones do not really require any actions. The other side is the need to understand the type of review comments. In view of the lack of publicly available datasets, we introduce a carefully annotated review comment dataset, *ReAct*. We analyze the properties of the dataset. We release the dataset to the research community along with some baseline systems for the two identified text classification tasks and analyze their performance.

Since the two tasks are not completely independent from each other, a multitask learning approach seems desirable. Instead of putting resources for crowdsourcing, an active learning approach may help in curating a better dataset. The small fraction of labeled data out of a large pool of unlabelled data also calls for a self-supervised learning algorithm using less data. The incorporation of continuous learning in this model may produce robust predictions with time.

References

1. Ahmadalinezhad, M., Makrehchi, M.: Detecting agreement and disagreement in political debates. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. pp. 54–60. Springer (2018)
2. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., et al.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)
3. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547 (2020)
4. Deng, Z., Peng, H., Xia, C., Li, J., He, L., Yu, P.S.: Hierarchical bi-directional self-attention networks for paper review rating recommendation. arXiv preprint arXiv:2011.00802 (2020)
5. Deshpande, S., Palshikar, G.K., Athiappan, G.: An unsupervised approach to sentence classification. In: COMAD. p. 88 (2010)
6. Dong, L., Wei, F., Duan, Y., Liu, X., Zhou, M., Xu, K.: The automated acquisition of suggestions from tweets. In: Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
8. Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., Schwartz, R.: A dataset of peer reviews (peerread): Collection, insights and nlp applications. arXiv preprint arXiv:1804.09635 (2018)
9. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
10. Leng, Y., Yu, L., Xiong, J.: Deepreviewer: Collaborative grammar and innovation neural network for automatic paper review. In: 2019 International Conference on Multimodal Interaction. pp. 395–403 (2019)
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

12. McCallum, A.: Openreview. `https://openreview.net/`
13. Meyers, B.S., Munaiah, N., Prud'hommeaux, E., Meneely, A., Wolff, J., Alm, C.O., Murukannaiah, P.: A dataset for identifying actionable feedback in collaborative software development. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 126–131 (2018)
14. Misra, A., Walker, M.: Topic independent identification of agreement and disagreement in social media dialogue. arXiv preprint arXiv:1709.00661 (2017)
15. Mukherjee, A., Liu, B.: Modeling review comments. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 320–329 (2012)
16. 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, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
17. Rao, A., Spasojevic, N.: Actionable and political text classification using word embeddings and lstm. arXiv preprint arXiv:1607.02501 (2016)
18. Sadvilkar, N., Neumann, M.: Pysbd: Pragmatic sentence boundary disambiguation. arXiv preprint arXiv:2010.09657 (2020)
19. Sancheti, A., Modani, N., Choudhary, G., Priyadarshini, C., Moparthi, S.S.M.: Understanding blogs through the lens of readers' comments. Computación y Sistemas **23**(3) (2019)
20. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
21. Simm, W., Ferrario, M.A., Piao, S., Whittle, J., Rayson, P.: Classification of short text comments by sentiment and actionability for voiceyourview. In: 2010 IEEE Second International Conference on Social Computing. pp. 552–557. IEEE (2010)
22. Spasojevic, N., Rao, A.: Identifying actionable messages on social media. In: 2015 IEEE International Conference on Big Data (Big Data). pp. 2273–2281. IEEE (2015)
23. Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., Goldstein, T.: An open review of openreview: A critical analysis of the machine learning conference review process. arXiv preprint arXiv:2010.05137 (2020)
24. Wang, W., Yaman, S., Precoda, K., Richey, C., Raymond, G.: Detection of agreement and disagreement in broadcast conversations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 374–378 (2011)
25. Wicaksono, A.F., Myaeng, S.H.: Mining advices from weblogs. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 2347–2350 (2012)
26. Wicaksono, A.F., Myaeng, S.H.: Automatic extraction of advice-revealing sentences foradvice mining from online forums. In: Proceedings of the seventh international conference on Knowledge capture. pp. 97–104 (2013)
27. Yin, J., Narang, N., Thomas, P., Paris, C.: Unifying local and global agreement and disagreement classification in online debates. In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. pp. 61–69 (2012)
28. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 26–32 (2003)