TalkDown: A Corpus for Condescension Detection in Context

Zijian Wang
Symbolic Systems Program
Stanford University
zijwang@stanford.edu

Christopher Potts
Department of Linguistics
Stanford University
cgpotts@stanford.edu

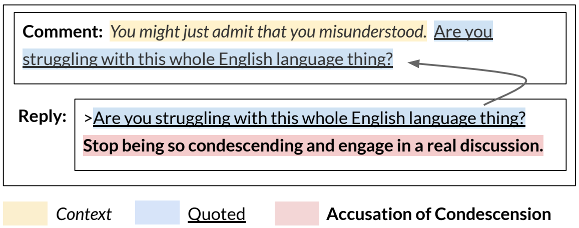
Abstract

Condescending language use is caustic; it can bring dialogues to an end and bifurcate communities. Thus, systems for condescension detection could have a large positive impact. A challenge here is that condescension is often impossible to detect from isolated utterances, as it depends on the discourse and social context. To address this, we present TALKDOWN, a new labeled dataset of condescending linguistic acts in context. We show that extending a language-only model with representations of the discourse improves performance, and we motivate techniques for dealing with the low rates of condescension overall. We also use our model to estimate condescension rates in various online communities and relate these differences to differing community norms.

1 Introduction

Condescending language use can derail conversations and, over time, disrupt healthy communities. The caustic nature of this language traces in part to the ways that it keys into differing social roles and levels of power (Fournier et al., 2002). It is common for people to be condescending without realizing it (Wong et al., 2014), but a lack of intent only partly mitigates the damage it can cause. Thus, condescension detection is a potentially high-impact NLP task that could open the door to many applications and future research directions, including, for example, supporting productive interventions in online communities (Spertus, 1997), educating people who use condescending language in writing, helping linguists to understand the implicit linguistic acts associated with condescension, and helping social scientists to study the relationship between condescension and other variables like gender or socioeconomic status.

Progress on this task is currently limited by a lack of high-quality labeled data. A deeper challenge is that condescension is often impossible to detect from isolated utterances. First, a characteristic of condescending language is that it is not overtly negative or critical – it might even include (insincere) praise (Huckin, 2002). Second, condescension tends to rest on a pair of conflicting pragmatic presuppositions: a speaker presumption that the speaker has higher social status than the listener, and a listener presumption that this is incorrect. For example, an utterance that is entirely friendly if said by one friend to another might be perceived as highly condescending if said by a customer to a store clerk. In such cases, the social roles of the participants shape the language in particular ways to yield two very different outcomes.



Comment: *You might just admit that you misunderstood.* Are you struggling with this whole English language thing?

Reply: >Are you struggling with this whole English language thing?
Stop being so condescending and engage in a real discussion.

Figure 1: In this example, the REPLY quotes from part of the COMMENT and says that this QUOTED text is condescending.

In this paper, we seek to facilitate the development of models for condescension detection by introducing TALKDOWN, a new labeled dataset of condescending acts in context. The dataset is derived from Reddit, a thriving set of online communities that is diverse in content and tone. We focus on COMMENT and REPLY pairs of the sort given in Figure 1, in which the REPLY targets a specific quoted span (QUOTED) in the COMMENT as being condescending. The examples were multiply-labeled by crowdsourced workers, which ensures high-quality labels and allows us to include nu-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3711–3719,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

anced examples that require human judgment.

Our central hypothesis is that context is decisive for condescension detection. To test this, we evaluate models that seek to make this classification based only on the QUOTED span in the COMMENT as well as extensions of those models that include summary representations of the preceding linguistic context, which we treat as an approximation of the discourse context in which the condescension accusation was made. Models with contextual representations are far superior, bolstering the original hypothesis. In addition, we show that these models are robust to highly imbalanced testing scenarios that approximate the true low rate of condescension in the wider world. Such robustness to imbalanced data is an important prerequisite for deploying models like this. Finally, we apply our model to a wide range of subreddits, arguing that our estimated rates of condescension are related to different community norms.

2 The TALKDOWN Corpus

We chose Reddit as the basis for our corpus for a few key reasons. First, it is a large, publicly available dataset from an active set of more than one million user-created online communities (subreddits).[1] Second, it varies in both content and tone. Third, users can develop strong identities on the site, which could facilitate user-level modeling, but these identities are generally pseudonymous, which is useful when studying charged social phenomena (Hamilton et al., 2017; Wang and Jurgens, 2018). Fourth, the subreddit structure of the site creates opportunities to study the impact of condescension on community structure and norms (Buntain and Golbeck, 2014; Lin et al., 2017; Zhang et al., 2017; Chandrasekharan et al., 2018).

The basis for our work is the Reddit data dump 2006–2018.[2] We first extracted COMMENT/REPLY pairs in which the REPLY contains a condescension-related word. After further filtering out self-replies and moderator posts, and normalizing links and references, we obtain 2.62M COMMENT/REPLY pairs.

Not all of these examples truly involve the REPLY saying that the QUOTED span is condescending. Our simple pattern-based extraction method is not sufficiently sensitive. To address this, we conducted an annotation project on Amazon Mechanical Turk. Our own initial assessment of 200 examples using a five-point Likert scale revealed two things that informed this project.

	Median	Mean	Std.	Max
QUOTED	18	22.79	18.40	399
REPLY	67	100.92	112.34	1,921
CONTEXT	47	98.40	154.59	2,136

Table 1: Basic statistics of the length of the examples in the corpus. CONTEXT is everything in the COMMENT before the QUOTED span.

First, we saw a clear split in the positive instances of condescension. In some, specific linguistic acts are labeled as condescending ("This is really condesending"), whereas others involve general user-level accusations that are not tied to specific acts ("You're so condescending"). We chose to focus on the specific linguistic acts. They provide a more objective basis for annotation, and they can presumably be aggregated to provide an empirically grounded picture of user-level behavior (or others' reactions to such behavior). Thus, for positive instances of condescension, we further limited our attention to COMMENT/REPLY pairs in which the REPLY contains a direct quotation from the COMMENT, using fuzzy match based on Levenshtein distance (Navarro, 2001), as illustrated in Figure 1. We extracted 66K such examples. Some statistics on these examples is given in Table 1.

Second, with the above ambiguity addressed, the signal of condescending or not is mostly clear. Thus, we designed the annotation project around a three-way multiple choice question: *condescending*, *not condescending*, and *cannot decide*. Each task began with instructions and two training questions, following by 10 different COMMENT/REPLY pairs to be labeled. Appendix A.2 provides screenshots of the annotation interface.

To process the annotations, we filtered out the work of annotators who did not correctly answer the training questions. The remaining annotators have moderate to substantial agreement (Fleiss $\kappa = 0.593$; Fleiss 1971; Landis and Koch 1977). We then used Expectation–Maximization, as in Dempster et al. 1977, to assign labels. This yields slightly better quality in our hand-inspection than labels by majority vote, presumably because it factors individual worker reliability into the decision making. In the end, we obtained 4,992 valid labeled instances: 65.2% labeled as *conde-*

[1] http://redditmetrics.com/history

[2] https://files.pushshift.io/reddit

	Positive	Negative
Balanced (1:1)	3,255	3,255
Imbalanced (1:20)	3,255	65,100

Table 2: Basic statistics of our dataset.

scending (henceforth, *positive*), and 34.8% as *non-condescending* (henceforth, *negative*).[3]

To fully balance the dataset, we pulled out one random month's data for each year in 2011 to 2017. We extracted instances using the same methods as described above, but we filtered out Comment/Reply pairs in which a condescension-related word appeared. Our final dataset thus consists of annotated positive and negative instances, with supplemental randomly-sampled negative instances. For our experiments, we partitioned the data into 80% train, 10% development, and 10% test splits. In addition, to simulate real-world situations, we built a dataset with a 1:20 ratio of positive to negative instances.[4] The basic statistics of the dataset are shown in Table 2.

3 Experiments

We now establish some baselines for the TalkDown Corpus and begin to test the hypothesis that contextual representations are valuable for this task. To do this, we use the BERT model of Devlin et al. (2019), which uses a Transformer-based encoder architecture (Vaswani et al., 2017) to learn word representations by training against a masked language modeling task and a next-sentence prediction task. Our models are initialized with the pretrained representations released by the BERT team and a fully connected layer on the top (Figure 4 in Devlin et al. 2019), which is then fine-tuned to our dataset (Peters et al., 2019). We explore both BERT Base ($BERT_B$) and BERT Large ($BERT_L$),[5] to determine whether the added expense of using $BERT_L$ is justified. Appendix B provides details on our process of hyperparameter tuning and optimization.

Input 1	Input 2	Model	Imb. F1	Bal. F1
Quoted	∧ Context	∧ $BERT_L$	**0.684**	**0.654**
Quoted	∧ Context	∧ $BERT_B$	0.657	0.596
Quoted		∧ $BERT_L$	0.650	0.640
	Context	∧ $BERT_L$	0.611	0.513
		random	0.371	0.500
		majority	0.488	0.333

Table 3: Performance (macro-F1) for predicting condescension on balanced and imbalanced versions of TalkDown. Model selection was done according to the procedure described in Appendix B.

	Quoted + Context		Quoted	
Ratio	Imb. F1	Bal. F1	Imb. F1	Bal. F1
1:1	0.542	**0.708**	0.554	**0.690**
1:20	0.670	0.574	0.620	0.518
2:20	0.682	0.619	0.640	0.554
3:20	**0.684**	0.654	0.646	0.585
4:20	0.678	0.632	**0.650**	0.640
5:20	0.668	0.626	0.645	0.582
10:20	0.672	0.656	0.641	0.640
15:20	0.665	0.641	0.641	0.593
20:20	0.674	0.621	0.645	0.597

Table 4: The impact of different train-set positive:negative ratios. All the models are $BERT_L$. The first row is based on the balanced dataset, and the rest are based on the imbalanced dataset with different oversampling ratios. Model selection again used the procedure in Appendix B.

3.1 Predicting Condescension

Table 3 summarizes the results of our core experiments. Input 1 and Input 2 describe the basis for the feature representations. Thus, for example, Quoted∧Context is a model that uses both the quoted span and the preceding linguistic context. We report two testing scenarios: *Balanced* and *Imbalanced*, in which there are 20 negative examples for each positive example.

The results clearly support our hypothesis that context matters; using the Quoted part and Context together give us 3–4% boost in macro-F1 using the same model architecture. In addition, we see that increasing the capacity of the model also helps, though more modestly. It's noteworthy that the performance of using the Quoted part is better than that of using Context alone, though the Quoted part is roughly three times shorter. Thus, there is a strong signal in the Quoted part – the replier chose this span for a reason – but the context contains a signal as well.

[3] There was just one case where *cannot decide* was the chosen label; it was in Spanish, so we excluded it and added a language classification step to our preprocessing pipeline.

[4] To the best of our knowledge, there is no prior work on what percentage of conversations on Reddit (or, more broadly, in daily conversations) are condescending. Thus, we chose the ratio based on informal observations on Reddit.

[5] The whole-word masking model was used as it performs better than the original one in multiple benchmarks.

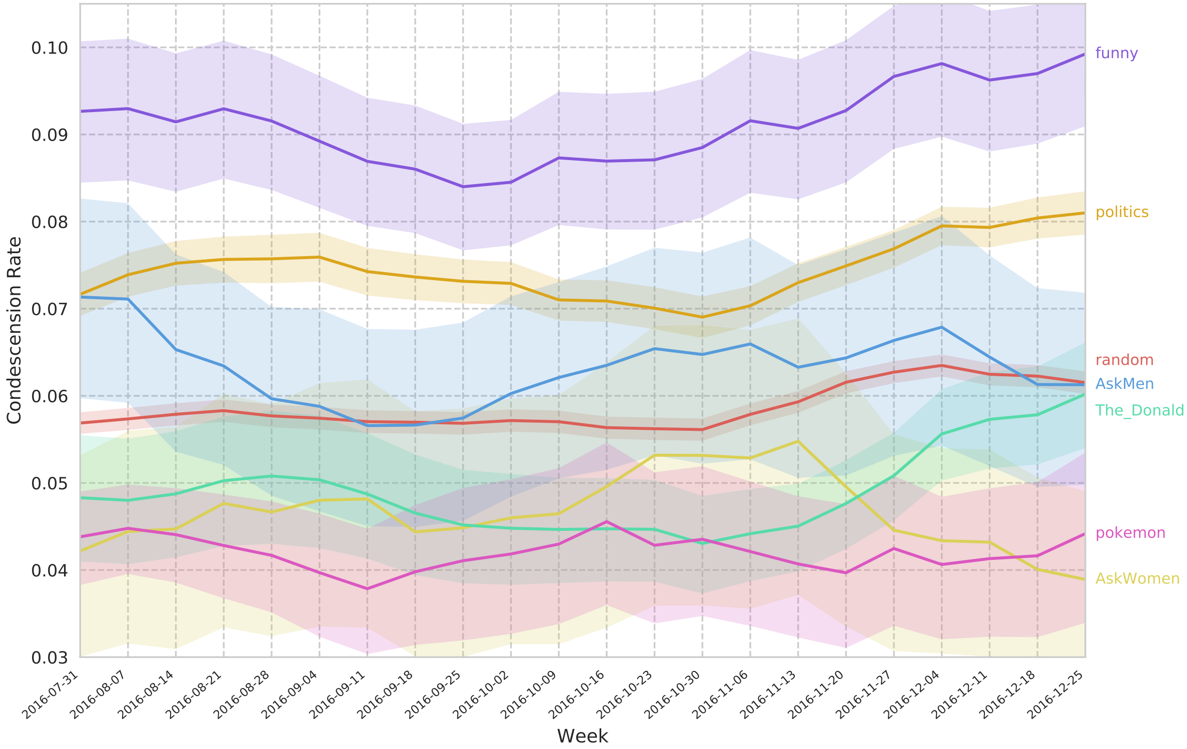


Figure 2: Condescension rates over time in a selection of subreddits, as predicted by our best model. The time window is centered around the 2016 U.S. Presidential Elections. We calculate a rolling mean with a window size of 5 and use July 2016 only for smoothing. We obtain 95% confidence intervals via bootstrapping.

3.2 Imbalanced Testing Scenarios

Imbalanced testing scenarios are more challenging, but they also better reflect usage rates of condescending language in public forums like Reddit. To further understand how best to get traction on this problem, we explored a range of different methods for creating training data. Our results are summarized in Table 4. As expected, the balanced problem is best addressed with a balanced dataset. For the imbalanced problem, we found that an oversampling ratio of 2 to 4 yielded the best performance. Our full QUOTED ∧ CONTEXT model is again clearly superior in these scenarios.

3.3 Condescension Rates Across Subreddits

Our hope for TALKDOWN is that it will play a role in developing systems that can help identify condescending acts on social media. This will depend on models trained on TALKDOWN being able to get an accurate read on condescension at scale. As a first step towards assessing this capability, we ran our models on 14 subreddits, over the time period of July 2016 to December 2016, which covers the 2016 U.S. Presidential Election, an event that we expect to influence condescension rates in various ways across Reddit. Appendix C lists these subreddits along with their post counts and estimated average rates of condescension. Figure 2 highlights a selection of them.

As a baseline, we include a 10% random sample from the top 100 most active subreddits.[6] Consistently above this baseline are ‘politics’ and ‘funny’. It makes sense that an overtly political subreddit would show a high rate of condescension (as do ‘news’ and ‘worldnews’; Appendix C): it’s a contentious topic in a contentious time period; see also the rising rate for ‘The_Donald’ in the post-election period. It is more surprising that ‘funny’ shows the highest rates. We do not have a deep understanding of why this is, but it could trace to our model confusing irony and sarcasm with condescension.

Below the baseline are ‘AskWomen’ and ‘pokemon’. We expect ‘pokemon’ to have low rates of condescension, as it strikes us as a supportive community. However, one might be surprised to see ‘AskWomen’ so low, especially as compared with ‘AskMen’, which has high rates in general.

[6] This is derived from the ‘subscribers’ section in http://redditlist.com/all, excluding ‘announcements’.

There is wide support for the idea that women experience more condescension than men do (Hall and Braunwald, 1981; Harris, 1993; McKechnie et al., 1998; Cortina et al., 2002; Trix and Psenka, 2003), as reflected in the recent lexical innovation *mansplaining*, which can be roughly paraphrased as 'a man condescending to a woman'.[7] However, community norms on 'AskWomen' and 'AskMen' are likely shaping these outcomes. Whereas the description for 'AskWomen' says it is "curated to promote respectful and on-topic discussions, and not serve as a debate subreddit", the description for 'AskMen' ends with "And don't be an asshole. Also, go away."

4 Conclusion

We introduced TALKDOWN, a new annotated Reddit corpus of condescending linguistic acts in context. Using BERT, we established baseline models that suggest this is a challenging task, and one that benefits from rich contextual representations. Finally, in qualitative analyses on diverse subreddits, we offered initial evidence that models trained on TALKDOWN generalize to new data, a prerequisite for using them to help improve online communities via condescension detection. The full dataset with the pretrained BERT model is available at `http://github.com/zijwang/talkdown`.

Acknowledgements

We would like to thank David Jurgens for his help with the Reddit dataset, Yulia Tsvetkov and Dan Jurafsky for their insightful early-stage suggestions, Lucy Li for detailed comments on an early draft of the paper, and our anonymous EMNLP–IJCNLP reviewers for their helpful feedback. We would also like to thank our Mechanical Turk contributors for their help in annotating the TALKDOWN Corpus. This research is based in part upon work supported by a Google Faculty Research Award and by the NSF under Grant No. BCS-1456077.

[7] https://en.wikipedia.org/wiki/Mansplaining

References

Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in Reddit using network structure. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 615–620, New York, NY, USA. ACM.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32:1–32:25.

Lilia M. Cortina, Kimberly A. Lonsway, Vicki J. Magley, Leslie V. Freeman, Linda L. Collinsworth, Mary Hunter, and Louise F. Fitzgerald. 2002. What's gender got to do with it? Incivility in the federal courts. *Law & Social Inquiry*, 27(2):235–270.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Marc A. Fournier, D.S. Moskowitz, and David C. Zuroff. 2002. Social rank strategies in hierarchical relationships. *Journal of Personality and Social Psychology*, 83(2):425.

Judith A. Hall and Karen G. Braunwald. 1981. Gender cues in conversations. *Journal of Personality and Social Psychology*, 40(1):99.

William L. Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Eleventh International AAAI Conference on Web and Social Media*.

Mary B. Harris. 1993. How provoking! What makes men and women angry? *Aggressive Behavior*, 19(3):199–211.

Thomas Huckin. 2002. Critical discourse analysis and the discourse of condescension. *Discourse Studies in Composition*, 155:176.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Zhiyuan Lin, Niloufar Salehi, Bowen Yao, Yiqi Chen, and Michael S. Bernstein. 2017. Better when it was smaller? community content and behavior after massive growth. In *Eleventh International AAAI Conference on Web and Social Media*.

Sally A. McKechnie, Christine T. Ennew, and Lauren H. Read. 1998. The nature of the banking relationship: A comparison of the experiences of male and female small business ownersi. *International Small Business Journal*, 16(3):39–55.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

Matthew Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Innovative Applications of Artificial Intelligence*, pages 1058–1065.

Frances Trix and Carolyn Psenka. 2003. Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, 14(2):191–220.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium. Association for Computational Linguistics.

Gloria Wong, Annie O. Derthick, E.J.R. David, Anne Saw, and Sumie Okazaki. 2014. The what, the why, and the how: A review of racial microaggressions research in psychology. *Race and Social Problems*, 6(2):181–200.

Justine Zhang, William L. Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Eleventh International AAAI Conference on Web and Social Media*.

A Data

A.1 In-house Annotation Analysis

Figure 3 shows the five-point Likert scale annotations between two in-house annotators. It can be seen that the signal of condescending or not is clear, and the agreement level between the two annotations is substantial: the Fleiss' κ is 0.613 for the five-point scale and 0.732 when normalized to three-point scale used in the paper (Fleiss, 1971; Landis and Koch, 1977).

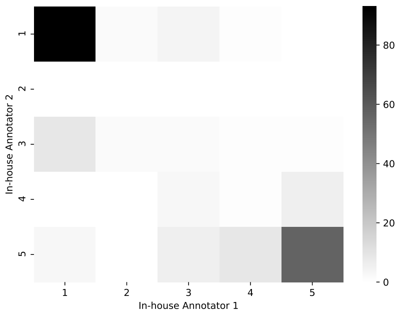


Figure 3: Heatmap for in-house initial assessment with a five-point Likert scale.

A.2 Annotation Interface

In this section, we show examples of the annotation interface we used on Amazon Mechanical Turk: Figure 4 and Figure 5.

Annotators were presented with the task name, the instructions, and two simple training questions, followed by a warning in red saying they needed to pass the training questions to proceed (Figure 4). They had unlimited trials for the training questions, and explanations (for both correct and incorrect answers) were presented directly after each trial. This helped the annotators learn how to approach the task.

After they passed the training questions, they were prompted that they could start to do the test questions (Figure 5). The interface of the test questions was similar to that of the training questions, but without explanations after selections. We explicitly checked that the annotators had made selections on each test question before submission, while this was not forced for training questions. This was to filter out possibly low-quality annotations, where the annotators did not pay attention to the instructions.

B Model Hyperparameters

Our BERT models were trained using a set of hyperparameters based on the recommendations in Devlin et al. 2019. Specifically, we set:

- Model Architecture:
  - $BERT_B$: Bert Base, Cased
  - $BERT_L$: Bert Large, Cased, with whole-word masking
- Learning rate: $\{0.5, 0.8, 1, 2, 3, 5\} \cdot 10^{-5}$
- Epoch: 2, 3
- Batch size: 32
- Max sequence length: 512

When optimizing these models, we set the batch size to 32 in order to ensure there was at least one positive instance per mini-batch. Grid search was performed with different learning rates and oversampling ratios, and best models were selected based on the best performance on the development set under the imbalanced setting. We found that oversampling 2 to 4 times the positive class (i.e., 10%–20% of the number of instances in the negative class) generally yielded good performance in all the experiments we ran. For all experiments, we used the HuggingFace PyTorch implementation of BERT.[8]

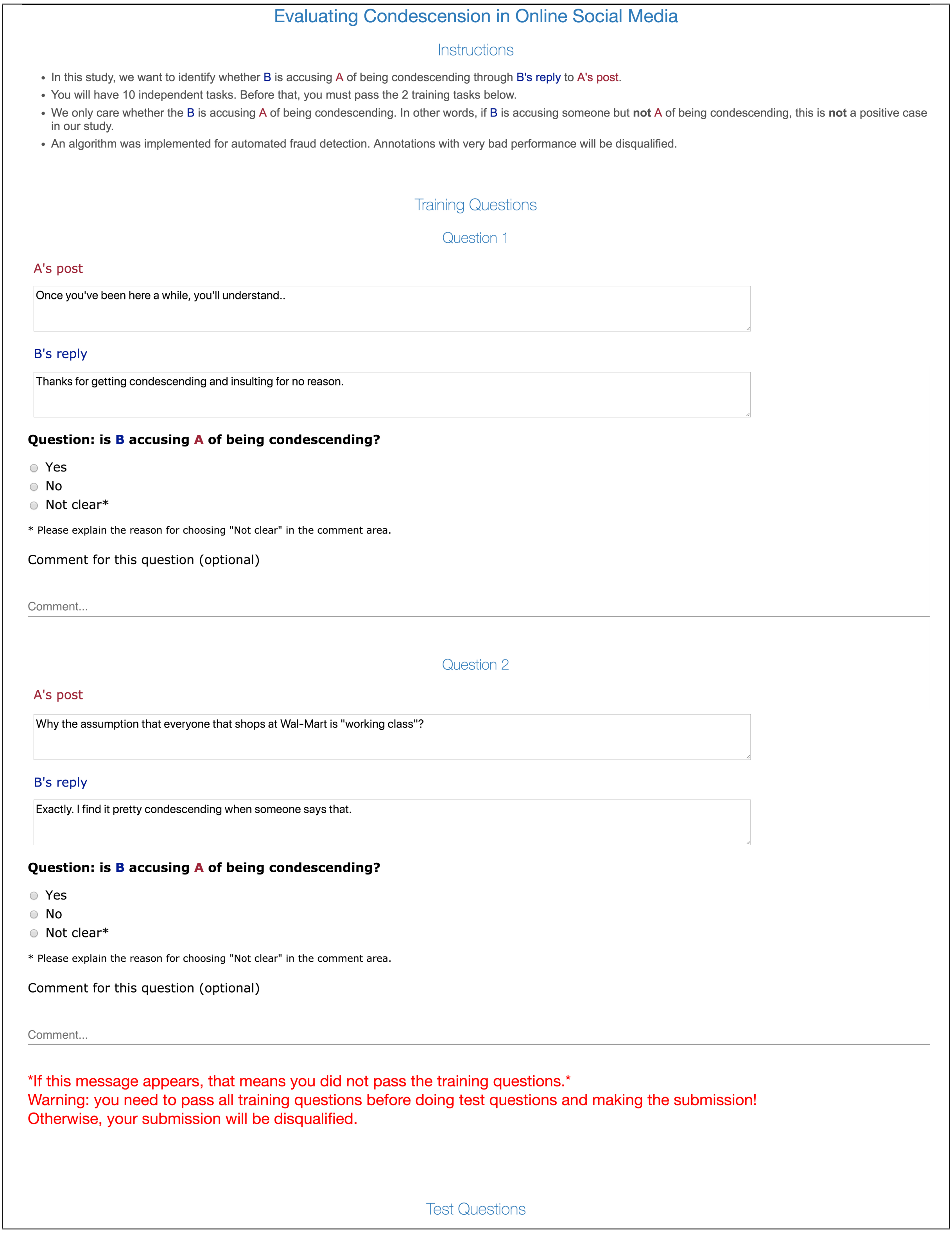
C Subreddit Condescension Rates

Table 5 shows basic statistics for all the subreddits we analyzed.

Subreddit	#Pairs	Mean rate	Std. err
4chan	6,099	0.072	0.189
AskMen	19,406	0.064	0.164
AskReddit	256,181	0.049	0.147
AskWomen	9,021	0.046	0.129
The_Donald	57,429	0.051	0.149
aww	8,727	0.059	0.168
funny	71,875	0.092	0.214
gaming	64,881	0.064	0.175
news	396,710	0.066	0.179
pokemon	38,353	0.043	0.142
politics	583,033	0.075	0.186
stopdrinking	910	0.038	0.112
tifu	35,443	0.068	0.177
worldnews	164,302	0.064	0.176
random	1,700,192	0.059	0.166

Table 5: Subreddit experiment statistics. The raw data are from the Reddit dump from July 2016 to December 2016. 'Pairs' are COMMENT/REPLY pairs as defined in the paper. 'Mean rate' is the mean rate of condescension as estimated by our best model, and 'Std. err' gives the associated standard error. '*random*' is a 10% random sample from the top 100 active subreddits over the same time period.

[8] https://github.com/huggingface/pytorch-transformers/



Evaluating Condescension in Online Social Media

Instructions

- In this study, we want to identify whether B is accusing A of being condescending through B's reply to A's post.
- You will have 10 independent tasks. Before that, you must pass the 2 training tasks below.
- We only care whether the B is accusing A of being condescending. In other words, if B is accusing someone but **not** A of being condescending, this is **not** a positive case in our study.
- An algorithm was implemented for automated fraud detection. Annotations with very bad performance will be disqualified.

Training Questions

Question 1

A's post

Once you've been here a while, you'll understand..

B's reply

Thanks for getting condescending and insulting for no reason.

Question: is B accusing A of being condescending?

○ Yes
○ No
○ Not clear*

* Please explain the reason for choosing "Not clear" in the comment area.

Comment for this question (optional)

Comment...

Question 2

A's post

Why the assumption that everyone that shops at Wal-Mart is "working class"?

B's reply

Exactly. I find it pretty condescending when someone says that.

Question: is B accusing A of being condescending?

○ Yes
○ No
○ Not clear*

* Please explain the reason for choosing "Not clear" in the comment area.

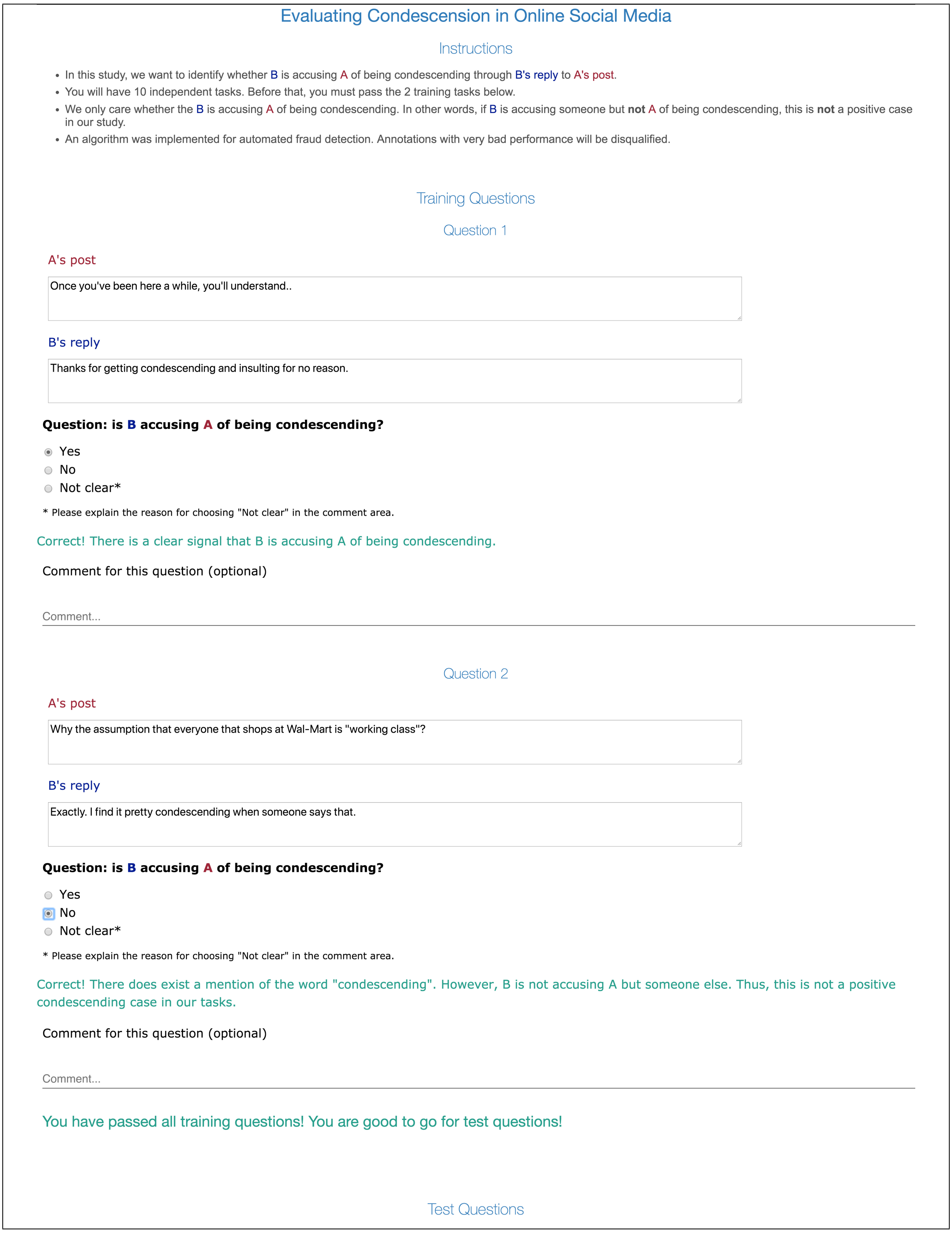
Comment for this question (optional)

Comment...

If this message appears, that means you did not pass the training questions.
Warning: you need to pass all training questions before doing test questions and making the submission!
Otherwise, your submission will be disqualified.

Test Questions

Figure 4: The initial view of instructions and training questions in the annotation interface.



Evaluating Condescension in Online Social Media

Instructions

- In this study, we want to identify whether B is accusing A of being condescending through B's reply to A's post.
- You will have 10 independent tasks. Before that, you must pass the 2 training tasks below.
- We only care whether the B is accusing A of being condescending. In other words, if B is accusing someone but **not** A of being condescending, this is **not** a positive case in our study.
- An algorithm was implemented for automated fraud detection. Annotations with very bad performance will be disqualified.

Training Questions

Question 1

A's post

Once you've been here a while, you'll understand..

B's reply

Thanks for getting condescending and insulting for no reason.

Question: is B accusing A of being condescending?

● Yes
○ No
○ Not clear*

* Please explain the reason for choosing "Not clear" in the comment area.

Correct! There is a clear signal that B is accusing A of being condescending.

Comment for this question (optional)

Comment...

Question 2

A's post

Why the assumption that everyone that shops at Wal-Mart is "working class"?

B's reply

Exactly. I find it pretty condescending when someone says that.

Question: is B accusing A of being condescending?

○ Yes
● No
○ Not clear*

* Please explain the reason for choosing "Not clear" in the comment area.

Correct! There does exist a mention of the word "condescending". However, B is not accusing A but someone else. Thus, this is not a positive condescending case in our tasks.

Comment for this question (optional)

Comment...

You have passed all training questions! You are good to go for test questions!

Test Questions

Figure 5: The view after the annotator passed the training questions.