I should not have said that: Measuring contextually inappropriate language in different social settings

Student: Jackson Sargent

Advisor: Dr. Jurgens

Teammates: Athena Aghighi, Michael Geraci

Introduction

Everyone has found themselves in a scenario where they think to themselves *should I say that?* The goal of this project is to investigate how context affects language appropriateness using computational methods. Involved in this is building a dataset of language-relationship-appropriateness sets for use in our work and for further research. We also are interested in what social information our model captures, and how it shows that appropriateness in common language is dependent on social context.

Advances in machine learning have pushed the cutting edge of natural language processing, which deals with how computers analyze and interpret written language. A part of this advancement is the use of language models, which learn to predict what word comes next given a sequence of words. Language models are trained on large datasets of text information to detect the patterns that appear in language. Models which utilize neural networks have performed well at various tasks attempting to measure language understanding, including question answering, summarizing, and detecting social phenomena like offensiveness, intimacy, or sentiment. We are building on this body of work, focusing on appropriateness and context understanding. Previous works have used context in their models by including previous parts of a conversation, whereas we will define context as a social relationship that accompanies a phrase spoken between the pair of people.

Whether or not saying something is appropriate depends on the context in which the interaction happens. This context exists along many different dimensions: physical location, intimacy, power, etc. Identifying these dimensions and how they manifest themselves is an area of active study among social scientists. Another related topic is how people approach taboo or possibly inappropriate topics in conversation, and what defines appropriateness in the first place. As machine learning advances, and more and more machine learning based language models interact with individuals and are deployed in research, researchers are investigating how they understand social information and whether or not they can be used to prevent abuse and foster positive communication online.

Outline

The main question we hope to address in this project can the social concept of contextual appropriateness be recognized by a language model trained from a dataset of phrases and relationships with labels of appropriate or not appropriate. To answer this question we first had to build a dataset, using human annotators and active learning. From this data and our model, we hope to gain further insights by probing the patterns that our language model exposes. First, what innate knowledge lies within language models about relationships and the appropriateness of language. Next, how appropriate does it consider language used in common discourse, and how would changing the context surrounding that discourse affect the model's decision making. Lastly, how does our analysis of contextual appropriateness relate to other social language problems like detecting hate speech, offensiveness, intimacy, or microaggressions.

Related Work

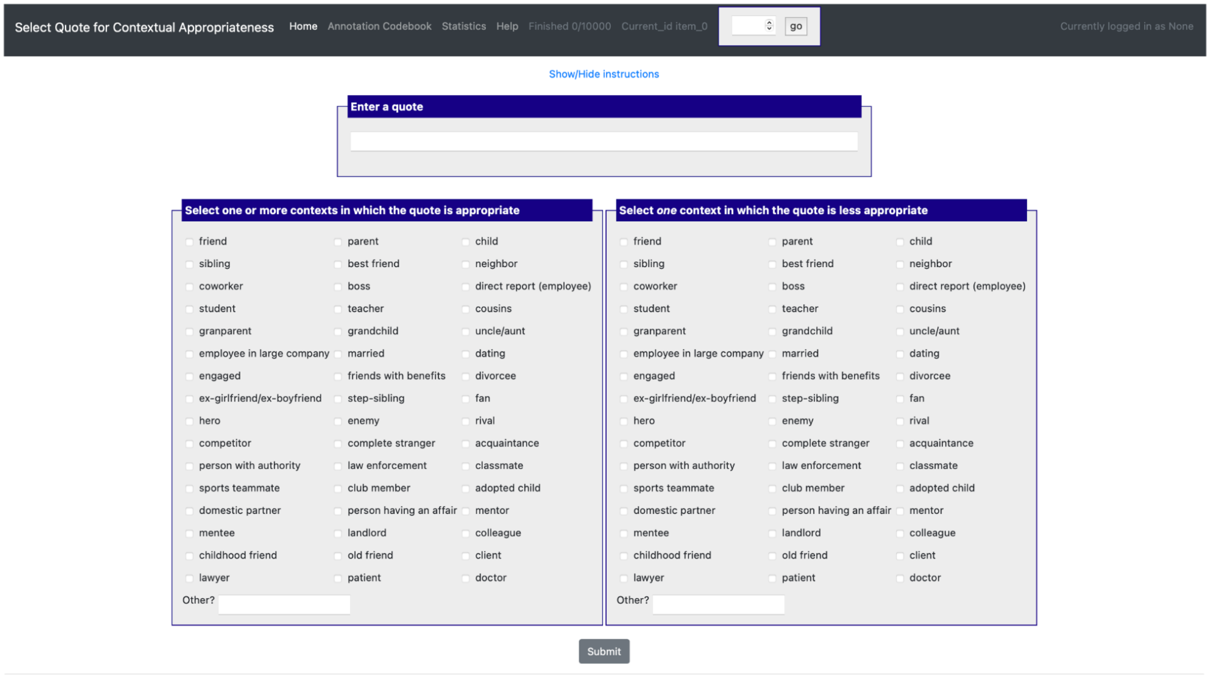
Understanding what makes language harmful or abusive and in what contexts this language appears has been a pursuit of NLP research for years. Computational language understanding has been used for years to detect hate speech and abusive language, and as the available language technologies have improved so too have the complexities of the systems deployed increased [1][2]. As models have gained the ability to tease out even more subtle differences in language, more complex tasks have emerged. One example is detecting microaggressions in social media posts, which requires language models to grasp the nuanced and inherently difficult to detect nature of the harmful text [3]. Another addition to the methodology of detecting and predicting characteristics of text is the use of additional context. Researchers have used email and chat message threads to understand whether or not additional information surrounding messages can affect model performance [4], and have used user information as context to improve model performance [5].

Alongside the computational research, there has been much work in understanding the interaction between social relationships and context and our interpretation of spoken language. Swear words and other taboo language have been studied in the context of different levels of relationships [6] [7], and others have researched how the intimacy of personal relationships affects what personal information people decide to share [8].

Methodology

To begin our investigation we first had to build a dataset. Our initial raw data came from annotation work done by the 4 researchers working on the project. Our lab has been developing an annotation tool, which we customized to fit our data annotation needs. When using the tool, the annotator would input a quote and select from a given list of relationships in which relationships they thought the quote they generated would be appropriate, and in which relationships they thought the quote would be inappropriate. We picked our relationships by focusing on relationships that have interesting social dynamics, where changing the relationship might change the appropriateness of what's being said. We made sure to include relationships from different areas of life, like working relationships, social relationships, and familial relationships. We excluded some of the most specific or uncommon relationships, and as we thought of more and more possible examples we settled on a list of 48 relationships to be as comprehensive as possible.

Alongside the quote generation process, we built a tool that generated quotes in order to give the annotator's possible inspiration for quotes to generate. These quotes were pulled from Reddit comments that were labeled as controversial by the website, and then passed through a classifier that labeled comments as either conversational or not. If the comment was conversational, it was used in our quote generator. The classifier that filtered the Reddit comments was trained on a dataset consisting of random reddit comments, a dataset of facebook empathetic dialogues, and a set of movie dialogues. Our initial data collection resulted in 4,096 examples. With this dataset, we built an initial model to be used in generating more data through annotation assisted by active learning.



Enter a quote

Select one or more contexts in which the quote is appropriate

☐ friend ☐ parent ☐ child
☐ sibling ☐ best friend ☐ neighbor
☐ coworker ☐ boss ☐ direct report (employee)
☐ student ☐ teacher ☐ cousins
☐ granparent ☐ grandchild ☐ uncle/aunt
☐ employee in large company ☐ married ☐ dating
☐ engaged ☐ friends with benefits ☐ divorcee
☐ ex-girlfriend/ex-boyfriend ☐ step-sibling ☐ fan
☐ hero ☐ enemy ☐ rival
☐ competitor ☐ complete stranger ☐ acquaintance
☐ person with authority ☐ law enforcement ☐ classmate
☐ sports teammate ☐ club member ☐ adopted child
☐ domestic partner ☐ person having an affair ☐ mentor
☐ mentee ☐ landlord ☐ colleague
☐ childhood friend ☐ old friend ☐ client
☐ lawyer ☐ patient ☐ doctor
Other?

Select *one* context in which the quote is less appropriate

☐ friend ☐ parent ☐ child
☐ sibling ☐ best friend ☐ neighbor
☐ coworker ☐ boss ☐ direct report (employee)
☐ student ☐ teacher ☐ cousins
☐ granparent ☐ grandchild ☐ uncle/aunt
☐ employee in large company ☐ married ☐ dating
☐ engaged ☐ friends with benefits ☐ divorcee
☐ ex-girlfriend/ex-boyfriend ☐ step-sibling ☐ fan
☐ hero ☐ enemy ☐ rival
☐ competitor ☐ complete stranger ☐ acquaintance
☐ person with authority ☐ law enforcement ☐ classmate
☐ sports teammate ☐ club member ☐ adopted child
☐ domestic partner ☐ person having an affair ☐ mentor
☐ mentee ☐ landlord ☐ colleague
☐ childhood friend ☐ old friend ☐ client
☐ lawyer ☐ patient ☐ doctor
Other?

Fig 1. This is what the initial set of annotators saw from POTATO, the annotation tool, while generating the initial set of examples

For our initial model, we used a PET Model [9], which utilizes language patterns and forms each problem as a fill in the blank task. Typical language model tasks work by giving the model some language input and producing a label output that abstractly represents the answer to some question. PET models instead utilize the fact that language models learn the structure and intuition of text to instead ask what the most likely word is to fill in the blank in a given sentence. An example might be, given a product review “It fell apart immediately”, to fill in the blank with “like” or “dislike”: It fell apart immediately, I [like/dislike] it. For our purposes, we constructed the following pattern, given two people and a quote said between them: [personA], to [personB], saying [quote] would be [very/not] appropriate. The PET model is built off of a pre-trained language model and fine-tunes to our specific task [10]. We used a PET model in our research because of how well it performs with the amount of data we had available after our initial annotation, and how it allows us to include the relationship context as well as the quote with relative ease. We trained our model on the 4096 initial data with X examples held out to test the model after it finished training. From this, we learned that the model is quite able to determine whether a quote is appropriate given the relationship between the speaker and listener. With this initial model, we are moving on to the next stage

Active learning is the process of utilizing initial annotated data and a trained model to seed the annotation of further data, increasing the efficiency of the data annotation process. By passing through new examples through our trained initial model, we can look and see what examples the model is most certain about, least certain about, and for what relationships it thinks certain quotes would be acceptable or unacceptable. This gives annotators a place to start when annotating and allows us to correct certain trends we see in the model's results if it is learning spurious correlations.

Another way we hope to annotate more data and gather a diverse set of perspectives on how relationships affect the appropriateness of our data is hiring additional annotators. To this end, we recently hired a few undergraduate Computer Science students from the University of Michigan to work alongside us and help annotating.

Results

Because we are still collecting data, all reports of results and experiments are preliminary, though we do expect our results to hold as we gather more data.

Modeling Results

With the data we initially collected, we can get some initial idea for how well the model performs on unseen data. We measure the success of the model over three metrics. Accuracy measures the number of guesses that are correct. F1-Score is another general metric for understanding model performance, which uses both Recall and Precision.

Training Data (4k Examples)	Accuracy	F1-Score
10% Data	0.742	0.826
50% Data	0.793	0.858
100% Data	0.790	0.849

Table 1: Measuring the models success over three different sizes of training data.

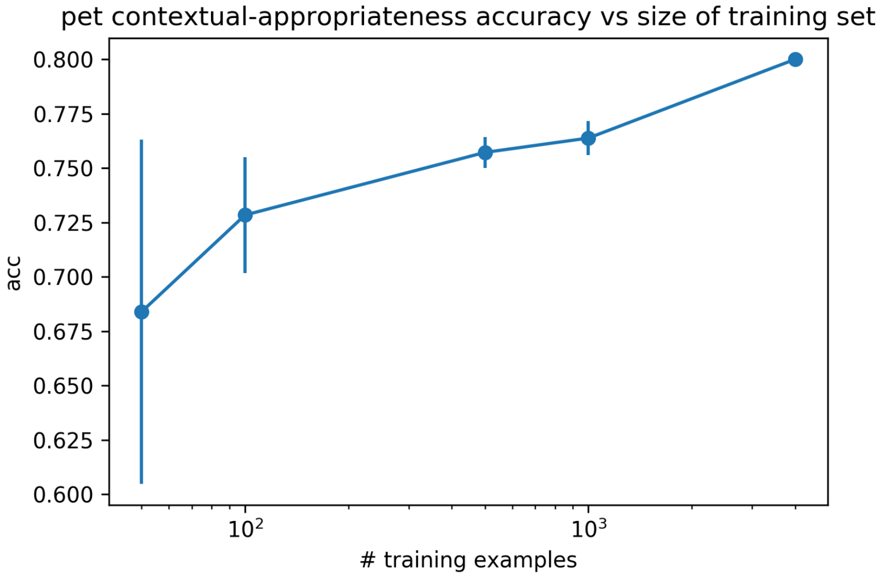


Fig 2. Test accuracy versus the number of training examples for the initial language model on our task, we can see that with all training examples used, we reached an accuracy of 0.8

Overall, the model performed quite well in both Accuracy and F1-Score. Interestingly, the evaluation showed that the model performed slightly worse when trained with 100% of the data than when trained with 50% of the data. This may because the training process has greater variability when trained with less data, and a good sample of data may have been selected for this specific run through. As we gather more data, however, we can expect these results to change. It will be interesting to where we begin to see diminishing returns as we add more and more training and evaluation data.

PRIDE Dataset

To investigate how our work can be applied to other datasets, we took a look at the PRIDE Dataset [11]. The PRIDE Dataset is a compilation of movie dialogue and the relationship between the people speaking. These relationships fit well within a subset of the relationships we were studying, so we decided to pass some of this data through our model and see what trends we can see within the dialogue and relationships from their data The relationships they used included those within family settings, social settings and professional settings. In their final dataset they used 18 relationships: parent, child, sibling, spouse, engaged, friend, enemy, lover, colleague, medical, commercial, boss, employee, teacher, student, and religious. We labeled all but the religious relationship in our annotation process, and they covered 35% of the relationships we annotated.

The first analysis we performed investigated how relationships differ in their general appropriateness. To do this, we selected a sample of dialogues and passed them through the model under all of the possible relationships. With these results in hand we could see how changing from an original relationship where a line of dialogue was appropriate to a different relationship affected whether or not that line of dialogue was appropriate. What we found was that there are clearly differences in general appropriateness across relationships as shown in Figure 3, where each row column pair represents a pair of relationships where the row is the original relationship and the column is the new one. Some relationships like Employee-to-Boss have clear trends. With this relationship, if you say something to your boss it's okay to say it to anyone else, but if you said something to someone else odds are it wouldn’t fly with your boss.

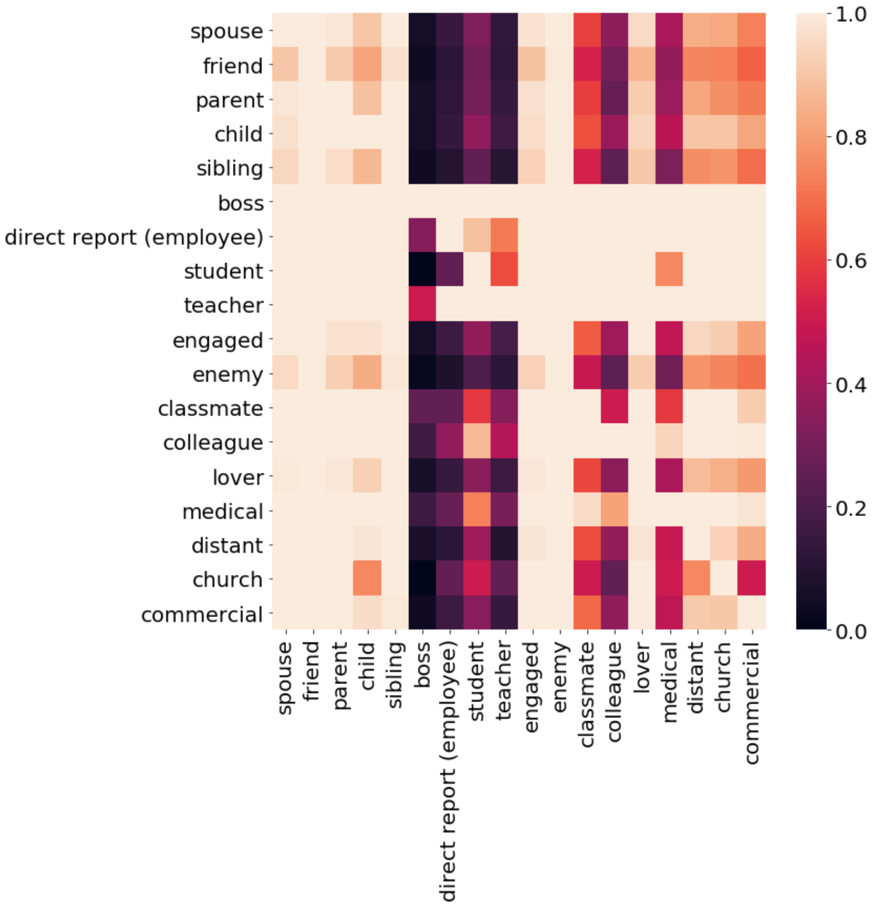


Fig 3. The probability that, given an utterance is appropriate in relationship Y, it will be appropriate in another relationship X.

We build upon this analysis in two ways. The first was trying to rank relationships on appropriateness. We calculated the probability that if you said something to an original relationship and it was appropriate, it would be appropriate over all of the other relationships, and ranked the relationships based on these probabilities. Table 2 shows these results, and the results largely follow common sense. Speaking to a boss, a teacher or a colleague requires a level of politeness not always present when talking to a sibling or a friend.

The last analysis we did was trying to cluster relationships based on the probabilities found in the row-column pairs of Figure 3. This clustering found 3 groups of relationships, the first being mostly familial and social relationships, the second being professional relationships, and the last being an odd trio, Student Colleague and Medical. While some clear similarities exist between the selected clusters, the odd few results mean that perhaps the model does not have enough data to distinguish between the different social groups that make up the various relationships studied.

As we gather more data, we plan to rerun our experiments to see what findings hold and what changes under more robust conditions. A model trained on more varied data captured by annotaters with a wider range of backgrounds might be able to pull out even more interesting insights than our initial results show.

Conclusions

Overall, our initial work shows promising results in the investigation of how large language models can understand and interpret social relationships in the context of language appropriateness. Using a novel data annotation tool, we were able to gather a new dataset for use in future machine social understanding tasks. Our next steps include gathering more data and performing more experimentation. From the completion of this project, further research could investigate language models to find what patterns in language they are learning that lead them to perform so well on tasks that require social understanding, or further investigate the differences that exist between relationships using the computational methods designed in this project.

Ry	P(Q,Rx$\mid$Q,Ry=1) $\forall Q \in Ry$
Boss	1
Teacher	0.97
Employee	0.94
Colleague	0.88
Student	0.87
Medical	0.85
Classmate	0.80
Commercial	0.72
Engaged	0.72
Distant	0.71
Child	0.70
Lover	0.69
Spouse	0.68
Church	0.68
Parent	0.66
Sibling	0.63
Friend	0.63
Enemy	0.62

Table 2. The probability that, for all quotes Q and all other relationships Rx, a quote will remain appropriate if it originated in relationship Ry

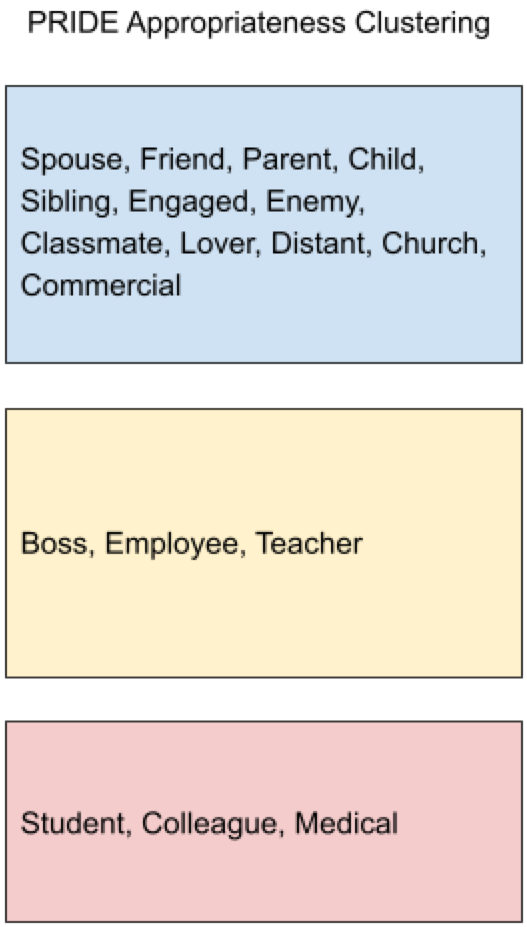

PRIDE Appropriateness Clustering
Spouse, Friend, Parent, Child, Sibling, Engaged, Enemy, Classmate, Lover, Distant, Church, Commercial
Boss, Employee, Teacher
Student, Colleague, Medical

Fig 4. Clustering relationships based on how similar their probabilities are from the analysis done shown in Figure 3.

Citations

[1] W. Warner και J. Hirschberg, ‘Detecting Hate Speech on the World Wide Web’, στο *Proceedings of the Second Workshop on Language in Social Media*, 2012, σσ. 19–26.

[2] D. Kumar, R. Cohen, και L. Golab, ‘Online abuse detection: the value of preprocessing and neural attention models’, στο *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2019, σσ. 16–24.

[3] L. Breitfeller, E. Ahn, D. Jurgens, και Y. Tsvetkov, ‘Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts’, στο *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, σσ. 1664–1674.

[4] S. Trajanovski, C. Atalla, K. Kim, V. Agarwal, M. Shokouhi, και C. Quirk, ‘When does text prediction benefit from additional context? An exploration of contextual signals for chat and email messages’, στο *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, 2021, σσ. 1–9.

[5] E. Mosca, M. Wich, και G. Groh, ‘Understanding and Interpreting the Impact of User Context in Hate Speech Detection’, στο *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 2021, σσ. 91–102.

[6] Baxter, L. A., & Wilmot, W. W. (1985). Taboo Topics in Close Relationships. *Journal of Social and Personal Relationships, 2*(3), 253–269. https://doi.org/10.1177/0265407585023002

[7] Kapoor, H. Swears in Context: The Difference Between Casual and Abusive

Swearing. *J Psycholinguist Res* 45, 259–274 (2016).

https://doi.org/10.1007/s10936-014-9345-z

[8] Greene, K., Derlega, V. J., & Mathews, A. (2006). Self-Disclosure in Personal Relationships. In A. L. Vangelisti & D. Perlman (Eds.), *The Cambridge handbook of personal relationships* (pp. 409–427). Cambridge University Press.

https://doi.org/10.1017/CBO9780511606632.023

[9] T. Schick en H. Schütze, "Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference", *arXiv e-prints*, bl arXiv:2001.07676, Jan 2020.

[10] J. Devlin, M.-W. Chang, K. Lee, και K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. arXiv, 2018.

[11] A. Tigunova, P. Mirza, A. Yates, en G. Weikum, "PRIDE: Predicting Relationships in Conversations", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, bll 4636–4650.