Fine-Tuning LLMs for Formality Style Transfer and Offensive Language Mitigation

Abhinav Jindal jindalab@usc.edu
Aryan Sagar Methil methil@usc.edu
Kartikeya Syal ksyal@usc.edu
Ojasvi Naik onaik@usc.edu
Vedant Singh vedantsi@usc.edu

Abstract

This paper addresses the need for formality style transfer and offensive language mitigation in professional communication, such as educational software, service chatbots, and professional email filters. Despite its importance, this problem has received little attention in the research community. To fill this gap, we present a study on fine-tuning large language models (LLMs), specifically GPT-3 Ada and Davinci models. As no suitable dataset was available, we created our own. We conducted a comparative analysis between the two models and achieved impressive results in addressing this challenge.

1 Introduction

This project uses Natural Language Processing (NLP) techniques, specifically Text Generation and Paraphrasing to convert informal language into formal, and remediate offensive language. There are several project domains that could incorporate both formality text transfer and offensive language mitigation. Service chatbots that must maintain a formal language and avoid inappropriate words, email filters that can transform messages into a professional tone as well as detect offensive language to block such messages, educational software used for teaching, and social media platforms that can automatically remove posts containing offensive language, are some of the major use cases of our project. Consider an example where a non-fluent person writes something like “Can u at least do this work by tom”, this doesn’t seem so formal and should be rephrased as “Could you please do this work by the end of day tomorrow?”.

Remarkably, this problem has not been extensively studied, and we were unable to locate freely accessible datasets to solve this problem. Hence, we propose to create our own custom dataset and fine-tune a LLM model like GPT-3 to adapt to this specific use case.

2 Related Work

For our project, we have not found any existing work that directly addresses this exact problem: Formality style transfer and eliminating offensive or toxic language. However, we did come across previous works that address each of these subtopics individually.

The paper (Mariano de Rivero and Ugarte, 2021) "FormalStyler: GPT based Model for Formal Style Transfer based on Formality and Meaning Preservation" presents FormalStyler, a model based on the GPT-2 architecture for informal to formal style transfer in natural language generation. The model ensures to generate formal text while retaining the original meaning of the input text.

The authors of the paper (Rao and Tetreault, 2018) introduce a new GYAFC dataset and benchmark for formal language style transfer. The GYAFC (Grammarly’s Yahoo Answers Formality Corpus) dataset consists of over 30,000 sentence pairs, each containing a formal and an informal version of the same sentence. They evaluate several state-of-the-art formality style transfer models on this dataset to get a more accurate evaluation of the models’ performance compared to existing metrics. The GYAFC dataset is one of the largest available datasets with informal-formal language pairs.

In the paper (Dale et al., 2021) "Text Detoxification using Large Pre-trained Neural Models", the authors propose a method for text detoxification which involves detecting and replacing toxic or offensive language in text with more neutral alternatives. They use large pre-trained neural language models, specifically BERT and GPT-2, to classify text as toxic or non-toxic, and to generate alternative non-toxic versions. The proposed method is evaluated on several benchmark datasets.

(Madaan et al., 2020) "Politeness Transfer: A Tag and Generate Approach" is a research paper that proposes an approach called Tag and Generate

involves first identifying inappropriate language using a set of predefined tags and then generating a new response that replaces the impolite language with more polite alternatives, using a deep neural network language model. The results show that their approach outperforms existing state-of-the-art methods.

The literature review provided us with a comprehensive overview of the current state of research in our project domain. Our methodology differs from the above-mentioned existing works as our model guarantees the rephrasing of text that is not only formal but also devoid of offensive language or toxicity.

3 Problem Description

The problem addressed in this paper is twofold: formality style transfer and offensive language mitigation. In professional settings, it is essential to communicate formally and respectfully, and the use of toxic language must be avoided. However, achieving this in practice can be challenging, especially when you want to send that frustrated email to your manager to get that long-pending raise.

The first part of the problem involves formalizing text to make it more appropriate for professional settings. This task involves transforming informal language into formal and structured language. For example, sentences with contractions, slang, and colloquialisms should be transformed into formal language, with longer and more complex sentence structures. This transformation must maintain the original meaning and sentiment of the text while improving its formality and professionalism.

The second part of the problem involves offensive language mitigation, which includes identifying and eliminating toxic language from text. This task involves detecting and removing offensive words, phrases, or tones from text. This is important to ensure that communication remains respectful and professional, without causing harm or offence to any individual or group. Our study demonstrates that formality style transfer and offensive language mitigation is not just about removing toxic words but infact about paraphrasing input sentences to formalize them and eliminate any offensive tonality.

Together, these tasks of formality style transfer and offensive language mitigation are essential for various applications like educational software, service chatbots or even professional email filters. However, despite its significance, this problem is not well-researched. Therefore, this paper presents a study on fine-tuning large language models (LLMs) for this specific task, aiming to provide insights and solutions for this important problem.

4 Methods

4.1 Dataset

Since there is no readily available dataset for this particular problem, we decided to create our own custom dataset, which can be used for our project and any future research in this particular domain. To maintain a consistent domain, we have used popular datasets from social media platforms like Twitter (Ganhotra et al., 2020), Reddit (red), Kaggle Jigsaw (jig), GYAFC (Rao and Tetreault, 2018), Wikipedia, and other popular toxic classification datasets. We have also included some custom instances to improve the overall scope and coverage of various cases in our dataset. Using these datasets have allowed us to maintain validity that such sentences are actually used in social media platforms and have scope of being formalized and intoxication, as shown by the target sentences in our dataset.

Our dataset handles long as well as short sentences, informal chat slangs, misspellings, grammatical errors, profanity, etc. Some of the examples of such informal and toxic cases are *"bro u knw what? am so done with dis job dude! I feel like I'm stuck here forever, and I can't stand the thought of coming in every day."*, *"I've had to sign in to my account multiple times, first time ever. It wouldn't take my password. Had to redo. Then creative cloud failed, had to sign back in. What's with the stupid text message and email codes I have to enter!!!? I already use a password for fuck sake!!!!"*.

A critical challenge in dataset creation was ensuring quality, as a poor quality data can result in inaccurate models and flawed decisions, ultimately jeopardizing project success. To maintain dataset quality, we utilized Grammarly and ChatGPT to meticulously review and refine our training dataset. This methodical approach enabled us to generate a reliable and accurate dataset, which ultimately enhanced our model's performance and the overall effectiveness of the project.

Our dataset and code is available publicly.(dat)

4.2 Fine-Tuning Model

4.2.1 Longer Inputs

The fine-tuned model struggled with longer sentences when using the default settings. Issues arose when input sentence length approached or exceeded the maximum token value, resulting in incomplete or illogical outputs. Setting the parameter value to an extreme is not advisable, as the model may fail to generalize effectively. This was demonstrated when our model encountered difficulties when we initially fine-tuned with a dataset of 80 examples and only trained on Ada. This limitation is illustrated in the following examples:

Consider this input which has 29 words that are more than the default max tokens value (16) :

So negative and critical all the time! You need to learn to be a positive and supportive fella if you want to be a good member of this team. -> If you seem at all down in the ongoing, constructive, and enjoyable interactions with

As seen above, the output comes out to be gibberish, and the sentence itself is incomplete. For something which is almost equal to the input length e.g. 30

So negative and critical all the time! You need to learn to be a positive and supportive fella if you want to be a good member of this team. -> It's important to remain positive and supportive in your work, but remember that you shouldn't be stuck in negative or critical emotional states every time you're

The output now comes out to be somewhat logical but it is still incomplete. Let's try even higher max tokens value such as 40.

So negative and critical all the time! You need to learn to be a positive and supportive fella if you want to be a good member of this team. -> The perceived negative behavior in your life is important to address. I worked with a colleague and we identified a number of areas in which to focus resources and improve employee experience in a positive manner. We do

Now the model generates a random story in the output. The output is also incomplete.

To optimize the performance of our fine-tuned GPT-3 models, we conducted experiments with different max tokens values to find the optimal parameter that would balance model generalization and output coherence. However, we faced challenges with this approach due to varying sentence lengths acros s different datasets. After conducting extensive research and trial and error, we discovered that we can dynamically calculate the max tokens for a particular test sentence. When we have data coming from multiple sources causing the data to have variable lengths, we cannot say ahead of time what constant value we should put for the max tokens. However, the output length seems to be similar to input length because even after paraphrasing, dropping the toxic words, replacing some words, the output length in terms of number of words remains similar. So we calculated the number of words in the inputs and outputs of the dataset and computed the absolute difference of the two quantities. The difference turns out to be small and does not vary much across the dataset. After observing the mean and standard deviation of this absolute difference quantity, we came to the conclusion that max tokens can be dynamically set to 10 + number of words in the input sentence. This allowed us to generate coherent and logically complete outputs for a wide range of input sentence lengths, improving the overall performance of the models.

4.2.2 Temperature

Temperature is a vital parameter in fine-tuning, controlling the deterministic nature of a model's output. After careful evaluation, we chose a temperature of 0 for our model, providing deterministic results and enabling simple output comparisons. By setting the temperature to 0, the model generates consistent responses aligned with desired objectives, allowing more reliable rephrasing of informal sentences into formal language, satisfying the majority of users.

Notably, a temperature of 0 limits creativity and variability in the model's output. However, in our case, prioritizing determinism proved more valuable for delivering reliable, context-appropriate responses. This temperature choice ensures our model generates consistent, accurate results, significantly enhancing its practical utility, overall performance, and facilitating reproducible evaluation outcomes.

5 Experimental Results

The first observation that we made was that dataset size matters. Initially, when we fine-tuned the Ada model on the initial 80 samples of our dataset, the GPT-3 Ada model was struggling with longer sequences as well as dealing with chat slangs.

Recall the case of longer inputs:
So negative and critical all the time! You need to

learn to be a positive and supportive fella if you want to be a good member of this team. -> If you seem at all down in the ongoing, constructive, and enjoyable interactions with.
As you can see that the output is incomplete.

Now, after fine-tuning the Ada and DaVinci models on the completed 400-sample dataset and finding the optimal hyperparameter values for the model, the model is giving good results that are grammatically correct and complete.

So negative and critical all the time! You need to learn to be a positive and supportive fella if you want to be a good member of this team. -> The impact of negative and critical behaviors on team members is important, so let's work to improve our communication and team dynamics, so we can all benefit from each other's strengths and experience.

Now dealing with the case of chat slang. With the 80-sample-sized dataset, the Ada model failed to deal with chat slang, which is using small informal words to represent the original English words, for example, using 'u' for 'you'. *u r too dumb to understand this small thing. -> I believe there might be some clarification needs to be provided.*

The fine-tuned Ada model on the final dataset gave this result:

u r too dumb to understand this small thing. -> I believe that you may be lacking in the understanding of this small detail.

The model understands that 'u' refers to 'you'. The DaVinci model gave the following output:

u r too dumb to understand this small thing. -> I recommend reviewing this information to better understand the concept at hand.

Sometimes DaVinci outperforms Ada in handling chat slang. Consider the example and the outputs from Ada and DaVinci respectively:

r u nuts? -> Would you like some nuts?

r u nuts? -> Is there something wrong with you?
Clearly, DaVinci understood the meaning better for this case. Apart from this, DaVinci is more robust than Ada when dealing with unseen and rare toxic ..-30–03–30phrases. In the following example + *Anybody having a colossal fuck on with Spotify ? Keep getting emails about dickheads in USA logging into my account. Songs pause half way through when played through google home device. Signed out of everywhere on multiple occasions and changed password. Still pausing.*
Comparing the outputs of Ada and DaVinci:

Ada -> *Spotify is getting emails about people being involved in a massive dickhead activity in the United States. Songs play half way through when played through Google Home, and the user is being logged into every time.*

DaVinci -> *I am facing issues with Spotify, like getting emails about unauthorized users from the USA trying to log into my account, and the songs are pausing half way through when played through the Google Home device. I have tried signing out of everywhere and changing the password multiple times*

As you can see, Ada is not able to remove the toxic word maybe because the dataset does not have such a word but DaVinci successfully dropped the toxic word.

After our experimentation with several kinds of prompts, we came to the conclusion that our fine-tuned GPT-3 models Ada and DaVinci are good at handling a variety of informal test cases, and as important as tuning the hyperparameters for the model turned out to be, equally important was making a dataset that had multiple types of data so that the model is able to handle whatever test case is thrown at it.

6 Conclusion and Future Work

We see that our created dataset works very well for our defined task and can be used for any future improvements in this domain. Fine-tuning GPT models on our dataset results in highly formalized and toxicity free rephrased sentences which can be used for various applications like customer services, datasets for text generation models, social media filters etc.

There are several potential avenues for future work on fine-tuning LLMs for formality style transfer and offensive language mitigation. This project could be extended to include more diverse datasets, including languages other than English, and to consider the impact of cultural and social factors on formality and offensive language. It could also be extended to explore the trade-offs between performance and interpretability. While LLMs are powerful tools for NLP, they can be difficult to interpret and may lack transparency. Exploring ways to improve the interpretability of LLMs could help increase trust and understanding of their use in professional settings.

7 Division of Labour

The team members, and their respective contribution is: Abhinav Jindal (100 data points generation, Data quality checks, Fine-tuning Ada, Comparative analysis), Aryan Sagar Methil (100 data points generation, Data quality checks, Fine-tuning Davinci, Comparative analysis), Kartikeya Syal (100 data points generation, Quality checks, Evaluation, Comparative analysis), Ojasvi Naik (100 data points generation, Quality checks, Evaluation, Comparative analysis), Vedant Singh (20 data points generation).

References

Formality-transfer-dataset-and-code.

Reddit's toxicity comments.

Toxic comment classification challenge.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jatin Ganhotra, Haggai Roitman, Doron Cohen, Nathaniel Mills, Chulaka Gunasekara, Yosi Mass, Sachindra Joshi, Luis Lastras, and David Konopnicki. 2020. Conversational document prediction to assist customer care agents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 349–356, Online. Association for Computational Linguistics.

Aman Madaan, Amrith Rajagopal Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Annual Meeting of the Association for Computational Linguistics*.

Cristhiam Tirado Mariano de Rivero and Willy Ugarte. 2021. Formalstyler: Gpt based model for formal style transfer based on formality and meaning preservation. *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 1.

Sudha Rao and Joel R. Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC corpus: Corpus, benchmarks and metrics for formality style transfer. *CoRR*, abs/1803.06535.