Offensiveness, Hate, Emotion and GPT: Benchmarking GPT3.5 and GPT4 as Classifiers on Twitter-specific Datasets

Nikolaj Bauer, Moritz Preisig, Martin Volk
University of Zurich, Department of Computational Linguistics
Andreasstrasse 15, 8050 Zurich, Switzerland
nikolaj.bauer@uzh.ch, moritz.preisig@uzh.ch, volk@cl.uzh.ch

Abstract

With the advent of transformer-based Large Language Models, GPT models have shown impressive performance on various NLP tasks without the need for domain-specific fine-tuning. In this paper, we extend the work of benchmarking GPT by turning GPT models into classifiers and applying them on three different Twitter datasets on Hate-Speech Detection, Offensive Language Detection, and Emotion Classification. We use a Zero-Shot and Few-Shot approach to evaluate the classification capabilities of the GPT models. Our results show that GPT models do not always beat fine-tuned models on the tested benchmarks. However, in Hate-Speech and Emotion Detection, using a Few-Shot approach, state-of-the-art performance can be achieved. The results also reveal that GPT-4 is more sensitive to the examples given in a Few-Shot prompt, highlighting the importance of choosing fitting examples for inference and prompt formulation.

Keywords: Classifier, Large Language Models, Social Media, Hate Speech detection, Emotion Classification, ChatGPT, GPT-4, Benchmarking

1. Introduction

With the publication of GPT-3 (Brown et al., 2020) the power of large generative language models and their applicability to a variety of Natural Language Processing (NLP) tasks without the need for fine-tuning has become apparent. The basic idea behind Generative Pre-trained Transformer (GPT) is taking the decoder part of a transformer (the architecture introduced by (Vaswani et al., 2017)) and thus creating a generative model. While earlier versions of GPT ((Radford et al., 2018), (Radford et al., 2019)) struggled to produce long and coherent paragraphs and still relied on fine-tuning in order to perform well on benchmark datasets, GPT-3 and subsequent models produce long and coherent texts and solve many tasks by only using a single prompt.

In continuing the development of GPTs the company OpenAI released GPT-3.5 (aka ChatGPT), an application of GPT-3, which has shown further improvements on a plethora of tasks (Mao et al., 2023) and lead to a revolution on the internet, becoming the fastest growing web platform ever [1]. Thanks to companies like OpenAI and Google generative models (not just for language, but also image and audio generation) are in the public eye, and pose challenges to research communities in various fields to test the limits, capabilities and ethical as well as societal implications of these models (Gozalo-Brizuela and Garrido-Merchan, 2023).

In NLP GPT's capabilities have already been tested on various established benchmarks. In an extensive comparison Laskar et al. (2023) find that GPT-3.5's performance is competitive, but ultimately worse than the state-of-the-art (SOTA) of single-task fine-tuned models. Similarly Kheiri and Karimi (2023) find that both GPT-3.5 and GPT-4 are still outperformed by specific fine-tuned models. However, the authors also show that fine-tuning to GPT-3.5 leads to massive improvements, achieving an increase of 22% in F1-score on sentiment analysis on Twitter. The newest edition of the GPT family, GPT-4, can also handle multi-modal input and has the ability to capture large contexts outperforming SOTA models in various tasks (Bang et al., 2023).

This paper presents an addition to benchmarking GPT by applying it as a classifier in hate-speech and offensiveness detection as well as emotion classification in the Twitter domain. Our results are in line with current literature, in that fine-tuned SOTA models still outperform GPT, although its performance is competitive. The main advantage of GPT is given when the training and test set are not optimally aligned: in the case of the hate-speech benchmark, where the training set is from a different time frame, GPT outperforms other models that rely on accurate training data. We provide all the technical implementations on Github [2].

[1] see https://time.com/6253615/chatgpt-fastest-growing/

[2] https://github.com/Boffl/gpt-classifier

TRAC-2024 Workshop, pages 126–133

2. Methodology

Since a generative model is per se not suited for classification, because it can create out of bounds responses (i.e. responses that are not one of the classes), we needed to modify it to turn it into a useful classifier. One possible way is the approach presented by (Winata et al., 2021). Simply put, they consider the probability over the vocabulary, given the prompt, and compare the probabilites of each of the class labels to find the most likely.

Unfortunately, OpenAI's API does not provide the whole probability distribution over the vocabulary. However, there is the possibility to change a parameter called `logit_bias`, which allows the user to artificially increase the probabilities of words prior to sampling. Thus, by setting the probabilities of the tokens representing the classes that one wishes to predict to a large number, one can make sure that the responses from the model only contain the predefined classes. [3]

When testing an Large Language Model (LLM) on benchmark datasets, it is important to test whether the model has seen the test set during its training. If this were the case, the interpretability of the performance scores on these datasets is difficult. Unfortunately, we were not able to do such a background check and thus our results on the benchmarking have to be interpreted with caution.

Finally it has to be noted, that we used OpenAI's API to access the models `gpt-3.5-turbo-0613` and `gpt-4-0613` in the period from October to December 2023 and January to February 2024, respectively. OpenAI updates the GPT models regularly, thus results might change in the future.

3. Tweeteval Data Sets

For the benchmarking of the GPT models we took the datasets from the Tweeteval Framework (Barbieri et al., 2020). This framework contains benchmark datasets for seven classification tasks on Twitter data in English. It was created in order to standardize the evaluation of current NLP tools, as the plethora of benchmark datasets made an overview of the state-of-the-art difficult. Many models have been tested on these datasets against which we will compare the performance of the GPT models as a classifier [4]. The datasets relevant to our tasks are described in the following subsections 3.1 - 3.3. A summary of the labelling and examples of the contained tweets can be found in Table 1.

3.1. Hate Speech Detection

The data for hate speech detection stems from Task 5 of SemEval-2019 (Basile et al., 2019). The collected data contains tweets from July to September 2018 in the test set, but a large part of the training data comes from a dataset collected for a previous task (Fersini et al., 2018). This, as both Basile et al. and Barbieri et al. (2020) mention, might be the main reason for the relatively low performance of SOTA models on this task and showcases one of the major advantages of using a large language model, as the need for training data falls away.

The task was specifically concerned with hate speech against women and immigrants. This was reflected in the data collection, in which the authors filtered for keywords that target these groups. The test set, against which we measure GPT's performance includes 3,000 tweets, where each target group is represented equally.

3.2. Detection of Offensive Language

The labeled dataset for tweets on offensive language stems from the SemEval 2019 Task 6 (Zampieri et al., 2019b). The dataset is labeled hierarchically, where on the first layer the presence of offensive language is detected, the second layer categorizes the type of offensive language and the third layer identifies the target of the offense (Zampieri et al., 2019a). For our purposes we only need the first layer, which represents a binary label *offensive* and *not offensive*. The test set, against which we test ChatGPT contains 860 tweets of which 28% have been manually labeled as offensive. The annotation process was carried out with a mixture of expert annotators and crowdsourcing. Fleiss' *kappa* among the expert annotators was high for the first hierarchical layer indicating that the annotation guidelines were clear and did not leave much room for ambiguity.

3.3. Emotion Detection

The data for the emotion detection is part of the SemEval 2018 Task 1 (Mohammad et al., 2018) with the name "Affect in Tweets", where the task is to identify the affectual state of a person from a specific tweet. The incorporation into TweetEval involves transforming this multi-label dataset into a

[3] The fact that GPT is working on subword tokens makes the process a bit more complicated. It is described well in this blogpost from which we took inspiration: `https://medium.com/edge-analytics/getting-the-most-out-of-gpt-3-based-text-classifiers-part-one-797460a5556e`. Note that the blog post is about GPT-3 and some adjustments have to be made to adapt to ChatGPT, for details see our github Repo.

[4] The framework's Github page includes a leaderboard showing the performance of models that have been tested (see `https://github.com/cardiffnlp/tweeteval`).

Task	Labels	Examples
Hate-Speech	hate	Whoever just unfollowed me you a bitch
	not-hate	@user You think bots can argue. You're so hysterical.
Offensiveness	offensive	@user And you're just another Twitter asshole. #Muted
	not-offensive	I'm starting to think these things are a cover for #maga
Emotion	anger	These nasty, common women who will bed another women's man [...]
	joy	Counting on you, Queensland. #StateOfOrigin #Broncos #maroons
	optimism	[...] I jumped in the pool of sharks a long time ago. #relentless #resilient
	sadness	All and boy play n0 no play dull and makes.

Table 1: Labels and example tweets from the datasets

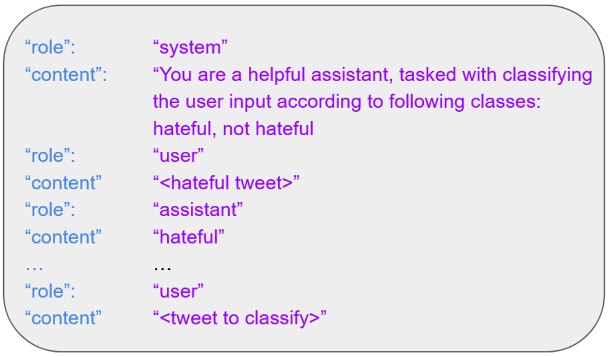
multi-class classification format. This was achieved by retaining only the tweets associated with a single emotion. Due to the limited number of tweets with single labels, Barbieri et al. opted for the four most prevalent emotions anger, joy, optimism and sadness.

4. Experiments with GPT as Classifier

We kept the prompt simple, expanding on OpenAI's default "you are a helpful assistant", by specifying that the task was to classify tweets and the desired labels. Since our approach does not depend on a particular way in which the model provides the answer, as it only considers which of the class labels is more likely according to the model's probability distribution, the specific prompt formulation is not relevant. Additionally, this simple style can be applied to all kinds of classification tasks.

We tested both a zero and a few-shot scenario. In the few-shot case we took 12 examples[5], with the same number of examples for each class, that were given to the model as a chat history in random order. The examples were randomly selected from the training set and both GPT-3.5 and GPT-4 were given the same examples. The text and the corresponding labels for the tweets used in the few-shot prompt can be found in tables 5 to 7 in the appendix. For an illustration of the prompting see Figure 1.

Since identifying hate-speech and offensiveness are related tasks, we also tested "switching the labels", i.e. asking ChatGPT to label the dataset on hate speech as offensive/not-offensive and vice versa. Then we checked with the gold label, if it would align (that is when a tweet is labeled "offensive" by one and "hateful" by the other).



```
"role":       "system"
"content":    "You are a helpful assistant, tasked with classifying
              the user input according to following classes:
              hateful, not hateful
"role":       "user"
"content"     "<hateful tweet>"
"role":       "assistant"
"content"     "hateful"
...           ...
"role":       "user"
"content"     "<tweet to classify>"
```

Figure 1: Prompt sent to OpenAI's API

5. Tweet Classification Results

Table 2 shows the overall performance of GPT-3.5 and GPT-4 in Zero- and Few-Shot setting over each task. To be able to compare our results with other top-performing models, we present the TweetEval-Score, which is based on the evaluations of the original papers of each of the subtasks and represents the macro-F1 score for the tasks Emotion detection, Hate detection, and Offensive language detection.

A first glance reveals that GPT-3.5 does not outperform SOTA models, except on the Hate-Speech dataset. This, however, should not be attributed to GPT's abilities in Hate-Speech detection but rather to the poor performance of the fine-tuned models. As mentioned in section 3.1, the training dataset was obtained at a different time period than the testset[6]. This shows the advantage of a large pre-

[5] The number of examples was chosen since we wanted to give the model the same amount of examples from each class, while also providing the same conditions for all four datasets. We originally had set out to test four datasets that contain 2-4 classes, 12 is the lowest common denominator.

[6] The best model that Barbieri et al. test on the hate benchmark achieves a macro f1 of almost 0.8 on the validation set, showing that the fine-tuned model still has an advantage, if the test data is relevant enough.

Model	Emotion	Hate	Offensive
GPT-3.5 (zero shot)	74.7	42.6	72.6
GPT-3.5 (12 shot)	75.7	**69.9**	67.4
GPT-4 (zero shot)	67.2	64.9	77.0
GPT-4 (12 shot)	**80.5**	62.8	69.9
SOTA	80.2	56.4	**82.2**

Table 2: GPT's performance (F1-Macro) on three datasets from Tweeteval (Barbieri et al., 2020) where GPT is prompted with a simple prompt as in Figure 1. SOTA refers to the current leaders on the Tweeteval leaderboard (see `https://github.com/cardiffnlp/tweeteval`), which at the time of writing are TimeLM (Loureiro et al., 2022) for Emotion and Offensiveness and BERTweet (Nguyen et al., 2020) for Hate-Speech.

trained language model compared to fine-tuned models in the case of data scarcity. Even if the performance in perfect conditions does not beat the best fine-tuned models, in a situation where the training material is not perfect (as it inevitably is in real world applications) ChatGPT outperforms.

Providing GPT-3.5 with examples in the Few-Shot setting does lead to a minor improvement on the Emotion task. In the Hate-Speech detection it leads to a big jump, which is caused by the fact that the dataset limits Hate Speech to women and immigrants, a crucial factor. In the Zero-Shot setting only 45% of the tweets that are labeled as hateful by GPT-3.5 are labeled as such in the dataset. This number jumps to over 60% after seeing the Few-Shot examples. Thus, providing just 12 examples is enough for GPT-3.5 to learn the intended target groups and distinguish Hate-Speech in general from Hate-Speech against these groups. On the Offensiveness dataset the provided examples lead to a decrease in performance. We suspect that the randomly chosen examples were more confusing than helpful. More carefully chosen or handcrafted Few-Shot prompts would have to be employed to check this hypothesis.

GPT-4 performs worse than GPT-3.5 on the emotion classification task in the Zero-Shot setting. However, the Few-Shot performance is on par with SOTA models. On the Offensivness task, GPT-4 outperforms GPT-3.5 in both settings. However, it is also confused by the provided examples. Since both models were given the same examples this is further evidence, that the examples were suboptimal in the case of Offensiveness. Additionally, GPT-4 is more sensible to the examples provided in the Few-Shot case. When the provided Few-Shot examples are helpful (as seems to be the case on the Emotion task), performance of GPT-4 increases over proportionally compared to GPT-3.5. On the other hand, when the provided examples are not fitting precisely (as in the Offensive task), GPT-4's performance is more strongly impaired than that of GPT-3.5.

GPT-4's Zero-Shot performance beats SOTA models on the Hate-Speech dataset. This runs against our expectation, as in the Zero-Shot scenario the model has no information about the specific definition of Hate-Speech in the dataset, with only women and immigrants as targets. This result is evidence of test set contamination. It is possible that GPT-4 has seen the test set during training or in the instruction fine-tuning. The examples that are added in the Few-Shot setting are confusing the model, as they did in the case of the offensiveness task. Since GPT-4 is sensitive to the provided examples, we suspect that peculiarities in the examples lead the model to misclassify the data.

We ran additional prompting GPT-3.5 to look for Offensiveness on the Hate-Speech dataset and vice versa. Interestingly there was not much difference in performance. The results of our GPT-3.5 experiments can be found in the Appendix.

6. Conclusion

Our results are mostly in line with the current literature on benchmarking GPT, showing that it performs well on classifying tweets as hate speech or offensive language. But it is not strictly better than SOTA fine-tuned models. ChatGPT's performance on the hate speech dataset compared to fine-tuned models is impressive. This case shows, how sensitive to training data such fine-tuned models can be. The training data in this case is from the same domain, just from a few years before, thus this setting simulates a real world scenario in which a model is used in an application. On top of that GPT-4's Few-Shot performance reaches SOTA, comparing it to task-specific fine-tuned models. However, our results have to be taken with a grain of salt, as we were not able to check if test set contamination was at play. In fact our results show evidence that GPT-4 has in fact seen the Hate-Speech dataset during training. One solution, short of OpenAI opening up their training datasets, would be to create a new test set, which neither GPT during training nor ChatGPT during instruction fine-tuning can possibly have seen.

Our experiments also show that the performance in a Few-Shot Scenario can be negatively influenced by the additional examples and that GPT-4 is more sensitive to the additional information provided in a Few-Shot prompt. Further work might also systematically explore the effects of different prompts on the performance as well as fine-tuning of GPT, to fully test its abilities as a classifier. For example, we will investigate the reformulation of the classification task as an inference task as proposed by Goldzycher and Schneider (2022).

Additionally, fine tuning generative LLMs on specific tasks and data might still be a fruitful approach. Kheiri and Karimi (2023) have shown massive improvements by fine-tuning ChatGPT on Sentiment Analysis. This indicates that this could also be applied to any other NLP task including the ones we used in our evaluation.

7. Bibliographical References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. ArXiv:2302.04023 [cs].

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *CEUR Workshop Proceedings*, volume 2263, pages 1–9. CEUR-WS.

Janis Goldzycher and Gerold Schneider. 2022. Hypothesis engineering for zero-shot hate speech detection. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 75–90, Gyeongju, Republic of Korea.

Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchan. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. ArXiv:2301.04655 [cs].

Kiana Kheiri and Hamid Karimi. 2023. SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning. ArXiv:2307.10234 [cs].

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. GPTEval: A Survey on Assessments of ChatGPT and GPT-4. ArXiv:2308.12488 [cs].

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI Blog.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI Blog.

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*, volume 30. Curran Associates, Inc.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1415–1420, Minneapolis, Minnesota.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

A. Appendix

Benchmark	metric	Zero-Shot	12-Shot	SOTA
EMOTION	macro F1	0.748	0.758	**0.802**
OFFENSIVE	macro F1	0.716	0.674	**0.822**
OFFENSIVE (offensive x hate)[1]	macroF1	0.706	-	**0.822**
HATE-SPEECH	macro F1	0.426	**0.699**	0.564
HATE-SPEECH (hate x offensive)[1]	macro F1	0.399	-	**0.564**
HATE-SPEECH (specific prompt)[2]	macro F1	**0.720**	0.708	0.564
SENTIMENT	macro Rec.	0.708	0.708	**0.737**

[1] Asking GPT to look for offensive tweets in the hate speech dataset and vice versa.

[2] The dataset labels only hate speech against women and immigrants. This prompt adds this information in the simple prompt

Table 3: ChatGPT's performance on four datasets from Tweeteval compared to the current leader on the dataset according to Barbieri et al. (2020) (see `https://github.com/cardiffnlp/tweeteval`). GPT is prompted with a simple prompt from figure 1 unless indicated otherwise.

Tweet	GPT-3.5	GPT-4	True Label
#Deppression is real. Partners w/ #depressed people truly dont understand the depth in which they affect us. Add in #anxiety &amp;makes it worse	Sadness	Sadness	Sadness
@user Is it just me that thinks it looks boring?	Sadness	Anger	Sadness
Comparing yourself to others is one of the root causes for feelings of unhappiness and depression.	Optimism	Sadness	Sadness
In need of a change! #restless	Optimism	Optimism	Sadness

Table 4: GPT Example Prediction: Examples of Predictions of GPT-3.5 and GPT-4 for the Emotion Subtask in a Zero-Shot setting

Tweet from the Emotion test set	Gold Label
"Some peoples thought process can be very alarming. These nasty, common women who will bed another women's man without conscience . . ."	anger
"When you have just about enough @user in your jar at work for 1/4 of a slice of toast #unhappy"	sadness
"Remember your identity is in Christ. Give the sting of rejection to Christ. He's been there he's done that and He has the scars to prove it."	optimism
"Counting on you, Queensland. #StateOfOrigin #Broncos #maroons #blues #NSWBlues #qld"	joy
"@user customer service is dreadful, phone bill is huge and get passed from person 2 person and keep taking money off my card #idiots"	anger
"I'm pre happy with my Arcadian run, beat a few people I was scared of"	joy
"I'm pale, I no longer wanna laugh, Or smile, All I wanna do is just fucking cry,"	sadness
"Not only has my flight been delayed numerous times, we have not been provided with a snack cart #horrific"	anger
"'When is it going to be that we start to define our own art?' Black music's relationship with literary tradition at Across Cultures #Mix2017"	optimism
"The greatest happiness is seeing someone you like stay happy - Daidouji Tomoyo [Cardcaptor Sakura]"	joy
"Shooting more than ever, making more mistakes than ever but I jumped in the pool of sharks a long time ago. #relentless *#resilient"	optimism
"All and boy play n0 no play dull and makes."	sadness

Table 5: FewShot Examples for the Emotion task that were sent to the model

Tweet from the Hate-Speech Test Set	**Gold Label**
"Austria proposes sending troops abroad to stop migrant movement"	hate
"@user @user Shut up, I heard you were a skank and totally let David M go all the way!"	hate
"@user @user Tell #Israel that Barack. @user #BuildThatWall Sir! Thank you Mr. Trump for fixing the #OBAMAMESS #failed policy. Hey Obama maybe you and #EricHolder can go sell guns to the cartels. Or Sell #UraniumOne Or pay for phony #R"	hate
"Dont forget to #ListenAndBelieve. I’m sure all those lynchings were justified. Women never lie about rape. It’s com. . . "	not-hate
"Money which are payed for social welfare - it increase. Together money for refugees will destroy the budget. People must decide: Welfare less and refugees or welfare without refugees. Both isnt possible."	hate
"@user @user @user When people say it’s the women’s fault that’s so wrong no women wants unwanted sexual advances nor rape."	not-hate
"shut the fuck up you stupud fucking edater , go get some fucking pussy irl you dumb muslim"	hate
"@user You think bots can argue. You’re so hysterical you can’t understand the basic meanings of words. And you’re still losing"	not-hate
"New Mexico Sheriffs arrest 2 NOT-WHITE men, Siraj Wahhaj & Lucas Morten, for holding at least 11 starving children in a heavily armed, pigsty Islamic compound. #EndALLImmigration#DeportThemAll #BuildTheWallhttps://t.co/tAIplXs7CL"	hate
"#CONGRATS @user When you sell out venues in California you have to buss a skank like this! It was only A Matter Of Time! Bless up theindiggnation ... the bands vibration is"	not-hate
"Angela Merkel precarious as Germany’s refugee row intensifies"	not-hate
"Immigrant Families Reunited In New York"	not-hate

Table 6: FewShot Examples for the Hate-Speech task that were sent to the model

Tweet from the Offeniveness Test Set	**Gold Label**
@user @user @user @user And you’re just another Twitter asshole. #Muted"	offensive
"I’m starting to think these things are a cover for #maga It distracts people from paying attention to trump."	not-offensive
"@user She is perfect"	not-offensive
"@user @user @user That’s correct. Talk to the trees and rocks about how great Antifa is. They cant give you any feedback. People can see how sick and demented Antifa is."	not-offensive
"#Trump #MAGA | @user : Jack Dorsey, to his credit, has openly acknowledged that he has a culture within his company that is hostile to conservatives. Now the question is, what is he going to do about it?"	not-offensive
"Honestly we all know he’s stupid but what is actually surprising is how dumb and easily persuaded the people of our country are .. he’ll say shit like this and every white redneck is screaming “YEAH MAGA BABY”"	offensive
"@user @user Question: Hows that gun control laws up the ass they have working out for them?"	offensive
"@user @user Seems like a cool guy"	not-offensive
"@user @user Mxm nigger thinks we give a fuck"	offensive
"@user @user OurCountry is being saved from evil slugs like Hillary Clinton. The Deep State is going down and I can’t wait for Hillary to be brought in front of a Military Tribunal. The penalty for treason is death I believe."	offensive
"@user If Kerry clown is actually doing it and admits he is and it’s a crime. Throw his ass in jail. He should be arrested...RIGHT ???"	offensive
"@user I thought this was more Antifa training at first..."	offensive

Table 7: FewShot Examples for the offensive task that were sent to the model