Hate Speech Detection With LLMs In a Low-Resource Setting

Anonymous ACL submission

Content Warning: This article contains descriptions and examples of hate speech, which some readers may find upsetting.

Abstract

This paper explores the ability of large language models (LLMs) to detect hate speech in low-resource settings with a focus on the Russian language. It specifically evaluates how well models like GPT-3.5 Turbo and LLaMA 2 can classify hate speech against LGBTQ+ individuals and Ukrainian war refugees. Zero-shot, few-shot, and fine-tuning methods are applied to assess model performance in non-English contexts. To address the lack of labelled hate speech data, high-quality data sets were created mainly sourced from Russian social media. While LLMs have some success, they struggle due to the dominance of English in their training data. (Heidloff, 2023) Fine-tuning and instruction-based methods show promise for improving classification accuracy. The study highlights the need for specialized data and training to boost performance in underrepresented languages.

1 Introduction

Online communication is everywhere, making it hard to maintain social harmony and affecting users' mental health due to hate speech (Spence et al., 2023). With the growth of online spaces and artificial intelligence (AI), LLMs have been evolving to handle complicated language tasks. These models mark a new era in addressing complex Natural Language Processing (NLP) tasks and represent a significant advancement, especially for traditionally underserved languages (Ramlochan, 2023). The recent geopolitical conflicts, such as Russia's invasion of Ukraine, have increased online hate speech, highlighting the importance of this research (Thapa et al., 2022; Rule of Law in Armed Conflicts Project (RULAC), 2024).

Attribute	Details
Language	Russian, Surzhik
Validation Method	Cross-labelling
Total Dataset Size	524
Fine-tuning Subset	368
Validation Subset	156
Data Sources	Social Media, AI Generation, News Platforms

Table 1: Dataset Characteristics for Hate Speech on War Victims and LGBTQ+ Community (262 entries each)

In this paper, we explore the ability of LLMs to detect hate speech in low-resource environments focusing on Russian hate speech against LGBTQ+ individuals and Ukrainian refugees. By conducting experiments using GPT-3.5 Turbo (further referred to as GPT-3.5) and LLaMA 2, this study entertains the theory that, despite their primary training in data-rich languages such as English, LLMs can significantly benefit settings considered to be low-resources (Maguerresse et al., 2020). This potential is attributed to methodologies, including few-shot and zero-shot in-context learning.

2 Datasets

The datasets used for the experiments were derived from selected samples from a large amount of online data from social media platforms, YouTube, Odnoklassniki, X (formerly known as Twitter), and VKontakte (Statista, 2024). These platforms were chosen due to their popularity in Russian-speaking communities and the prevalence of hate speech content.

Instead of directly using all collected data, we adopted a hybrid approach: relevant data were gathered from these platforms using carefully curated keywords associated with hate speech, and this data was then manually streamlined and adapted. This approach ensured high-quality data, providing

greater control over contextual relevance and diversity. The context of the ongoing conflict between Russia and Ukraine, coupled with the introduction of anti-LGBTQ+ legislation in Russia, provided a rich source of contemporary examples for the database (Thapa et al., 2022; Trevelyan, 2024).

The datasets also contain partially synthetic data generated by LLMs such as Gemini and ChatGPT-3.5 to enhance and diversify the datasets. This approach was used only for neutral and positive samples since usually the policies of publically available LLMs do not allow the creation of harmful content like hateful sentences (OpenAI, 2024). Additionally, there are some Surzhik examples present to provide more complex cases for the analysis, since it is also used in the context of Russian hate speech against Ukrainians (Andrusyak et al., 2018).[1] The datasets are divided into three main categories: `positive`, `negative`, and `neutral`, each further classified based on the presence or absence of profanity, resulting in six subsets total. Examples of the datasets entries can be found in the Table 2

[2] It is important to acknowledge that annotating such datasets comes with its challenges, particularly due to the emotional complexities inherent in hate topics. The data was annotated during the collection and organized into 6 groups: neutral, positive, or negative and whether a sentence is profane. The Russian-speaking project members extra cross-annotated a subset of 90 entries from both datasets, which were then used for experiments as high-quality samples (Spence et al., 2023).

The datasets have the potential to prove that a small, well-chosen set of examples could significantly improve a model's capability to generate relevant and high-quality content (Hennings, 2023). This is based on the concept of LIMA (Less is More for Alignment), which challenges the common misconception that extensive data is necessary for fine-tuning LLMs. It demonstrates that modern LLMs can be significantly improved for specific tasks with just a few high-quality examples (Zhou et al., 2023).

[1] Surzhik refers to a range of mixed sociolects of Ukrainian and Russian languages used in certain regions of Ukraine (Andrusyak et al., 2018).

[2] The datasets generated in the paper will be published alongside other similar datasets as part of a larger research project.

3 Methods

3.1 Experimental Setup

We utilize two classification tasks in this study: a binary task, where the labels are *hateful* and *not hateful*, and a ternary task, where the labels are *neutral*, *negative*, and *positive*. These tasks are common for hate speech research and form the basis for evaluating the models' performance in detecting and classifying hate speech (Thapa et al., 2022; Pronoza et al., 2021).

Two large language models, GPT-3.5 and LLaMA 2, were used for the experiments. They were evaluated using zero-shot, few-shot, and instruction-based fine-tuning approaches. Zero-shot and few-shot setups were used to assess the generalizability of the models without extensive training. At the same time, fine-tuning was conducted to adapt the models to the specific nuances of Russian hate speech. Both models underwent instruction-based fine-tuning using the mentioned dataset. For GPT-3.5, fine-tuning involved using instructions (prompts) and examples to guide the model's learning, while leveraging OpenAI's user-friendly interface. LLaMA 2, on the other hand, was used with parameter-efficient fine-tuning (PEFT) with Low-Rank Adaptation (LoRA), which reduced computational requirements while maintaining performance (Hu et al., 2021).

For the fine-tuning of LLaMA 2, we employed Low-Rank Adaptation (LoRA) with the following hyperparameters: 4 epochs, a learning rate of 0.0002, batch size of 1, LoRA rank of 32, LoRA alpha of 64, and a dropout rate of 0.05.

For GPT-3.5, fine-tuning was conducted over 6 epochs with a batch size of 4 and a learning rate of 1.00e-5.

3.2 Prompts

[3] Two types of prompts were tested for the experiments: zero- and few-shot prompts.

In the zero-shot setup, the models were tested on their ability to classify hate speech without any prior exposure to the labelled dataset. Example:

> *I'll present a sentence, please label it as 'negative', 'neutral', or 'positive' towards any specific group.*

[3] The severity level prompt, also extra developed for this research, asked the model to rate hate speech on a scale from 1 to 10. This approach faced challenges due to the subjectivity of hate speech perception and inconsistencies across model outputs. This is why it was not considered in further and final experiments.

Few-shot learning involved providing the models with a few annotated examples to improve their understanding of the task. Example:

> *"I'll present a sentence; please label it as 'negative', 'neutral', or 'positive' with a focus on negative sentiments towards the LGBTQ+ community. Respond with one word. Here are some examples: 1. "I hate them!." Label: Negative 2. "This is normal" Label: Neutral 3. "I love those people" Label: Positive*

The full version of the prompts can be found in Appendix A.1.

4 Results

Few-shot and Zero-shot Learning: The experiments assessed the models' ability to detect hatefulness and negative sentiment toward vulnerable groups in zero-shot and few-shot setups. Both LLMs were tested with prompts tailored to improve their understanding of the context of hate speech. The results show that both models perform significantly better in the few-shot setup compared to zero-shot.

GPT-3.5: In particular, fine-tuned GPT-3.5 achieved a 5% higher Accuracy score in the few-shot setting. In Figure 4 (Appendix A.3), we present the test results for both datasets using few-shot and zero-shot prompts on the ternary task, comparing performance before and after fine-tuning.

Few-shot learning effectively improved the recognition of nuanced hate speech, such as subtle derogatory terms often misclassified as neutral in zero-shot setups. Well-crafted examples likely provided the context needed to interpret ambiguous language more accurately. Similarly, fine-tuned models demonstrated greater sensitivity to implied hate speech, enhancing both accuracy and nuanced understanding—key factors for effective content moderation.

These findings highlight the value of high-quality, domain-specific examples for tasks like hate speech detection, particularly when less direct or explicit language is involved (Brown et al., 2020).

For binary classification, the fine-tuned model achieved 0.87 accuracy, slightly improving over the base model's 0.86. It performed well in distinguishing non-hateful content, making it effective for content moderation. In English tasks, the model sustained 0.87 accuracy, with gains in recall for non-hateful content, showing adaptability across languages (see Appendix A.2, Figure 5).

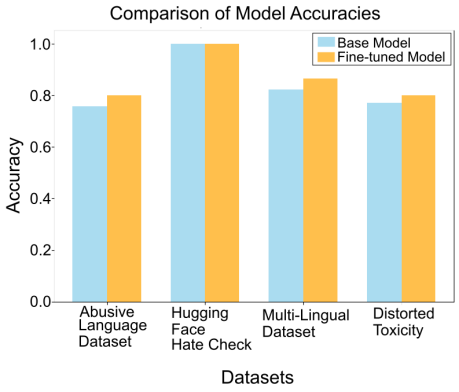

Figure 1: Comparison of Fine-tuned and base Model GPT-3.5 on Outsource Datasets

The fine-tuned GPT model was tested across several datasets (Appendix A.3, Figure 1), showing improved accuracy from 75.5% to 80.0% on the Abusive Language Dataset and perfect accuracy of 100% on the HateCheck dataset (Andrusyak et al., 2018; Röttger et al., 2021). On the Kaggle Multi-Lingual Dataset, accuracy increased from 82.4% to 86.4%, and from 77.0% to 80.0% on the Distorted Toxicity dataset (Gorbunova, 2022). These results highlight the effectiveness of fine-tuning for adapting models to different linguistic environments and improving their ability to handle nuanced hate speech, especially in multilingual settings.

LLaMA 2: The LLaMA 2 model was evaluated on both zero-shot and few-shot learning scenarios for binary and ternary classification tasks. In the zero-shot setup, LLaMA 2 achieved an accuracy of 70% for three-label classification, while it performed slightly better at binary classification, surpassing an accuracy of 0.8 (see 2). However, the few-shot setup showed a drop in accuracy for ternary classification, showing the model's struggle with more nuanced classification tasks. Prompts with more examples, using 6 instead of 3 sentences, show an increase in the accuracy of the model. This suggests that LLaMA 2 benefits from more contextual information, allowing it to better understand patterns and nuances in the data, leading to improved performance.

The results that are seen in Figure 2 suggest that LLaMA 2 maintained consistent performance

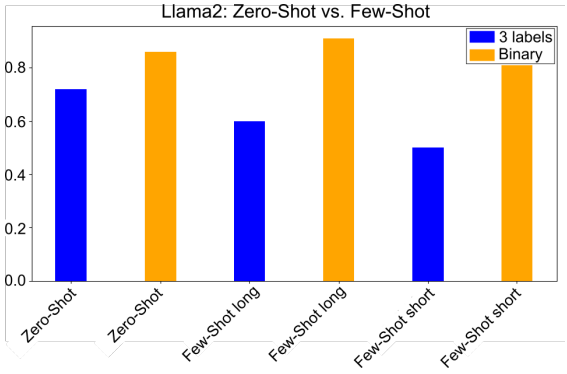

Figure 2: Comparison of Zero-Shot and Few-Shot Accuracy for LLaMA 2 in Binary and Ternary Classification Tasks

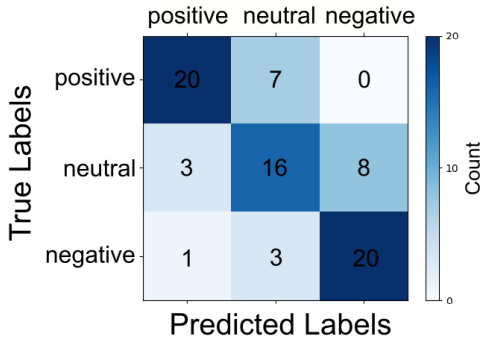

Figure 3: Confusion Matrix based on Test Results for LLaMA 2 with Zero-shot Prompt for Ternary Task

for simpler binary tasks, but its effectiveness diminished in the more complex ternary classification, especially without additional contextual information provided by examples. This emphasizes the need for targeted examples to improve the model's ability to understand complex linguistic nuances. Moreover, the few-shot learning scenario highlighted that providing a small number of high-quality examples can significantly influence the model's understanding and decision-making process, particularly when dealing with subtle language variations.

5 Conclusion

This study found that while models like LLaMA 2 and GPT-3.5 demonstrate impressive language processing capabilities, accurately detecting hate speech in low-resource settings remains a challenging task due to linguistic complexities like context, tone, and intent.

Few-shot vs. Zero-shot Learning A comparison between few-shot and zero-shot learning demonstrates the superior performance of few-shot learning, emphasizing the effectiveness of example-based learning in improving model performance.

Impact of Profanity Profane language significantly affects the ability of base models to distinguish between hate speech and non-hate speech accurately. Although fine-tuning reduces some of this effect, it does not eliminate it. Particularly in three-label classification, models frequently mislabel neutral examples as negative, even after fine-tuning.

Challenges in Identifying Neutral Content Assigning a 'neutral' label remains particularly challenging for both LLMs (See Figure 3). This highlights a specific area where current models struggle, emphasizing the need for future improvements in distinguishing content that falls between explicitly harmful and explicitly benign categories.

Contribution of New Dataset We introduce unique, high-quality Russian datasets focused on underexplored hate speech targets, laying the groundwork for further academic investigation and improvements in automated hate speech detection.

6 Future Work

Despite the availability of hate speech datasets in Russian, significant gaps remain in coverage for specific groups. While some data addresses hate speech targeting Ukrainians and other nationalities, representation of LGBTQ+ issues within these datasets is notably limited. This underscores a broader gap in Russian-language hate speech research. Future investigations should focus on expanding datasets for low-resource contexts, where some target groups are underrepresented. Increasing dataset size could enhance the accuracy of hate speech detection in smaller models.

Further exploration could also examine the performance of other LLaMA 2 variants or LLaMA 3, specifically the 13 billion and 70 billion parameter versions. Additionally, future work could explore alternative fine-tuning techniques beyond LoRA (Low-Rank Adaptation) to improve model performance while maintaining computational efficiency.

7 Limitations

The research faced two main limitations. First, the choice of the LLaMA 2 model with 7 billion parameters was due to limited computational re-

sources, which restricted the model's potential compared to larger alternatives. Second, due to the sensitive nature of hate speech data, not all data was cross-labelled to protect the mental well-being of labellers, reducing the amount of labelled data available for analysis. Additionally, certain examples were excluded to comply with OpenAI's usage policies, further limiting the data used for model fine-tuning. Also, some examples of hate speech identified in the datasets were not acceptable under OpenAI's usage policies and therefore could not be used in the fine-tuning process of the models. This further limited the range of data that could be effectively incorporated into the model training.

References

Bohdan Andrusyak, Mykhailo Rimel, and Roman Kern. 2018. Detection of abusive speech for mixed sociolects of russian and ukrainian languages. In *The 12th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018*, pages 77–84, Karlova Studanka, Czech Republic. Tribun EU.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *Preprint*, arXiv:2005.14165.

Alla Gorbunova. 2022. Automatic toxic comment detection in social media for russian.

Niklas Heidloff. 2023. Language support for large language models. Accessed 11.04.2024.

Mark Hennings. 2023. Fine-tune llama 2.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

OpenAI. 2024. Usage policies. Accessed: 2024-12-26.

Ekaterina Pronoza, Polina Panicheva, Olessia Koltsova, and Paolo Rosso. 2021. Detecting ethnicity-targeted hate speech in russian social media texts. *Information Processing & Management*, 58(6):102674.

Sunil Ramlochan. 2023. How does llama-2 compare to gpt-4/3.5 and other ai language models. Accessed 08.04.2024.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, misc. Association for Computational Linguistics.

Rule of Law in Armed Conflicts Project (RULAC). 2024. International armed conflict in ukraine. https://www.rulac.org/browse/conflicts/international-armed-conflict-in-ukraine. Accessed: 29.11.2024.

Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 17(4):Article 8.

Statista. 2024. Top active social media platforms in russia. https://www.statista.com/statistics/867549/top-active-social-media-platforms-in-russia/. Accessed: 2024-11-29.

Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multimodal dataset for hate speech detection on social media: Case-study of russia-Ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6. Association for Computational Linguistics.

Mark Trevelyan. 2024. Russia adds 'lgbt movement' to list of extremist and terrorist organisations. Accessed 20.04.2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

A Appendix

A.1 Used prompts:

An example of the severity-level prompt:

Rate the severity of hate speech in each sentence on a scale from 1 to 10, where 1 indicates no hate speech and 10 signifies extremely severe hate speech.

Zero-shot prompt:

"I'll present a sentence; please label it as either 'negative', 'neutral' or 'positive' with a focus on negative sentiments towards some group. Respond with one word."

Short few-shot prompt:

"I'll present a sentence; please label it as either 'negative', 'neutral' or 'positive' with a focus on negative sentiments towards the LGBTQ+ community. Respond with one word. Here are some examples:
1. Sentence: "Заебали своим ЛГБТ, не нужна эта зараза тут!. (You are so fucking annoying with your LGBT-shit, we don't need this dirt here!)"
Label: negative

2. Sentence: "Для асексуалов понимание собственных границ и коммуникация с партнерами - охуенно важный аспект отношений.(It is very fucking important for asexuals to communicate and understand their boundaries in their relationships.)"
Label: neutral 3. Sentence: "Отъебитесь уже от людей, это же так охуенно жить в гармонии со своей ориентацией. (Can you just leave people the fuck alone? It is so god-damn amazing to live in harmony with yourself.)"
Label: positive"
Sentence:

Long few-shot prompt:

I'll present a sentence; please label it as either 'negative', 'neutral' or 'positive,' with a focus on negative sentiments towards some specific community or a group of people. Respond with one word. Here are some examples:

1. Sentence: "Заебали своим ЛГБТ, не нужна эта зараза тут!. (You are so fucking annoying with your LGBT-shit, we don't need this dirt here!)"
Label: negative

2. Sentence: "Для асексуалов понимание собственных границ и коммуникация с партнерами - охуенно важный аспект отношений.(It is very fucking important for asexuals to communicate and understand their boundaries in their relationships.)"
Label: neutral

3. Sentence: "Отъебитесь уже от людей, это же так охуенно жить в гармонии со своей ориентацией. (Can you just leave people the fuck alone? It is so god-damn amazing to live in harmony with yourself.)"
Label: positive

4. Sentence: "Беженцам надо как-то блять помочь. (We need to fucking help the refugees somehow)"
Label: neutral

5. Sentence: "И ведь все эти охуевшие мрази, так или иначе, будут выдворены из ЕЭС.(And all these fucking scum will be kicked out of the EEC one way or another.)"
Label: negative

6. Sentence: Беженцы не сидят на жопе ровно блять, а ищут новые возможности для образования и развития в новых странах. (Refugees don't sit on their fucking asses, they look for new opportunities for education and development in new countries.)
Label: positive
Sentence:

A.2 Examples of datasets' entries:

Text	Labels
Беженцам надо как-то помочь.	Neutral, Not profane
Украинцы - это символ надежды, блять	Positive, Profane
Опять АБВГДейку+-= обижают -	Negative, Not Profane

Table 2: Examples of Dataset Entries. Translation: (1) We have to help the refugees somehow. (2) Ukrainians are the symbol of hope, for fuck's sake. (3) Someone offended the ABCDies+-= again.

A.3 Figures:

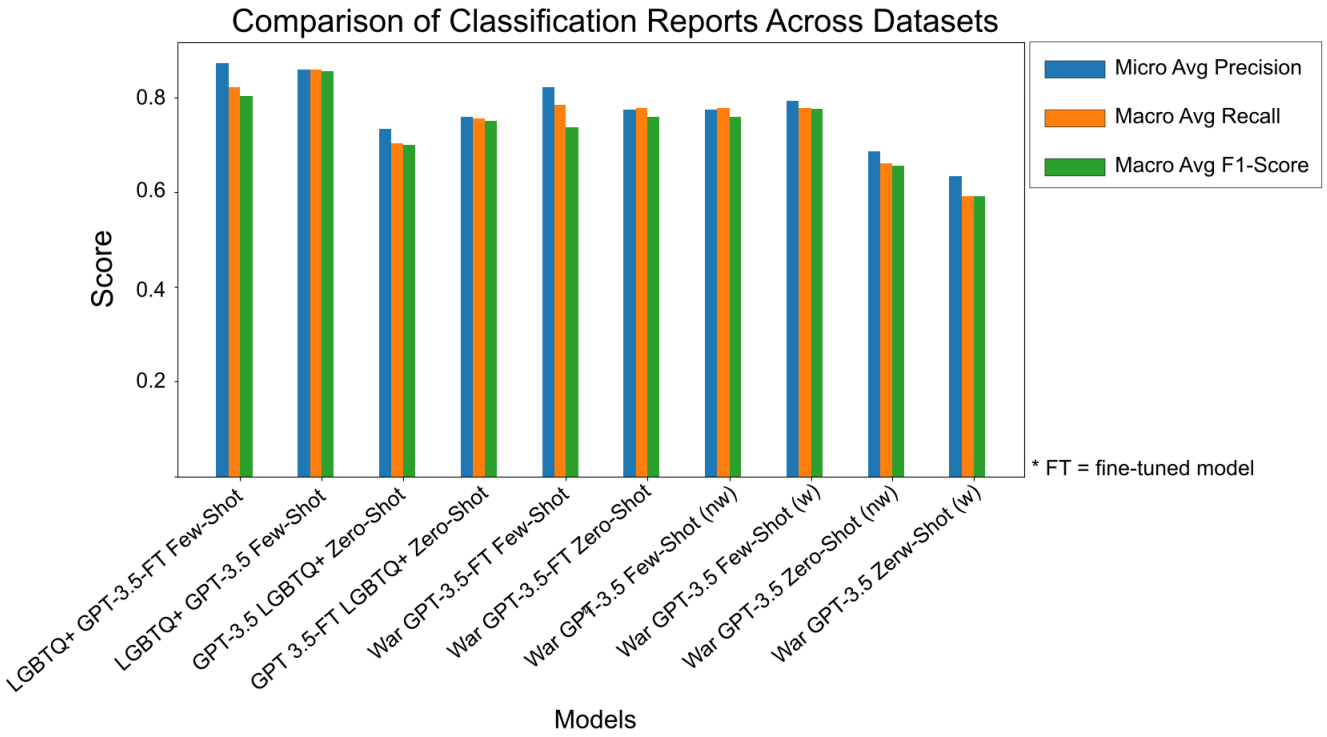

Figure 4: GPT-3.5 Turbo Results for both Datasets on ternary task before and after Fine-tuning

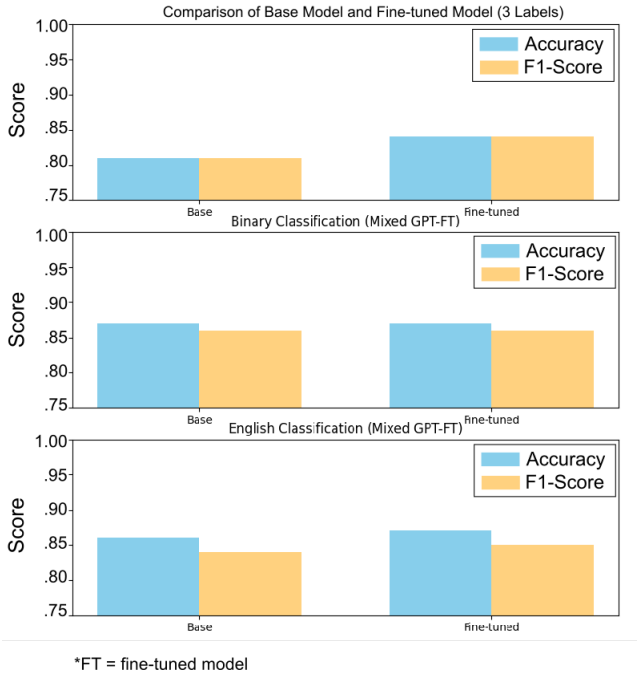

Figure 5: GPT-3.5 Turbo Results for Different tasks