

Fine-Tuning GPT-3.5 for Hate Speech Detection in Albanian YouTube Comments: Challenges and Analysis

Hersi Kopani¹, Rovena Llapushi²

¹Department of Statistics and Applied Informatics, Faculty of Economy, University of Tirana, Albania,
hersi.kopani@gmail.com

²Department of Mathematics "Tullio Levi-Civita" University of Padua, Italy
Rovena.llapushi@studenti.unipd.it

Abstract: Hate speech on social media appears to be inevitable in our era of ever-more connectivity. This presents a distinct challenge because it is difficult to design automated techniques to identify such speech, especially for low-resource languages. We introduce the PAR database, derived from YouTube comments associated with politics, actuality, and reality show content. Our study compares the performance of the GPT 3.5 model applied to these comments, which are written in Albanian, which is considered a low-resource language for NLP models and predicts the performance. We assess the model's ability to detect hate speech across three topics, analyze the performance on each individual topic, and appraise the effect of translating comments from Albanian jargon to standard Albanian on the model's performance. This analysis is crucial because the model utilizes the translated English version as the target language for distinguishing such comments. Lastly, we debate whether large language models like GPT 3.5 are required for this kind of task. Is utilizing a model of this scale for transfer learning that beneficial, even if the model has no knowledge of the Albanian language? Or should we focus more of our attention on engineering and text annotation?

Keywords: *hate speech; social media; low-resource language; fine-tuning LLMs.*

1. Introduction

The number of active users on social media has increased rapidly. Because of that, comments posted on social media platforms like YouTube or Meta are often considered offensive comments or languages that violate the identity or origin of categories. Violating these ethical standards affects the lives of people psychologically or emotionally. In Albania, numerous instances of hate speech have been reported by the media, with several suspected cases of suicide allegedly linked to this phenomenon, as per police investigations.

Coming up with a clear definition of hate speech is a challenge, as human behaviors are complex to analyze. Also, there are many factors, such as way of expression, dialects, writing style, contextual circumstances, and some trending terms, that are present on high-audience TV programs. Moreover, labeling some comments as "hate speech," which may not necessarily be accurate, infringements upon the principles of freedom of expression. The concept of hate speech has been the subject of extensive debate and scrutiny within both political and academic spheres, yet a universally accepted definition has remained elusive (Stakic, 2011), despite the United Nations Strategy and Plan of Action on Hate Speech (2022) defining it as "any kind of communication in speech, writing, or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender, or other identity factor." Within the realm of artificial intelligence (AI) models, particularly in the domain of natural language processing (NLP), the Albanian language is frequently categorized as a low-resource language. Low-resource languages can be understood as languages that are less studied, lacking in resources, less computerized, less privileged, less commonly taught, or characterized by low density, among other denominations. This classification presents numerous challenges due to the limited availability of algorithms and training data resources specific to the Albanian language. These components are crucial for the development of successful AI models capable of preventing hate speech. Currently, social media algorithms prioritize languages such as English or French for the detection of hate speech, resulting in difficulties in identifying and moderating hate comments in the Albanian language. Attempting to translate

Albanian content into English for moderation purposes may introduce inaccuracies, thereby underscoring the complexities associated with addressing hate speech in low-resource languages. This paper provides a comprehensive review of current research efforts aimed at identifying hate speech in the Albanian language. Additionally, it examines the performance of the GPT 3.5 API when applied to the PAR dataset, achieving an accuracy of 83% with a standard deviation of 0.02. Consequently, based on the results obtained, the recommendations for the Albanian language are as follows: 1) construct a high-quality dataset that can serve as a benchmark for identifying hate speech in Albanian social media; 2) involve researchers and Albanian linguistic experts to contribute to the development of a set of rules for the identification of hate speech statements; and 3) develop an efficient tokenizer, which is essential for a robust language model. Working with a large, well-constructed dataset is crucial for producing a high-quality tokenizer, which in turn enables the full utilization of the potential inherent in even large models through fine-tuning processes.

2. Related work

Based on research results and our knowledge, there have been some publications analysing hate speech on social media, including Instagram, Facebook, and YouTube, in the Albanian language. These studies focus on creating datasets by collecting comments from a variety of social media sites, like Instagram or YouTube, running different models, and explaining accuracy by interpreting the results. Referring to the first publicly available SHAJ dataset annotated using OffensEval schema, the accuracy of different models such as NaïveBayes as baseline, BiLSTM and BERT shows that BERT performs better with an F1 score of 0.77. Nurçe, Keci, and Derczynski (2021) concluded that it is difficult to detect hate speech in instances of irony, sarcasm, or when there is a lack of context.

The same database SHAJ is extended and used to build a set of rules applied to BERT models using the ERASER benchmark when Kacuri (2023) explained that the focus of the research was more on explainability and less on calculation and concludes that BERT models trained with rationales performed slightly better in terms of comprehensiveness but were weaker in terms of sufficiency.

In this paper, the dataset used is called PAR, and it is only fine-tuned to address the awareness of the lack of annotated data and the deficiency in knowledge due to limited data availability for language model training. This is particularly crucial because large language models (LLMs) often lack sufficient data for effective application in low-resource languages like Albanian.

3. Dataset

3.1 Purpose

In this paper, the dataset used is called PAR, and it is only fine-tuned to address the awareness of the lack of annotated data and the deficiency in knowledge due to limited data availability for language model training. This is particularly crucial because large language models (LLMs) often lack sufficient data for effective application in low-resource languages like Albanian.

3.2 Data Collection

Dataset PAR¹ focuses on three main topics: politics, reality, and reality shows. The goal of selecting these three topics is to provide the variety of data associated with a large audience of TV shows. Comments are coming from YouTube platforms downloaded from the three main TV channels: TopChannel, KlanTV, and News24. According to Social Blade² results for 2024, two of these TV channels, in particular TopChannel and KlanTV, have the highest number of subscribers on YouTube, and the contents that they share generate high interactivity between users. Meanwhile, after our investigation of the topic that was selected for politics, News24 had the third most commented video. For politics and actuality, it uses the same topic and downloads all comments written on YouTube; however, for reality shows, due to the high number of

¹ <https://github.com/HersiKopani/NLP-HateSpeechAlb>

² <https://socialblade.com/youtube/top/country/al/mostsubscribed>

comments, it is used only on TopChannel, one uploaded series of a dating show, and extracts all comments. After downloading these comments using the external tool Export Comment, the dataset contains 450 rows of data related to politics, 527 rows of data related to reality, and 475 rows of data related to reality shows. uploaded on YouTube, which is the social media platform where the comments are extracted. Regarding ethics and privacy, even though these comments are public and readable by all, on the dataset, the authors are hidden by respecting the privacy of each YouTube user. During data collection, it became evident that numerous comments in the dataset were written in dialects, with a writing style that highlighted various grammar mistakes and dialectical nuances.

3.3 Data Annotation

Before annotating the dataset, first authors wrote the comments grammatically correct, then they followed up the process by annotation rows. Each author has labeled the row T to identify a comment as hate speech and F for the opposite. After annotating all comments from two authors, the results are compared, and discrepancies are verified with the assistance of an Albanian dictionary. Below there are the category records based on annotation results if the comments were hating speech or not.

Table 1. Aggregate results of dataset based on category.

Category	Hate Speech T	Hate Speech F
Politics	185	265
Actuality	188	339
Reality Show	236	239

4. GPT 3.5 Architecture and Methodology

4.1 GPT-n Overview

GPT-n is a language model that operates from left to right, utilizing only the decoder portion of the transformer architecture, without the cross-attention layer. The process begins with word embeddings and positional encodings. Attention mechanisms are then applied to these initial inputs. Following this, several feed-forward layers are used, with regularization steps incorporated at the end. This entire process is encapsulated within a 12-layer transformer decoder. The model is trained using the standard method of cross-entropy loss.

4.2 Fine-tuning

The fine-tuning loss is composed of two parts: the loss specific to the task at hand, and the loss from language modeling. During the fine-tuning phase, the model's structure remains unchanged, apart from the last linear layer. The input format varies depending on the task. Our focus is on single sentence classification. To classify individual sentences, the data is fed into the model as it was during training, and the label is predicted based on the final representation of the last input token.

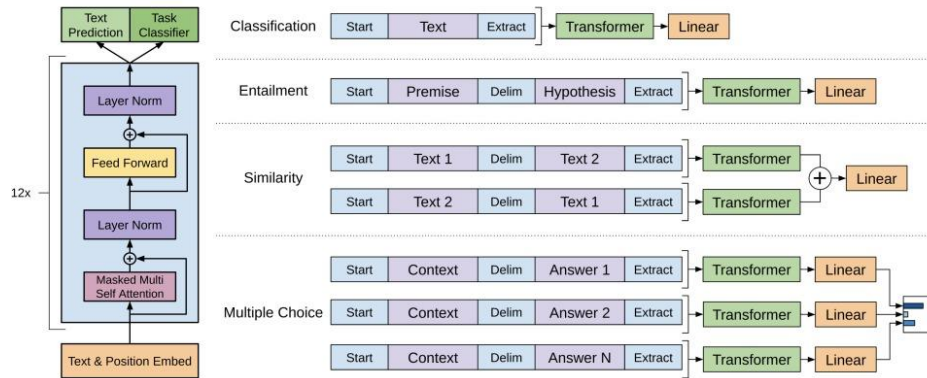


Figure 1. The architectural structure of GPT 3.5 model (Randford, 2018)

4.3 The revolution of GPT-3

The GPT-3 paper by OpenAI³ for 2023, presents a novel concept known as in-context learning, which deviates from traditional artificial intelligence practices. In the conventional supervision paradigm, if we want to classify text as a binary problem, we first create a dataset of positive and negative examples and then train a custom model to make the binary distinction. While this model can be powerful, it may not scale to the complexity of human experience. In-context learning promises that a single, large, frozen language model can serve all these purposes. We provide the model with examples of truthful and deceptive opinions, hoping that it will learn in-context about the distinction we're trying to make. The types of n-shot learning that can be introduced to the GPT-3 model include:

- Zero-shot learning: The model can predict the answer given only the task name with no examples.
- One-shot learning: In addition to the task name and description, we provide the model with one example, and the model will be able to predict the answer.
- Few-shot learning: A few examples are introduced to the model along with the task description.

Another significant innovation is the concept of "self-supervision", a powerful mechanism for acquiring rich representations of form and meaning. In self-supervision, the model's objective is to learn co-occurrence patterns in the sequences it is trained in. The model is simply learning to assign high probability to attested sequences. These models are thought of as generators, but the generation involves sampling from the model. This is powerful because self-supervision requires minimal human effort and has facilitated the rise of another important mechanism, large-scale pretraining.

³ <https://openai.com/index/chatgpt/>

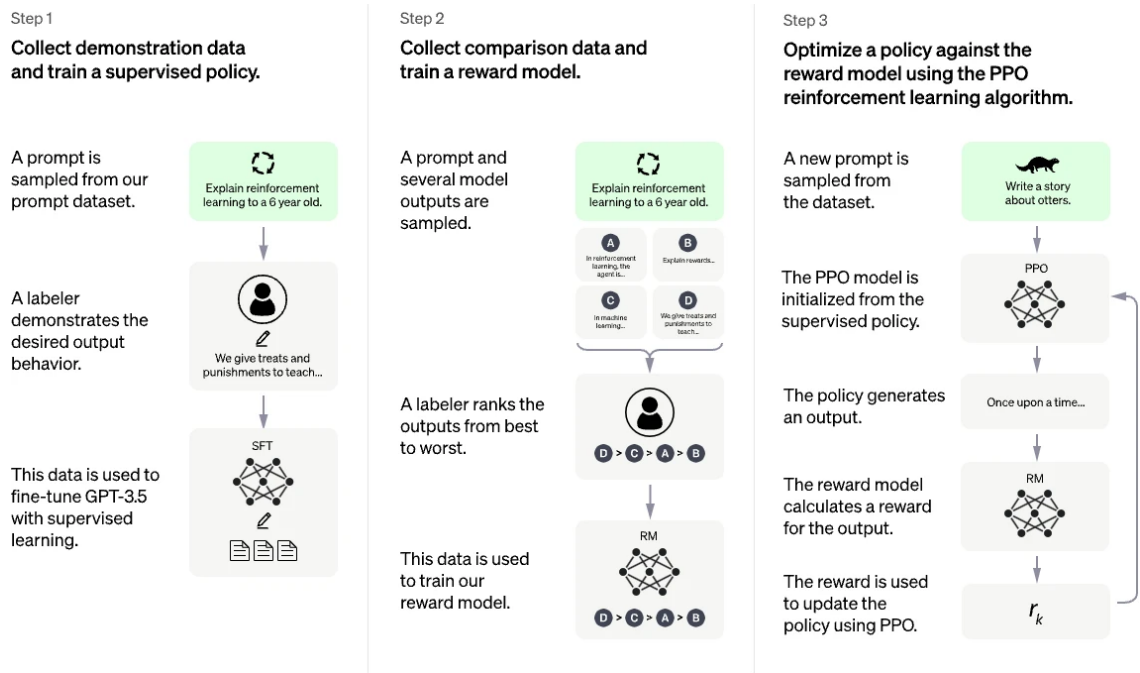


Figure 2. Open AI's schematic representation of the GPT-n LLM (OpenAI, 2023)

The final component is the role of human feedback. The best models, referred to as Instruct models by OpenAI, are trained with more than just self-supervision. Two important aspects from the ChatGPT blog post diagram include:

- The language model is fine-tuned on human-level supervision that represents step 2 on figure, making binary distinctions about good and bad generations.
- In a second phase, the model generates outputs and humans rank all the outputs the model has produced (Step 3). This feedback goes into a lightweight reinforcement learning mechanism. In both phases, there are important human contributions that take us beyond the self-supervision step.

4.3 Methodology

As we mentioned in the section of Dataset, the amount of data is considerably low for the task at hand. To assess the performance of the model, we decided to conduct 4-fold cross validation, splitting the data set into 75% for training and 25% for testing each time, considering a trade-off between the generalization and the cost of fine-tuning such a big model. For the sake of clarity, we use the term best and worst performing model to indicate the fold for which the fine-tuned model shows higher accuracy. Each fold is created using a custom function built for this purpose benefiting from the StratifiedKFold⁴ method published in 2024, ensuring that each fold has roughly the same proportions of each class. To configure GPT-3.5-turbo-0125⁵, we need to specify the following parameters in the API: System, User, and Assistant. These parameters are defined as follows:

- **System role** - "Për një fjali të dhënë, detyra juaj është të përcaktoni nëse komenti i bërë nga një përdorues ka përmbajtje ofenduese (T) ose jo (F)."
- **User** - the comment text
- **Assistant** - the label, True (T) or False (F)

We convert the dataset to a .json file where each row, which contains one opinion, has the following format:

```
{
```

⁴ https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

⁵ <https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset>

```

"messages": [
    {"role": "system", "content": System_role},
    {"role": "user", "content": Comment},
    {"role": "assistant", "content": Labels}
]
}

```

We run 3 epochs for each cross validation. Firstly, we fine-tune the model with the training dataset. In addition, we test the fine-tuned model on the testing set. Last step consists of calculating the metrics: accuracy, precision, recall and F-score.

After these steps are completed, we performed an explainability analysis of False Positive and False Negative. We show the ROC curve for the best and worst performed model. To get insights related to the understanding of the model for Albanian language we used different prompts combined with the tokenization obtained from the algorithm used by GPT-3.5-turbo-0125⁶.

5. Experiment

5.1 Modeling Politics data

As stated, we tried, to the best of our knowledge, to rewrite the comments in a standard written language. Doing so, we could test whether a better written comment could make it understandable for the model and make the right classification. We critically analyzed the dataset to understand which of the categories contained complex comments that would make the decision to use the model more difficult. As we can imagine, the comments under the Politics topic are complicated, as the level of irony and sarcasm is at a higher level compared to the other topics. We used the row data from Politics to test our hypothesis. In the following table, we present the results of each fold, changing only the input we give to the model: first, the comment as it was written, and second, the comment corrected by us, again to the best of our knowledge. Notably, the model has a higher performance when receiving corrected versions of the comment, having as an average accuracy over the fourth folds of 87% and the low standard deviation of 0.01 shows the model's consistency across different folds in the cross-validation process. Shortly, we will report the meaning of the metrics obtained for the first fold of original comments:

- Precision - the model predicts a positive class of benign cases in 68% of the cases. So, there's a 32% chance that a positive prediction is a false positive.
- Recall - the model correctly identifies 70% of all actual positive cases.
- F1-Score - this is the harmonic mean of precision and recall. An F1- Score of 0.69 suggests that the model is reasonably balanced in terms of precision and recall, but there is still room for improvement.

Table 2. Results of GPT 3.5 on Politics dataset

Metric	Original Comments				Corrected Comments			
	1-fold	2-fold	3-fold	4-fold	1-fold	2-fold	3-fold	4-fold
Accuracy	0,74	0,77	0,79	0,83	0,80	0,8	0,87	0,89
Precision	0,68	0,72	0,69	0,78	0,73	0,73	0,82	0,87
Recall	0,70	0,72	0,87	0,83	0,78	0,78	0,87	0,87
F1 Score	0,69	0,72	0,77	0,80	0,76	0,76	0,84	0,87
FP	15	13	18	11	10	16	6	6

⁶ https://github.com/openai/openai-cookbook/blob/main/examples/How_to_count_tokens_with_tiktoken.ipynb

FN	14	13	6	8	13	14	9	6
Avg Accuracy	0.78				0.87			
Std	0.03				0.01			

Out of expectation, we see that the model performs better with the original comments in the second fold. Before moving on with the results obtained by the linguistic analysis, we will plot the ROC curve for the worst and best performing model. The AUC⁷ value reported represents how much the model is capable of distinguishing between classes. It's a statistical measure that we can use to evaluate the model predictions using a probabilistic framework. Again, for the best and worst model we have 89% and 74% respectively, chance of correctly distinguishing between positive and negative class.

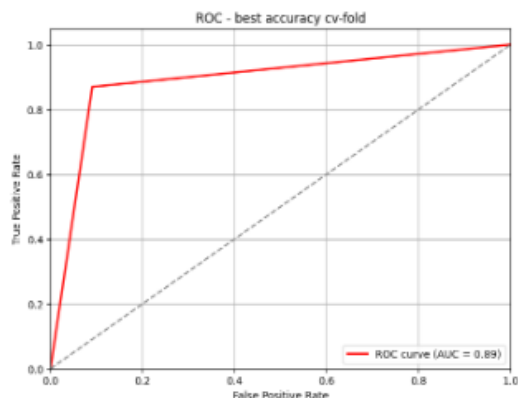


Figure 3. ROC Best Accuracy cv-fold

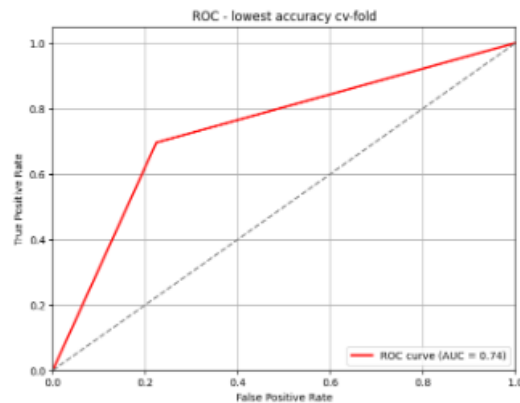


Figure 4. ROC lowest accurate cv-fold

5.1.1 Linguistic analysis of FP and FN

To understand the reason behind this result, we made a check between FP and FN of the original comments and corrected one, to find out the differences. At this point, our investigation proceeds in two steps. Firstly, investigate further the comment for which the model gave the correct labeling to the original comment but the wrong one for the corrected comment. To fulfill this, there are a couple of actions implemented. Primary step, prompt the fine-tuned model to get the labeling. In this way, we can understand if the model has just memorized the example or if it can “reason” by itself. Then analyze the results after applying the tokenization. From this action, we can understand if the specific words are presented differently in the embedding space, depending on the way they are written. The last one asks chatGPT to give an explanation related to the understanding the model has about the embeddings obtained. Investigate the comment for which the model gave the wrong labeling to both versions of the comment. Secondly, investigate the comment for which the model gave the wrong labeling to both versions of the comment. Starting with the analysis of the first comment:

Table 3. The comment details and labels for analyzing step.

Comment	Corrected Comment	Label	Model's Label for Original Comment	Model's Label for corrected comment
---------	-------------------	-------	------------------------------------	-------------------------------------

⁷ <https://www.baeldung.com/cs/ml-accuracy-vs-auc>

MOJ XHULALJA NE SPAKE MALOKU SALI BERISHA KA ME SHUME SE 5 PROKURORE SPIUNE TE TIJE TE FSHEHTE PERSNALE QE SHERBIMET E DINE MOJ MALOKE DIPLOMATE E SALI BERSHURES	Moj, Xhulalja ne SPAK, maloku Sali Berisha ka me shume se 5 prokurore spiune te tij te fshehte personale qe sherbimet e dine, moj maloke diplomatiqe e Sali Berishes.	T	T	F
--	--	---	---	---

The first step consists of using the prompt “Përcakto nëse fjalitë e mëposhtme kanë përmbajtje ofenduese ose shprehin urrejtje”, giving as input the two sentences to the fine-tuned and gpt-3.5-turbo-0125 model, we obtain the correct labeling. Using the prompt “Bëj një analizë të fjalive dhe për secilën listo elementët që mund të kenë përmbajtje ofenduese”, we obtain the main word that triggers the offensiveness “MALOKU”. The second one relies on using the encoding function for the words “MALOKU” and “maloku” we obtain a different tokenization, which could be the main reason why the model didn’t categorize this sentence in a correct way when receiving as input it’s corrected version .

Table 4. Token information of word "maloku"

Example string: “MALOKU”	Example string: “maloku”
Token_integers: [89899, 4012,52]	Token_integers: [14991, 16900]
Token bytes: [b ‘MAL’, b ‘OK’, b‘U’]	Token bytes: [b ‘MAL’, b ‘OKU’]

The last step focuses of using the prompt in English “What can you infer from the example below? Please pay attention to tokenization.”, our aim is to gather as much as possible hints on how this model interacts with the tokens, what meaning gives, and hopefully this could be useful to give a reason why these two words are interpreted in a different word. Here is the answer: “MAL OK U: MAL likely represents the Albanian word for mountain (mal), OK could be an abbreviation or a colloquial term, possibly meaning “okey” or indicating approval”, U could be a pronoun indicating “you in English. Corrected version: mal oku: mal remains the same, meaning mountain, oku is potentially replacing OK, but it might be a form of the verb “oku” meaning “to read”, or it could be a variation or slang term”. We realize that the model is mixing the knowledge from the different languages and lacks understanding of the real meaning of our word. A potential explanation for the misclassification of the corrected comment would be the interpretation of the word as “mal” and “lexo”.

Second comment analysis:

Table 5. The comment details and labels for analyzing step.

Comment	Corrected Comment	Label	Model's Label for Original Comment	Model's Label for corrected comment
Love z Lubonja ♥ ♥ ♥ Intelektual, origjinal, me kurajo civile, Hila nje intelektual provincial qe I pelqen vetja□	Love z.Lubonja ♥ ♥ ♥ Intelektual, origjinal, me kurajo civile! Hila nje intelektual provincial qe i pelqen vetja□	F	T	T

Using the same prompt as in the first case, again we obtain correct labeling of row data, potentially, the model has memorized the examples. The tokenization is done in the same way. No need to make further analysis. We changed our approach in the analysis of this sentence. The approach is explained in the next point.

Using the prompt “Bëj një analizë të fjalisë dhe listo të gjithë elementët që mund të kenë përmbajtje ofenduese.”, in this way the model it is built, will “think” that potentially there should be an offensive element and it tries to guess, ending up in what is called hallucination (Huang et al., 2022). Here is the

answer: “Në fjalitë e dhëna, elementët që mund të kenë përmbajtje ofenduese janë: 1. “Hila” në këtë kontekst, kjo fjalë përdoret si një shprehje me karakter ofendues për të përshkruar dikë si rëndomtë, i pafuqishëm, ose i padëshiruar. 2. “Që i pëlqen vetja” edhe kjo pjesë e fjalisë mund të përmbajë një konotacion negativ, duke implikuar arrogancë ose egoizëm në personin që merret në pyetje”. Unfortunately, we were not able to identify the meaning of this word in other languages. The model is clearly capturing the offensive word in a wrong way.

5.2 Modeling all topics

Moving to the next experiment, we pass as input to the model only the corrected comment, this for two reasons. First, we saw from the first experiment that the model outperforms when having as input the corrected comments. Second, is related to the increased cost, as the tokens when using row data from all the categories are greater in number. In this experiment, we assess the performance of the model when having comments from different topics. The main question is related to the fact that the model can understand the offensive terms used for each of the topics and its ability to gain “general knowledge” for our language. Furthermore, we compare the model’s performance for each category within each fold, emphasizing its performance on Politics row data. The results are reported below:

Table 6. Metrics values of 4 folds on all categories

Metric	Corrected Comments			
	1-fold	2-fold	3-fold	4-fold
Accuracy	0,83	0,85	0,86	0,79
Precision	0,78	0,79	0,84	0,71
Recall	0,82	0,86	0,81	0,87
F1 Score	0,80	0,82	0,83	0,78
FP	35	34	23	55
FN	28	22	29	20
Average Accuracy	0,83			
Std of Accuracy	0,02			

Combining all the topics together leads to an average accuracy 3% less than when having a specific topic with a standard deviation of 0.02. This small standard deviation suggests that there is minimal fluctuation in the model’s performance from one subset to another, reinforcing the reliability of the model. The best model performs on the third fold and the less performant in terms of accuracy is the forth model. Below we report for each fold the metrics related to the topic category:

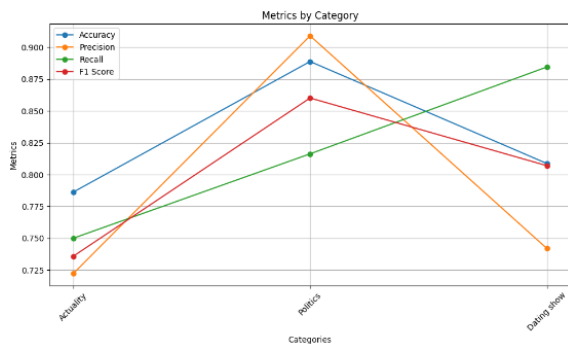


Figure 3. Metrics of 1-fold of all topics

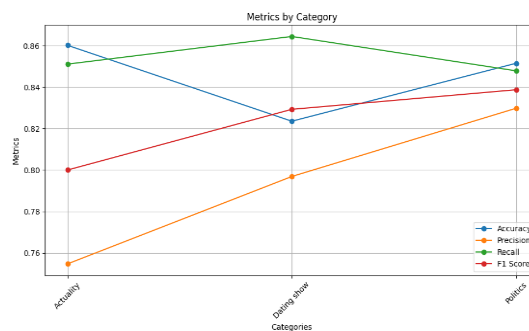


Figure 4. Metrics of 2-fold of all categories.

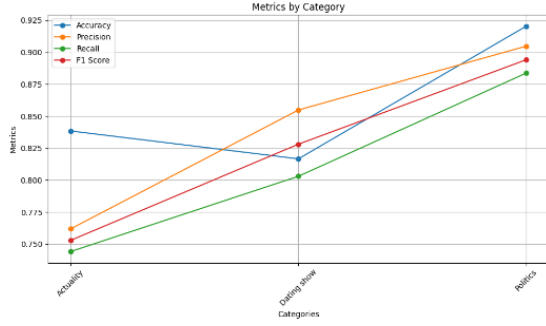


Figure 7. Metrics of 3-fold of all categories.

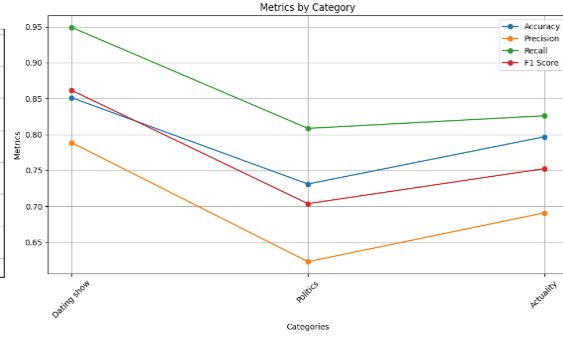


Figure 8. Metrics of 4-fold of all categories.

Overall, we can say that the model performs quite in the same way for the comments related to the Actuality, shows a tendency of low performance related to Dating comments that is related to the nature of the comments that according to us have a tendency of bullying. On the other hand, it shows a good performance for the topics related to Politics, reaching on the second fold an accuracy of 92%. This could be related to the fact that we first fine-tuned the model only on Politics data, and looking at the diagram of how the model works, somehow it is benefiting from the reinforcement learning scheme. Looking at the ROC curve for the best and worst model, we can say that the latter model has a higher degree of class separability, making it the more accurate model by 5%.

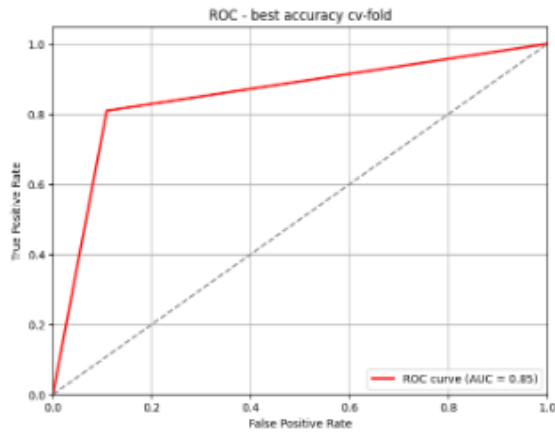


Figure 5. ROC best accuracy cv-fold

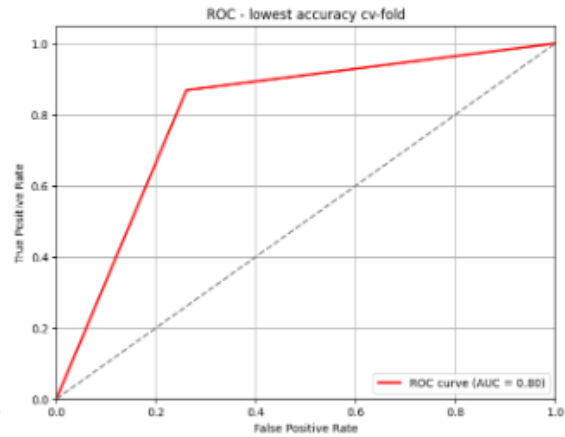


Figure 6. ROC lowest accuracy cv-fold

5.2.2 Linguistic analysis of FP and FN

We analyze the FP and FN in a slightly different way in this experiment. The plots show the number of each class compared to the comment's category:

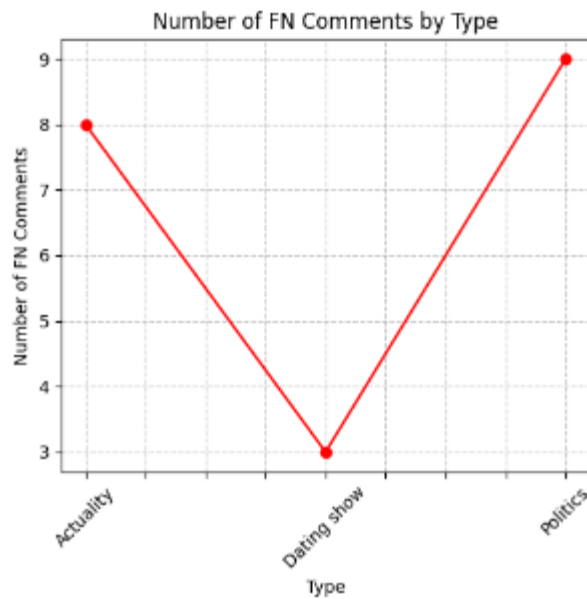


Figure 7. Number of FP Comments by Topics

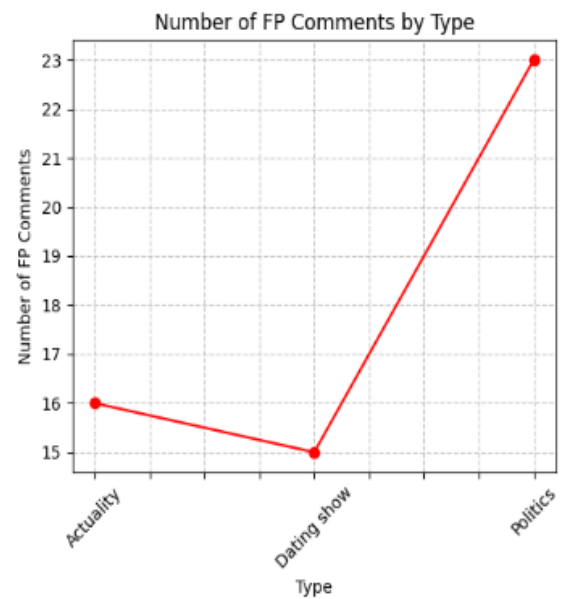


Figure 8. Number of FN Comments by Category

The model has a lower number of FP/FN for the Dating show category and a higher number of FP/FN for the comments related to Politics, as expected from our analysis. Overall, the model tends to predict a positive comment incorrectly, resulting in a higher number of predictions as Hate Speech when they should be labeled as False. Depending on the use case in which we use the model to classify such comments, one can prefer to have more FP than FN or vice versa. If we consider the case of filtering out the offensive comments from social networks, then missing a hate speech is more costly, then we might tolerate more FP to ensure hate speech is not overlooked. As we are dealing only with the corrected comments, we will use a different approach when analyzing the comments under these categories. Below is the list of some of the comments labeled in a wrong way. Note that we selected the ones that are easy to make a distinction between HS and not. In addition, we consider the comments of the split for which the model has a low performance in terms of accuracy.

Table 7. False positives comment sample.

False Positives
Avokati i doktorit tha qe doktori eshte nje hero dhe sulmohet kot nga konkurentet! Hajde shtet, hajde! Ky duhet te ishte ne burg qe para 5 vjetesh.
Çfare fitimi more avokat, e mori ferra kete vend...😞
Taku eshte gjigand! 😁

Table 8. False Negative Comments sample

False Negatives
Shume e neveritshme kjo Enisa
O Hila po ve dhe mendjen ndonjehere ne perdorim, jo gjithmone vetem llapen
Puuuu sa ta shpif ti o Grida me kete te folur. Fol o grua mire, thjeshte dhe qarte.

This investigation follows these steps:

Using the prompt “Përcakto nëse fjalitë e mëposhtme kanë përmbajtje ofenduese ose shprehin urrejtje”, passing the list of the sentences reported above to the fine-tuned and gpt-3.5-turbo-0125 model, we obtained a new label for our sentences. In summary, the second model tends to mark all the sentences as non-offensive, while the first model correctly classified the sentences of the FP class. Still no improvement on the FP class.

Using the prompt “Bëj një analizë të fjalive dhe për secilën listë elementët që mund të kenë përmbajtje ofenduese”, we want to understand what makes the model “to think” in a specific way. The reason given for the second sentence caught our attention: “Ofendimi drejtuar ndaj personit, duke përdorur fjalë vulgare”.

Using the prompt “Analizo në detaje fjalinë 2 dhe listë elementët ofendues, gjithmonë nëse ka të tillë”, we can get the specific word that leads the model to identify this sentence as offensive. The response is as follows “Elementët ofenduese në fjalinë 2 janë: - Fjala vulgare “ferra”.

Using the function available on OpenAI cookbook we tried to analyze how this word is projected into the embedding space. The word “ferra” is a root word that does not contain suffixes. In an optimal tokenization algorithm for our language, this word should be treated as a unique one. The word is tokenized into:

Example string: “ferra”

Token_integers: [69, 14210]

Token bytes: [b ‘f’, b ‘erra’]

Using the prompt “Pse fjala “ferra” konsiderohet vulgare?”, we aimed to get the insight from which language the meaning of this word is taken. Considering the response “Fjala “ferra” konsiderohet vulgare për shkak të përmbajtjes së saj seksuale dhe shprehjes negative që ka në shqip”, we use another prompt in English language “Can you please explain what the meaning of this word is?” to finally get to the desired goal “In Albanian, the word “ferra” is a derogatory term that translates to “bitch” in English”. We found out that the word “Perra” in Spanish has the same meaning and is being tokenized in the same way as the word “Ferra”. In a multilingualism framework, this is a well-known problem for low-resource languages. As these languages benefit when the model is trained in multi-corpora, according to Chang, Arnett, Tu, and Bergen (2023), it can also lead to what is known as the curse of multi-lingually. As dataset sizes increase, adding multilingual data begins to hurt performance for both low-resource and high-resource languages, likely due to limited model capacity. That is exactly what is happening here. As the words have the same token bytes, they share the same position in the embedding, making them have the same meaning, which is right for one language and totally wrong for the other language.

6. Conclusion

Our study's experiments show that significant improvements can be obtained when the GPT-3.5 API is fine-tuned for the purpose of detecting hate speech on Albanian social media. In particular, the accuracy rises from 78% (with a standard deviation of 0.03) to 87% (with a standard deviation of 0.01) when comments are updated to align with the standard Albanian language. The model performs better when focusing just on comments related to the Politics category rather than when it includes comments from three distinct topic areas (83% of accuracy with std of 0.02). Since the model must comprehend and adjust to the subtleties of each issue, it becomes more difficult to detect hate speech effectively as it is exposed to a greater range of language uses, contexts, and patterns. The study reveals that the fine-tuning of the GPT-3.5 API for detecting hate speech in Albanian social media can lead to mixed interpretations of tokens, causing potential misclassifications. It also found that the model's construction might predispose it to anticipate offensive elements, leading to hallucinations or incorrect identifications. Lastly, the research highlighted the challenges of multilingualism, where words with similar token bytes in different languages can confuse the model, a phenomenon known as the curse of multilingualism.

References

- [1] Stakic Isidora, “Homophobia and hate speech in Serbian public discourse: how nationalist myths and stereotypes influence prejudices against the LGBT minority. Journal of Balkan and Near

Eastern Studies”, The Equal Right Trust (ERT), pp.44-462, Vol. 7, May 2021, published by The Equal Right Trust (ERT) Available:

www.equalrightstrust.org/ertdocumentbank/ERR7_isidora.pdf

- [2] United Nations, “What is hate speech?”, September 2022, Published by United Nations, Available: <https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech>
- [3] Erida Nurce, Jorgel Keci and Leon Derczynski, “Detecting abusive Albanian”, 10th May 2022, Published by Cornell University, DOI: :2107.13592,2021, Available: <https://arxiv.org/ftp/arxiv/papers/2107/2107.13592.pdf>
- [4] Kaçuri Muhamet, “Explainability of hate speech classification for Albanian language using rule-based systems and neural networks”, September 2023, Published by Vienna University of Technology (TU), DOI: 10.34726/hss.2023.105780 Available: <https://repositum.tuwien.at/handle/20.500.12708/188586?mode=full>
- [5] Radford, Alex, Narasimhan Karthik, Salimans Tim and Sutskever Ilya, “Improving Language Understanding by Generative Pre-Training” , pg. 4, 2018, Published by OpenAI, Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- [6] Huang Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng et al., “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions” pg.6-12, 19th February 2022, Published by Cornell University, DOI: 0.48550/arXiv.2311.05232, Available: <https://arxiv.org/abs/2201.12345>
- [7] Chang Tyler A., Arnett Catherine, Tu Zhuowen and Bergen K. Benjamin, “When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages”, 15th November 2023, Published by Cornell University, DOI:10.48550/arXiv.2311.09205, Available: <https://arxiv.org/pdf/2311.09205>