1. Summarize the main idea of the project in your own words and explain if the suggested approaches and outcomes seem justified to solve the problem (100- 150 words).

The project aims to identify abusive, toxic, racist, or sexist language on online forums and social media platforms using text classification models, sentiment analysis, named entity recognition, topic modeling, and sequence labeling. The goal is to flag abusive comments and phrases and reduce derogatory language online, which could benefit groups such as women, children, teenagers, people of the LGBTQ+ community, journalists, and activists. The suggested approaches and outcomes seem justified to solve the problem of identifying abusive and toxic language online. The use of various NLP techniques such as text classifiers, sentiment analysis, named entity recognition, and topic modeling are the correct methods which can help in identifying and flagging abusive comments, words, and phrases used on various online platforms.

1. How would you grade the novelty of the project in terms of the data, models, or problem? (Select one option from below)

Highly innovative/ Neutral/ Not innovative

This project is neutral in terms of the models because there has already been a lot of Github models ready to solve your problem. This is the URL: <https://github.com/topics/offensive-language>.

One example is to use different kinds of ensemble models with voting systems to identify offensive language in Twitter. This is the URL: <https://github.com/quinnbp/WT2018>.

This project is neutral in terms of the problem as well because it is a relatively common application of NLP techniques to identify abusive phrases online. There are many similar projects and research studies in this area.

One real-life example towards this problem is the use of AI to detect hate speech on social media platforms using Google’s BERT. The example talks about developing algorithms to detect and alert us to toxic and inflammatory comments and flag them for removal. The algorithms use machine learning techniques such as natural language processing, such as BERT, to identify patterns in the text that are indicative of toxic content. The algorithms are trained on large datasets of labeled data to learn what constitutes toxic content and what does not. This is the URL: <https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/>.

1. Provide suggestions/ feedback to improve the project outcomes or the writing (50 words minimum)

Here are some suggestions to improve the project outcomes:

1. It is important to ensure that the data used in the project is of high quality and representative of the target population. Consider using multiple datasets from different sources to ensure diversity and avoid bias. Additionally, it is important to ensure that the data is properly labeled, which is essential for the success of supervised learning models.
2. While the proposed NLP techniques such as text classification, sentiment analysis, and named entity recognition are useful, it may be beneficial to explore additional techniques such as machine translation, word embeddings, and attention-based models. These techniques may provide additional insights into identifying abusive texts online and improve the innovation.
3. Consider using pre-trained models or transfer learning techniques to leverage existing knowledge and improve the model's accuracy. You can use pre-trained models such as BERT, ELMO, or GPT, which have achieved state-of-the-art performance in various NLP tasks.
4. I do not understand how feature engineering is able to evaluate and compare the models, however one of the common practices in classification is using the confusion matrix for model evaluation. You can also consider using other metrics such as precision, recall, F1 score, ROC curve, and AUC score, which can provide a more comprehensive evaluation of the model's performance. You can also perform cross-validation or hyperparameter tuning to ensure that the model is robust and generalizes well to unseen data.
5. To make the tool more user-friendly and accessible to the public, you can consider developing a web or mobile application that allows users to input their suspected text and get the result. You can also provide visualizations or explanations that help users understand how the model works and what factors affect the prediction.
6. While it is important to identify abusive comments and language, there is a risk of false positives and false negatives, which can result in unintended consequences. False positives may flag innocuous comments as abusive, leading to censorship and a chilling effect on free speech, while false negatives may fail to flag abusive language, allowing it to continue unchecked. It is important to consider these potential harms and develop methods to minimize them.
7. If the project involves scraping online comments or other data from users, it is important to ensure user privacy and obtain their consent. It is particularly important to read through the terms and conditions of Twitter about scraping to ensure you do not violate the terms or even the law.

Regarding research article to review, I suggest this one for your project on detection to trigger a brainstorm of an innovative idea that nobody has done yet:

Alrashidi, Jamal, A., Khan, I., & Alkhathlan, A. (2022). A review on abusive content automatic detection: approaches, challenges and opportunities. PeerJ. Computer Science, 8, e1142–e1142. https://doi.org/10.7717/peerj-cs.1142