Hello everyone! We're students from the Department of Artificial Intelligence and Data Science at B V Raju Institute of Technology, and we're excited to share our project: Real-Time Hate Speech Detection with Robust DistilBERT-SVM. Social media platforms like Twitter are incredible for connecting people across the world, letting us share ideas, opinions, and stories instantly. But there's a darker side—hate speech has become a growing problem, spreading harmful content that can deeply hurt people's feelings, cause mental stress, and even lead to real-world conflicts. For example, hateful comments targeting someone's race, religion, or gender can make online spaces feel unsafe, discourage people from expressing themselves, and sometimes even fuel violence or discrimination in communities. We were motivated to tackle this issue because we believe everyone deserves a safe space to connect online, and technology can play a big role in making that happen. Our project aims to detect hate speech quickly and accurately, helping to create a more inclusive and respectful digital world.

We developed a smart system that can identify hate speech in tweets almost instantly, which is so important because harmful content needs to be caught early before it spreads and causes more damage. One challenge we faced was that hate speech is often rare compared to other types of posts, making it harder to spot, but we worked hard to ensure our system could handle this and fairly detect harmful content across different situations. Another challenge was ensuring the system was fast enough for real-world use, because moderators need quick results to take action right away. Our system delivers results in just 1.5 seconds, making it practical for live moderation! It's also highly effective, correctly identifying hate speech 94 out of 100 times, which is better than many other approaches out there. Plus, it improved at catching even the rarest harmful content by 10%, ensuring we don't miss anything that could hurt someone or disrupt online harmony.

This project supports SDG 16: Peace, Justice, and Strong Institutions by reducing online violence and helping create safer digital communities. It's a step toward using technology responsibly to protect people from harm and build a more peaceful online world. For example, by catching hate speech early, we can stop it from spreading and causing bigger problems like bullying, discrimination, or even violence in communities. Beyond that, our system has real-world applications—it can help social media platforms improve their moderation, support schools in preventing cyberbullying, and even assist organizations working to reduce online radicalization. We also thought a lot about the ethical side of our work, making sure our system is fair and doesn't accidentally target innocent users, because trust and fairness are key in technology like this. Looking ahead, we want to expand our system to work in multiple languages, so it can help people all over the world, and improve its ability to understand tricky things like sarcasm or cultural differences. We're also eager to collaborate with social media companies, researchers, or policymakers to make an even bigger impact. Thank you for listening—we'd love to hear your questions!

Cg

Good [morning/afternoon] everyone, I am [Your Name], and I am excited to present our research on **real-time hate speech detection using a hybrid DistilBERT-SVM model**. Social media has become a major platform for discussions, but unfortunately, it is also a space where hate speech spreads rapidly. Detecting and moderating harmful content is a major challenge due to **the complexity of language, context dependence, and large volumes of data**. Traditional methods, such as keyword-based filtering or basic machine learning models, often fail to capture **subtle and implicit hate speech**, making moderation ineffective. Our research aims to **build an intelligent, real-time detection system that is both highly accurate and computationally efficient**.

To tackle this challenge, we propose a **hybrid model combining DistilBERT and SVM**. **DistilBERT**, a lightweight transformer model, is used for **context-aware feature extraction**, allowing it to understand the deeper meaning of words in different contexts. These embeddings are then passed to an **SVM classifier**, which is computationally efficient and robust for text classification. One key issue in hate speech detection is **class imbalance**—hate speech is much less frequent compared to neutral or offensive speech, leading to biased predictions. To overcome this, we use **MarianMT-based data augmentation** to create balanced datasets and apply **adversarial training (FGSM)** to enhance model robustness.

Our system is designed for **real-time performance** and is deployed as a **Flask-based web application**, allowing users to classify tweets instantly. The model achieves an accuracy of **94%**, significantly outperforming traditional approaches such as **Logistic Regression (90%) and Naïve Bayes (85%)**. The model is also capable of detecting **subtle hate speech that may not contain explicit offensive words but carries harmful intent**, making it a valuable tool for automated moderation.
Beyond accuracy, interpretability is a crucial aspect of our research. We are incorporating **SHAP and LIME analysis** to provide explainable AI (XAI) insights, ensuring that the model's decisions are transparent and trustworthy. This is especially important for **content moderators and policymakers**, who need to understand why a particular piece of content was flagged. Furthermore, our future work focuses on **multilingual support**, enabling the system to detect hate speech across different languages and cultural contexts, making it more globally applicable.

In conclusion, our research presents an **accurate, real-time, and explainable** hate speech detection system that bridges the gap between **cutting-edge NLP and practical implementation**. By leveraging the power of **transformers, machine learning, and adversarial robustness**, we provide a solution that can be integrated into social media platforms to **create a safer online space**.
Thank you for your time, and I look forward to any questions!