GenAI- Unit1-Submission 2-Project1

NAME:DHRUV HEGDE	SRN: PES2UG23CS172	SECTION: C

PROBLEM STATEMENT

Privacy Policy TL;DR
Goal: Summarize the 50-page Terms of Service into "What data do they steal?".
Tech: Summarization.

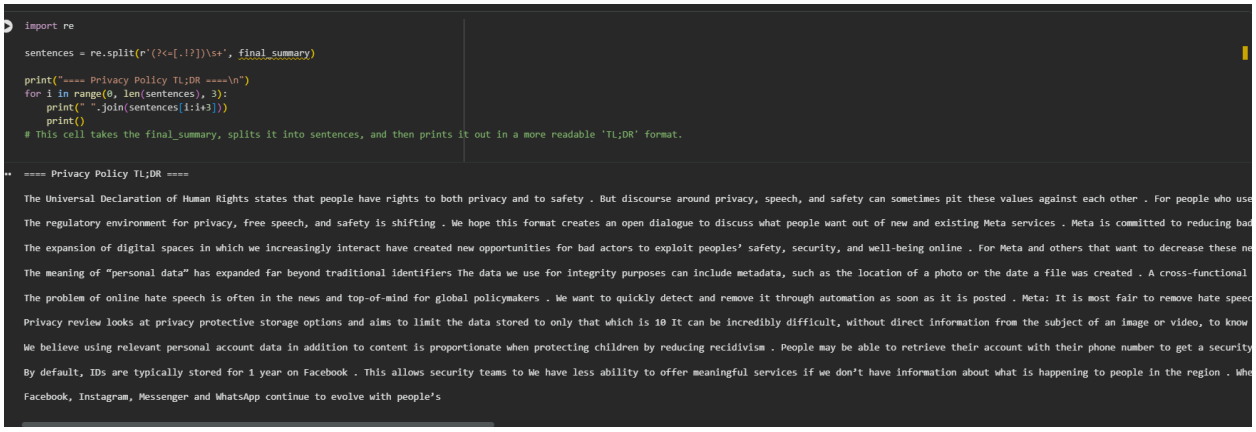
ABSTRACT

This project presents a Privacy Policy TL;DR system that automatically summarizes long and complex privacy policy documents into concise, readable text. Using transformer-based abstractive summarization, the system reduces lengthy Terms of Service or Privacy Policy PDFs into key points that highlight what data is collected, how it is used, and how users can control it. The goal is to help users quickly understand privacy implications without reading the entire document.

PROJECT DESCRIPTION

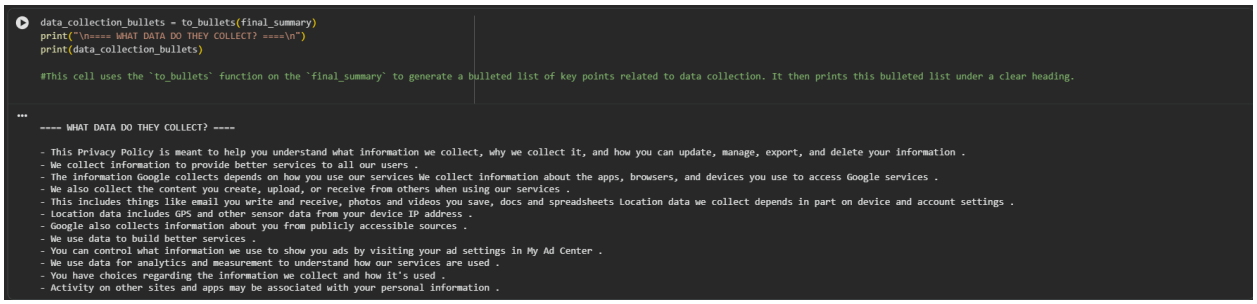
Privacy policies are often long and difficult for users to read and understand. In this project, I built a system that takes a privacy policy document as input and first cleans the extracted PDF text by removing extra newlines and unnecessary spaces. The cleaned text is then split into manageable chunks, since summarization models have a maximum input length. I used the summarization pipeline from the Hugging Face Transformers library with the shleifer/distilbart-cnn-12-6 model. The system dynamically adjusts the summary length based on the document size to preserve important information. The final output is a readable TL;DR summary, which can optionally be formatted into paragraphs or bullet points for better clarity.

SAMPLE OUTPUT



```
import re

sentences = re.split(r'(?<=[.!?])\s+', final_summary)

print("==== Privacy Policy TL;DR ====\n")
for i in range(0, len(sentences), 3):
    print(" ".join(sentences[i:i+3]))
    print()
# This cell takes the final_summary, splits it into sentences, and then prints it out in a more readable 'TL;DR' format.
```

```
==== Privacy Policy TL;DR ====

The Universal Declaration of Human Rights states that people have rights to both privacy and to safety . But discourse around privacy, speech, and safety can sometimes pit these values against each other . For people who use

The regulatory environment for privacy, free speech, and safety is shifting . We hope this format creates an open dialogue to discuss what people want out of new and existing Meta services . Meta is committed to reducing bad

The expansion of digital spaces in which we increasingly interact have created new opportunities for bad actors to exploit peoples' safety, security, and well-being online . For Meta and others that want to decrease these ne

The meaning of "personal data" has expanded far beyond traditional identifiers The data we use for integrity purposes can include metadata, such as the location of a photo or the date a file was created . A cross-functional

The problem of online hate speech is often in the news and top-of-mind for global policymakers . We want to quickly detect and remove it through automation as soon as it is posted . Meta: It is most fair to remove hate speec

Privacy review looks at privacy protective storage options and aims to limit the data stored to only that which is 10 It can be incredibly difficult, without direct information from the subject of an image or video, to know

We believe using relevant personal account data in addition to content is proportionate when protecting children by reducing recidivism . People may be able to retrieve their account with their phone number to get a security

By default, IDs are typically stored for 1 year on Facebook . This allows security teams to We have less ability to offer meaningful services if we don't have information about what is happening to people in the region . Whe

Facebook, Instagram, Messenger and WhatsApp continue to evolve with people's
```



```
data_collection_bullets = to_bullets(final_summary)
print("\n==== WHAT DATA DO THEY COLLECT? ====\n")
print(data_collection_bullets)

#This cell uses the `to_bullets` function on the `final_summary` to generate a bulleted list of key points related to data collection. It then prints this bulleted list under a clear heading.
```

```
==== WHAT DATA DO THEY COLLECT? ====

- This Privacy Policy is meant to help you understand what information we collect, why we collect it, and how you can update, manage, export, and delete your information .
- We collect information to provide better services to all our users .
- The information Google collects depends on how you use our services We collect information about the apps, browsers, and devices you use to access Google services .
- We also collect the content you create, upload, or receive from others when using our services .
- This includes things like email you write and receive, photos and videos you save, docs and spreadsheets Location data we collect depends in part on device and account settings .
- Location data includes GPS and other sensor data from your device IP address .
- Google also collects information about you from publicly accessible sources .
- We use data to build better services .
- You can control what information we use to show you ads by visiting your ad settings in My Ad Center .
- We use data for analytics and measurement to understand how our services are used .
- You have choices regarding the information we collect and how it's used .
- Activity on other sites and apps may be associated with your personal information .
```