TnCSimplifier - Summarizes Terms and Conditions/ Privacy policies

Devansh Goswami
devansh21460@iiitd.ac.in

Ritisha Singh
Ritisha21089@iiitd.ac.in

Samanyu Kamra
samanyu21487@iiitd.ac.in

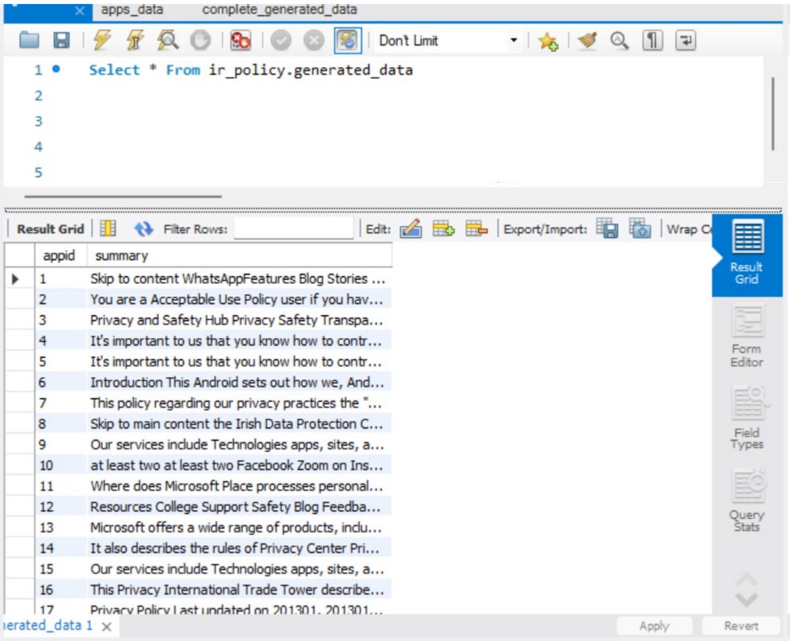
Shriya Verma
shriya21490@iiitd.ac.in

1 Updated Baseline Results

We evaluated our dataset using five different LLM models: Pegasus, BART, and three variants of T-5. After assessing the cosine similarity scores, we determined that Pegasus outperformed the other models for summarizing privacy policies and terms and conditions.

1.1 Updation of database:

Following the selection of the Pegasus model, we utilized it to condense the privacy policy texts contained within our database. These condensed summaries were then incorporated into a new table named 'generateddata'. This table serves as a valuable resource for future use, allowing users to access pre-generated summaries without the need for the model to regenerate the output each time.



```
Select * From ir_policy.generated_data
```

appid	summary
1	Skip to content WhatsAppFeatures Blog Stories ...
2	You are a Acceptable Use Policy user if you hav...
3	Privacy and Safety Hub Privacy Safety Transpa...
4	It's important to us that you know how to contr...
5	It's important to us that you know how to contr...
6	Introduction This Android sets out how we, And...
7	This policy regarding our privacy practices the "...
8	Skip to main content the Irish Data Protection C...
9	Our services include Technologies apps, sites, a...
10	at least two at least two Facebook Zoom on Ins...
11	Where does Microsoft Place processes personal...
12	Resources College Support Safety Blog Feedba...
13	Microsoft offers a wide range of products, inclu...
14	It also describes the rules of Privacy Center Pri...
15	Our services include Technologies apps, sites, a...
16	This Privacy International Trade Tower describe...
17	Privacy Policy Last updated on 201301, 201301...

Figure 1: Generated Summaries using Pegasus

Preprint. Under review.

2 Data Pre-processing

Our preprocessing function aims to prepare text for input into the PEGASUS summarization model, with a focus on handling named entities and facilitating intelligent chunking for simplifying terms and conditions.

The preprocessing includes:

1. Removal of URLs, HTML tags, and decoding of HTML entities.
2. Named entities are temporarily replaced with placeholders to prevent them from being split during subsequent processing steps.
3. All Special characters, excluding common punctuation marks, and the placeholders for entities are removed.
4. The text is split into sentences and rejoined to ensure proper sentence boundaries.

This preprocessing technique is optimal for inputting terms and conditions into PEGASUS because the removal of extra elements and simplified structure of the text allows the model to focus more effectively on summarizing the essential content of the terms and conditions text.

3 Pegasus Model

Pegasus is a specific model that gives generated abstracts based on the article which means the model reads the article text and writes a suitable headline. This abstractive text summarization is one of the most challenging tasks,involving understanding long passages, information compression, and language generation.

The Pegasus model is impressive symbolic reasoning, as researchers concluded based on performance and evaluation results. Some variations show not just the human-level of natural language understanding, even beating the human performance in generating summarized output.

3.1 Architecture:

The current practice for this task would be to train a language model by predicting the masked out token at the end of the sequence. The task is known as self-supervised autoregressive language modeling. Next, fine-tune it on a dataset of labeled summaries and have the generalized output of an article or a document that hasn't seen on the training test, and contains the most information from the source documents. The aim of fine-tuning this kind of architecture is to gain the best results on downstream NLP tasks such as yelp-review, natural language inference, and question-answering.

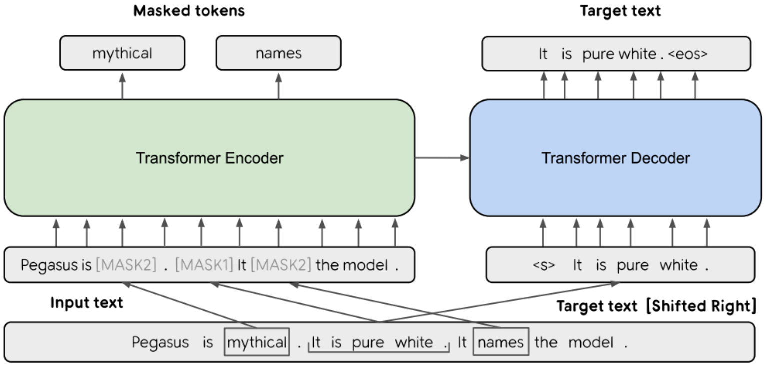

Masked tokens
mythical
names
Target text
It is pure white . <eos>
Transformer Encoder
Transformer Decoder
Pegasus is [MASK2] . [MASK1] It [MASK2] the model .
<s> It is pure white .
Input text
Target text [Shifted Right]
Pegasus is mythical . It is pure white . It names the model .

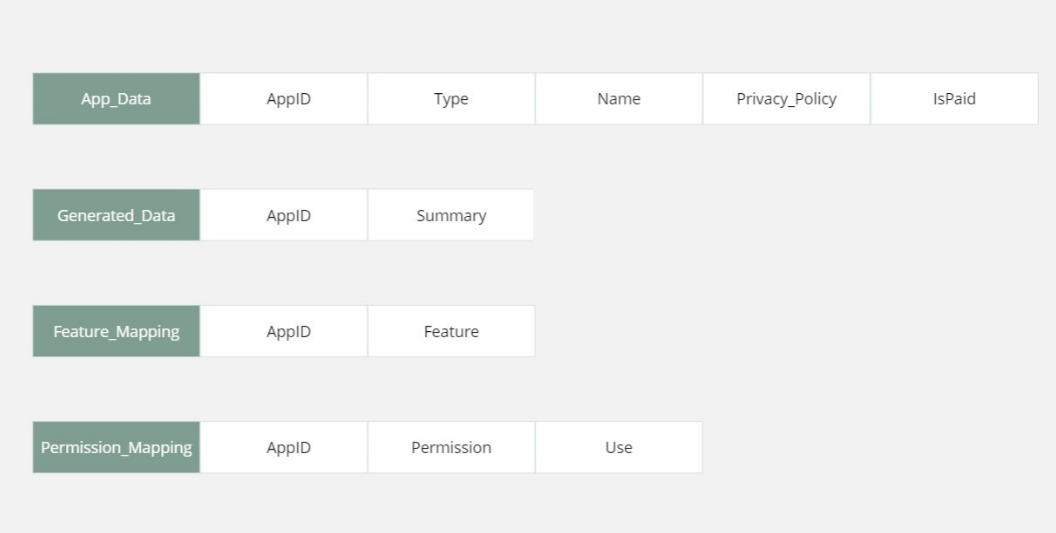
Figure 2: Pegasus Model Architecture

In our code we used the pre-processed tokens as inputs to generate the summary using the Pegasus pre-trained model.

4 Mid Term - 1 Results

In the baseline section, we had curated the data that we used to train and test our summarization models.In addition to this, We tested a number of summarization models and assessed their performance based on their cosine similarity scores, leading us to the selection of Pegasus.

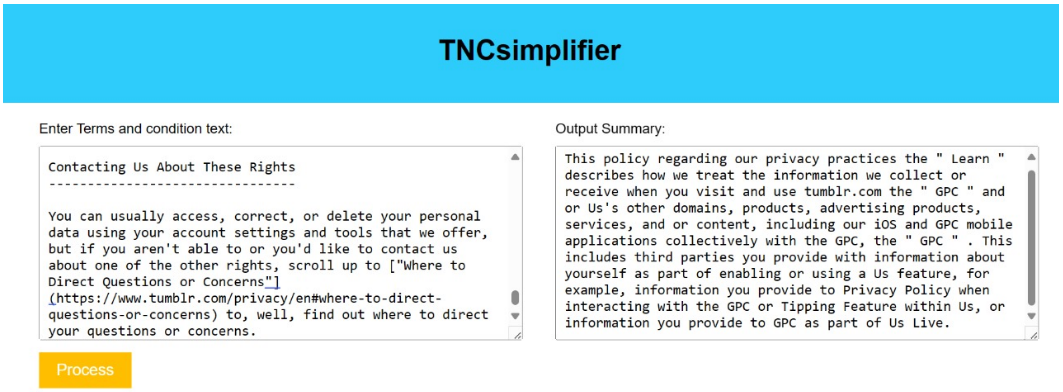
In this segment of the project, We utilized the pretrained pegasus model to generate summaries of the privacy policies we had available in our database.



App_Data	AppID	Type	Name	Privacy_Policy	IsPaid
Generated_Data	AppID	Summary			
Feature_Mapping	AppID	Feature			
Permission_Mapping	AppID	Permission	Use		

Figure 3: Updated Database Schema

In addition to this, We developed a basic web application prototype that takes input text and returns summary back to user.



```
Contacting Us About These Rights
--------------------------------

You can usually access, correct, or delete your personal
data using your account settings and tools that we offer,
but if you aren't able to or you'd like to contact us
about one of the other rights, scroll up to ["Where to
Direct Questions or Concerns"]
(https://www.tumblr.com/privacy/en#where-to-direct-
questions-or-concerns) to, well, find out where to direct
your questions or concerns.
```

```
This policy regarding our privacy practices the " Learn "
describes how we treat the information we collect or
receive when you visit and use tumblr.com the " GPC " and
or Us's other domains, products, advertising products,
services, and or content, including our iOS and GPC mobile
applications collectively with the GPC, the " GPC " . This
includes third parties you provide with information about
yourself as part of enabling or using a Us feature, for
example, information you provide to Privacy Policy when
interacting with the GPC or Tipping Feature within Us, or
information you provide to GPC as part of Us Live.
```

Figure 4: Website Prototype