

NLP Project: Text Summarization App

Introduction:

My app, Terms and Conditions Text Summarizer is specifically designed and trained on The purpose of this application is to summarize the Terms and Conditions into an easy-to read format and extract all the key words and relevant information. This app aims to make it more efficient to scan the Terms and Conditions section and decipher the main topics being presented within it. The inspiration behind this application is to make people aware of the agreements and contracts that they sign up for. Oftentimes, many people tend to glance at the terms and conditions without fully reading through it. This is primarily due to how lengthy the Terms and Conditions tend to be. What's more, is that many companies also tend to update their Terms and Conditions periodically, making it difficult for people to keep up with the contracts. Nonetheless it is essential to understand what we're signing up for. We utilize the power of AI and NLP to generate the Term and Condition summaries.

Usage:

In order to run this application, copy and paste the "Terms and Conditions" text into the input textbook. Then click the Submit button. This will begin generating the summary and will output it in the left text box entitled "Summary"

Documentation:

The frameworks used to build this application include:

- Hugging Face's Transformer's library, specifically the Summarization pipeline
 - This pipeline makes it very convenient to perform zero shot learning. I utilized this model to gain an insight to the outputs of zero-shot learning and identify areas of enhancements for the model. Since Transformer models are able to also perform text summarization due to their attention blocks, I used this as a baseline to identify the improvements that further refining of the model lead to. I also use this to explore analyze the various methodologies that can help
- Hugging Face's Pegasus Model: (<https://huggingface.co/ksi319/legal-pegasus>)
 - The reason why I chose this specific Pegasus model is because it's a Transformers-based Abstractive Summarization model and this model has been fine tuned for the legal domain (ksi319/legal-pegasus). Furthermore, the "Terms and Conditions" fall under the legal category and thus will gain more customized training experience when
- RAKE
 - RAKE (Rapid Automatic Keyword Extraction) is a library that automatically calculates the most important words within a corpus based on the frequency and how often it show up alongside other words. This library was used to generate the keywords and most common phrases within the terms and conditions Text. This way, users can glimpse at this list and get a gist of the main topics that the terms and conditions text is focused on.
 - Furthermore, the application also uses RAKE to also retrieve keyword phrases ranked highest to lowest with scores. This helps understand the most essential

keywords within the text and what users should pay attention to within the terms and conditions text.

- OpenDatasets (retrieve the dataset from Kaggle that'll be used to finetune the model)
 - Using the OpenDatasets library, I imported the training data that can be used to further fine-tune the model and perform additional training.
- Gradio
 - Used to develop an intractable and user-friendly interface for users to summarize text with the click of a button

Below are sample outputs to my screen:

This is a snippet of the summary generated from Instagram's terms and conditions

Terms and Conditions Text Summarizer

The interface consists of two main panels. The left panel, titled 'Input text', contains a text area with the following text: 'Since it takes such a large amount of data to teach effective models, a combination of sources are used for training. These sources include information that is publicly available online and licensed information, as well as information from Meta's products and services. There are more details on how we use information from Meta's products and services in our . When we collect public information from the internet or license data from other providers to train our models, it may include personal information. For example, if we collect a public blog post it may include the author's name and contact information. When we do get personal information as part of this public and licensed data that we use to train our models, we don't specifically link this data to any Meta account.' Below the text area are two buttons: 'Clear' and 'Submit'. The right panel, titled 'Summarized Text', contains a text area with the following summarized text: 'When we collect public information from the internet or license data from other providers to train our models, it may include personal information.<n>When we do get personal information as part of this public and licensed data, we don't specifically link this data to any Meta account.<n>We also care about your physical safety while using our Services. So do not use our Services in a way that would distract you from obeying traffic or safety laws. For example, never use the Services while driving. And never put yourself or others in harm's way just to capture a Snap or to engage with other Snapchat features. You can find out more about how we use information from Meta's products and services in our privacy policy. You may also contact us at privacy@meta.com.' Below the text area is a 'Flag' button.

Here is a snippet of the summarization generated from SnapChat's terms and conditions

Terms and Conditions Text Summarizer

The interface consists of two main panels. The left panel, titled 'Input text', contains a text area with the following text: 'we reserve the right to remove any offending content; terminate or limit the visibility of your account, and retain data relating to your account in accordance with our data retention policies; and notify third parties — including law enforcement — and provide those third parties with information relating to your account. This step may be necessary to protect the safety of our users, and others, to investigate, remedy, and enforce potential Terms violations, and to detect and resolve any fraud or security concerns. We also care about your physical safety while using our Services. So do not use our Services in a way that would distract you from obeying traffic or safety laws. For example, never use the Services while driving. And never put yourself or others in harm's way just to capture a Snap or to engage with other Snapchat features.' Below the text area are two buttons: 'Clear' and 'Submit'. The right panel, titled 'Summarized Text', contains a text area with the following summarized text: 'If you fail to comply, we reserve the right to remove any offending content; terminate or limit the visibility of your account; and retain data relating to your account in accordance with our data retention policies.<n>This step may be necessary to protect the safety of our users, and others, to investigate, remedy, and enforce potential Terms violations, and to detect and resolve any fraud or security concerns.<n>We also care about your physical safety while using our Services. So do not use our Services in a way that would distract you from obeying traffic or safety laws. For example, never use the Services while driving. And never put yourself or others in harm's way just to capture a Snap or to engage with other Snapchat features. We encourage you to follow these instructions.' Below the text area is a 'Flag' button.

Here is an example of a larger text retrieved from Amazon AWS's service terms

Terms and Conditions Text Summarizer

Input text

1. Universal Service Terms (Applicable to All Services)

The Service Terms below govern your use of the Services. Capitalized terms used in these Service Terms but not defined below are defined in the AWS Customer Agreement or other agreement with us governing your use of the Services (the "Agreement"). For purposes of these Service Terms, "Your Content" includes any "Company Content" and any "Customer Content," and "AWS Content" includes "Amazon Properties."

1.1. You may not transfer outside the Services any software (including related documentation) you obtain from us or third party licensors in connection with the Services without specific authorization to do so.

1.2. You must comply with current technical documentation applicable to the Services (including applicable user, admin, and developer guides) posted on the AWS Site at <https://docs.aws.amazon.com/index.html> (and any successor or related locations designated by us).

1.3. You will provide information or other materials related to Your Content (including copies of any client-side applications) as reasonably requested by us to verify your compliance with the Agreement. You will reasonably cooperate with us to identify the source of any problem with the Services that we reasonably believe may be attributable to Your Content or any end user materials that you control.

Clear

Submit

Summarized Text

The Amazon Web Services (AWS) Customer Agreement (the "Agreement") sets out the terms and conditions under which you can use the Services. The AWS Service Terms govern your use of the Services, including, but not limited to, the use of Company Content and any "Customer Content," and "AWS Content" includes "Amazon Properties"

You may not transfer outside the Services any software (including related documentation) you obtain from us or third party licensors in connection with the Services without specific authorization to do so. You must comply with current technical documentation applicable to the Services (including applicable user, admin, and developer guides) posted on the AWS Site at <https://docs.aws.amazon.com/index.html> (and any successor or related locations designated by us). You will provide information or other materials related to Your Content (including copies of any client-side applications) as reasonably requested by us to verify your compliance with the Agreement. You will reasonably cooperate with us to identify the source of any problem with the Service that we reasonably believe may be attributable to your Content or any end user materials that you control.

Flag

Finally, here is some summarization from Google's terms of service

Terms and Conditions Text Summarizer

Input text

Permissions page.

Other content

Finally, some of our services give you access to content that belongs to other people or organizations — for example, a store owner's description of their own business, or a newspaper article displayed in Google News. You may not use this content without that person or organization's permission, or as otherwise allowed by law. The views expressed in other people or organizations' content are theirs, and don't necessarily reflect Google's views.

We built our reputation on providing useful, reliable services like Google Search and Maps, and we're continuously improving our services to meet your needs. However, for legal purposes, we offer our services without warranties unless explicitly stated in our service-specific additional terms. The law requires that we explain this using specific legal language and that we use capital letters to help make sure you see it, as follows:

TO THE EXTENT ALLOWED BY APPLICABLE LAW, WE PROVIDE OUR SERVICES "AS IS" WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT. FOR EXAMPLE, WE DON'T MAKE ANY WARRANTIES ABOUT THE CONTENT OR FEATURES OF THE SERVICES, INCLUDING THEIR ACCURACY, RELIABILITY, AVAILABILITY, OR ABILITY TO MEET YOUR NEEDS.

We also care about your physical safety while using our Services. So do not use our Services in a way that would distract you from obeying traffic or safety laws. For example, never use the Services while driving. And never put yourself or others in harm's way just to capture a Snap or to engage with other Snapchat features.

Clear

Submit

Summarized Text

You have no obligation to provide content to our services and you're free to choose the content that you want to provide.

You may use Google's content as allowed by these terms and any service-specific additional terms, but we retain any intellectual property rights that we have in our content. Don't remove, obscure, or alter any of our branding, logos, or legal notices.

Some of our services give you access to content that belongs to other people or organizations, for example, a store owner's description of their own business, or a newspaper article displayed in Google News. You may not use this content without that person or organization's permission, or as otherwise allowed by law. The views expressed in other people's content are theirs, and don't necessarily reflect Google's views.

Flag

Use via API · Built with Gradio

Transformer Model:

- Our model utilizes the transformer model as well as performing additional finetuning to ensure that the best summaries are generated.

Contributions:

In order to develop this model and tailor it specifically for “terms and conditions” I implemented a pre-trained Transformer model, the Pegasus. This Pegasus also came in a fine-tuned format

which was fine-tuned on legal datasets, which is what I ultimately used for my text summarization. Next, I incorporated

My original goal for this project was to start off with a pre-trained model, and then fine-tune the model using the legal text dataset. Then, provide additional training to the model by using a corpus of Terms and Conditions text and their summarized versions and performing sequence classification.

I also explored whether it was possible to further fine tune the already-fine tuned model, I've learned the hard way that using a pre-trained model that has already been fine-tuned and attempting to add another layer of fine-tuning to this required large amounts of GPU and compute power. This resulted in Collab's runtime disconnecting and crashing and I lost all of my session data. To combat this, I tweaked the model and adjusted the number of epochs, batch sizes, length of the input text, etc. Additionally, I experimented using multiple Transformer models, especially ones which were smaller in than others including distilled-bert, so that the layering did not result in an enormous model. When running on GPU, the model was still struggling to process all of the finetuning and crashed. One of the greatest takeaways from this project was that the be

Limitations:

Some of the limitations of this app include size restrictions. Currently the text summarization tool can only process 1024 words at a single time. To combat this, we encourage users to work in batches when entering their Terms and Conditions into our application. Another limitation of the application is that it takes longer time to load when it encounters large portions of text, thus, it is more time consuming to add in a large amount of text at a single time.

As described above, the physical hardware and memory complexities made additional training and enhancements to the model difficult and in most cases impossible. Since this is hosted on Google Colab it has a RAM limit of 32 GB using the free version, I was unable to successfully run my sequence-to-sequence training algorithm. I removed the further fine-tuning implementation using a Terms of Service dataset out of my code (so that anyone running my model will not experience any session crashes or failures), however there may still be remnants of it within the text.

Link to the demo:

<https://github.com/Khossain01/TxtSummarizationTermsConditions>