Terms of Services Summarization

Team Caffeine: Chieng Chang, Tongshu Yang, Ziang Ren
Georgia Institute of Technology

Goal

- Gather user opinions on terms of services and privacy policies from major social media platforms
- Improve the readability of Terms of Services by generating extractive summaries

Importance

- Users could read the part of the Term of Services and privacy policy where they are interested the most in a shorter time period.
- Protect users' right for knowing how software or applications using their data.
- Increase readability of Term of Services and privacy policy so that users from different backgrounds would be able to understand the contents.

Data

Twitter: Query tweets from official Twitter API using Tweepy based on keywords in the content and filter again with Hashtags. The data size is limited by the official API querying rate. Once the querying size increased to 2000, the API will force the program into sleep.

Reddit: a third-party database (pushshift.io) is used to perform keyword search for relevant submission IDs. These IDs are used to query submission details using the official Reddit API.

PESC: The summarized data for term of services segments have human-written reference summaries.

Change of Dataset

Twitter: Before midterm, tweets are ordered based on the tweets retweet counts and then ordered again based on favorite number. After midterm, we changed the tweets to be ordered based on the sum of number of retweets and favorite. Then we only keep top ***n*** number of tweets while querying ***1.5n*** number of tweets.

Data Source	Time Period	Data Size
Twitter	2020-2022	1000 Tweets
Reddit	2019-2021	22767 Posts
Plain English Summarization of Contracts	2020s-2020s	446 sets of Terms of Services segments

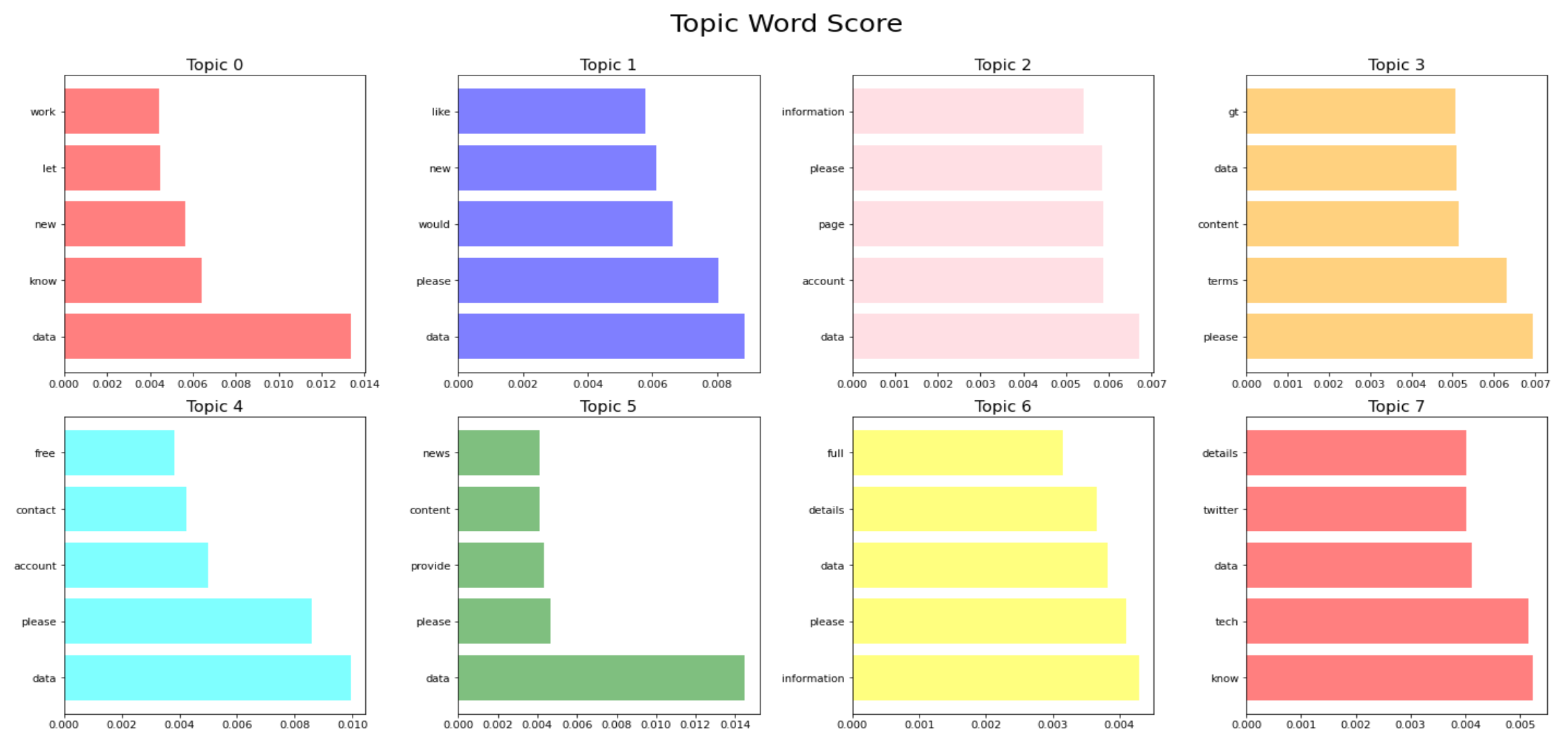
Change of Dataset (Cont.)

Reddit: the dataset is filtered for unrelated submissions, submissions that are deleted or removed, or has no text (example: a submission containing only a link to an external media). This leaves us with 9946 of 22767 submissions.

Method & Results

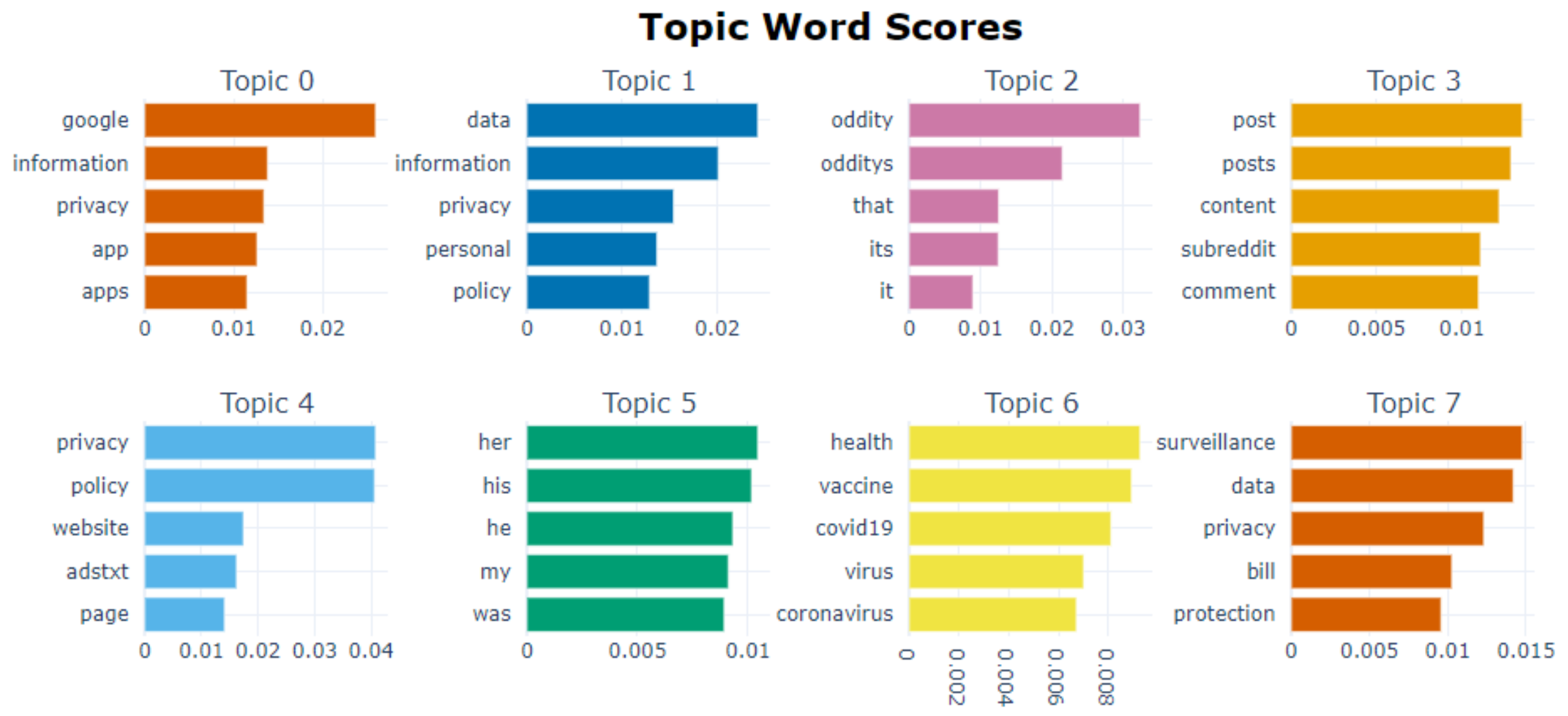
Twitter Topic Model

Pre-process the tweets contents by removing nonsense words and special characters. Then we use LDA model to get the top 8 topic words with scores.



Reddit Topic Model

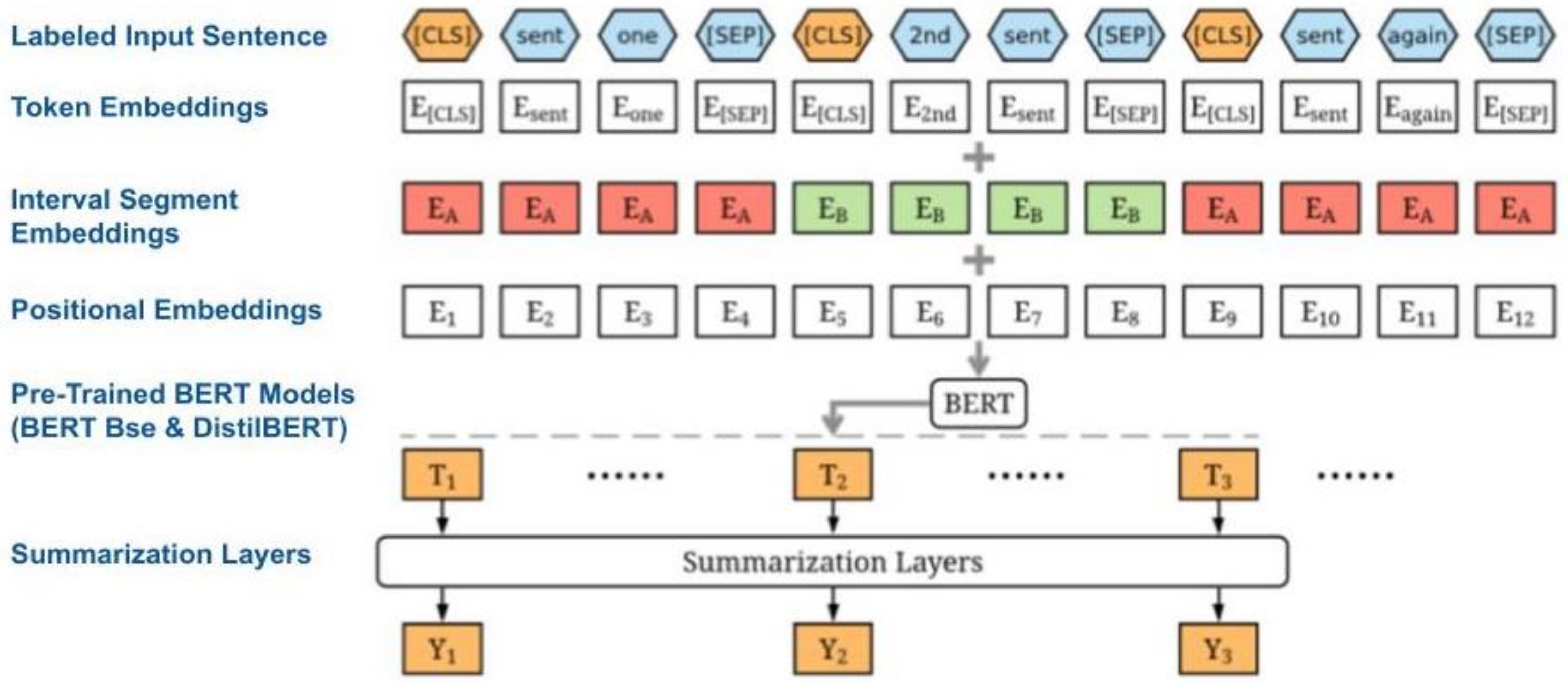
Reddit submission texts are stemmed using a standard SpaCy pipeline. The resulting text is passed to a BERTopic model to cluster for the top 10 topics.



Extractive Summarization Model

We developed an extractive summarization model by adding summarization layers on top of BERT Base and DistilBERT. We also made use of Legal-BERT to tokenize and pre-process input text. Specifically, we used the CONTRACTS-BERT-BASE model that was pre-trained on United States contracts corpora. During the prediction process, we used Trigram Blocking to reduce redundancy in summaries.

Model Architecture

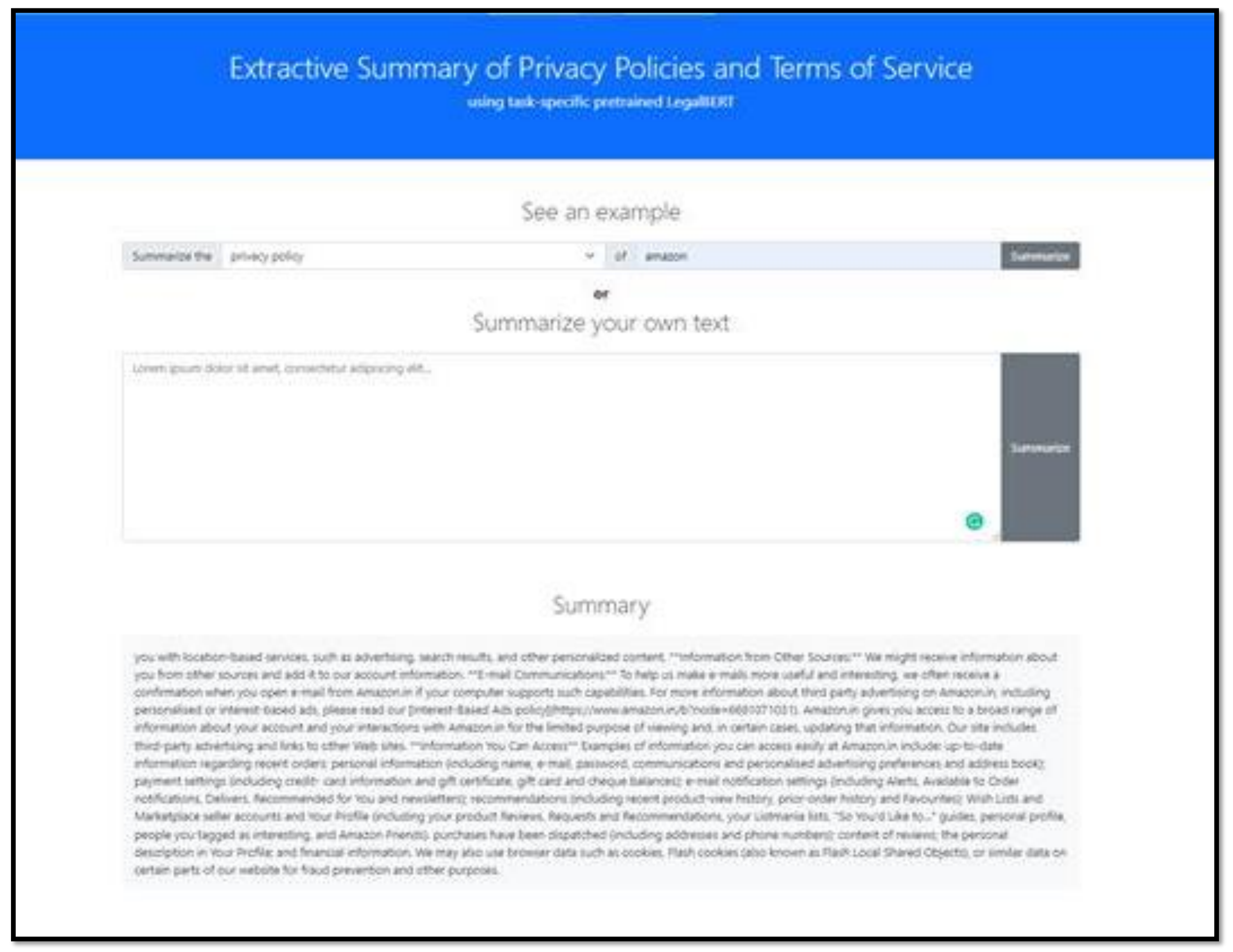


Model Result

Model	ROUGE-1	ROUGE-2	ROUGE-L
BERT Base	**38.49**	**12.82**	**31.50**
DistilBERT	37.19	11.15	29.72
TextRank	24.03	7.16	17.10
KLSum	23.56	6.94	16.93
Lead-1	23.87	7.47	17.19
Lead-K	26.38	7.52	17.63

Web App Demo

A simple web interface is designed for users to interact with our extractive summarization models. The user can select a privacy policy or terms of service from a range of companies and organizations and request a summary.



Extractive Summary of Privacy Policies and Terms of Service

using task-specific pretrained LegalBERT

See an example

Summarize the | privacy policy | of | amazon | Summarize

or

Summarize your own text

Lorem ipsum dolor sit amet, consectetur adipiscing elit...

Summarize

Summary

you with location-based services, such as advertising, search results, and other personalized content. **Information from Other Sources:** We might receive information about you from other sources and add it to our account information. **E-mail Communications:** To help us make e-mails more useful and interesting, we often receive a confirmation when you open e-mail from Amazon.in if your computer supports such capabilities. For more information about third party advertising on Amazon.in, including personalised or interest-based ads, please read our [Interest-Based Ads policy](https://www.amazon.in/b?node=6601071031). Amazon.in gives you access to a broad range of information about your account and your interactions with Amazon.in for the limited purpose of viewing and, in certain cases, updating that information. Our site includes third-party advertising and links to other Web sites. **Information You Can Access:** Examples of information you can access easily at Amazon.in include: up-to-date information regarding recent orders; personal information (including name, e-mail, password, communications and personalised advertising preferences and address book); payment settings (including credit card information and gift certificate, gift card and cheque balances); e-mail notification settings (including Alerts, Available to Order notifications, Delivers, Recommended for You and newsletters); recommendations (including recent product-view history, prior-order history and Favourites); Wish Lists and Marketplace seller accounts and Your Profile (including your product Reviews, Requests and Recommendations, your Listmania lists, "So You'd Like to..." guides, personal profile, people you tagged as interesting, and Amazon Friends). purchases have been dispatched (including addresses and phone numbers); content of reviews; the personal description in Your Profile; and financial information. We may also use browser data such as cookies, Flash cookies (also known as Flash Local Shared Objects), or similar data on certain parts of our website for fraud prevention and other purposes.