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 (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 Result

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT Base</td>
<td>38.49</td>
<td>12.82</td>
<td>31.50</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>37.19</td>
<td>11.15</td>
<td>29.72</td>
</tr>
<tr>
<td>TextRank</td>
<td>24.03</td>
<td>7.16</td>
<td>17.10</td>
</tr>
<tr>
<td>KLSum</td>
<td>23.56</td>
<td>6.94</td>
<td>16.93</td>
</tr>
<tr>
<td>Lead-1</td>
<td>23.87</td>
<td>7.47</td>
<td>17.19</td>
</tr>
<tr>
<td>Lead-K</td>
<td>26.38</td>
<td>7.52</td>
<td>17.63</td>
</tr>
</tbody>
</table>

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.