The Creation and Analysis of a Website Privacy Policy Corpus

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh

Motivation

Privacy policies are pervasive online, but they are long and complex documents that Internet users rarely read. Internet users who are unaware of what happens to their data cannot make meaningful privacy choices.

We built an annotated corpus of privacy policies to enable NLP efforts to automate their interpretation. This was the first large-scale effort to annotate privacy policies at such a fine level of detail.

Corpus Composition

The annotations consist of data practices which fall into ten different categories. Each data practice is grounded in spans of policy text and has a category-specific set of attributes.

Below: Data practice statistics for the entire corpus (frequency) and per policy (median). Fleiss’ Kappa is calculated at the segment level.

<table>
<thead>
<tr>
<th>Data Practice Category</th>
<th>Freq.</th>
<th>Median</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Party Collection/Use</td>
<td>8,956</td>
<td>74</td>
<td>.76</td>
</tr>
<tr>
<td>Third Party Sharing/Collection</td>
<td>5,230</td>
<td>39</td>
<td>.76</td>
</tr>
<tr>
<td>Other</td>
<td>3,551</td>
<td>25</td>
<td>.49</td>
</tr>
<tr>
<td>User Choice/Control</td>
<td>1,791</td>
<td>13</td>
<td>.61</td>
</tr>
<tr>
<td>Data Security</td>
<td>1,009</td>
<td>7</td>
<td>.67</td>
</tr>
<tr>
<td>Int’l and Specific Audiences</td>
<td>941</td>
<td>6</td>
<td>.87</td>
</tr>
<tr>
<td>User Access, Edit, and Deletion</td>
<td>747</td>
<td>5</td>
<td>.74</td>
</tr>
<tr>
<td>Policy Change</td>
<td>550</td>
<td>4</td>
<td>.73</td>
</tr>
<tr>
<td>Data Retention</td>
<td>370</td>
<td>2</td>
<td>.55</td>
</tr>
<tr>
<td>Do Not Track</td>
<td>90</td>
<td>0</td>
<td>.91</td>
</tr>
</tbody>
</table>

Segment-Category Prediction

Predicting the practice categories of policy segments is a first step to simplifying or automating the annotation process.

<table>
<thead>
<tr>
<th>Data Practice Category</th>
<th>Log. Reg.</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P  R  F</td>
<td>P  R  F</td>
</tr>
<tr>
<td>First Party Collection/Use</td>
<td>.73  .67  .70</td>
<td>.76  .73  .75</td>
</tr>
<tr>
<td>Third Party Sharing/Collection</td>
<td>.64  .63  .63</td>
<td>.67  .73  .70</td>
</tr>
<tr>
<td>User Choice/Control</td>
<td>.45  .62  .52</td>
<td>.65  .58  .61</td>
</tr>
<tr>
<td>Across All 10 Categories</td>
<td>.53  .65  .58</td>
<td>.66  .66  .66</td>
</tr>
</tbody>
</table>

The Annotation Process

We worked with legal experts to develop an annotation scheme for privacy policy text. Skilled annotators used a web-based annotation tool (below) to apply it to policy text.

Policy text was divided into segments which were roughly equivalent to paragraphs. Annotators worked on one segment at a time.

A Resource for the Research Community

Researchers can explore the corpus on our interactive website.

You can download the corpus at data.usableprivacy.org

Future Directions

The corpus enables research in several directions, including:

- Automated extraction of data practices from text
- Cohesive interpretation of data practices in a privacy policy
- Identification of sectoral norms and outliers
- User-oriented summarization of privacy policies