Privacy policies on websites are based on the **notice-and-choice principle**: They notify Web users of their privacy choices. However, many users do not read privacy policies or have difficulties understanding them. The resulting **information asymmetry** leaves users uninformed about their privacy choices, can lead to **market failure**, and calls the notice-and-choice principle into question altogether.

Our classifiers (Naive Bayes and/or rules) have an overall F-1 score of **90%** (when compared to human annotators trained in privacy law). The baseline accuracy consists of always selecting the classification that occurred the most for the respective category in our training set.

To ensure the reliability of annotations we calculated the inter-annotator agreement by Krippendorff’s alpha, which indicated for all categories fair or good agreement (except for Ad Disclosure). It is striking that performance (F-1 score) correlates to agreement (Krippendorff’s alpha).

The baseline accuracy consists of always selecting the classification that occurred the most for the respective category in our training set.

When a user requests a privacy policy analysis, the program checks whether analysis results are available at a crowdsourcing repository (to which crowd contributors can submit analysis results of policies). If results are available, they are returned and displayed to the user (I. Crowdsourcing Analysis). If no results are available, the policy text is fetched from the policy website, analyzed by automatic classifiers on the client machine, and the analysis results are displayed to the user (II. Classifier Analysis).

Our classifiers (Naive Bayes and/or rules) have an overall F-1 score of **90%** (when compared to human annotators trained in privacy law). The baseline accuracy consists of always selecting the classification that occurred the most for the respective category in our training set.

To ensure the reliability of annotations we calculated the inter-annotator agreement by Krippendorff’s alpha, which indicated for all categories fair or good agreement (except for Ad Disclosure). It is striking that performance (F-1 score) correlates to agreement (Krippendorff’s alpha).

<table>
<thead>
<tr>
<th>Section</th>
<th>Mean Sem. D.</th>
<th>Significance (P)</th>
<th>Odds Ratio (Z)</th>
<th>95% C. Int (Z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extr. Text</td>
<td>1.87</td>
<td>0.02</td>
<td>2.07</td>
<td>1.12-3.81</td>
</tr>
<tr>
<td>Policy Storage</td>
<td>2.08</td>
<td>0.04</td>
<td>1.51</td>
<td>1.02-2.22</td>
</tr>
</tbody>
</table>

We measured this ambiguity in form of **semantic diversity**. The less ambiguity in the extracted text for the classifier to analyze and in the section for the annotators to read, the fewer misclassifications and disagreements occurred, respectively.

**5. Semantic Diversity**

Our experimental results suggest that classifier performance is inherently limited as it correlates to the same variable to which human interpretations correlate—the ambiguity of natural language. We measured this ambiguity in form of **semantic diversity**. The less ambiguity in the extracted text for the classifier to analyze and in the section for the annotators to read, the fewer misclassifications and disagreements occurred, respectively.

**6. Reference**