Multimedia Appendix 1. The sample size simulation for human coding

Case 1: We conducted simulation in order to determine sample size for each stratum. Data were generated assuming the population size of 4 million, retrieval precision of 95%, and retrieval recall of 84%. The retrieved ($n_1$) to the unretrieved ($n_2$) ratio was 1 to 39, and prevalence was set 2.8%. Let $m$ be the sample size for retrieved data and $k$ be the sample size for unretrieved data. We randomly sampled $m$ out of $n_1$ and $k$ out of $n_2$, computed precision and recall on the sampled data, and checked whether their 95% confidence intervals included the true values. We replicated this many times to obtain the average confidence intervals.

The variability of retrieval recall estimate is affected by the size $k$. Therefore, we repeated the simulation by increasing $k$ at a fixed value of $m$. Figure 2 displays how the average confidence intervals for recall estimates change as $k$ increases from 1,000 to 8,000 while $m$ is fixed at 3,000. The gain in variability reduction for recall estimate is small when $k$ is above 6000. The simulation results for $m=2,000$ are presented in the table below; they show a similar pattern. The coverage probability of all precision estimates was satisfactory.

Case 2: We also considered another scenario that the population size was 10 million, retrieval precision and recall were 92% and 85% respectively. The $n_1$ to $n_2$ ratio was approximately 1 to 815, and prevalence was set 0.1%. We considered $m=600$ and $k=8,000$ to 30,000 because of the large $n_2$. The coverage probability for interval estimates of recall does not reach the desired level 95% in many cases, the variability around the estimate is still high even when $k=30,000$. The simulation suggests that the ratio of $n_1$ to $n_2$ affects accuracy of the recall estimates; when the ratio is tiny, taking a sizeable sample of the unretrieved tweets may compensate for inaccuracy. But human coders can code only so much before fatigue interferes. To determine an appropriate sample size for human coding, a balance between the desired level of statistical precision and feasibility should be considered.
**Case 1:** population size of 4 million, the retrieved to the unretrieved ratio = 1:39, prevalence = .028, precision = .95, recall = .84

**Case 2:** population size of 10 million, the retrieved to the unretrieved ratio = 1:815, prevalence = .001, precision = .92, recall = .85

<table>
<thead>
<tr>
<th></th>
<th>Retrieved</th>
<th>Unretrieved</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m</td>
<td>k</td>
<td>Mean</td>
<td>95% L</td>
</tr>
<tr>
<td><strong>Case 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>3</td>
<td>1,000</td>
<td>0.03</td>
<td>0.9500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,000</td>
<td>0.05</td>
<td>0.9501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3,000</td>
<td>0.08</td>
<td>0.9499</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,000</td>
<td>0.10</td>
<td>0.9501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5,000</td>
<td>0.13</td>
<td>0.9500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6,000</td>
<td>0.15</td>
<td>0.9501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7,000</td>
<td>0.18</td>
<td>0.9502</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8,000</td>
<td>0.21</td>
<td>0.9500</td>
</tr>
<tr>
<td><strong>Case 2</strong></td>
<td>2</td>
<td>1,000</td>
<td>0.03</td>
<td>0.9500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,000</td>
<td>0.05</td>
<td>0.9499</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3,000</td>
<td>0.08</td>
<td>0.9498</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,000</td>
<td>0.10</td>
<td>0.9501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5,000</td>
<td>0.13</td>
<td>0.9502</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6,000</td>
<td>0.15</td>
<td>0.9499</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7,000</td>
<td>0.18</td>
<td>0.9497</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8,000</td>
<td>0.21</td>
<td>0.9499</td>
</tr>
</tbody>
</table>

m = the sample size of retrieved data, k = the sample size of unretrieved data, % = sampling fraction in percentage. Each scenario was repeated 3000 times. The mean of point estimates and mean of 95% confidence limits are reported. The coverage probability C(%) shows how many times the 95% confidence intervals contain the true value.