CLARA

Confidence of Labels and Raters


Facebook

Dimitris Kalimeris

Harvard University
Automation is essential for large-scale systems

99.5% of fake accounts actioned automatically

0.5%
However, there are many areas where human decisions are needed.
It is sometimes hard for humans to make decisions.
What makes people noisy decision makers?
Individual bias

More Bias

More Variance

Less Bias

Less Variance
Ambiguity of the guidelines
Subjectivity in the decision
Simple mistakes
Let’s consider a few examples where noisy decisions have negative outcomes.
PREVALENCE

The percentage of policy-violating content out of all content seen by Facebook users.
ENFORCEMENT

Taking down content or entities that violate the community standards.
TRAINING MODELS

Using human-generated labels as “ground truth” for training ML models.
So how should we deal with noisy decision making?
Hire experts

Ask several people

Leverage Machine Learning
But how do we aggregate multiple labels into a single decision?

...and can we quantify the certainty of the decision?
CLARA: Confidence of Labels and Raters
CLARA: Confidence of Labels and Raters

![Diagram of CLARA model with nodes and edges representing the relationship between labels, raters, and scores.](image)

The diagram shows the following variables:

- \( \theta \)
- \( y_i \) (Labels)
- \( s_{i,c} \) (Scores)
- \( \psi_{a,k} \) (Raters)
- \( r_{i,j} \)
- \( a_{i,j} \)
- \( \mu_{c,k} \)
- \( \Sigma_{c,k} \)

The scores are connected to the labels, indicating the confidence of the raters. The scores are also connected to the variables \( \theta \) and \( \psi_{a,k} \), suggesting the influence of the model parameters and raters on the confidence levels.

The score of 0.8 indicates the confidence level of the raters in their assignments.
CLARA: Confidence of Labels and Raters

Diagram showing variables and relationships:
- $\theta$
- $y_i$
- $r_{i,j}$
- $a_{i,j}$
- $S_{i,c}$
- $\psi_{a,k}$
- $\mu_{c,k}$
- $\Sigma_{c,k}$

Scores and Labels representation:
CLARA: Confidence of Labels and Raters

<table>
<thead>
<tr>
<th>Labels</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer Confusion Matrix</td>
<td>Overall Prevalence</td>
</tr>
<tr>
<td>$\psi_{a,k}$</td>
<td>$\psi_{a,k}$</td>
</tr>
<tr>
<td>$r_{i,j}$</td>
<td>$a_{i,j}$</td>
</tr>
<tr>
<td>$y_i$</td>
<td>$s_{i,c}$</td>
</tr>
<tr>
<td>$\theta$</td>
<td>$\mu_{c,k}$</td>
</tr>
<tr>
<td>$\Sigma_{c,k}$</td>
<td>$\psi_{a,k}$</td>
</tr>
<tr>
<td>$A$</td>
<td>$K$</td>
</tr>
<tr>
<td>$N_i$</td>
<td>$I$</td>
</tr>
<tr>
<td>$C$</td>
<td>$K$</td>
</tr>
<tr>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
</tr>
</tbody>
</table>

0.8
CLARA: Confidence of Labels and Raters

Overall Prevalence

Item True Label

Reviewer Confusion Matrix

Score Mixture
CLARA: Confidence of Labels and Raters

<table>
<thead>
<tr>
<th>Labels</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item True Label</td>
<td>Reviewer Confusion Matrix</td>
</tr>
<tr>
<td>Overall Prevalence</td>
<td>Score Mixture</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\psi_{a,k} & \\
r_{i,j} & \\
a_{i,j} & \\
y_i & \\
\theta & \\
\Sigma_{c,k} & \\
\mu_{c,k} & \\
\end{align*}
\]

\[
\begin{align*}
A & \\
N_i & \\
I & \\
K & \\
C & \\
\end{align*}
\]
Applications

1. Prevalence Measurement
2. Reviewer Performance Measurement
3. Labeling Efficiency
Prevalence Measurement

Measure the percentage of policy-violating content out of all content seen by Facebook users
Prevalence Measurement

Population of content
Prevalence Measurement

Population of content

Sample

Sampled subset of content
Prevalence Measurement

Population of content → Sampled subset of content → Send for review → Labels
Prevalence Measurement

Population of content

Sampled subset of content

Send for review

Labels

Prevalence estimate

Prevalence Measurement Estimate
Prevalence Measurement

Population of content

Sampled subset of content

Labels

Estimate

- Majority Vote with bootstrapping
- CLARA

Send for review

Prevalence estimate
Prevalence Estimates (Violation Type A)
Reviewer Performance Measurement

Measure the performance of reviewers in labeling violating content
Measuring Reviewer Performance

Content for review
Measuring Reviewer Performance

Reviewers

Send for review

Content for review

Single labels by reviewers
Measuring Reviewer Performance

Content for review

Reviewers

Send for review

Single labels by reviewers

Auditors

Sample for audit

Multiple labels by auditors
Measuring Reviewer Performance

Content for review → Send for review → Reviewers

Single labels by reviewers → Sample for audit → Auditors

Multiple labels by auditors → Estimate → Reviewer Precision/Recall Estimates
Measuring Reviewer Performance

- Reviewers
- Auditors

- Majority Vote
- CLARA

Content for review → Send for review → Sample for audit → Multiple labels by auditors → Estimate

Single labels by reviewers → Reviewer Precision/Recall Estimates
Reviewer Precision/Recall Estimates

![Graph showing precision and recall estimates over time.]

- **MV: Precision**
- **MV: Recall**
- **CLARA: Precision**
- **CLARA: Recall**

Y-axis: Absolute error
X-axis: Day
Labeling Efficiency

Improve labeling efficiency by only sending content for additional review if the confidence is low.
Improving Labeling Efficiency

Send for review

Content for review

Reviewer
Improving Labeling Efficiency

Content for review

Reviewer

Send for review

Estimate label confidence

Is the confidence sufficiently high?
Improving Labeling Efficiency

Send for review → Estimate label confidence → Is the confidence sufficiently high?

- Yes → Stop reviewing
- No → Reviewer
Improving Labeling Efficiency

- Content for review
- Reviewer

Snippet of the diagram:

- Send for review → Estimate label confidence
- Is the confidence sufficiently high?
  - Yes → Stop reviewing
  - No → Send for additional review
Improving Labeling Efficiency

Is the confidence sufficiently high?

- Yes
- No

Send for additional review

Estimate label confidence

Content for review

Reviewer

Stop reviewing

- Stratified sampling
- Snorkel
- CLARA
Cost/Accuracy Tradeoff Curve

Accuracy vs. % of Total Labels

- CLARA
- Snorkel
- Sampled
System Overview

- **Reviewing Tool**: Sends content for an additional review.
- **Realtime Prediction Service**: Checks if confidence is sufficiently high.
- **Hadoop**: Stores labels and fits a model.
- **FB Learner Flow**: Makes online inference and stores model.
- **Thrift**: Sends results.
- **Dashboard**: Visualizes, monitors, and analyzes.
System Overview

Reviewing Tool

Realtime Prediction Service

Confidence estimate
- No
- Yes

Is the confidence sufficiently high?

Dashboard

ONLINE

OFFLINE

Hadoop

Fit model

FB Learner Flow

Store model

Thrift

Take action

Store results

Make online inference

OFFLINE

ONLINE

Sample

Content

Store labels

Store labels
Conclusion

• CLARA, a system developed and deployed at Facebook to estimate the uncertainty of human labels

• Extensive simulations and comparison with state-of-the-art

• Results on real Facebook deployment

And the source code

https://github.com/facebook/clara