We explore both replacement and enrichment Hebrew training data trained using annotated samples in other languages. In most cases, weights taken from particular groups in Facebook:

- Ynet, the shadow, 0404
- תושביה דובס, בוט, "גאומרים" בישנה, מלין
- A list of Hebrew keywords was used to find offensive comments.

Our contributions:
- A new annotated dataset of Facebook comments written in Hebrew.
- Monolingual evaluation of multiple supervised models and text representations for a task of offensive language detection.
- Cross-lingual and multilingual evaluations of the explored methods with Semitic languages as target languages.

The pipeline:

- Hebrew dataset: 5,217 comments, taken from particular groups in Facebook:
  - ynet, the shadow, 0404
  - תושביה דובס, בוט, "גאומרים" בישנה, מלין
  - A list of Hebrew keywords was used to find offensive comments.

- Arabic dataset: 9,000 comments, written in Arabic:
  - A list of Arabic keywords was used to find offensive comments.

Our contributions:
- A new annotated dataset of Facebook comments written in Hebrew.
- Monolingual evaluation of multiple supervised models and text representations for a task of offensive language detection.
- Cross-lingual and multilingual evaluations of the explored methods with Semitic languages as target languages.

The data:
- Hebrew: OLA (Litvak et al., 2021)
- Hebrew: OLA + Liebeskind (Liebeskind et al., 2017)
- English: OLID (Zampieri et al., 2019)

The pipeline:

- Hebrew dataset: 5,217 Facebook comments, additional manual labeling
  - Kappa agreement 0.8
  - 44,100 tweets
  - 33,240 used after filtering
  - Kappa agreement 0.6

- Arabic dataset: 9,000 tweets
  - Kappa agreement 0.75
  - 11,140 tweets

Models:
- 1-3. Random Forest (RF)
- 4-6. Support Vector Machine (SVM)
- 7-9. Logistic Regression (LR)
- 10. Fine-tuned mBERT

Evaluation:
- Scenarios:
  - Monolingual learning: each model is trained and tested on the same language.
  - Cross-lingual learning: aims at checking whether missing training data in a target language can be compensated by training a model on the same language.
  - Multilingual learning: is performed for testing whether one joint multilingual model can be trained using annotated samples in multiple languages.

Results:
- Train/test data split
- Train: 80% of A
- Test: 20% of A
- Train: 80% of B
- Test: 20% of B
- Train: 80% of A + 80% of B
- Test: 20% of B

Discussion:
- mBERT model is superior for most cases, especially in cross-lingual and multilingual experiments.
- Weak evidence approving a possible advantage of mBERT vectors as a representation model in monolingual setup.
- All the results achieved in the cross-lingual settings for Semitic languages are significantly lower than their monolingual results:
  - except Recall in Hebrew.
- Multilingual data augmentation performs well in most cases:
  - extending the Hebrew training set with the data in Arabic results in the same accuracy score.

Error analysis:

<table>
<thead>
<tr>
<th>Language</th>
<th>Sample size</th>
<th>Missing annotation</th>
<th>Word-based classification</th>
<th>Interann</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>30</td>
<td>10 (20%)</td>
<td>11 (33%)</td>
<td>23 (76.67%)</td>
</tr>
<tr>
<td>Hebrew</td>
<td>30</td>
<td>7 (23.33%)</td>
<td>7 (23.33%)</td>
<td>6 (53.33%)</td>
</tr>
</tbody>
</table>

The dataset can be downloaded from: [https://github.com/rezek/HebrewDataset](https://github.com/rezek/HebrewDataset)