Social Media Content Analysis for Inappropriate Content
Akansha Sharma (ads8486@rit.edu)                 Advisor: Dr. Leon Reznik
Rochester Institute of Technology

Objective
- Posting on Social Networking websites to express feelings and opinions about daily issues and concerns has seen an exponential increase.
- Such communication between individuals may contain sharing of inappropriate content which can emerge threats like cyberbullying.
- Project goal is to design and build a classifier to detect such abusive text messages and images with inappropriate content.

Background
- The text datasets used to train and test the model are from the Social Networking websites MySpace and Formspring.
- The datasets are of the format bio, date, location, userid, gender, location and body text and classified as Yes or No depending on whether it contains inappropriate content.
- Text dataset for each website consists of about 60,000 xml files depicting the conversation between users on these websites.
- Image dataset consists of about 100 images wherein 50 images are appropriate and 50 are inappropriate.

Design
- Text and image data are fed to a text and image analysis module respectively.
- Determine the accuracy of the modules.

Data Cleaning and Pre-Processing
- For Text Analysis, data cleaning involved removing the irrelevant attributes such as bio, date, location and gender.
- Data preprocessing involved removing punctuation & special characters, and the commonly used stop words.
- For Image Analysis, each image was converted to .jpg format.
- Each image is then resized.

Text Analysis
- Created a vocabulary of about 350 inappropriate and abusive words.
- Each text message represented as a feature vector using the sklearn feature extraction module.
- If word present in the vocabulary appears, corresponding attribute set to 1, else set to 0
- Used the Naive Bayes algorithm for final classification of the text data.

Classification Results
1. Formspring Dataset

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate</td>
<td>0.85</td>
<td>0.92</td>
<td>0.88</td>
</tr>
<tr>
<td>Inappropriate</td>
<td>0.91</td>
<td>0.83</td>
<td>0.87</td>
</tr>
<tr>
<td>Avg/Total</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
</tr>
</tbody>
</table>

2. Myspace Dataset

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate</td>
<td>0.89</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>Inappropriate</td>
<td>0.93</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td>Avg/Total</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Image Analysis
- Images analysis: Detect interest points of an image using sklearn feature extraction.
- Split the image into cells, compute histogram over each cell and concatenate after normalization. Results in feature vector for input image.
- Map vectors to visual words in the vocabulary. Words mapped using the K-means clustering algorithm.
- Used Support Vector Machines to classify the data.
- The accuracy of the model is approximately 88%.

Conclusion and Future Work
- Implemented Text classification module using the Bag of Words approach and the Naive Bayes algorithm. Achieved 90% accuracy.
- Future work can include analyzing images which have text embedded in them. In text analysis, include sentiment or context related knowledge to enhance the detection.

References