Predicting Age-Appropriateness of Fanfiction Short Stories

Oishi Banerjee, Minh Nguyen, Christine Phan
oishib@stanford.edu, minh084@stanford.edu, cxphan@stanford.edu

Summary
In order to address a need in fanfiction to properly identify sensitive content, we built a multi-class model that classifies stories. The result was a bidirectional LSTM-RNN model with a 0.59 accuracy that included the features story text and tags.

Background
In fanfiction, writers have the ability to label their writings for sensitive content, such as drug use, violence, and sexuality explicit content. This helps readers identify a work’s target or appropriate audience. However, not all stories are fully labeled.

Problem
Build a model labelling stories by age-appropriateness to help readers choose appropriate reading material.

Dataset & Features
- 13,242 total stories scraped from archiveofourown.org
- Dataset is split evenly among the four categories
  - General Audiences
  - Teen and Up
  - Mature
  - Explicit
- Each story contained the following features:
  - Text of the story
  - Summary
  - Tags (themes, important characters, keywords, etc.)
  - Kudos

Model Overview
The data was split into 80-10-10 for our train-dev-test set. Used GloVe embedding for text. Models built and tuned:
- Logistic Regression (Baseline Model)
- CNN Model
- LSTM-RNN Model

Results

Results (continued)

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
<th>Model Accuracy</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.39</td>
<td>RNN w/text &amp; tags</td>
<td>0.66</td>
</tr>
<tr>
<td>CNN w/text</td>
<td>0.81</td>
<td>RNN w/text, summary, tags</td>
<td>0.67</td>
</tr>
<tr>
<td>RNN w/text</td>
<td>0.69</td>
<td>RNN w/text, tags, kudos</td>
<td>0.67</td>
</tr>
<tr>
<td>RNN w/text &amp; summary</td>
<td>0.83</td>
<td>RNN w/text, summary, tags, kudos</td>
<td>0.67</td>
</tr>
<tr>
<td>(LSTM-RNN w/text &amp; tags)</td>
<td>0.39</td>
<td>(not implemented)</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Discussion & Conclusions
- Performance: Neural networks models significantly outperformed the baseline model. For RNN, adding tags in addition to text provided the most significant accuracy increase, likely because tags provide clear labels on what content is expected in the text itself.
- Although low performance, consistent across different architectures.
- Over-fitting on training set: add more data, different regularization techniques, cross-validation
- Common sources of errors: included the accuracy of authors’ own tags and misunderstanding context (some LGBT+ stories classified as ‘mature’).

Future Work
- With a larger dataset, we can experiment with Facebook “fastText”, a library for efficient learning of word representations and sentence classification.
- Investigate and expand architecture search to more complex models, use of transformers such as Google BERT.
- Spending more time getting high-quality labeled data to validate model performance.

References
