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 sexually 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

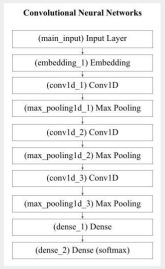
Models

Logistic Regression (Baseline)

- One fully connected layer perceptron using linear activation
- Output softmax layer for multi-class classification

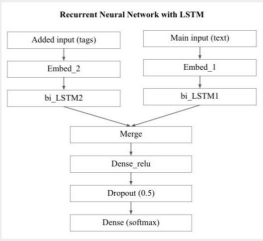
CNN Model:

- Consists of 3 of the ID convolutional layers, each followed by a max pooling layer, flattened, and then connected to a fully connected layer
- 256 filters of size 5
- Pooling windows of size 2
- ReLu activation
- Softmax for multi-class classification

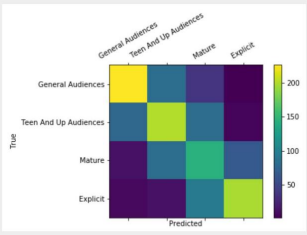


RNN Model:

- Bidirectional LSTM layer for text
- Tested sub-models by adding one extra input (summary, tags, kudos) at a time, then combined → assess how the features affect model performance
- Concat another bidirectional LSTM layer output as needed
- 256 hidden units
- ReLu activation
- Softmax for multi-class classification
- A dropout rate of 0.5 for regularization



Results



Results (continued)

Model	Accuracy	Model	Accuracy
Logistic Regression	**0.27**	RNN w/ text & kudos	**0.50**
CNN w/ Text	**0.51**	RNN w/ text, summary, tags	**0.57**
RNN w/ Text	**0.49**	RNN w/ text, tags, kudos	**0.57**
RNN w/ text & summary	**0.53**	RNN w/ text, summary, tags, kudos	**0.57**
RNN w/ text & tags	***0.59***	*State-of-the-art (not implemented)*	*reported >0.9*

(LSTM-RNN with story text and tags)

Class	Accuracy	Precision	Recall	F1-score
General Audience	0.66	0.74	0.64	0.69
Teen and Up	0.55	0.70	0.66	0.68
Mature	0.48	0.41	0.48	0.44
Explicit	0.64	0.53	0.55	0.54

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 LGBTQ+ 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

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543)
Yoon Kim. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882
Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.
Graves, A. (2012). Long short-term memory. In Supervised sequence labelling with recurrent neural networks (pp. 37-45). Springer, Berlin, Heidelberg.
Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.
Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555
Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.
Lee, J. Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827.
Wang, X., Jiang, W., & Luo, Z. (2016). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2428-2437).