A Genre-Aware Attention Model to Improve the Likability Prediction of Books

Suraj Maharjan[⋆] **Manuel Montes-y-Gómez**[†] **Fabio A. González**[‡] **Thamar Solorio**[⋆]
[⋆]Department of Computer Science, University of Houston
[†]Instituto Nacional de Astrofisica Optica y Electronica, Puebla, Mexico
[‡]Systems and Computer Engineering Department, Universidad Nacional de Colombia
{smaharjan2,tsolorio}@uh.edu
mmontesg@ccc.inoep.mx, fagonzalezo@unal.edu.co

Abstract

Likability prediction of books has many uses. Readers, writers, as well as the publishing industry, can all benefit from automatic book likability prediction systems. In order to make reliable decisions, these systems need to assimilate information from different aspects of a book in a sensible way. We propose a novel multimodal neural architecture that incorporates genre supervision to assign weights to individual feature types. Our proposed method is capable of dynamically tailoring weights given to feature types based on the characteristics of each book. Our architecture achieves competitive results and even outperforms state-of-the-art for this task.

1 Introduction

Book likability prediction is an important but challenging task. It can be a valuable resource for supporting buying decisions. The experience of choosing a book can be daunting for readers, considering the overwhelming number of books being published. On the other hand, being able to predict how a book will fare in the market has relevant economic value for the publishing industry in order to increase their revenue. The current process is guided by humans, but this is error-prone, very subjective, and a non-scalable process.

An alternative to the human-guided process is to design a reliable automatic system that predicts the likability of books. Such a system, we argue, must be able to take into account all of the many aspects involved in the eventual success of a book. These include not only the topic of the book and the writing style of the author, but in the case of creative writing, also include elements such as creativity, plot structure, and the flow of sentiments (Hall, 2012; Archer and Jockers, 2016; Maharjan et al., 2018; Kar et al., 2018). Other relevant aspects influencing readers' interest for a book could be the cover and the title of the book.

We believe that in addition to the ability to incorporate the different aspects, it is equally important to have a robust mechanism that gives higher weight to the most relevant aspects, while at the same time disregards the noisy or redundant aspects. Traditionally, this is achieved by searching through multiple feature combination experiments for an optimal combination of different feature types (Yang and Pedersen, 1997; Forman, 2003). The main problem with these methods is that they are time-consuming and too rigid. The resulting feature types are fixed for every document. In some books, the style of the author may contribute more than the specific topic, whereas the reverse may be true for other books. These methods lack the ability to dynamically assign weights to different features based on the characteristics of a particular test instance. Most likely, a more flexible scheme that adjusts feature weights based on the current book, can lead to better results.

This paper attempts to solve this problem by introducing a novel method that is capable of automatically combining information from different aspects and learning to weight them dynamically for each book in order to improve likability prediction. Our method also extends the attention model to incorporate domain specific information like the genre of books. As far as we know, we are the first to use genre supervision while computing attention weights and to use them in the field of feature importance. There are many potentially relevant aspects of books that make them likable by readers. Here we focus on different textual modalities, like the lexical, stylistic, syntactic, and neural representations, along with the visual modality from book covers. Our main contributions in this paper are as follows:

- We propose a novel neural architecture,

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3381–3391
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

which incorporates genre supervision for computing attention weights to learn the importance of hand-engineered and deep learning features coming from different modalities for predicting the likability of books.

- We show through our results that an adaptive combination of features with the genre-aware attention model performs better than strong baselines and also outperforms state-of-the-art.
- We present visualizations that increase interpretability of our results and also demonstrate the advantages of our model.

Along with these contributions, we also show that book cover images contain sufficient information by themselves to perform likability classification, although their contribution becomes negligible in the presence of strong textual features.

2 Methodology

We propose a model that we call Genre-Aware Attention model (GA), which dynamically weights features coming from different aspects of a book by using genre supervision. We first feed our textual and visual features through a non-linear layer to train higher feature representations. We then use our genre-aware attention model to compute appropriate weights for these feature representations. The motivation to add genre information comes from our previous work showing that adding genre classification as an auxiliary task to success prediction improved results (Maharjan et al., 2017). Moreover, it is also reasonable to expect that different genres should have different sets of features that are more relevant when trying to predict whether readers will like the book. For instance, in *Science Fiction*, the theme may be more relevant than say, in *Drama*, where the characters and their interactions or their struggles might be more relevant for likeability prediction.

2.1 Features

For our features, we build on the work by Maharjan et al. (2017) that provides a comprehensive exploration of different hand-crafted features and neural representations. They showed that a combination of writing density (WR) (distribution of word, character, sentences, and paragraphs), Book2Vec, and recurrent neural network representations (RNN) works well for books. Similar to their work, our textual features consist of word, character, and typed character n-grams (Sapkota et al., 2015), syntactic features, sentiment and sentic concepts and scores (SCS) (Cambria et al., 2014), style-related WR and readability (R), and neural representations learned using Word2Vec (Mikolov et al., 2013), Doc2Vec and RNN. We consider these categories of the textual features as different modalities or sources since they capture different aspects of a book and are generated by different processes. In addition to these features, we also add visual information extracted from the book covers. To extract the visual features, we rely on state-of-the-art visual feature extractor methods like VGG (Simonyan and Zisserman, 2014) and Resnet (He et al., 2016), initialized with the weights trained on the Imagenet dataset.

2.2 Genre-Aware Attention Model

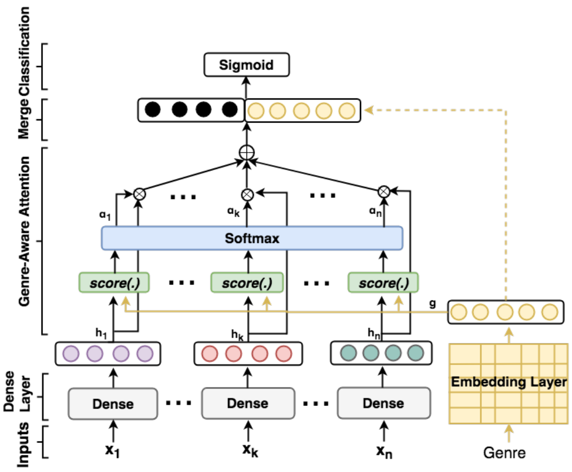


Figure 1: Genre-Aware Attention Model.

Figure 1 shows the overall architecture of our Genre-Aware Attention model. Let X be a collection of books. For a book $x \epsilon X$, let $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ be the feature representations from the different textual modalities and the visual modality. Since these features have different dimensions, we first pass them through a non-linear layer to project them into a space with the same dimension using Equation 1. This will allow us to perform a weighted average of features from different modalities according to their importance:

$$\mathbf{h_i} = selu(\mathbf{W_h x_i} + \mathbf{b_h}) \qquad (1)$$

The source code and data for this paper can be downloaded from https://github.com/sjmaharjan/genre_aware_attention where i is the index of the modality whose feature representation is fed into the network, $\mathbf{W_h}$ is the weight matrix, $\mathbf{x_i}$ is the input feature vector for the ith modality, $\mathbf{b_h}$ is the bias, and $selu$ (Klambauer et al., 2017) is the activation function. All of these feature vectors from different modalities may not be equally important to the final representation and in turn to the likability prediction task. We use the genre-aware attention mechanism to learn the importance of each of these features towards our task and aggregate them to get the final representation. The final book representation $\mathbf{r}$ is the weighted sum of $\mathbf{h_i}$ vectors:

$$\mathbf{r} = \sum_i \alpha_i \mathbf{h_i} \tag{2}$$

where α_i are the weights measuring the importance of the different modalities. The GA model combines the genre vector $\mathbf{g}\epsilon\mathbb{R}^{d_g}$ (d_g being the dimension of the genre vector) while computing the α weights. The α_i weights are computed as follows:

$$\alpha_i = \frac{\exp(score(\mathbf{h_i}, \mathbf{g}))}{\sum_{i'} \exp(score(\mathbf{h_{i'}}, \mathbf{g}))} \tag{3}$$

and the $score(.)$ function is defined as:

$$score(\mathbf{h_i}, \mathbf{g}) = \mathbf{v}^T selu(\mathbf{W_a h_i} + \mathbf{W_g g} + \mathbf{b_a}) \tag{4}$$

where, $\mathbf{W_a}$ and $\mathbf{W_g}$ are the weight matrices and $\mathbf{v}$ is the weight vector. The addition of $\mathbf{W_g g}$ incorporates genre supervision. These parameters are shared across all modalities. This will prevent parameter explosion that is likely to occur when the number of modalities is high, which is the case for us. To further investigate the effect of the genre, we also experiment by concatenating the genre vector $\mathbf{g}$ to the final weighted averaged vectors from different modalities $\mathbf{r}$ to obtain $\mathbf{r}; \mathbf{g}$. The dotted line from genre vector $\mathbf{g}$ represents this in Figure 1. We then use a non-linear layer with sigmoid activation to project the book representation (either $\mathbf{r}$ or the concatenation $\mathbf{r}; \mathbf{g}$) to class probabilities.

$$\hat{p} = \sigma(\mathbf{W_c r} + \mathbf{b_c}) \tag{5}$$

where, $\mathbf{W_c}$ is the weight matrix and $\mathbf{b_c}$ is the bias vector. Finally, we train the network by minimizing the binary cross entropy loss using Adam (Kingma and Ba, 2015).

$$L = -\sum_i p_i \log \hat{p}_i \tag{6}$$

where, p_i and $\hat{p}_i$ are true labels and predictions, respectively.

3 Dataset

We experiment with the dataset collected by Maharjan et al. (2017). The dataset consists of books from eight different genres: *Detective Mystery*, *Drama*, *Fiction*, *Historical Fiction*, *Love Stories*, *Poetry*, *Science Fiction*, and *Short Stories*. These books have been reviewed by at least ten reviewers. Based on the average rating received by the books on Goodreads[1], they labeled the books into two categories: *Successful* and *Unsuccessful*. The collection has a total of 1,003 books. However, the dataset did not include book covers. We augmented this dataset by downloading the covers from Goodreads. Since this dataset only contains publicly available books, all of them were published over 100 years ago. Some of the books only had the title of the book on a plain background as their cover images on Goodreads. We manually searched for these books with Google Image Search and found the actual covers for most of them. However, even after an exhaustive search, we were unable to obtain proper covers for 21 books. We did not remove these books from the dataset for the sake of comparison with Maharjan et al. (2017).

4 Experiments and Results

We used the same train and test folds as used by Maharjan et al. (2017) for all of our experiments. The dataset consists of 349 books belonging to the *Unsuccessful* class and 654 books belonging to the *Successful* class. Since the dataset is imbalanced, they as well as we use weighted F1-score to evaluate the performance.

4.1 Baselines

The most naive baseline will be to predict the majority class for all test instances. This majority class baseline yields a weighted F1-score of 50.6% for the likability classification task. This baseline will help to understand whether our proposed model is actually learning from the data at all. Apart from this, we compare with the results from Maharjan et al. (2017) and we also define several other baselines to validate the superiority of our proposed model. All of the

[1] https://www.goodreads.com/

baseline methods are listed below:

Mah'17: The current state-of-the-art for this dataset by Maharjan et al. (2017). They have several results on various combinations of textual features.

Mah'17+Vis: This method is the extension of the Mah'17 method with the addition of visual features. Similar to them, we use the SVM classifier under two settings: Single-task (ST) and Multitask (MT). In ST, we simply predict the likability of books. In MT, along with predicting likability, we also predict genre simultaneously. This experiment will allow us to make a direct comparison with Mah'17 regarding the effect of adding visual modalities.

Concatenation: Similar to GA, we first feed the features from different modalities through a nonlinear layer each having the same number of neurons. We then concatenate them to obtain the final representation for a book. We send this representation to a sigmoid layer for success prediction.

Average Pooling: Instead of concatenation, we take an average of the features after passing them through the non-linear layer. This is also comparable to an attention model assigning equal weights to all modalities.

Attention: We use a multilayer perceptron to learn the appropriate weights for each of the features from different modalities. This method is similar to our proposed method, except that we do not use genre information for computing the attention weights. We compute the $score(.)$ as $\mathbf{v}^T selu(\mathbf{W_a h_i} + \mathbf{b_a})$, without the genre information. This experiment will help us understand the importance of genre in computing weights for the feature types.

Bilinear Model: We combine the non-linear transformed modalities $\mathbf{h_1}, \ldots, \mathbf{h_n}$ using a bilinear form $(\mathbf{h_i}^T \mathbf{W_b h_j} + \mathbf{b_b})$, where $\mathbf{W_b} \epsilon \mathbb{R}^{k \times d_{h_i} \times d_{h_j}}$ is the weight tensor and $\mathbf{b_b}$ is the bias vector (Socher et al., 2013; Laha and Raykar, 2016; Fukui et al., 2016; Gao et al., 2016). This operation gives us a k-dimensional vector. In the case of more than two modalities, we first create $\binom{n}{2}$ pairs of these modalities and combine each of them using a bilinear form. The final book vector is the concatenation of the resulting vectors from each of these pairs. Bilinear models are used in the visual question answering community to fuse visual and textual information (Fukui et al., 2016). This experiment will help us understand how our proposed model compares with other state-of-the-art multimodal approaches.

For all these models as well, we also performed additional experiments by concatenating the genre vector $\mathbf{g}$ with the final representations $\mathbf{r}$ obtained from each of these models to study the significance of including genre explicitly for likability prediction.

4.2 Experimental Settings

For the experiments involving the SVM classifier, we tuned the C hyper-parameter with values {1e-4, . . . , 1e4} by performing three-fold grid search over the training data and then used the best hyperparameters to train the final model. For the neural network experiments, we first separated 20% of the training data as a validation set and tuned dropout rates {0.2, 0.4, 0.5}, different weights initialization schemes {Glorot Uniform (Glorot and Bengio, 2010), LeCun Uniform (LeCun et al., 1998)}, learning rate with Adam {1e-4, . . . , 1e-1}, number of hidden neurons in different layers {100, 200}, and batch size {1, 4, 8} with early stopping criteria. We initialized the genre embeddings with orthogonal vectors.

4.3 Results

Table 1 shows and compares our results with different baselines. We experimented with both low performing as well as high performing features and their combinations as found by Maharjan et al. (2017). We obtained the best weighted F1-score of 75.4% with our proposed GA+Genre concatenation model. This is 4.2% and 8.7% above the corresponding results reported by Mah'17 with their MT and ST settings, respectively. We also see a significant* improvement of 6.5% (over MT) and 22.2% (over ST) when using *RNN* features with our proposed method as compared to Mah'17. These results support the superiority of our method in learning high-quality book representations than Mah'17's state-of-the-art methods.

The results also show that it is beneficial to use at least some form of attention over just *Average Pooling*. This suggests that using all available features without regards to their individual contribution towards the task at hand can actually worsen the performance. Our proposed model is capable

*We used the McNemar significance test.

	Mah'17		**Mah'17+Vis**		**Concatenation**		**Average Pooling**		**Bilinear**		**Attention**		**Genre Attention**	
Features	**ST (SVM)**	**MT (SVM)**	**ST (SVM)**	**MT (SVM)**	**-**	**+ Genre**	**-**	**+ Genre**	**-**	**+ Genre**	**-**	**+ Genre**	**-**	**+ Genre**
Bigram	65.9	68.5	68.8	65.6	67.6	60.3	61.9	59.1	66.9	67.9	62.1	69.2	65.9	**70.2**
Clausal	50.6	55.8	59.0	53.9	57.0	58.2	60.7	60.1	55.9	**67.4**	50.6	66.5	50.7	**67.4**
Readability	50.6	63.4	60.8	58.8	54.3	59.3	59.3	58.3	60.8	60.8	53.6	65.7	52.1	**67.6**
RNN	52.9	68.6	68.9	68.9	71.9	69.3	67.6	68.4	69.5	71.9	71.4	70.7	**75.1**	73.7
Book2Vec (DBoW+DMM)	69.5	72.9	69.5	72.9	66.1	65.4	64.4	63.9	61.9	66.5	68.7	68.2	70.5	**73.4**
SCS,WR,Typed n-gram	72.0	71.0	69.2	68.8	67.9	60.8	63.1	66.7	67.6	69.2	69.2	69.1	72.5	**73.0**
WR,Book2Vec,RNN	70.1	73.5	66.1	71.6	70.6	69.0	66.7	70.1	70.1	66.4	70.0	70.5	**73.7**	73.1
All best handcrafted + RNN	66.7	71.2	69.3	72.2	68.1	67.4	66.6	65.7	50.6	50.6	69.8	70.6	70.2	**75.4**

Table 1: Weighted F1-scores(%) for different multimodal methods for books' likability classification task (ST=Single Task, MT=Multitask, SCS=Sentic Concepts and Scores, WR=Writing Density, RNN=Recurrent Neural Network Representations, + Genre= genre embedding $\mathbf{g}$ concatenated with the final book vector $\mathbf{r}$). Our baselines and proposed method include visual features as well.

of assigning importance to these features and the results clearly show that this works to our benefit. The results also demonstrate the added advantage of using genre supervision while computing feature weights. There is a considerable improvement in the performance over the *Attention* method after taking the genre information into account using our GA method. We suspect that the genre meta-information is helping to learn more specialized weights based on the genre of the books.

With the neural baseline methods like *Concatenation* and *Average pooling*, we do not always see improvement in performance after combining the genre information with the final book representation. Apart from these two, the combination of genre information does improve the results for other methods. The *Bilinear* and *Attention* methods seem to be able to utilize this information well. However, none of these methods are capable of doing better than our method. GA and GA+Genre concatenation models always achieve the best performance for all experiments. This also illustrates the latent power of our method to better exploit domain information like genre for performance improvement.

Another interesting finding is that with the addition of multiple modalities, the performance of Bilinear methods degrades to the majority class baseline (Table 1, last row). This may be due to parameter explosion with the increase in the number of modalities. However, our method is able to selectively weight the feature sources and discount the effect of redundant and irrelevant features to obtain the best performance, even with a larger number of modalities. In short, we see that our proposed method is able to cope with feature pollution and parameter explosion.

Next, we investigate the addition of visual information with the textual information for the likability prediction of books. Under the ST setting with SVMs, we see that the low performing textual features are benefited significantly by the addition of visual features, sometimes even outperforming the MT setting (Table 1, rows 1-4). However, the visual features are not able to contribute much when combined with strong textual features that were already performing well. On the other hand, for the MT setting, the performance decreases for most of the feature combinations with the addition of the visual modality. We suspect that book covers are not very helpful at predicting genre and thus the MT setting does not do well with additional visual features.

Visual Results: Our next set of experiments considers only the visual information for books' likability prediction. Even though we do believe that this current corpus might not be ideal for using cover features, we believe it is still interesting to explore whether the current book covers have sufficient information to perform likability classification with reasonable accuracy. We used VGG and Resnet to extract features from book cover images. We replaced the top layers by a dense layer of 256 neurons, and a classification layer (eight neurons with softmax for genre classification and one neuron with sigmoid activation for success classification). We also added a dropout layer in between the dense and the classification layer. The layers were initialized with weights trained on the Imagenet dataset.

Tasks	**Likability**		**Genre**	
Features	**ST (F1)**	**MT (F1)**	**ST (F1)**	**MT (F1)**
VGG	59.9	**61.8**	**24.7**	24.1
Resnet	58.7	60.0	24.6	24.0
VGG + SVM	58.8	57.7	25.4	19.6
Resnet + SVM	**59.5**	54.5	**25.9**	19.7

Table 2: Weighted F1-scores(%) for visual features for likability and genre classification of books with Single Task (ST) and Multitask (MT) settings.

Table 2 shows the results with only the visual features for likability and genre classification of books under the ST and MT settings. We obtain the highest weighted F1-score of 61.8% and 25.9% for likability and genre classification tasks, respectively. With the neural experimental setup, we get similar performance under the ST and the MT settings for both tasks. We also experimented with transferring the visual feature vectors to the SVM classifier under the ST and the MT settings. We saw a decrease in performance under the MT settings with both the VGG and Resnet features (Table 2, last two rows). This is the opposite of the Mah'17 results for the textual features as seen in Table 1. The reason behind this may be due to the fact that the textual features are better at both the likability and the genre classification tasks individually, whereas the visual features are not as good as the textual features for the genre classification task. Iwana et al. (2016) also concluded that genre classification with book covers is a difficult task as book covers have images with few visual features or ambiguous features.

These results also empirically verify the decrease in performance for the MT settings with the addition of visual features for likability prediction. Although these results are significantly lower (p<0.001*) than our best results, they are still better than the majority baseline (50.6% and 10.7% for success and genre classification tasks, respectively). These results support our hypothesis that the books' cover images correlate with the likability of books. Also, they dictate for the need of extracting other features that consider different aspects of books.

5 Attention Weights Visualization

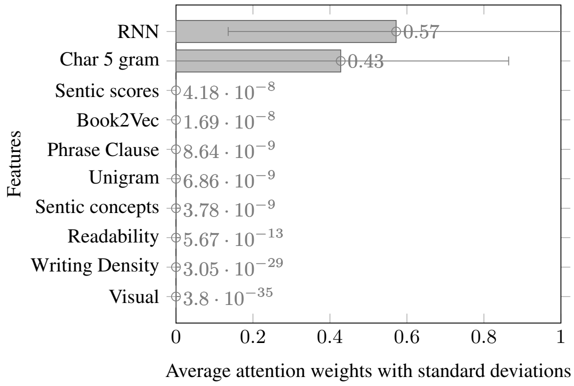


Figure 2: Feature importance for the feature combination: All best handcrafted, RNN, and visual (RNN = Recurrent Neural Networks).

Figure 2 shows the average attention weights given by the best model to the different feature types for the books in the test set. The purpose of this visualization is to understand which aspects of a book are deemed to be more important by the model. The figure shows that most of the weights are assigned to the *Char* 5-*gram* and the *RNN* representations. The results in Table 1 also support that *RNN* features are indeed one of the most important features. The contribution of the visual representations is negligible in the presence of strong textual features. The results in Table 1 also validate this finding. These two textual features also dominate over the other weaker textual features. In the same way, as for the visual features, we see negligible weights assigned to the other textual features as well. Our model seems to have learned that the *Char* 5-*gram* and the *RNN* features can cover the information given by the rest of the features. The *Char* 5-*gram* feature is capable of capturing the content, topic, and style of a text and as such might be able to cover the *Unigram* and *Sentic Concepts* features. Likewise, the *Book2Vec* features may be non-essential in the presence of the *RNN* representations. The model is reducing redundant information that does not aid the classification task and instead might just add noise.

In order to validate that features given the top weights by our model are indeed the best features for the task, we ran an experiment with only the *Char* 5-*gram* and the *RNN* features. We were able to obtain a weighted F1-score of 73.6% with just these two features. This score is close to the best score of 75.4%, showing that these features are indeed good features for the task. Also, note that our model was able to figure out this feature set automatically, while using traditional methods would have entailed performing multiple experiments ($2^n - 1$ experiments, where n is the number of feature types) which is often times not possible to do exhaustively. There is still an extra boost when using the whole feature set rather than using just the *Char* 5-*gram* and the *RNN* features. Since our method tailors the feature weights to each book and its genre as well, the boost likely comes from the presence of other visual and textual features, which at least for some books must be informative.

We just saw that only two out of all feature types are given most of the weights. However, the results in Table 1 show that even without these fea-

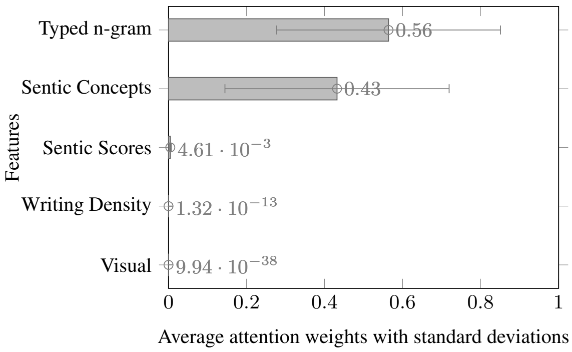


Figure 3: Feature importance for feature combination Sentic Concepts and Scores, Writing Density, Typed n-gram, and Visual.

tures, we are able to get good performances. To understand this, we analyze a model that does well without these two features. Figure 3 plots the average attention weights for a model with *Sentic Concepts and Scores*, *Writing Density*, *Typed n-grams*, and *Visual* features' combination. We see that the weights now shift to *Typed n-grams*, and *Sentic Concepts and Scores*. The topic and content captured through *Sentic concepts* and the style with *Typed n-grams* prove important. These features capture different aspects of books and are not strongly correlated with one another. Our model is capable of figuring out that in the absence of *Char* 5-*gram*, which encompasses all this information, these other features need to be made more prominent. We can also see that the model knows three different feature types to capture the same amount of information as captured by the two best ones from before.

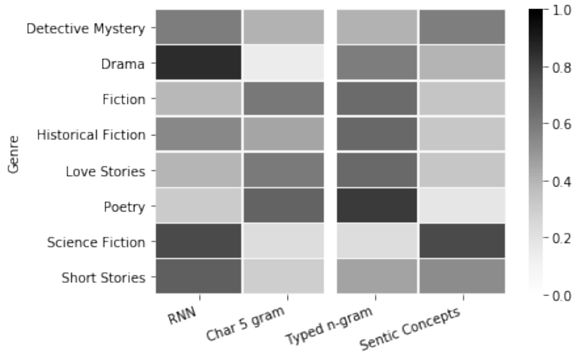


Figure 4: Average attention weights with respect to genre for the best features from two models.

Figure 4 further breaks down the attention weights by genre for *RNN* and *Char* 5-*gram*, and *Typed n-grams* and *Sentic Scores and Concepts*. From the figure, it is evident that different genres respond differently to each feature type. Comparing the two models, we see that *Char* 5-*gram* activates similarly to *Typed n-gram*, and *RNN* similarly to *Sentic concepts* for different genres.

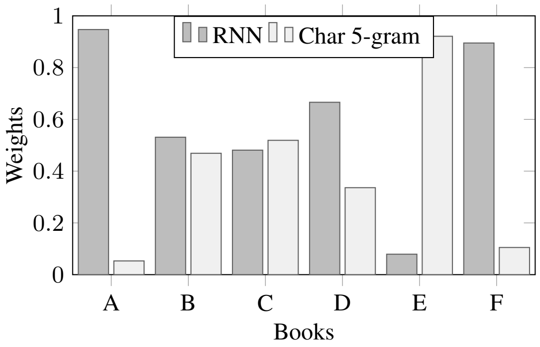
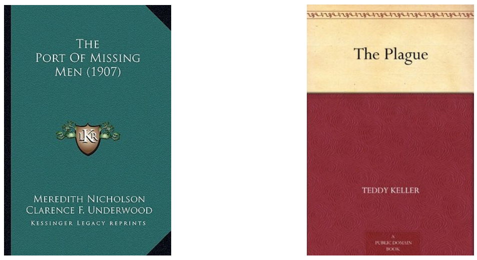


Figure 5: Feature importance as assigned by attention weights for two most important features for six different books: *A=The Count of Monte Cristo, B=The Scouts of the Valley, C=The Daughter of the Commandant, D=The Northern Light , E=The Great Secret, F=House of the Seven Gables.*

Figure 5 shows the feature importance for the *Char 5-gram* and *RNN* feature types for six different books having different attention weights for the two features. This validates our assumption that the model is able to dynamically learn and assign weights to different modalities, not only according to the genre but also according to the characteristics of each book. The high variance of attention weights for the top features in Figures 2 and 3 also support this claim. This gives an edge to our model and helps it excel over all other methods.

6 Error Analysis

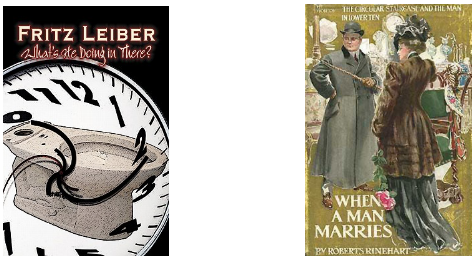


(a) The Port of Missing Men

(b) The Plague

Figure 6: Books misclassified by visual features but correctly classified when textual features are added.

We took the books that were misclassified when we used the visual features only but were correctly predicted after the combination with the textual dimensions. As expected, we found that the books without proper covers were misclassified by visual features. But upon addition of other textual



(a) What's He Doing in There? (b) When a Man Marries

Figure 7: Books misclassified by visual and text features' combination but correctly classified when only visual features are used.

features, they were correctly classified. Figure 6 shows the cover image of two of such books. The fact that the cover has no images with just plain background, and title, leaves little information for the visual modality. Similarly, we also analyzed the books that were correctly classified by visual features only and misclassified when textual features were added. Figure 7 shows two such books. Both the cover image and the title (present in the cover) of these two books seem to be interesting and are very likely to attract a reader's attention.

7 Related Work

Prior works have shown that stylistic traits to be useful features to predict success of books (Ashok et al., 2013; Underwood and Sellers, 2016; Maharjan et al., 2017). Ashok et al. (2013) used stylistic features extracted using the first 1K sentences from books to classify highly successful literature from less successful literature. van Cranenburgh and Bod (2017) used lexical and rich syntactic tree features to distinguish the degrees of high and less literary novels. Louis and Nenkova (2013) defined genre-specific and general features to predict the article quality in science journalism domain. Maharjan et al. (2017) compared their work with Ashok et al. (2013) and presented a new dataset for the book success prediction task. Their multitask approach with the combination of deep representations and hand-crafted features improved the classification results. Maharjan et al. (2018) also showed that modeling sequential flow of emotions across entire books improves likability prediction of books. Iwana et al. (2016) used neural networks to learn relationships between book covers and genre. They showed that book covers tend to have carefully designed color and tone, objects, and text. Our work relies on prior works' hand-engineered and deep learning features but differs in a way how these features are combined to produce a meaningful book representations.

The attention mechanism (Bahdanau et al., 2014) has been successfully applied in enhancing the document representation for several text classification (Zhang et al., 2016; Wang et al., 2016b), sentiment classification (Kar et al., 2017; Nguyen and Shirai, 2015; Wang et al., 2016a), question answering (Tan et al., 2015; Chen et al., 2016a; Hermann et al., 2015), named entity recognition (Bharadwaj et al., 2016; Aguilar et al., 2017), summarization (Rush et al., 2015), image-captioning (Xu et al., 2015) tasks. Zhang et al. (2017) used summary vectors and position vectors while computing the attention weights for the slot filling problem. Chen et al. (2016b) applied user preferences and product characteristics as attentions to words and sentences in reviews to learn the final representation for the sentences and reviews. They used these representation to do the sentiment classification task and showed that adding user information was much more effective in enhancing the document representations than the product information. Similar to their idea, we fuse the genre information while computing attention weights.

8 Conclusions and Future Work

We present a novel method to fuse the information coming from different modalities using a genre-aware attention mechanism to predict the likability of books. We showed that our proposed method outperforms strong baselines and state-of-the-art by learning to distinguish the important features from irrelevant or redundant ones. Other methods either suffered from feature pollution or parameter explosion and yielded low performance. Along with this, our results also showed that the book cover images by themselves also have sufficient information to perform success prediction. However, the difficulty in predicting genre from book covers decreased the performance in multi-task settings with additional visual features. We also used different visualizations to support our findings and improve interpretability of our model. As future work, we will extend the proposed method to include components that learn weights for individual feature elements and not only the entire feature type. This could likely result in higher quality multimodal representations.

Acknowledgments

This work was partially funded by the National Science Foundation under award 1462141. We would like to thank the anonymous reviewers, John Arevalo, Deepthi Mave, and Prasha Shrestha for reviewing the paper and providing helpful comments and suggestions.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A multitask approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Jodie Archer and Matthew L Jockers. 2016. *The Bestseller Code: Anatomy of the Blockbuster Novel*. St. Martin's Press.

Vikas Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764, Seattle, Washington, USA. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. Senticnet 3: A common and commonsense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1515–1521. AAAI Press.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016a. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016b. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659, Austin, Texas. Association for Computational Linguistics.

Andreas van Cranenburgh and Rens Bod. 2017. A data-oriented model of literary language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1228–1238, Valencia, Spain. Association for Computational Linguistics.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas. Association for Computational Linguistics.

Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2016. Compact bilinear pooling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

James W Hall. 2012. *Hit Lit: Cracking the Code of the Twentieth Century's Biggest Bestsellers*. Random House.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2016. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*.

Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2017. Ritual-uh at semeval-2017 task 5: Sentiment analysis on financial data using neural networks. In *Proceedings of the 11th International Workshop on*

Semantic Evaluation (SemEval-2017), pages 877–882, Vancouver, Canada. Association for Computational Linguistics.

Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2018. Folksonomication: Predicting tags for movies from plot synopses using emotion flow encoded neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2879–2891, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. *CoRR*, abs/1706.02515.

Anirban Laha and Vikas Raykar. 2016. An empirical evaluation of various deep learning architectures for bi-sequence classification tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2762–2773, Osaka, Japan. The COLING 2016 Organizing Committee.

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. 1998. Efficient backprop. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, pages 9–50, London, UK, UK. Springer-Verlag.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A. González, and Thamar Solorio. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, Valencia, Spain. Association for Computational Linguistics.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. Gonzalez, and Thamar Solorio. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations (ICLR), Workshop*.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514, Lisbon, Portugal. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–102, Denver, Colorado. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 926–934. Curran Associates, Inc.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*.

Ted Underwood and Jordan Sellers. 2016. The longue durée of literary prestige. *Modern Language Quarterly*, 77(3):321–344.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. 2016a. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Zhongqing Wang, Yue Zhang, Sophia Lee, Shoushan Li, and Guodong Zhou. 2016b. A bilingual attention network for code-switched emotion prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1634, Osaka, Japan. The COLING 2016 Organizing Committee.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81.

Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *Icml*, volume 97, pages 412–420.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.