Investigating the Relationship between Literary Genres and Emotional Plot Development

Evgeny Kim, Sebastian Padó and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{evgeny.kim,sebastian.pado,roman.klinger}@ims.uni-stuttgart.de

Abstract

Literary genres are commonly viewed as being defined in terms of content and style. In this paper, we focus on one particular type of content feature, namely lexical expressions of *emotion*, and investigate the hypothesis that emotion-related information correlates with particular genres. Using genre classification as a testbed, we compare a model that computes *lexicon-based emotion scores* globally for complete stories with a model that tracks *emotion arcs* through stories on a subset of Project Gutenberg with five genres.

Our main findings are: (a), the global emotion model is competitive with a large-vocabulary bag-of-words genre classifier (80 % F_1); (b), the emotion arc model shows a lower performance (59 % F_1) but shows complementary behavior to the global model, as indicated by a very good performance of an oracle model (94 % F_1) and an improved performance of an ensemble model (84 % F_1); (c), genres differ in the extent to which stories follow the same emotional arcs, with particularly uniform behavior for anger (mystery) and fear (adventures, romance, humor, science fiction).

1 Introduction and Motivation

Narratives are inseparable from emotional content of the plots (Hogan, 2011). Recently, Reagan et al. (2016) presented an analysis of fictional texts in which they found that there is a relatively small number of universal plot structures that are tied to the development of the emotion *happiness* over time ("emotional arcs"). They called the arcs "Rags to riches" (rise), "Tragedy" (fall), "Man in a hole" (fall-rise), "Icarus" (rise-fall), "Cinderella" (rise-fall-rise), and "Oedipus" (fall-rise-fall). They also clustered fictional texts from Project Gutenberg[1] by similarity to emotion arc types, suggesting that their arc types could be useful for categorizing literary texts. At the same time, their analysis suffered from some limitations: it was mostly qualitative and limited to the single emotion of happiness. Crucially, they did not investigate the relationship between emotions and established literary classification schemes more concretely.

The goal of our study is to investigate exactly this relationship, extending the focus beyond one single emotion, and complementing qualitative with quantitative insights. In this, we build on previous work which has shown that stories from different literary genres tend to have different flows of emotions (Mohammad, 2011). The role of emotion has been investigated in different domains, including social media (Pool and Nissim, 2016; Dodds et al., 2011; Kouloumpis et al., 2011; Gill et al., 2008), chats (Brooks et al., 2013), and fairy tales (Alm et al., 2005).

As the basis for our quantitative analysis, we adopt the task of genre classification, which makes it possible for us to investigate different formulations of emotion features in a predictive setting. Genres represent one of the best-established classifications for fictional texts, and are typically defined to follow specific communicative purposes or functional traits of a text (Kessler et al., 1997), although we note that literary studies take care to emphasize the role of artistic and aesthetic properties in genre definition (Cuddon, 2012, p. 405), and take a cautious stance towards genre definition (Allison et al., 2011; Underwood et al., 2013; Underwood, 2016).

Traditionally, computational studies of genre classification use either *style-based* or *content-*

[1] https://www.gutenberg.org

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings, pages 17–26, Vancouver, BC, August 4, 2017.

based features. Stylistic approaches measure, for instance, frequencies of non-content words, of punctuation, part-of-speech tags and character n-grams (Karlgren and Cutting, 1994; Kessler et al., 1997; Stamatatos et al., 2000; Feldman et al., 2009; Sharoff et al., 2010). Content-aware characteristics take into account lexical information in bag-of-words models or build on top of topic models (Karlgren and Cutting, 1994; Hettinger et al., 2015, 2016). A precursor study to ours is Samothrakis and Fasli (2015), who assess emotion sequence features in a classification setting. We extend their approach by carrying out a more extensive analysis.

In sum, our contributions are:

1. We perform genre classification on a corpus sampled from Project Gutenberg with the genres *science fiction*, *adventure*, *humor*, *romantic fiction*, *detective and mystery stories*.

2. We define two emotion-based models for genre classification based on the eight fundamental emotions defined by Plutchik (2001) – *fear*, *anger*, *joy*, *trust*, *surprise*, *sadness*, *disgust*, and *anticipation*. The first one is an *emotion lexicon model* based on the NRC dictionary (Mohammad and Turney, 2013). The second one is an *emotion arc model* that models the emotional development over the course of a story. We avoid the assumption of Reagan et al. (2016) that absence of happiness indicates fear or sadness.

3. We analyze the performance of the various models quantitatively and qualitatively. Specifically, we investigate how *uniform* genres are with respect to emotion developments and discuss differences in the importance of lexical units.

2 Experimental Setup

To analyze the relationships between emotions expressed in literature and genres, we formulate a genre classification task based on different emotion feature sets. We start with a description of our data set in the following Section 2.1. The features are explained in Section 2.2 and then how they are used in various classification models (in Section 2.3).

2.1 Corpus

We collect books from Project Gutenberg that match certain tags, namely those which correspond to the five literary genres found in the Brown corpus (Francis and Kucera, 1979): adventure (Gutenberg tag: "Adventure stories"), romance ("Love stories" and "Romantic fiction"), mystery ("Detective and mystery stories"), science fiction ("Science fiction"), and humor ("Humor"). All books must additionally have the tag "Fiction". We exclude books which contain one of the following tags: "Short stories", "Complete works", "Volume", "Chapter", "Part", "Collection". This leads to a corpus of 2113 stories. Out of these, 94 books (4.4 %) have more than one genre label. For simplicity, we discard these texts, which leads to the corpus of 2019 stories with the relatively balanced genre distribution as shown in Table 1.

Genre	Count
adventure	569
humor	202
mystery	379
romance	327
science fiction	542
$\sum$	2019

Table 1: Statistics for our Gutenberg genre corpus.

2.2 Feature Sets

We consider three different feature sets: bag-of-words features (as a strong baseline), lexical emotion features, and emotion arc features.

Bag-of-words features. An established strong feature set for genre classification, and text classification generally, consists of bag-of-words features. For genre classification, the generally adopted strategy is to use the n most frequent words in the corpus, whose distribution is supposed to carry more genre-specific rather than content- or domain-specific information. The choice of n varies across stylometric studies, from, *e.g.*, 1,000 (Sharoff et al., 2010) to 10,000 (Underwood, 2016). We set $n = 5,000$ here. We refer to this feature set as BoW.

Lexical emotion features. Our second feature set, EmoLex, is a filtered version of BoW, capturing lexically expressed emotion information. It consists of all words in the intersection between the corpus vocabulary and the NRC dictionary (Mohammad and Turney, 2013) which contains 4,463 words associated with 8 emotions. Thus, it incor-

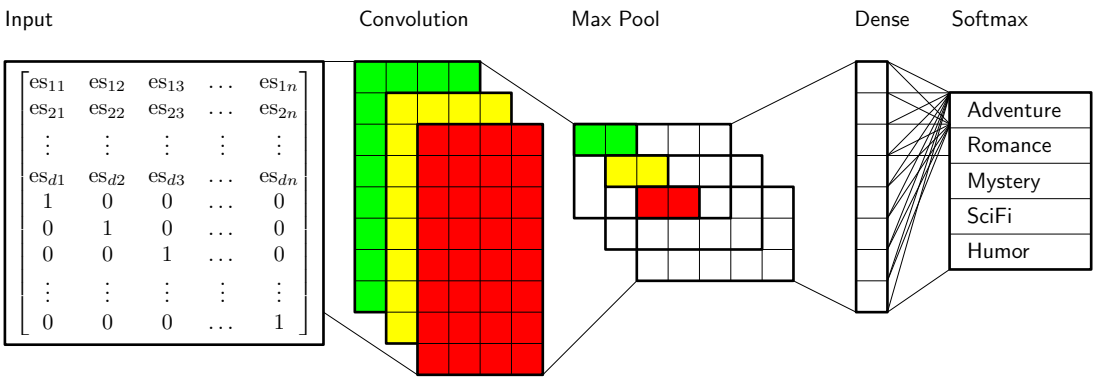

Figure 1: Architecture of CNN model

porates the assumption that words associated with emotions reflect the actual emotional content (Bestgen, 1994). We do not take into account words from "positive"/"negative" categories or those that are not associated with any emotions. This model takes into account neither emotion labels nor position of an emotion expression in the text.

Emotion arc features. The final feature set, EMOARC, in contrast to the lexical emotion features, takes into account both emotion labels and position of an emotion expression. It represents an emotion arc in the spirit of Reagan et al. (2016), but considers all of Plutchik's eight fundamental emotion classes. We split each input text into k equal-sized, contiguous segments S corresponding to spans of tokens $S = \langle t_n, \ldots, t_m \rangle$. We treat k as a hyper-parameter to be optimized (*cf.* Section 2.4).

We define a score $\text{es}(e, S)$ for the pairs of all segments S and each emotion e as

$$\text{es}(e, S) = \frac{c}{|D_e| \cdot |S|} \sum_{t_i \in S} \mathbf{1}_{t_i \in D_e} ,$$

where D_e is the NRC dictionary associating words with emotions, c is a constant set for convenience to the maximum token length of all texts in the corpus C ($c = \max_{S \in C} |S|$), and $\mathbf{1}_{t_i \in D_e}$ is 1 if $t_i \in D_e$ and 0 otherwise. This score, which makes the same assumption as the lexical emotion features, represents the number of words associated with emotion e per segment, normalized in order to account for differences in vocabulary size and book length.

The resulting features form an $8 \times k$ "emotion-segment" matrix for each document that reflects the development of each of the eight emotions throughout the timecourse of the narrative (*cf.* Section 2.3).

2.3 Models for Genre Classification

In the following, we discuss the use of the feature sets defined in Section 2.2 with classification methods to yield concrete models.

We use the two lexical feature sets, BOW and EMOLEX, with a random forest classifier (RF, Breiman (2001)) and multi-layer perceptron (MLP, Hinton (1989)). RF often performs well independent of chosen meta parameters (Criminisi et al., 2012), while MLP provides a tighter control for overfitting and copes well with non-linear problems (Collobert and Bengio, 2004).

The emotion arc feature set (EMOARC) is used for classification in a random forest, multi-layer perceptron, and a convolutional neural network (CNN). For the first two classification methods, we flatten the emotion-segment matrix into an input vector. From these representations, the classifiers can learn which emotion matters for which segment, like, *e.g.*, "high value at position 2", "low value at position 4" and combinations of these characteristics. However, they are challenged by a need to capture interactions such as "position 2 has a higher value than position 3", or similar relationships at different positions, like "highest value at position around the middle of the book".

To address this shortcoming, we also experiment with a convolution neural network, visualized in Figure 1. The upper part of the input matrix corresponds to the emotion-segment matrix from Section 2.2. Below, we add k one-hot row vectors each of which encodes the position of one segment. This representation enables the CNN with EMOARC features to capture the development of different emotions between absolute segment positions – it can compare the "intensity" of different emotions over time steps. By considering all emotions through time steps in a text, the CNN can model patterns

		Genre														
		adventure			humor			mystery			romance			sci-fi			Micro-Av.		
Model	Features	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
RF	BoW	75	89	81✔○	83	51	63✔○	82	78	80✔○	73	74	74✔○	88	87	87✔○	80✔○	80✔○	80✔○
MLP	BoW	73	75	74	69	58	63	70	70	70	66	71	68	85	84	85	74	74	74
RF	EmoLex	69	88	78	91	39	54	81	74	78	75	73	74	83	84	84	77	77	77
MLP	EmoLex	80	79	80✔✖✱	78	78	78✔✖	80	79	79✔✖✱	71	76	73✔✖✱	91	89	90✔✖✱	81✔✖✱	81✔✖✱	81✔✖✱
RF	EmoArc	51	72	59	70	27	39	55	33	41	59	63	61	70	73	71	58	58	58
MLP	EmoArc	55	60	57	56	36	44	49	47	48	57	71	63	72	65	68	58	58	58
CNN	EmoArc	56	60	58○✖	56	43	48○✖	49	46	48○✖	57	70	63○✖	74	68	71○✖	59○✖	59○✖	59○✖
SVM	Ensemble	80	86	83✱	80	78	79	87	79	83✱	79	78	78✱	90	91	91✱	84✱	84✱	84✱

Table 2: Results for genre classification on the Gutenberg corpus (percentages). We use bootstrap resampling (Efron, 1979) to test for significance of differences ($\alpha = 0.05$) (a), pairwise among the best models for each feature set (RF BoW, MLP EmoLex, CNN EmoArc) and (b), between the best individual model (MLP EmoLex) and the SVM Ensemble model. Legend: ✔ MLP EmoLex vs. RF BoW, ✖ CNN EmoArc vs. MLP EmoLex, ○ CNN EmoArc vs. RF BoW, ✱ Ensemble SVM vs. MLP EmoLex.

outside the expressivity of the simpler classifiers.

Formally, the CNN consists of an input layer, one convolutional layer, one max pooling layer, one dense layer, and an output layer. The convolutional layer consists of 32 filters of size $(8 + k) \times 4$. The max pooling layer takes into account regions of size 1×2 of the convolutional layer and feeds the resulting matrices to the fully connected dense layer with 128 neurons.

2.4 Meta-Parameter Setting

We choose the following meta-parameters: For RF, we set the number of trees to 250 in BoW and EmoLex and to 430 in EmoArc. In MLP, we use two hidden layers with 256 neurons each, with an initial learning rate of 0.01 that is divided by 5 if the validation score does not increase after two consecutive epochs by at least 0.001. Each genre class is represented by one output neuron. For the number of segments in the text, we choose $k = 6$.

3 Genre Classification Results

Table 2 shows the main results in a 10-fold cross-validation setting. The BoW baseline model shows a very strong performance of 80 % F_1. Limiting the words to those 4,463 which are associated with emotions in EmoLex significantly improves the classification of humorous and science fiction books, which leads to a significant improvement of the micro-average precision, recall, and F_1 by 1 percentage point. This result shows that emotion-associated words predict genre as well as BoW model even though fewer words, and particularly less content-related words are considered. This aspect is further discussed in the model analysis in Section 4.3 and Table 7. We test for significance of differences ($\alpha = 0.05$) using bootstrap resampling (Efron, 1979), see the caption of Table 2 for details.

Among the EmoArc models, we find the best performance (59 % F_1) for the CNN architecture underlining the importance of the model to capture emotional *developments* rather than just high or low emotion values. The EmoArc models significantly underperform the lexical approaches. At the same time, their results are still substantially better than, *e.g.*, a most frequent class baseline (which results in 12 % F_1). Thus, this result shows the general promise of using emotion arcs for genre classification, even though the non-lexicalized emotion arcs represent an impoverished signal compared to the lexicalized BoW and EmoLex models.

This raises the question of whether a model combination could potentially improve the overall result. Table 3 quantifies the complementarity of the models: Its diagonal shows true positive counts for each model. The other cells are true positive hits for the column models which were *not* correctly classified by the row model. Therefore, the additional contribution, *e.g.*, by MLP EmoArc over MLP EmoLex consists in 123 additional correctly classified texts. Conversely, 586 texts are correctly classified by MLP EmoLex, but not by MLP EmoArc.

Model	Features	RF BoW	MLP BoW	RF EmoLex	MLP EmoLex	RF EmoArc	MLP EmoArc	CNN EmoArc
RF	BoW	**1616**	110	38	184	73	98	103
MLP	BoW	228	**1498**	215	298	172	182	176
RF	EmoLex	99	158	**1555**	240	72	111	114
MLP	EmoLex	161	157	156	**1639**	133	123	131
RF	EmoArc	503	484	441	586	**1186**	194	197
MLP	EmoArc	536	502	488	584	202	**1178**	100
CNN	EmoArc	520	475	470	571	184	79	**1199**

Table 3: Model comparison. Numbers on the diagonal show the numbers of overall true positives for the respective model. Numbers in other cells denote the number of instances correctly classified by the column model, but not by the row model.

These numbers indicate that our models and feature sets are complementary enough to warrant an ensemble approach. This is bolstered by an experiment with an oracle ensemble. This oracle ensemble takes a set of classifiers and considers a classification prediction to be correct if at least one classifier makes a correct prediction. It measures the upper bound of performance that could be achieved by a perfect combination strategy. Taking into account predictions from all the models in Table 2 yields a promising result of 94 % F_1 (precision=recall=94 %), an improvement of 14 percentage points in F_1 over the previous best model.

Following this idea of a combination strategy, we implement an ensemble model that is an L1-regularized L2-loss support vector classification model that takes predictions for each book from all the models as input and performs the classification via a 10-fold cross-validation. The results for this experiment are given in Table 2 in the last row. Overall, we observe a significant improvement over the best single model, the MLP EmoLex model.

As the results show, the outcome of our ensemble experiment is still far from the upper bound achieved by the oracle ensemble. At the same time, even the small, but significant, improvement over the best single model provides a convincing evidence that further improvement of the classification is possible. However, finding a more effective practical combination strategy presents a multiaspect problem with vast solution space which we leave for future work. We now proceed to obtaining a better understanding of the relationship between emotion development and genres.

	Genre				
Emotion	Adv.	Humor	Myst.	Rom.	Sci-fi
Anger	0.21	0.20	**0.25**	0.28	0.18
Anticipation	0.12	0.10	0.17	0.15	0.16
Disgust	0.17	0.22	0.14	0.21	0.14
Fear	**0.28**	**0.22**	0.19	**0.32**	**0.19**
Joy	0.15	0.09	0.14	0.19	0.16
Sadness	0.21	0.18	0.12	0.25	0.15
Surprise	0.17	0.16	0.19	0.23	0.17
Trust	0.16	0.17	0.07	0.07	0.13

Table 4: Average uniformity of emotion-genre pairs measured by Spearman correlation. Highest uniformity per genre marked in bold.

4 Model and Data Analysis

4.1 Uniformity of Prototypical Arcs

The results presented in the previous section constitute a mixed bag: even though overall results for the use of emotion-related features are encouraging, the specific EmoArc model was not competitive. We now investigate possible reasons.

Our first focus is the fundamental assumption underlying the EmoArc model, namely that *all works of one genre develop relatively uniformly with respect to the presence of individual emotions over the course of the plot*. We further concretize this notion of *uniformity* as correlation with the *prototypical emotion development for a genre* which we compute as the average vector of all emotion scores (*cf.* Section 2.2) for the genre in question.

We formalize the *uniformity* of a emotion arc of a text with scores $\langle \text{es}_1, \ldots, \text{es}_k \rangle$ as the Spearman rank correlation coefficient with the prototypical vector $\langle \overline{\text{es}}_1, \ldots, \overline{\text{es}}_k \rangle$. Spearman coefficients range between -1 and 1, with -1 indicating a perfect inverse correlation, 0 no correlation, and 1 perfect correlation. In contrast to, *e.g.*, a Euclidean distance, this measures the emotion arc in a similar manner to the CNN.

Figure 2 shows the results in an emotion-genre matrix. Each cell presents the emotion scores for the six segments, shown as vertical dotted lines. The thick black line is the prototypical development, and the grey band around it a 95% confidence interval. We see the three most correlated (*i.e.*, most prototypical) books in blue, and the curves for the three least correlated (*i.e.*, most idiosyncratic) books in dashed red.

The figure shows that there are considerable differences between emotions-genre pairs: some of them have narrow confidence bands (*i.e.*, more uni-

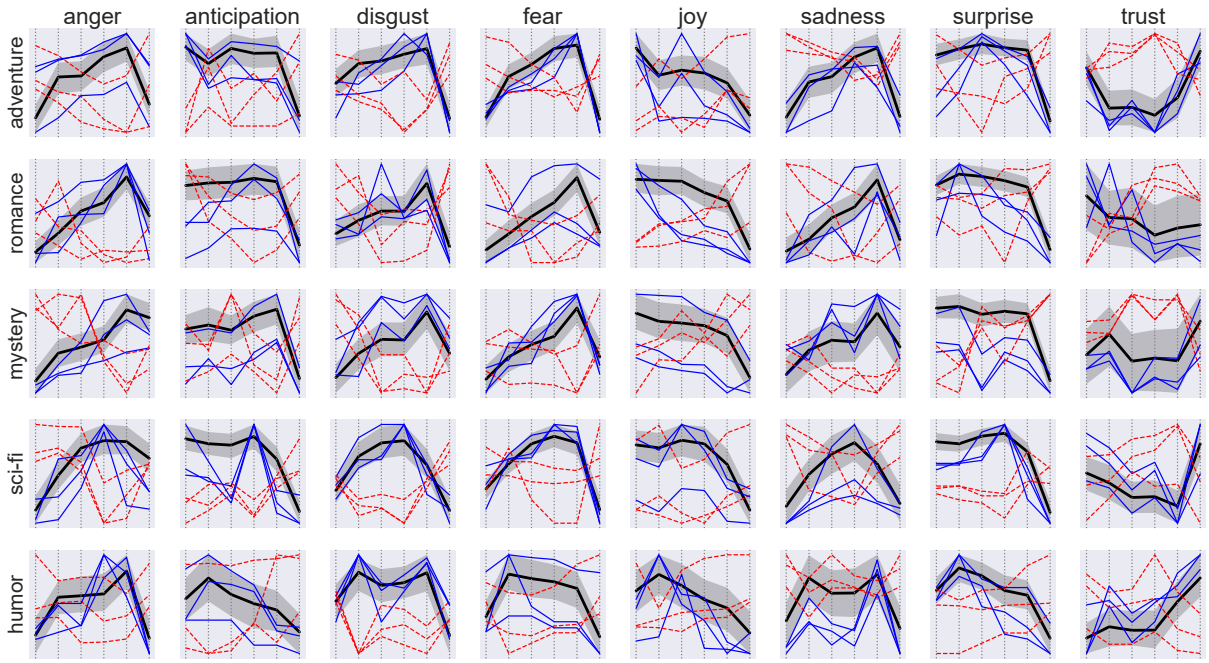

Figure 2: Emotion developments per genre. Thick black line: prototypical development. Grey band: 95% confidence interval. Blue lines: 3 most correlated books within each emotion-genre pair. Red dashed lines: 3 least correlated books within each emotion-genre pair.

form behavior), such as *fear*, while others have broad confidence bands (*i.e.*, less uniform behavior), such as *trust* and *anticipation*. Table 4, which lists the average uniformity (Spearman correlation) for each genre-emotion pair, confirms this visual impression: the emotions that behave most consistently within genres are *fear* (most uniform for four genres) and *anger* (most uniform for *mystery*). In contrast, the emotions *anticipation* and *trust* behave nonuniformly, showing hardly any correlation with prototypical development.

These findings appear plausible: *fear* and *anger* are arguably more salient plot devices in fiction than *anticipation* and *trust*. More surprisingly, *happiness/joy* is not among the most uniform emotions either. In this respect, our findings do not match the results of Reagan et al. (2016): according to our results, *joy* is not a particularly good emotion to base a genre classification on. We discuss reasons for this discrepancy below in Section 5.

At the level of individual books, Figure 2 indicates that we find "outlier" books (shown in dashed red) with a development that is almost completely inverse compared to the prototype for essentially *all* emotion-genre pairs, even the most uniform ones. This finding can have two interpretations: either it indicates unwarranted variance in our analysis method (*i.e.*, the assignment of emotions to text segments is more noisy than we would like it to be), or it indicates that the correlation between the emotional plot development and the genre is weaker than we initially hypothesized.

As a starting point for a close reading investigation of these hypotheses, Table 5 lists the three most and least prototypical books for each genre, where we averaged the books' prototypicality across emotions. We cannot provide a detailed discussion here, but we note that the list of least prototypical books contains some well-known titles, such as *La dame aux Camilias*, while the top list contains lesser known titles. A cursory examination of the emotion arcs for these works indicates that the arcs make sense. Thus, we do not find support for noise in the emotion assignment; rather, it seems that more outstanding literary works literally "stand out" in terms of their emotional developments: their authors seem to write more creatively with respect to the expectations of the respective genres.

4.2 Emotion Arcs and Genre Classification

Above, we have established that arcs for some emotions are more uniform than others, and that there are outlier texts for every emotion and genre. But does the degree of uniformity matter for classification? To assess this question, we analyze the average prototypicality among books that were classi-

Genre	Most prototypical	Least prototypical
adventure	*Bert Wilson in the Rockies*, Duffield, J. W. *The Outdoor Girls of Deepdale; Or, camping and tramping for fun and health*, Hope, L. *Blown to Bits; or, The Lonely Man of Rakata*, Ballantyne, R. M.	*Chasing the Sun*, Ballantyne, R. M. *The Bronze Bell*, Vance, L.J. *Chester Rand; or, The New Path to Fortune*, Alger, H.
romance	*The Girl in the Mirror*, Jordan, Elizabeth Garver *The Unspeakable Perk*, Adams, Samuel Hopkins *The Maid of Maiden Lane*, Barr, Amelia	*La Dame aux Camilias*, Dumas, A. *Through stained glass*, Chamberlain, George *Daddy-Long-Legs*, Webster, Jean
mystery	*The Woman from Outside [On Swan River]*, Footner, H. *The Old Stone House and Other Stories*, Green, A.K. *In Friendship's Guise*, Graydon, W.M.	*The Grell Mystery*, Froest, F. *My Strangest Case*, Boothby, G. *The Treasure-Train*, Reeve, A. B.
scifi	*The Great Drought*, Meek, S. P. *The Finding of Haldgren*, Diffin, Charles *The Tree of Life*, Moore, C. L.	*Looking Backward, 2000 to 1887*, Bellamy, E. *Let 'Em Breathe Space!*, Del Rey, L. *The Second Deluge*, Serviss, Garrett P.
humor	*Captains All and Others*, Jacobs, W. *The Rubáiyát of a Bachelor*, Rowland, H. *The Temptation of Samuel Burge (Captains All, Book 8)*, Jacobs, W. W.	*Just William*, Crompton, Richmal *Baby Mine*, Mayo, Margaret *Torchy and Vee* , Ford, Sewell

Table 5: Most and least prototypical books regarding overall emotional development in each genre

Model Family	Classification	Avg. Spearman on +	Avg. Spearman on −	Δ between + and −
BoW	RF	0.185	0.164	0.021
	MLP	0.184	0.170	0.014
EmoLex	RF	0.182	0.176	0.006
	MLP	0.181	0.179	0.002
EmoArc	RF	0.193	0.162	0.031
	MLP	0.206	0.144	0.062
	CNN	0.205	0.145	0.060

Table 6: Average prototypicality (measured as correlation with prototypical emotion arc) for books that are correctly (+) and incorrectly (−) predicted by each model. Positive Δ means higher prototypicality for correct classifications.

fied correctly and incorrectly for each classification model from Section 2.3.

The results in Table 6 show that the average prototypicality is always higher for correctly than for incorrectly classified books. That being said, there appears to be a relationship between the feature set used and the size of this effect, Δ. This size is smallest for the BoW models and not much larger for the EmoLex models. It is considerably larger for the EmoArc models and particularly higher for the MLP EmoArc model.

We draw three conclusions from this analysis: (1), EmoArc features and models based on them are meaningful for the task of literary genre classification, as evidenced by higher correlation coefficients in the correctly predicted instances. (2), since emotion arcs are exactly the type of information that the CNN EmoArc model bases its classification decision on, emotional uniformity is indeed a prerequisite for successful classification by EmoArc, and its lack for some genres and emotions explains why EmoArc does not do as well as the more robust BoW and EmoLex models. (3), the difference in correlation ranks between correct and incorrect predictions validates the idea of an ensemble classification scheme and may serve as a starting point for deeper investigation of differences between models in future work.

4.3 Feature Analysis of Lexical Models

After having considered EmoArc in detail, we now complete our analysis in this paper by a more in-depth look at the feature level. We focus on features that are most strongly associated with the genres, using a standard association measure, pointwise mutual information (PMI), which is considered to be a sensible approximation of the most influential features within a model.

Table 7 shows that most strongly associated features with each genre differ in their linguistic status between BoW and EmoLex. For example, for the genre *romance*, most BoW features are infrequent words like specific character names which do not generalize to unseen data (*e.g.*, *Gerard, Molly*).

BoW					EmoLex				
Adv.	Humor	Mystery	Romance	SciFi	Adv.	Humor	Mystery	Romance	SciFi
tarzan	ses	coroner	gerard	planet	hermit	wot	murderer	sally	projectile
damon	iv	kennedy	molly	solar	hut	wan	jury	mamma	rocket
canoes	sponge	detective	willoughby	planets	fort	comrade	attorney	marry	beam
blacks	ay	inspector	fanny	projectile	lion	rat	robbery	tenderness	scientist
indians	says	detectives	clara	mars	tribe	bye	police	loving	blast
ned	wot	trent	maggie	rocket	spear	beer	crime	charity	bomb
savages	wan	scotland	eleanor	rip	jungle	idiot	criminal	love	emergency
spain	mole	murderer	cynthia	jason	swim	jest	murder	marriage	system
whale	ha	rick	yo	phone	rifle	school	suicide	passionate	center
eric	ma	scotty	jill	globe	don	mule	clue	holiday	pilot

Table 7: Top ten EmoLex and BoW features by pointwise mutual information values with each genre.

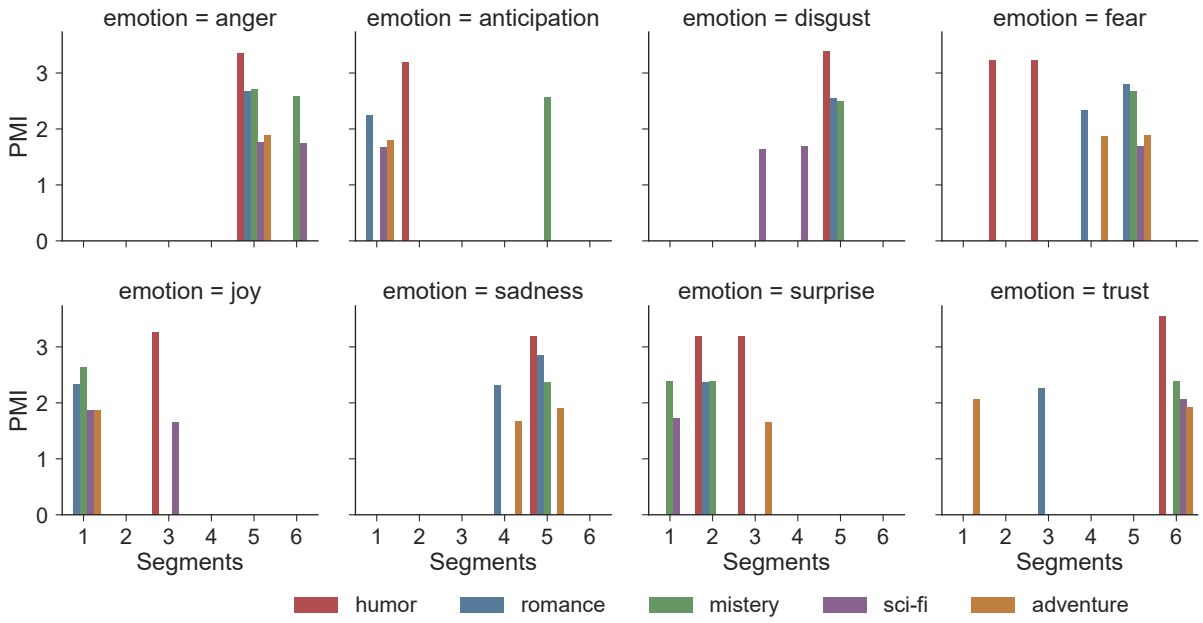

Figure 3: Top EmoArc features for each genre ranked according to their PMI values.

The EmoLex features consist of words related to emotions (*e.g.*, *mamma, marry, loving*). In *mystery*, the most important BoW features express typical protagonists of crime stories (*e.g.*, *coroner, detective, inspector, Scotland*). For EmoLex, we see similar results with a stronger focus on affect-related roles (*e.g.*, *murderer, jury, attorney, robbery, police, crime*). In sum, we observe that the feature sets pick up similar information, but from different perspectives: the BoW set focusing more on the objective ("what") and the EmoLex set more on the subjective ("how") level.

As a combination of the analysis in Section 4.2 with the PMI approach, Figure 3 visualizes the EmoArc features as "peak" features that fire when an emotion is maximal in one specific segment (*cf.* Section 3). The results correspond well to the prominent maxima of emotion arcs shown in Figure 2. For the genre of adventure, *e.g.*, *trust* and *anticipation* peak at the beginning. *Sadness, anger*, and *fear* peak towards the end, however, the very end sees a kind of "resolution" with trust becoming the dominating emotion again. At the same time, *anger* and *sadness* seem to be dominating all genres towards the end, and *joy* plays an important role in the first half of the books for most genres.

5 Discussion and Conclusion

In this paper, we analyzed the relationship between emotion information and genre categorization. We considered three feature sets corresponding to three levels of abstraction (lexical, lexical limited to emotion-bearing words, emotion arc) and found interesting results: classification based on emotion-words performs *on par* with traditional genre feature sets that are based on rich, open-vocabulary

lexical information. Our first conclusion is therefore that emotions carry information that is highly relevant for distinguishing genres.

A further aggregation of emotion information into emotion arcs currently underperforms compared to the lexical methods, indicating that relevant information gets lost in our current representation. We need to perform further research regarding this representation as well as the combination of different feature sets, since these appear to contribute complementary aspects to the analysis of genres, as the excellent performance of an oracle shows. Our ensemble approach significantly outperforms the best single model but still outperforms the oracle result.

Our subsequent, more qualitative analysis of the uniformity of emotion arcs within genres indicated that some, but not all, emotions develop moderately uniformly over the course of books within genres: *Fear* is most uniform in all genres except mystery stories, where *anger* is more stable. Unexpectedly, *joy* is only of mediocre stability. At the same time, our study of outliers indicates that this conforming to the prototypical emotion development of a given genre appears to be a *sufficient, but not necessary* condition for membership in a genre: we found books with idiosyncratic emotional arcs that were still unequivocally instances of the respective genres. As with many stylistic properties, expectations about emotional development can evidently be overridden by a literary vision.

This raises the question of what concept of genre it is that our models are capturing. Compared to more theoretically grounded concepts of genre in theoretical literary studies, our corpus-based grounding of genres is shaped by the books we sampled from Project Gutenberg. Many of these are arguably relatively unremarkable works that exploit the expectations of the genres rather than seminal works trying to redefine them. The influence of corpus choice on our analysis take may also explain the apparent contradictions between our by-emotion results and the ones reported by Reagan et al. (2016), who identified *happiness/joy* as the most important emotion, while this emotion came out as relatively uninteresting in our analysis. Our observations about the influence of individual artistic decisions have, however, made us generally somewhat hesitant regarding Reagan et al.'s claim about "universally applicable plot structures".

In future work, we want to pursue (a) the close reading direction and analyse a relatively small number of classical works for each genre with respect to their prototypicality in more detail, as well as (b) the distance reading direction, investigating the potential for a better combination of the different classification schemes into an ensemble model.

References

Sarah Danielle Allison, Ryan Heuser, Matthew Lee Jockers, Franco Moretti, and Michael Witmore. 2011. *Quantitative formalism: an experiment*. Stanford Literary Lab.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, BC, pages 579–586.

Yves Bestgen. 1994. Can emotional valence in stories be determined from words? *Cognition & Emotion* 8(1):21–36.

Leo Breiman. 2001. Random forests. *Machine learning* 45(1):5–32.

Michael Brooks, Katie Kuksenok, Megan K Torkildson, Daniel Perry, John J Robinson, Taylor J Scott, Ona Anicello, Ariana Zukowski, Paul Harris, and Cecilia R Aragon. 2013. Statistical affect detection in collaborative chat. In *Proceedings of the Conference on Computer-Supported Cooperative Work*. pages 317–328.

Ronan Collobert and Samy Bengio. 2004. Links between perceptrons, MLPs and SVMs. In *Proceedings of the Twenty-first International Conference on Machine Learning*. New York, NY, USA, ICML.

Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. 2012. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision* 7(2–3):81–227.

John Anthony Cuddon. 2012. *Dictionary of literary terms and literary theory*. John Wiley & Sons.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one* 6(12):e26752.

Bradley Efron. 1979. Bootstrap methods: another look at the jackknife. *The annals of Statistics* 7(1):1–26.

Sergey Feldman, Marius Marin, Julie Medero, and Mari Ostendorf. 2009. Classifying factored genres with part-of-speech histograms. In *Proceedings of*

Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Boulder, Colorado, pages 173–176.

W. Nelson Francis and Henry Kucera. 1979. Brown corpus manual. Online: http://clu.uni.no/icame/brown/bcm.html.

Alastair J. Gill, Robert M. French, Darren Gergle, and Jon Oberlander. 2008. Identifying emotional characteristics from short blog texts. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. pages 2237–2242.

Lena Hettinger, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2015. Genre classification on german novels. In *Proceedings of the 26th International Workshop on Database and Expert Systems Applications*. pages 249–253.

Lena Hettinger, Fotis Jannidis, Isabella Reger, and Andreas Hotho. 2016. Classification of literary subgenres. In *Proceedings of DHd 2016*. Leipzig, Germany.

Geoffrey E Hinton. 1989. Connectionist learning procedures. *Artificial intelligence* 40(1):185–234.

Patrick Colm Hogan. 2011. *What literature teaches us about emotion*. Cambridge University Press.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*. Kyoto, Japan, pages 1071–1075.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain, pages 32–38.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *International AAAI Conference on Web and Social Media*. Barcelona, Spain, pages 538–541.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. pages 105–114.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29(3):436–465.

Robert Plutchik. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist* 89(4):344–350.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. *arXiv preprint arXiv:1611.02988* .

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5(1):31.

Spyridon Samothrakis and Maria Fasli. 2015. Emotional sentence annotation helps predict fiction genre. *PloS one* 10(11):e0141922.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of babel: evaluating genre collections. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics* 26(4):471–495.

Ted Underwood. 2016. The life cycles of genres. *Cultural Analytics* 1. https://doi.org/doi:10.7910/DVN/XKQOQM.

Ted Underwood, Michael L Black, Loretta Auvil, and Boris Capitanu. 2013. Mapping mutable genres in structurally complex volumes. In *Proceedings of the IEEE International Conference on Big Data*. pages 95–103.