Significance testing for the classification of literary subgenres

Lena Hettinger, Fotis Jannidis, Isabella Reger, Andreas Hotho
University of Würzburg

1. Introduction

The automatic classification of literary genres, especially of novels, has become a research topic in the last years (Underwood 2014, Jockers 2013). In the following we report on the results from a series of experiments using features like most frequent words, character tetragrams and different amounts of topics (lda) for genre classification on a corpus of German novels. Two problems will be the main focus of this paper and they are both caused by the same factor: there are only few labeled novels available. So how can experiments be designed and evaluated reliably in a setting like this. We are especially interested in testing results for significance to get a better understanding of the reliability of our research. The scarcity of labeled data is also one of the reasons some researchers segment novels. We will show that without a test for significance it would be easy to misunderstand our results and we will also show that using segments of the same novel in the test and the training data leads to overestimation of the predictive capabilities of the approach.

2. Setting

In the following we will describe our corpus and feature sets. Our corpus consists of 628 German novels mainly from the 19th century obtained from sources like TextGrid Digital Library[1] or Projekt Gutenberg[2]. Novels have been manually labeled according to their subgenre after research in literary lexica and handbooks. The corpus contains 221 adventure novels, 57 social novels and 55 educational novels; the rest belongs to a different or more than one subgenre.

Features are extracted and normalized to a range of [0,1] based on the whole corpus consisting of 628 novels. We have tested several feature sets beforehand and found stylometric and topic based to be the most promising (c.f. Hettinger et al. 2015). To represent stylometric features we employ 3000 most frequent words (mfw3000) and top 1000 character tetragrams (4gram). Topic based features are created using Latent Dirichlet Allocation (LDA) by Blei et al. (2003). In literary texts topics sometimes represent themes, but more often they represent topoi, often used ways of telling a story or parts of it (see also Underwood 2012, Rhody 2012). For each novel we derive a topic distribution, i.e. we calculate how strongly each topic is associated with each novel. We try different topic numbers and build ten models for each setting to reduce the influence of randomness in LDA models. We remove a set of predefined stop words as well as Named Entities from the novels as we have shown before that the removal of Named Entities tends to improve results.

[1] textgrid.de/digitale-bibliothek
[2] gutenberg.spiegel.de

3. Evaluation

Classification is done by means of a linear Support Vector Machine (SVM) as we have already shown in Hettinger et al. (2015) that it works best in this setting (see also Yu 2008). In each experiment we apply stratified 10-fold cross validation to the 333 labeled novels and report overall accuracy and F1-Score (c.f. Jockers 2013). The majority vote (MV) baseline for our genre distribution yields an accuracy score of 0.66 and F1 score of 0.27 (see fig. 1).

	adventure	educational	social		precision
adventure	221	55	57	333	66%
educational	0	0	0	0	0%
social	0	0	0	0	0%
	221	55	57	333	
recall	100%	0%	0%		Acc: 66%
f1	80%	0%	0%		F1: 27%

Fig. 1: Cross table for majority vote baseline.

In the cross tables of Figure 1 and 2 each column represents the true class and each row the predicted genre. Correct assignments are shaded in grey, average accuracy in green and average F1 score in red.

	adventure	educational	social		precision
adventure	218	3	5	226	96%
educational	1	41	14	56	73%
social	2	11	38	51	75%
	221	55	57	333	
recall	99%	75%	67%		Acc: 89%
f1	98%	74%	70%		F1: 81%

Fig. 2: Cross table for mfw 3000 as an example for classification results.

Because there are not many labeled novels in the domain of genre classification we expanded our corpus by splitting every novel into ten equal segments. Features are then constructed independently for the resulting 3330 novel segments. To test the influence of the LDA topic parameter *t* in conjunction with having more LDA documents we evaluate topic features for *t* =100, 200, 300, 400, 500 (see figure 3 and 4).

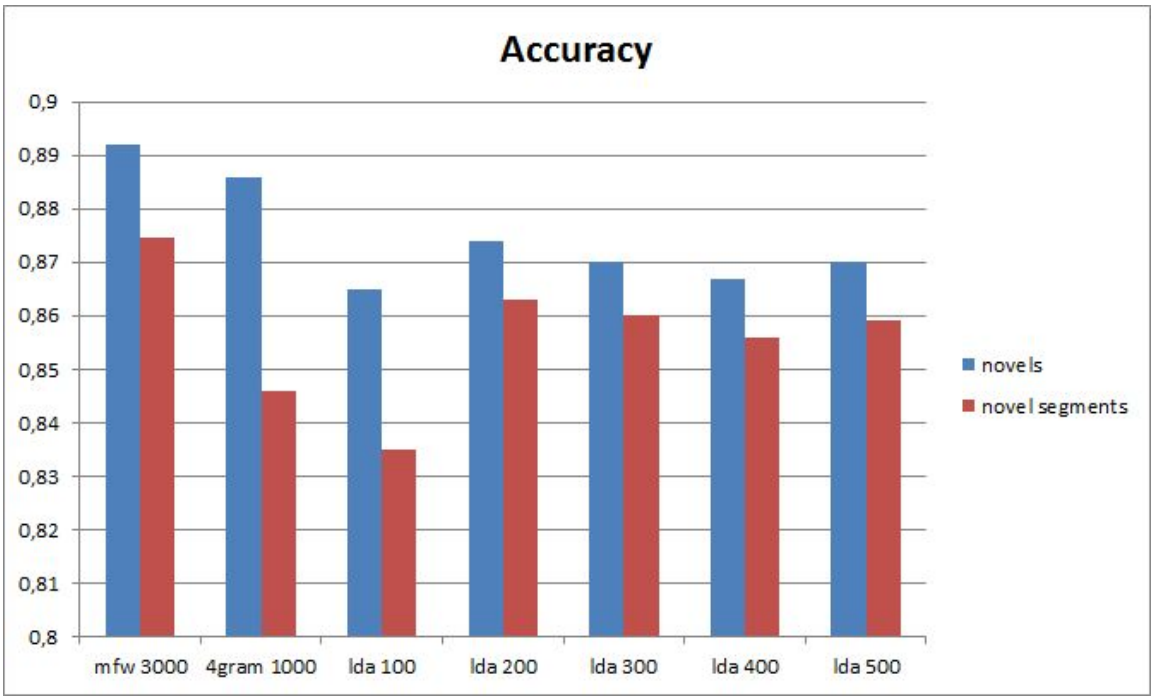

Fig. 3: Accuracy scores for novels and novel segments and different feature sets.

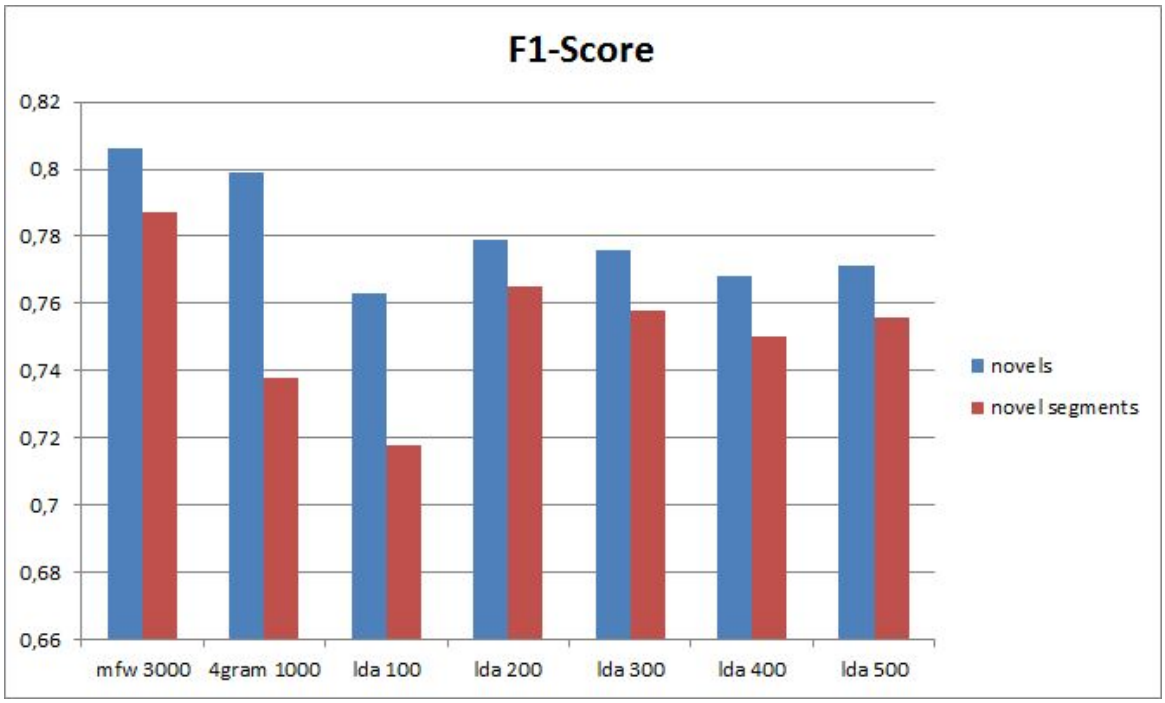

Fig.4: F1 scores for novels and novel segments and different feature sets.

Results show that our evaluation metrics tend to drop if novels are segmented. This could mean that genre is indeed a label for the whole literary work and not parts of it. On the other hand many differences are pretty small. Therefore we would like to test if these differences are statistically significant or if they should be attributed to chance.

4. Tests of statistical significance

When working with literary corpora there are few genre labels available for two reasons. First, the task of labeling the genre of a novel is strenuous; second, literary studies have mostly concentrated on a rather small sample, the canonical novels. Another issue is the creation of a balanced corpus, because for historical reasons the distribution of literary genres is not uniform and also the process of selecting novels for digitization has made the situation even more complicated. This generally results in data sets of less than 1000 items or even less than 100, see for example Jockers (2013) where 106 novels form a corpus or Hettinger et al. (2015) where we evaluate on only 32 novels.

The problem arising from small corpora is that small differences in results may originate from chance. This can be investigated by using statistical tests (c.f. Kenny 2013 and Nazar and Sánchez Pol 2006). A standard tool to detect if two data sets are significantly different is Student's t-test which we will use in the following to control the results of our experiments. We use two variations of Student's t-test with $\alpha = 0.05$:

- the one-sample t-test to compare the accuracy of a feature set against the baseline
- the two-sample t-test to compare accuracy results for two feature sets

In both cases the data set considered consists of ten accuracy results from ten-fold cross validation and accordingly 100 data points for LDA from its ten models. Due to the small sample size we drop the assumption of equal variance for the two-sample t-test. The results for the one-sample t-tests show that every single feature set yields significantly better accuracy than the baseline (66.4%). We can therefore conclude that feature sets classify novels not randomly and that they do incorporate helpful genre clues.

	4gram	4gram parts	lda100	lda100 parts	lda200	lda200 parts	lda300	lda300 parts	lda400	lda400 parts	lda500	lda500 parts	mfw 3000
4gram parts	0,0934												
lda100	0,2881	0,1998											
lda100 parts	0,0208	0,4426	0,0000										
lda200	0,5508	0,0661	0,1381	0,0000									
lda200 parts	0,2492	0,2494	0,7641	0,0000	0,0837								
lda300	0,4178	0,1194	0,4866	0,0000	0,5328	0,3450							
lda300 parts	0,1994	0,3320	0,4665	0,0002	0,0316	0,6780	0,1852						
lda400	0,3590	0,1531	0,6949	0,0000	0,3506	0,5152	0,7803	0,3010					
lda400 parts	0,1393	0,4887	0,1553	0,0013	0,0040	0,2824	0,0512	0,5182	0,0976				
lda500	0,4269	0,1125	0,4393	0,0000	0,5553	0,3045	0,9607	0,1560	0,7368	0,0387			
lda500 parts	0,1795	0,3607	0,3231	0,0001	0,0106	0,5277	0,1114	0,8612	0,2005	0,5935	0,0877		
mfw3000	0,7714	0,0190	0,0577	0,0009	0,1929	0,0450	0,1182	0,0305	0,0891	0,0166	0,1217	0,0251	
mfw3000 parts	0,5836	0,0713	0,2289	0,0001	0,9397	0,1607	0,5655	0,0875	0,4113	0,0293	0,5870	0,0572	0,2332

Fig.5: Two sided t-test for $\alpha = 0.05$ on accuracy of genre classification on 333 German novels.

P-values for the two-sided t-tests are reported in Figure 5. As we have used $\alpha = 0.05$ as significance threshold every p-value smaller than 0.05 is statistically significant (shaded in grey). From Figure 5 it follows that differences between segmented and not-segmented novels are **not** statistically significant in most cases except for LDA, *t*=100. Besides results do not differ significantly for different topic numbers *t*=100,200,300, 400,500 apart from lda100 parts, which performs significantly worse than any other LDA feature.

An important assumption of the two-sample t-test is that both samples have to be independent. This is the case here as each time we do a cross validation we split the data independently from any other cross validation run. Thus, even if we repeat our experiments for a number of iterations (see e.g. Hettinger et al. 2015) we still get independent evaluation scenarios. Therefore we can apply the two-sided t-test in our setting to support our claims. In case of dependency of samples we could instead use paired t-tests on accuracy per novel.

5. Novel segmentation

A crucial factor when segmenting novels is how to distribute the segments between test and training data set. We decided that in our case we have to put all of the ten segments a novel was divided into either in the test or in the training data set as we want to derive the genre of a novel not seen before. Another possibility which Jockers (2013) exploited is to distribute segments randomly between training and test set. In his work “Macroanalysis” Jockers investigates how function words can be used to research aspects of literary history like author, genre etc. In the following we want to replicate the part concerning genre prediction using German novels.

When segments of one novel appear in both test and training data we achieve an accuracy of 97.5% and F1 score of 95.9% - that is close to perfect (see fig. 6). Such a partitioning of the novels dramatically overestimates predictive performance on unseen texts. In comparison, Jockers (2013) achieves an average F1 score of 67% on twelve genre classes. His results are worse because we are only using three different genres while he is doing a multiclass classification with 12 classes. But nevertheless 67% probably still overestimates the real predictive power of this approach, because in our setup using the segments in both, test and training data, increased F1 by than 17%.

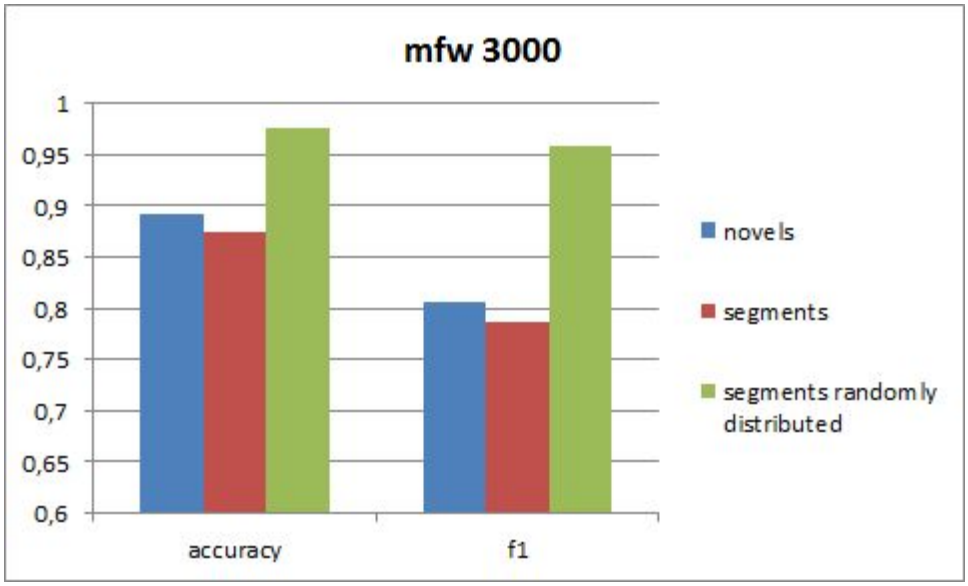

Fig.6: Results for different partitioning strategies.

6. Conclusion

In this work we looked at the methodology and evaluation of genre classification of German novels and discussed some of the methodical pitfalls of working with data like this. We discovered that only some of our results turned out to be statistically significant whereas for example the statement, that stylometric perform better than topic-based features, could not be fortified. Therefore our opinion is that research findings on small data sets should be scrutinized especially carefully for example by using statistical tests.

References

Blei D., Ng A. & Jordan M. (2003). “Latent dirichlet allocation”. In: *The Journal of machine Learning research* 3, p. 993-1022.

Hettinger L., Becker M., Reger I., Jannidis F. & Hotho A. (2015). “Genre classification on German novels”, In: *Proceedings of the 12th International Workshop on Text-based Information Retrieval*.

Jockers M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Kenny A. (1982). *The computation of style: An introduction to statistics for students of literature and humanities*. Elsevier.

Nazar R. & Sánchez Pol M. (2006). "An extremely simple authorship attribution system." In: *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics/Language and the Law*.

Rhody L. M. (2012). “Topic Modeling and Figurative Language”. *Journal of Digital Humanities*, 2(1), http://journalofdigitalhumanities.org/2-1/.

Underwood T. (2012). “Topic modeling made just simple enough”. Blog post 7.4.2012 http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Underwood T. (2014). “Understanding Genre in a Collection of a Million Volumes”, Interim Report. http://dx.doi.org/10.6084/m9.figshare.1281251 (26.8.2015)

Yu B. (2008). “An Evaluation of Text Classification Methods for Literary Study”. In: *Literary and Linguistic Computing* 23 (2008): 327-343.