Classification of Literary Subgenres

,
lena.hettinger@uni-wuerzburg.de
Universität Würzburg, Deutschland

,
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

,
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

,
hotho@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Introduction

Literary scholars and common readers use labels like educational novel, crime novel or adventure novel to organize the large domain of fiction. In both discourses the use of these categories is well-established even though they are evolving and tend to be inconsistent. The classification of genres is one of the standard tasks in document classification and has been researched intensively (cf. Biber 1989; Santini 2004; Freund et al 2006; Sharoff et al. 2010). Some results seem impressive, for example distinguishing clear-cut genres like poetry from fiction (Underwood 2014), but most texts on literary genre classification emphasize, as the literature on genre classification in general, the variability of genre signals (Allison et al. 2011: 19; Underwood et al. 2013; Underwood 2014). The scores for genre classification over all categories are therefore often not very high. Jockers for example reports an accuracy of 67% (Jockers 2013: 81). Genre classification in general works best with most frequent words, all words or character tetragrams (Freund et al. 2006; Sharoff et al. 2010) and most of the reported experiments for literary genre classification also use all words or only the *n* most frequent word (sometimes including punctuation) as features. In a series of experiments we examine whether it is possible to enhance these results for the classification of subgenres of novels. Our research is motivated by an understanding of novel genres as concepts which are differentiated by style, settings, character constellations and plots. We use most frequent words as an indicator for style and network characteristics as an indicator for character constellations. Setting is partially covered by topic models which also represent information on typical ways of telling a story, narrative *topoi*. We have to omit plot, as we don't have a reliable way to represent plot by any indicators yet.

Setting

In the following we will describe the corpus and the features we use for the task of subgenre classification.

Corpus

Our corpus consists of 628 German novels mainly from the 19th century (roughly 1745 to 1935) obtained from sources like the or the German Projekt Gutenberg . The novels have been manually labeled according to their subgenre after research in literary lexica and handbooks. The corpus contains 221 adventure novels, 88 social novels and 86 educational novels; the rest are novels from different subgenres.

Features

As mentioned in section 1 we use three types of features (stylometric, topic based and network) that are described in more detail in Hettinger et al. (2015). Features are extracted and normalized to a range of [0,1] based on the whole corpus consisting of 628 novels.

Stylometric features

We use word frequencies as well as character tetragrams to represent stylometric features. We tested different amounts of most frequent words and decided to work with the top 3000 (mfw3000). Additionally we use the top 1000 character tetragrams (4gram).

Topic features

We use Latent Dirichlet Allocation (LDA) by Blei et al. (2003) to extract topics from our data. In literary texts topics sometimes represent themes, but more often they represent topoi, often used ways of telling a story or parts of it. For each novel we derive a topic distribution, i.e. we calculate how strongly each topic is associated with each novel. We try different preprocessing approaches and topic numbers and build ten models for each setting to reduce the influence of randomness in LDA models. In every setting we first remove a set of predefined stop words from the novels and then use LDA on our corpus of 628 novels. The different forms of preprocessing we use are:

- no additional preprocessing
- removal of Named Entities (Jannidis et al. 2015)
- word stemming
- word lemmatization

- word lemmatization + removal of Named Entities

Network features

We use the character recognition system described in Jannidis et al. (2015) to identify the characters of each novel. Although the NER tool may be employed with co-reference resolution we do not make use of this option here. We extract proper names to build a network where each node is a character and the number of co-occurrences of two characters in the same paragraph is the weight of the edge between these two. The network of each novel is reduced to the most central characters and the most frequent interactions in order to bring out their basic shape. The network feature set consists of the total number of characters in a novel and six network measures: maximum degree centrality, global efficiency, transitivity, average clustering coefficient, central point dominance and density.

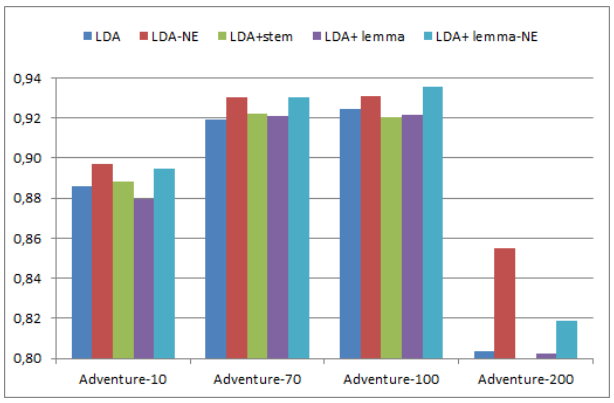
Evaluation

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 100 iterations of 10-fold cross validations to account for the small data sets. The depicted results are the average over 1000 classification accuracy values. We want to investigate the following subgenre constellations:

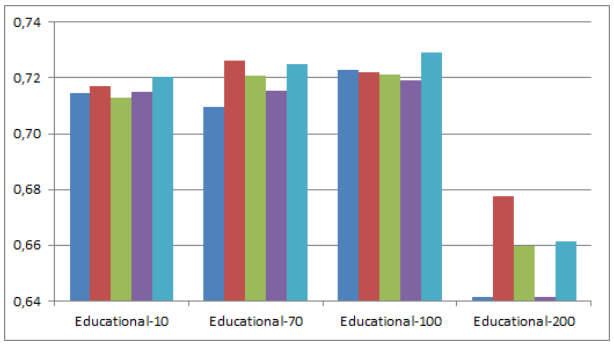
- Adventure versus non-Adventure novels
- Educational versus non-Educational novels
- Social versus non-Social novels
- Adventure versus Educational novels and
- Educational versus Social novels

Depending on the setting the label distribution is often imbalanced. To make results comparable we use undersampling where in each of the 100 iterations a new sample is drawn from the larger class while all instances of the smaller class are used. This accounts for a majority vote (MV) baseline that always yields an accuracy score of 0.50 as both classes have equal size.

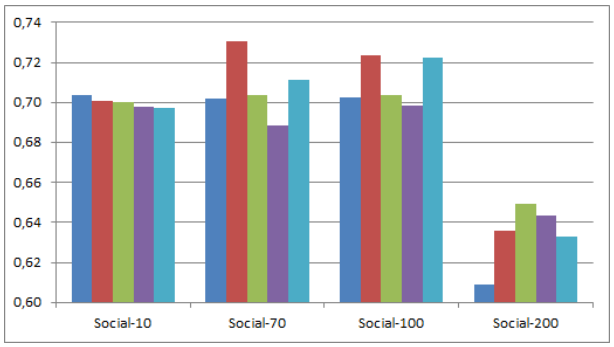
To determine the influence of the LDA topic parameter *t* and different preprocessing procedures we report accuracy for *t* =10, 70, 100 and 200 (see figure 1). Differences between different preprocessing categories are minimal, but removal of named entities seems to improve results overall. We observe comparable result for *t* =10, 70 and 100 and a drop in performance for *t* = 200. Therefore, we will use LDA with lemmatized words and named entities removed for *t* = 100 in the following experiments.



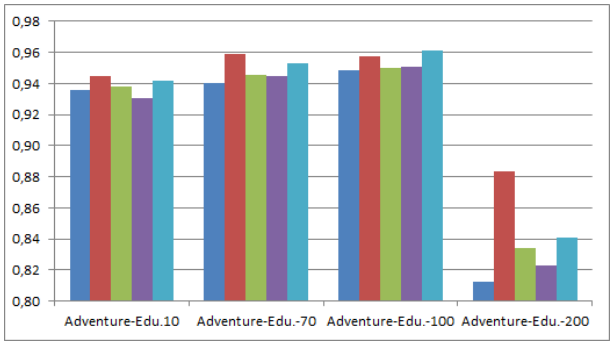
a) adventure/non-adventure



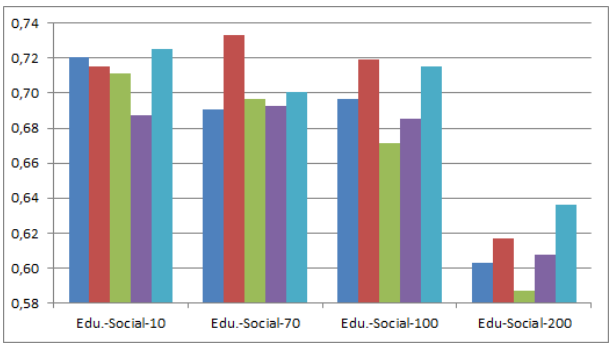
b) educational/non-educational



c) social/non-social



d) adventure/educational



e) educational/social

***Fig. 1**: Classification results for different subgenre settings in terms of accuracy using LDA with topic size t = 10, 70, 100, 200 and five different preprocessing procedures on 628 German novels*

When comparing different feature sets across our subgenre constellations we can see that semantic based features (mfw, 4grams, lda) all perform quite good while network features perform rather poorly (see figure 2). With an accuracy score of more than 90% adventure novels seem to be fairly easy to differentiate from other genres. In contrast, the other genres don't show such a distinct signal using surface features.

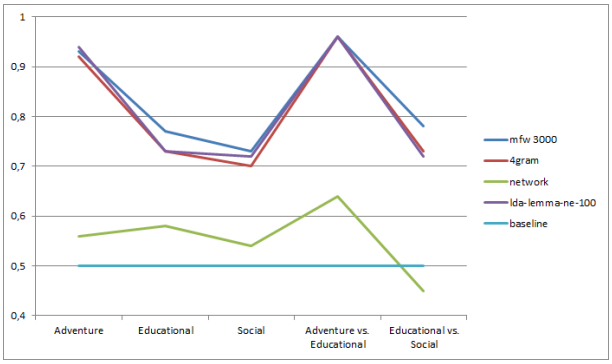


***Fig. 2**: Accuracy for different scenarios and feature sets including the majority vote baseline.*

As the classification performance of adventure/ educational is quite impressive we take a closer look at the discriminating words of these genres (Figure 3). Some of the most typical words of adventure novels include *captain, shore, (on) board, help, danger, immediately*. On the other hand words like *soul, love, world, heart, (at) home, mother, joy, often, father, emotion, human, to love* are characteristic for educational novels.

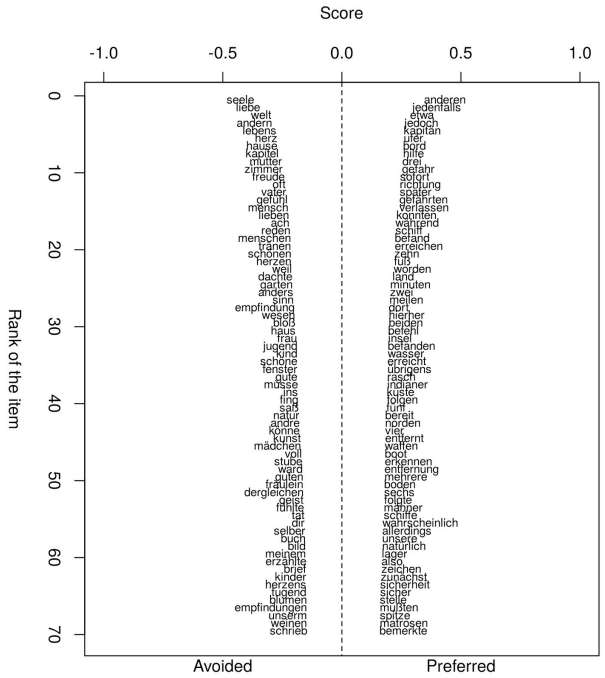


***Fig. 3**: Discriminating words for adventure (preferred) and educational (avoided) novels (Craig's Zeta)*

To test whether authorship of novels influences our results we removed the author signal by allowing only one document per author. In this way, we construct a new dataset called 'uni'. The sampling is done once so that the same novels are used in each setting. As shown in figure 4, we observe a much lower quality after removing the authorship information indicating an overemphasized focus of features and models on the hidden authorship signal. This varies for different settings as adventure/educational shows a loss of 0.09 (blue lines) and educational/social loses 0.23 (red lines). The relatively small loss in the first setting is remarkable as it contains 8 novels per author on average. One would expect the opposite given the weaker author signal of two novels per author for the other categories.

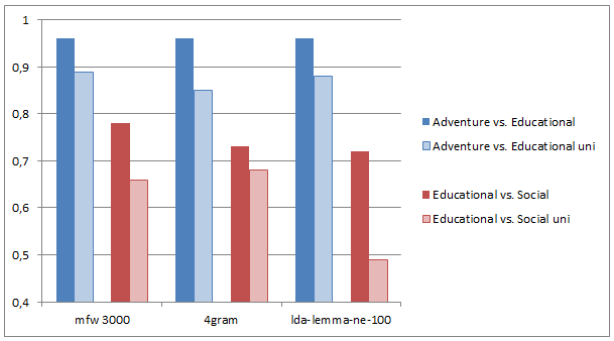


***Fig. 4**: The influence of authorship*

In the next experiment we test whether the combination of feature sets changes our classification results. To balance

the size of the feature sets we use Principal Component Analysis (PCA) and construct 100 features from the 3000 mfw and 1000 4gram features each. As shown in figure 5 some feature sets improve when combined (e.g. 4gram100 and lda100) and for others (e.g. lda100 and network) performance decreases. But classification results vary greatly in this setting as signaled by standard deviation bars so these differences should not be overrated.

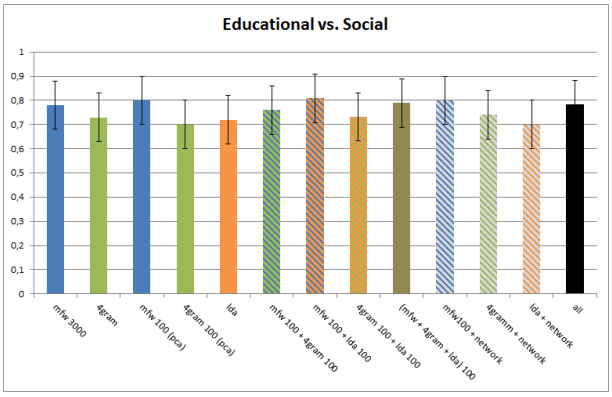


***Fig. 5**: Classification results for combinations of different feature sets including bars for standard deviation*

Conclusion

In this work we classified subgenres of German novels using different feature sets (mfw 3000, 4gram, lda etc.). Some subgenres, like adventure novels, are much easier to classify than others. Most of the applied feature sets showed a varying but comparable performance except the network features. The weak performance of network features might be caused by the weak link between the novel genre and character constellation. The variability of the subgenre signal could not be countered by using higher level features like topics and network characteristics. Interestingly, the author signal has a strong influence on the classification quality. The strength of influence seems to depend on category but is visible in all experiments. In the future, we would like to extend our work by using different network features, work on advanced topic models and find a reliable indicator for plot. Another challenge we have not faced yet is the development of subgenres over time.

Bibliographie

Allison, Sarah / Heuser, Ryan / Jockers, Matthew / Moretti, Franco / Witmore, Michael (2011): *Quantitative Formalism*. An Experiment (= Stanford Literary Lab Pamphlet 1) http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf [letzter Zugriff 09. Februar 2016].

Biber, Douglas (1989): "A typology of English texts", in: *Linguistics* 27: 3-43.

Blei, David / Ng, Andrew / Jordan, Michael (2003): "Latent Dirichlet allocation", in: *The Journal of Machine Learning Research* 3: 993-1022.

Finn, Aidan / Kushmerick, Nicholas (2006): "Learning to classify documents according to genre", in: *Journal of the American Society for Information Science and Technology (JASIST)*. Special Issue on Computational Analysis of Style 57, 11: 1506-1518.

Freund, Luanne / Clarke, Charles L. A. / Toms, Elaine G. (2006): "Towards genre classification for IR in the workplace", in: *Proceedings of the 1st international conference on Information interaction in context* (IIiX). New York, NY: ACM 30-36. http://dx.doi.org/10.1145/1164820.1164829 [letzter Zugriff 09. Februar 2016].

Hettinger, Lena / Becker, Martin / Reger, Isabella / Jannidis, Fotis / Hotho, Andreas (2015): "Genre classification on German novels", in: *Proceedings of the 12th International Workshop on Text-based Information Retrieval*.

Jannidis, Fotis / Krug, Markus / Reger, Isabella / Toepfer, Martin / Weimer, Lukas / Puppe, Frank (2015): "Automatische Erkennung von Figuren in deutschsprachigen Romanen", in: *DHd-Tagung 2015.*. Von Daten zu Erkenntnissen, 23. bis 27. Februar 2015, Graz.

Jockers, Matthew L. (2013): *Macroanalysis*. Digital Methods and Literary History. Champaign: University of Illinois Press.

Krug, Markus (2015): *NERDetection* https://github.com/MarkusKrug/NERDetection [letzter Zugriff 09. Februar 2016].

Petrenz, Philipp / Webber, Bonnie (2011): "Stable classification of text genres", in: *Computational Linguistics* 37, 2: 385-393.

Porter, Martin / Boulton, Richard (2001-2014): *Snowball* http://snowball.tartarus.org/ [letzter Zugriff 09. Februar 2016].

Projekt Gutenberg-DE (1994-): *Projekt Gutenberg-DE* http://gutenberg.spiegel.de/ [letzter Zugriff 09. Februar 2016].

Santini, Marina (2004): "State-of-the-art on Automatic Genre Identification", in: *Technical Report ITRI-04-03, ITRI, University of Brighton (UK)*.

Schmid, Helmut(1994-): *TreeTagger*. A language independent part-of-speech tagger http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [letzter Zugriff 09. Februar 2016].

Sharoff, Serge / Wu, Zhili / Markert, Katja (2010): "The Web Library of Babel: evaluating genre collections", in: *Proceedings of the conference on Language Resources and Evaluation (LREC), Malta* 3063-3070.

TextGrid (2006-2015): *Die Digitale Bibliothek bei TextGrid* [letzter Zugriff 09.Februar 2016].

Underwood, Ted (2014): *Understanding Genre in a Collection of a Million Volumes*. White Paper Report 29.12.2014 http://files.figshare.com/1857045/UnderstandingGenreInterimReport.pdf [letzter Zugriff 09. Februar 2016].

Underwood, Ted / Black, Michael L. / Auvil, Loretta / Capitanu, Boris (2013): "Mapping Mutable Genres in Structurally Complex Volumes", in: *2013 IEEE International Conference on Big Data* http://arxiv.org/abs/1309.3323v2 [letzter Zugriff 09. Februar 2016].

Yu, Bei (2008): "An Evaluation of Text Classification Methods for Literary Study", in: *Literary and Linguistic Computing* 23: 327-343 http://dx.doi.org/10.1093/llc/fqn015 [letzter Zugriff 09. Februar 2016].