Confounding variables in Sub-Genre classification: instructive problems

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Universität Würzburg

Leinen, Peter
P.Leinen@dnb.de
Deutsche Nationalbibliothek

This paper started out as a report on the state of the art in text classification, but over time it became much more a reflection on the pitfalls in modeling genre using classification. The start of our research was motivated by developments in text classification: Recent years have seen new approaches like gradient boosting and deep neural networks. Our initial goal was to inform about these approaches, which are seldom used yet in the digital humanities. But this proved to be only a starting point for a deeper exploration of genre structures of our collection of dime novels (‘Heftromane’, ‘Groschenromane’).

Most research on genre classification has been looking into what you could call ‘high level classes’ like newspaper genres (news, editorials etc.; e.g. Frank and Bouckaert, 2006) or web genres (blog, personal website etc.; e.g. Eissen and Stein, 2004). Under this perspective all texts we are looking at belong to one genre: the novel. The subgenres are types of love stories like the doctor novel (‚Arztroman‘) or the country novel (‚Heimatroman‘) and types of adventure novels, mainly distinguished by the setting: the war novel (‚Kriegsroman‘) or the science fiction novel. These novels are cheap (‘dime novels’) and published in a booklet format and are usually distributed via magazine kiosks and not book shops (Stockinger 2018). From the very beginning it was clear to us, that they don’t contain a random collection of each genre. On the contrary, the crime novels for example are just a small and very specific subsection of crime novels in general. But nevertheless we assumed that genre is the main aspect to group novels - for publishers and readers.

Our dataset consists of 11,600 dime novels from 12 different genres (see Fig.1). The genre label come from the four publishers who divide the market among themselves. (Bastei, Martin Kelter, Pabel Moewig and Cora). The corpus has been documented in previous studies such as Jannidis et. al. (2019a) and Jannidis et. al (2019b).

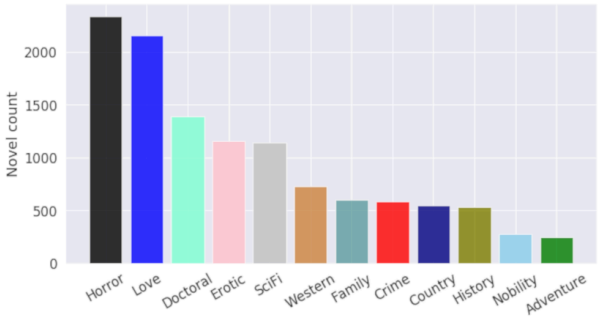

Figure 1: Novels per Genre

We have employed three groups of methods: traditional feature-based classifiers (Group A), modern feature-based classifiers (Group B) and deep learning (Group C). While Group A and B are based on document-term-matrix (20,000 most-frequent-words, tf-idf-weighted, stopwords removed, dimensionality reduction with LSI to 1000 features) as input, Group C works with unprocessed text. Named entities are removed completely. Hyperparameter optimization was done by sampling from the space of values recommended by the documentation of the libraries and (Olson et al. 2017) using Optuna (Akiba et al. 2019): In table 1 we report the best performance. We evaluated the performance of the deep learning approaches in advance on a smaller dataset, so that later only the best architecture had to be extensively tested (table 1). To increase speed initialized with pre-trained (wikipedia.de+30.000 novels) fasttext embeddings (Bojanowski et al. 2016). As a compromise between performance and speed we used the BiRNN architecture for all following experiments.

Table 1: Prestudy of deep learning architectures (4 subgenres, 800 novels)

	Fasttext	Flair	CNN	CNN +BiRNN	BiRNN	HATTN
f1-score	.886	.931	.925	.935	.923	.926
Time per epoch (seconds)	<5	288	210	190	90	215
Time to converge (minutes)	3	48	28	25	6	21

Table 2: Results of subgenre classification[1]

	Multi. NaiveBayes[2]	Logistic Regression	SVM (svc)	K-Nearest Neighbors
f1-score	.932	0.940	0.948	0.915
	XGBoost	LightGBM	CatBoost	BiGRU
f1-score	*	.878	*	.907

As was to be expected from the experience of previous studies on genre classification, the results were initially very good (Jannidis et. al. 2019a). They decreased slightly (~ 2 %) when we added novels from the publisher "CORA". With this addition our collection contains almost all dime novels published in recent years. Table 2 shows the classification results for this collection.

The decrease of our F1-score alone wasn't a great surprise, as the addition of new data is expected to increase diversity within groups and complicates classification. But two observations were irritating: First, we noticed classification results were improved when we included stopwords. Usually removing stopwords improves classification performance (Toman et al. 2006; Gonzales and Quaresma 2014). As most stopwords are typical function words which are used in stylometric research, this indicated that authorship information was used in the classification. Secondly, we noticed strong fluctuations between cross-validation folds, which seemed to indicate a very uneven class distribution.

To understand the first phenomenon better, we plotted the distribution of the authors across the genres (see Fig. 2): Many authors write exclusively within a genre. The greatest overlap can be found in the genres *Love* and *Family*.

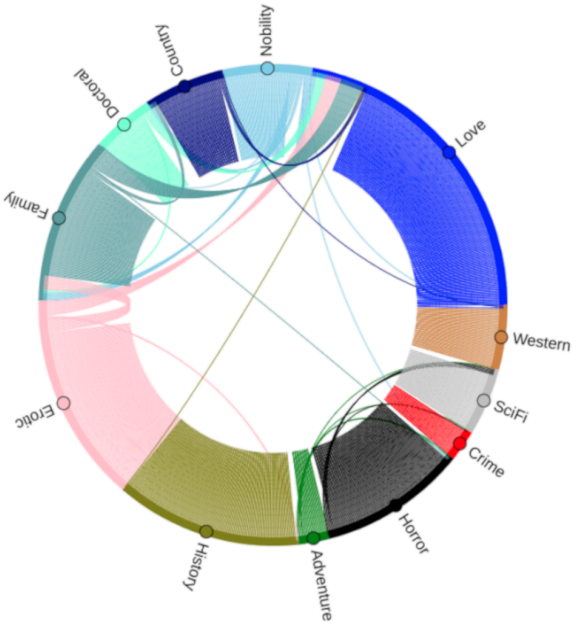

Figure 2: Inter-genre authorship

So, indeed, the authorship information could be used to identify the genre of text, but not in all genres equally.

In order to gain an insight into the influence of genre and publisher on the text form, we use Ivis (Szubert 2019) for unsupervised dimensionality reduction. The coloring of the data points according to publisher (figure 3) and genre (figure 4) shows the strong influence of these variables on the texts. It is also clear that Cora Verlag allows less variance among genres and thus becomes the most discriminatory factor. Figure 5 shows a detail of the previous plot, but focuses on microstructures. Theses structures indicate, that on this level genre and publisher are not enough to explain the distribution and that something else – author or series – comes into play.

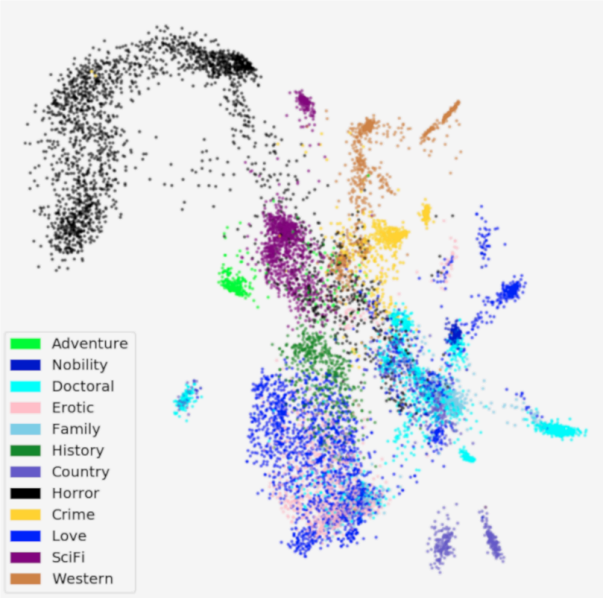

Figure 3: Ivis dimension reduction based on 20.000 mfw. Colors indicate genre.

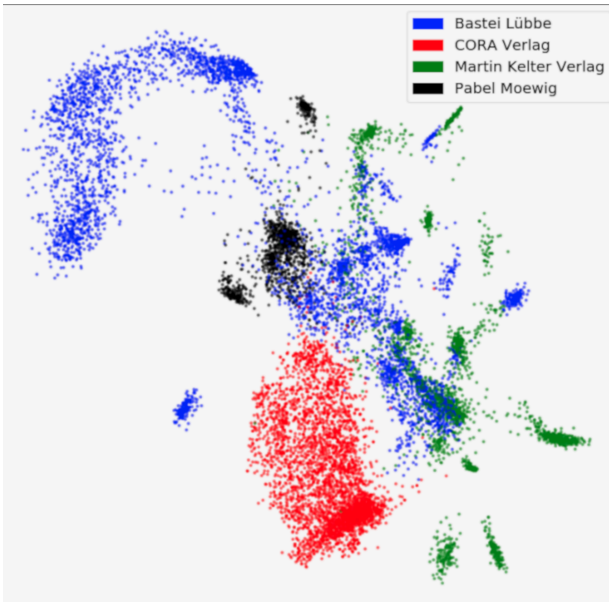

Figure 4: Ivis dimension reduction based on 20.000 mfw. Colors indicate publishers.

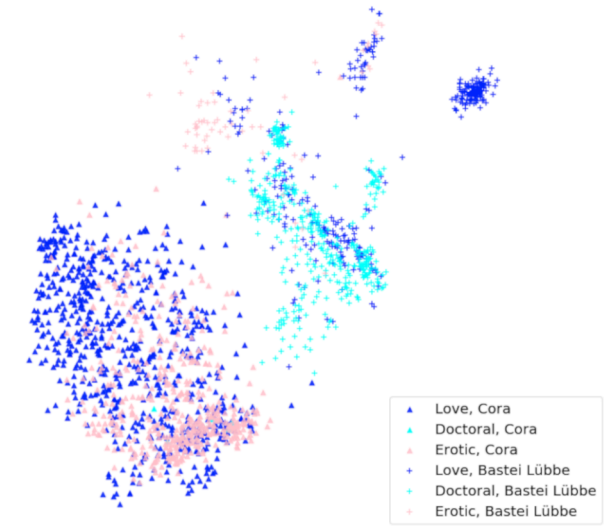

Figure 5: Detail of fig. 4, showing genres from publishing houses Bastei and Cora.

Obviously the variables publisher, author and series are influencing the distributions of our features, the words of the texts, and the variable, we want to predict, the genre. In a classical scientific model publisher, author and series would be called *confounding variables*, but in text classification the role of confounding has been mostly overlooked, probably because usually the main goal is prediction and not causal inference (Landeiro / Culotta 2016). Confounding variables are those factors in statistical models, that lead to false correlations or bias. For example, in an experiment that investigates the relationship between age of a person and the tendency to drive fast, the car would be a confounding variable. Because older people have probably a higher income and own faster cars. Something very similar is happening here. In the next section, we will apply a standard measure to control for confounding variables (restriction), while keeping the machine learning setup.

We created a restricted setup with a clear separation of authors, series and publishers between training and test data (i.e. authors which were in the training data, were not included in the test data etc.), and tested the subgenres in an one-vs-rest scheme. Figure 6 shows the results of this setup with at least 30 different combinations of test and training data per genre and a sample size of 200 novels split in half for training and test data.

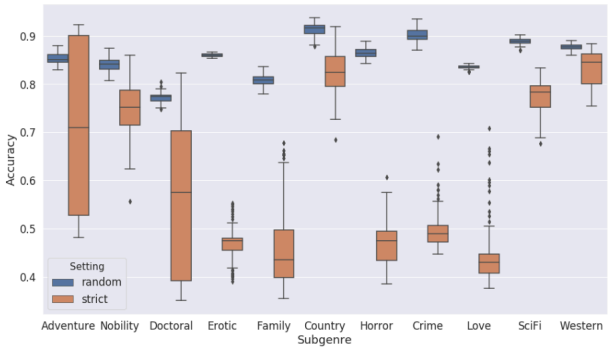

Figure 6: Binary Classification of Genres (Logistic Regression). Strict: No shared authors, series or publishers in training and testing dataset. Random: random sample to compare performances. Historic novels are excluded due to insufficient data.

The performance of the ‘strict setups’ is lower, sometimes even below 50%. This behavior is the result of negative examples in the training data being more similar to the positive examples of the test data, for example in love and doctoral novels of Cora.

Though now we control for confounding variables, it is less clear, what it implies for the genre model. It is not unusual in genre theory to conceptualize genre in an ideal way as independent of other factors like authorship, time, publisher etc. which corresponds to the ‘strict’ version of splitting train and test data. But at the same time, these factors may be so intertwined with the genre features, that it is difficult, if not impossible to separate them at all (Hempfer 2010). Under this perspective our attempt to construct a ‘clean’ and strict model of genre, independent of publishers etc. is a misguided attempt.

Looking back we now see that we started our research with some assumptions which seem to be unfounded for this part of the literary market which is dominated by four publishers: We assumed that the genre labels have the same function as in the rest of the literary market. But the small number of publishers seems to create a different situation. We assume now, that at least in some instances combinations of genre names with publisher

names (loves stories from Cora vs. love stories from Bastei-Lübbe) describe the clusters best. To start to evaluate this hypothesis, we trained the corpus on label combinations: 1) Genre and Publisher, e.g. 'Cora-Love', 2) Genre and Series. Figure 6 shows, that in many, but not all cases these combinations achieve very good results, which indicates that a clear-cut set of features corresponds these combinations. In some genres the same is true for series, for example doctoral or horror, while in others the series have no clear feature set (erotic, love).

Figure 7: Classification of series and publisher within a genre (one-vs-rest scheme). Points of single observations are colored by the series publishers.

To explore this in more detail, we looked at those genres where the values for a randomized and a strict setup in figure 4 are markedly different, which we see as a sign of a heterogeneity of the genre which was masked in the random setup. In this experiment we classified each of these six genres, using different setups for the separation of training and test set in order to control for the confounding variables. For the love novels this shows for example, that separating cleanly between the authors didn't reduce the performance, while doing the same with the series results in a drastic drop (figure 7), showing again, that in this genre, the genre cohesion is quite low, while the publishers and even the series have distinctive features.

Following up the indications for confounding variables we uncovered the complicated situation of genre in this subfield of the literary market. We succeeded to explore some of its substructures which haven't been described yet in literary studies, though it has been always one of its topics that this kind of literature is a commodity (Nusser 1973, Nusser 1991, Nutz 1999, Stockinger 2018). It is quite astonishing that almost every genre behaves differently, but this may be the result of a decades-old competition between this small number of publishers. Probably the different structures correspond to different strategies of each publisher. Bastei-Lübbe for example seems to follow a strategy where each series has a distinct profile, while Cora is focussing more on the publisher name as brand (Fig. 7 and Fig. 4) - though the clustering may also be influenced by the fact that Cora translates many novels from English. It would be an interesting follow-up-project, to find out, whether the readers of these genres know about these structures and how this knowledge directs their choices. Last but not least, we think that the strategies to control for known and unknown confounding variables in text classification, especially if it is done to understand existing structures and not so much to predict really new data, needs to be explored in more detail.

Acknowledgements

We like to thank Reviewer 2 for providing detailed and very informative feedback especially on the relation between data leakage and confounding variables as well as on the evaluation of dimension reduction techniques.

Fußnoten

1. Our code can be found: https://github.com/LeKonArD/info_leakage
2. For Multinomial Naive Bayes, Logistic Regression, SVM and K-NN we used the library Scikit-Learn (Pedregosa 2011). For the new gradient boosting approaches we used XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017), CatBoost (Dorogush et al. 2017).

Bibliographie

Akiba, Takuya / Shotaro Sano / Toshihiko Yanase / Takeru Ohta / Masanori Koyama (2019): „Optuna: A Next-generation Hyperparameter Optimization Framework". *CoRR* abs/1907.10902. http://arxiv.org/abs/1907.10902.

Bojanowski, Piotr / Edouard Grave / Armand Joulin / Tomas Mikolov (2016): „Enriching Word Vectors with Subword Information". *CoRR* abs/1607.04606. http://arxiv.org/abs/1607.04606.

Chen, Tianqi / Carlos Guestrin (2016): „XGBoost: A Scalable Tree Boosting System". *CoRR* abs/1603.02754. http://arxiv.org/abs/1603.02754.

Dorogush, Anna Veronika / Andrey Gulin / Gleb Gusev / Nikita Kazeev / Liudmila Ostroumova Prokhorenkova / Aleksandr Vorobev (2017): „Fighting biases with dynamic boosting". *CoRR* abs/1706.09516. http://arxiv.org/abs/1706.09516.

Eissen, Sven Meyer zu / Stein, Benno (2008): „Retrieval models for genre classification". *Scandinavian Journal of Information Systems.*

Frank, Eibe / Remco R. Bouckaert (2006): „Naive Bayes for Text Classification with Unbalanced Classes". In *Knowledge Discovery in Databases: PKDD 2006* , herausgegeben von Johannes Fürnkranz, Tobias Scheffer, und Myra Spiliopoulou, 503–510. Berlin, Heidelberg: Springer Berlin Heidelberg.

Goncales, T. / Quaresma, P. (2014): Evaluating preprocessing techniques in a Text Classification problem. In Information Processing & Management 50. Jg., Nr. 1, S. 104-112.

Hempfer, Klaus W. (2010): „Zum begrifflichen Status der Gattungsbegriffe: Von 'Klassen' zu 'Familienähnlichkeiten' und 'Prototypen'." *Zeitschrift Für Französische Sprache Und Literatur* 120, 1: 14-32. http://www.jstor.org/stable/40619075 .

Jannidis, Fotis, / Konle, Leonard / Leinen, Peter (2019a): Thematic Complexity. DH 2019 in Utrecht. Conference Abstracts.

Jannidis, Fotis / Konle, Leonard / Leinen, Peter (2019b): Makroanalytische Untersuchung von Heftromanen. DHd 2019. Conference Abstracts.

Kaufman, Shachar / Saharon Rosset / Claudia Perlich / Stitelman (2012): „Leakage in Data Mining: Formulation, Detection, and Avoidance". *ACM Trans. Knowl. Discov. Data* 6 (4): 15:1–15:21. https://doi.org/10.1145/2382577.2382579 .

Ke, Guolin / Qi Meng / Thomas Finley / Taifeng Wang / Wei Chen / Weidong Ma / Qiwei Ye / Tie-Yan Liu (2017): „LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In *Advances in Neural Information Processing Systems 30* , herausgegeben von I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, und R. Garnett, 3146–3154. Curran Associates, Inc. http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf .

Landeiro, V. / Culotta, A. (2016): „Robust Text Classification in the Presence of Confounding Bias". AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence , 186–193.

Nusser, Peter (1973): *Romane für die Unterschicht. Groschenhefte und ihre Leser.* Stuttgart: Metzler.

Nusser, Peter (1991): *Trivialliteratur*. Stuttgart: Metzler.

Nutz, Walter (1999): *Trivialliteratur und Populärkultur.* Opladen: Wiesbaden.

Olson, Randal S. / William La Cava / Zairah Mustahsan / Akshay Varik / Jason H. Moore (2017): „Data-driven Advice for Applying Machine Learning to Bioinformatics Problems". *arXiv:1708.05070 [cs, q-bio, stat]*, August. http://arxiv.org/abs/1708.05070.

Pedregosa, F. / G. Varoquaux / A. Gramfort / V. Michel / B. Thirion / O. Grisel / M. Blondel u. a. (2011): „Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12: 2825–2830.

Ribeiro, Marco Tulio/ Sameer Singh / Carlos Guestrin (2016): „'Why Should I Trust You?': Explaining the Predictions of Any Classifier". *arXiv:1602.04938 [cs, stat]*, Februar. http://arxiv.org/abs/1602.04938 .

Stockinger, Claudia (2018): „Das All dort draußen zeigt uns, wer wir sind. Die Leseuniversen der Groschenhefte". In Steffen Martus / Carlos Spoerhase (ed.): *Gelesene Literatur: Populäre Literatur im Medienwandel.* Text und Kritik. edition text und kritik.

Szubert, Benjamin, et al. (2019): "Structure-Preserving Visualisation of High Dimensional Single-Cell Datasets." *Scientific Reports*, vol. 9, no. 1, June, p. 8914, doi: 10.1038/s41598-019-45301-0.

Toman, M. / Tesar, R. / Jezek, K. (2006): "Influence in Word Normalization on Text Classification." Proceedings of InSciT 4 (2006): 354-358.