Genre classification on German novels

Lena Hettinger*, Martin Becker*, Isabella Reger†, Fotis Jannidis† and Andreas Hotho*‡
*Data Mining and Information Retrieval Group, University of Würzburg
Email: {hettinger, becker, hotho}@informatik.uni-wuerzburg.de
†Institut für Deutsche Philologie, University of Würzburg
Email: {isabella.reger, fotis.jannidis}@uni-wuerzburg.de
‡L3S Research Center, Leibniz Universität

Abstract—**The study of German literature is mostly based on literary canons, i.e., small sets of specifically chosen documents. In particular, the history of novels has been characterized using a set of only 100 to 250 works. In this paper we address the issue of genre classification in the context of a large set of novels using machine learning methods in order to achieve a better understanding of the genre of novels. To this end, we explore how different types of features affect the performance of different classification algorithms. We employ commonly used stylometric features, and evaluate two types of features not yet applied to genre classification, namely topic based features and features based on social network graphs and character interaction. We build features on a data set of close to 1700 novels either written in or translated into German. Even though topics are often considered orthogonal to genres, we find that topic based features in combination with support vector machines achieve the best results. Overall, we successfully apply new feature types for genre classification in the context of novels and give directions for further research in this area.**

I. Introduction

The study of German literature is mostly based on literary canons, i.e., small sets of specifically chosen documents [1]. In particular, the history of novels has been characterized using a set of only 100 to 250 works. Using small sets inherently limits the information conveyed, for example regarding the variety of different published works, development of genres, narrative techniques or themes. In order to address this issue, more works have to be included. However, the number of novels to inspect is large and the amount of information to process is not easy to handle. The lexicon of German poets and prosaists from 1800 to 1913 alone lists over 9200 works [2]. Yet, as the number of digitally available historic texts increases, this issue can be addressed using computational approaches.

Problem setting. We focus on genre classification in order to be able to extend existing literary canons with previously uncategorized novels using machine learning techniques. In particular, we focus on German novels written from the 16th to the 20th century and their classification into the subgenres educational and social novel. Yet, a major problem in this context is that there are differing opinions on how to identify genres. While most works on automatic genre classification use stylometric features like common word frequencies [3] or different kinds of markers (e.g. noun, POS tag, or punctuation mark counts) [4], [5], literary research distinguishes genres on a different level. It is mentioned, for example, that different genres elicit different social interactions between characters [6]. Also, topics often used for text categorization [7] are considered orthogonal to genres because documents addressing the same topic can be of a different genre [8], [9], [4], [5] and, in contrast to topics, a genre "describes something about what kind of document it is rather than what the document is about" [10]. Devitt [11] criticizes that genres are not about formal features or text classification and proposes a notion based on how humans experience the written text. This implicates that a broad variety of features has to be considered for genre classification. Genre concepts differ with respect to their level of generality, for example ”literature” (superordinate), ”novel” (basic), ”social novel” (subordinate) [12]. While most research on automatic text classification concentrates on the basic level, we are focussing on the subordinate level, also called subgenre.

Approach. To address the issue of genre classification on German novels, we are going to explore different features and classification algorithms. This is a first attempt to combine different genre facets. Overall, we categorize these facets and their corresponding features into three categories: features based on stylometrics ([3], [4], [5]), features based on content, and features based on social networks ([13], [6]), aiming to cover "human experience" [11] as completely as possible. For textual statistics, we use word frequencies as similarly proposed in [3]. For content based features, we are applying Latent Dirichlet Analysis [14] and automatically extract topic distributions. In order to cover features dealing with social networks, we extract character and interaction graphs from the novels and use them to build corresponding features. We compare these features using different classification algorithms, including rule based approaches, support vector machines and decision trees. Feature extraction is based on a set of 1682 novels written in or translated into German from the early 16th to the 20th century. Training and testing is executed on a subset of 132 labeled samples.

Contribution and Findings. To the best of the authors' knowledge, we are the first to quantitatively compare different feature types and classification algorithms on German novels. Also, as far as the sighted literature suggests, content specific features based on statistical topics as well as features based on social networks have not been applied to genre classification in general. Even though our testing data set is limited, our experiments indicate that topic-based features are a good feature for sub-genre classification. This is an interesting result, since literature research as well as other works in the field of genre classification emphasize that topics and genres are orthogonal concepts. Our results implicate that this orthogonality does not necessarily diminish the value of topics for genre classification. In contrast and against points raised by literature research ([13], [6]), classification on features based on social networks perform worse than the baseline. Yet this might be due to error prone named entity recognition or a mismatch of the extracted features. Overall, we successfully apply new feature types for genre classification in the context of novels and are able to give directions for further research in this area.

Structure. The rest of the paper is structured as follows: In Section II we describe works related to this paper. In Section III we introduce the features we are using for classification and in Section IV we give a short overview of our data set, the classification algorithms we apply and the corresponding results. We finish the paper with a discussion of our results in Section V and a short conclusion in Section VI.

II. Related Work

Genre classification is broadly applied in different areas, i.e., music [15], movies [16] and text based documents. For text based documents, there are several subcategories. Prominent examples are web pages [8], news paper articles ([3], [7]) or English prose [5]. There is also work crossing these subcategories as by Lee et. al [17] who take into account English and Korean research papers, homepages, reviews and more. In our paper we focus on a subset of genre classification in literature, namely classifying genres of novels, all of them either written in or translated into German language. Specifically, we investigate social and educational novels as sub-genres of novels.

Biber [9] introduces five dimensions of English texts and applies them to text categorization. He defines this approach using textual statistics like the frequency of nouns, present tense verbs, or word lengths and argues that such dimensions are context and domain specific. He does not apply them to automatic genre categorization. However, Karlgren and Cutting use some of Bibers features for recognizing text genres using discriminant analysis [4] on the Brown corpus [18] which represents a sample of English prose. They achieve an error rate of about 4% on the classes "Informative" and "Imaginative" and error rates up to 48% on more fine-grained genres. Building on research from Kessler et al. [5] and Burrows [19], Stamatatos et al. [3] use most frequent words for genre classification. They emphasize that frequencies are easy to compute and do not rely on the performance limits of external tools. They achieved error rates as low as 2.5% on newspaper articles using discriminative analysis. [8] use similar text statistics to classify genres of web pages, achieving an accuracy of 70%.

Jockers [20] reports on using successfully 42 high frequency tokens for the clustering of novel genres but cannot separate the authorial signal from the genre and time period signals. Underwood et al. [21] correctly point out two problems for the classification of novel genres: Historical heterogeneity because of the development of novel genres over time and feature heterogeneity because of the length of novels in contrast to articles, web pages etc. often used in genre classifications. Rosen-Zvi et al. [22] recognize the close relationship of words, topics and authors. They build a joint author-topic model based on Latent Dirichlet Analysis (LDA) [14]. This method directly or indirectly uses the notion of "topics" for author attribution. In this paper, we pick up this idea and use topics inferred using LDA as features for genre classification on German novels.

Another dimension that characterizes genre are social networks and interaction graphs between characters [6]. Extracting social networks or interactions from literary work has been addressed before [23] but was not applied to genre classification. Thus, in this paper, we build social network and interaction graphs and use different graph metrics as building blocks for new features.

TABLE I. Top tokens for 1682 German novels

rank	words	frequency	rank	words	frequency
1	.	7.003.325	6	sie	2.312.731
2	,	5.026.058	7	er	2.188.019
3	und	4.050.873	8	zu	2.168673
4	die	3.745.769	9	≫	2.131.584
5	der	2.626.422	10	ich	1.945.178

III. Features

In this section, we describe the features we investigate for genre classification and how we derive them from the data set. As mentioned before we look at three types of features: stylometric, content-based and social features. The stylometric as well as content-based features are extracted and normalized based on the whole corpus consisting of 1682 novels. Social features on the other hand are extracted from each novel separately. Thus, no inter-novel dependencies are present. Overall we obtain 216 features, all normalized to a range of [0,1], i.e., 101 stylometric, 70 content-based and 45 social feature as described in the following.

A. Stylometric features

In literary studies stylometry denotes the statistical analysis of texts to distinguish between authors or genres, e.g. by looking at word distributions. As mentioned by Stamatatos et al. [3] there is a wide range of stylometric features, including punctuation frequencies. Along the same lines, we also focus on word and punctuation frequencies to represent stylometric features. Note, that Stamatatos et al. use word frequencies over the whole English language and report this approach to perform better than just using word frequencies calculated from their corpus consisting of relatively short news paper articles. However, in contrast the average word count in our corpus is about 100,000 words over almost 1700 novels. This results in enough data to assume that we have a rather representative sample of the German language. We determine the most common words and punctuation marks and use the overall first hundred of them for classification. The ten most common tokens are depicted in Table I. To reduce bias stemming from differing text lengths, the features are normalized by dividing by the sum of the top hundred frequencies. As a small addition we also add the length of the text as a feature resulting in overall 101 features for this feature type.

B. Content-based features

In contrast to stylometric features which focus on features of writing style, content-based features capture the content of the corresponding novels. One particular way to represent content in the form of word distributions are topics as for example used in newspaper categorization [7]. For example, the top words associated with topics "emotions" and "formal society" are shown in Table III. In our work, we first remove predefined stop words from the novels and then use Latent Dirichlet Allocation (LDA) [14] on all 1682 works to extract topics in the form of word distributions. As LDA is an unsupervised topic model we can use it to automatically build topics, some if which might correspond to core vocabularies of genres. This is in contrast to core vocabularies which need genre information and manual selection at some point. These topics are then used to derive a topic distribution for each novel, i.e., we calculate how strongly each topic is associated with each novel. We interpret each topic association to a novel as a feature. In particular, we use LDA to extract 70 topics and set both required parameters α and β to 0.01 where each novel represents one document used as input for LDA, resulting in 70 content-based features. In this first study on combining different feature spaces, we did not optimize parameters.

C. Social features

Literary studies suggest that the genres we look at, namely social and educational novels, can be characterized by the number of protagonists and their interactions [13]. Thus, we aim to capture such characteristics using features derived from character and interaction graphs.

Character graph: Each character forms a node. Whenever two characters appear in one sentence we draw an edge.

Interaction graph: All characters in one sentence form a node representing an interaction. When the next interaction is described we draw an edge between successive interactions.

In order to create these graphs, we need to identify characters for each novel. We extract them by using the named entity recognition tool[1] by Jannidis et al. [24] which adapts well to the domain of German novels. Social features are then based on the most important characters and interactions in each novel, identified by using standard centrality measures, namely degree, closeness, betweenness and eigenvector centrality as for example defined by Noori [25]. High centrality values are supposed to correspond to important characters.

Based on these characters and interactions we attempt to model two important aspects: On one hand, there are novels which focus on one protagonist and one important interaction. On the other hand, there are novels which show a broader variety of characters and interactions, all equally important for the plot. Of course there are also novels which are a mixture of both aspects. We try to model these aspects by comparing the different centralities and derive features on i) how the centralities for the single most important character and the single most important interaction agree across different centrality measures, ii) how centrality measures differ across the most important four characters and interactions and finally, iii) we try to characterize the importance across the ten most important characters and interactions by modelling and comparing corresponding centrality distributions. In the following, we describe the resulting 45 social features in detail.

i) For centrality agreement on the most important characters and interactions, we take the nodes with the highest centrality values for each of the four centrality measure from both, the character and interaction graph, and calculate the average of the corresponding eight centrality values (ac). We calculate the same average for each type of graph separately summing up four centrality values each, i.e., one for each centrality measure, resulting in an average for the character graph (acf) and the interaction graph (aci). Additionally, we construct two features measuring whether the average for the character graph (acf) and the interaction graph (aci) agree (sd) or disagree (dd). Novels with one protagonist and one important interaction should have high scores for (ac), (acf), (aci) and (sd). Overall this results in 5 features.

ii) For centrality difference on the most important nodes, our motivation is that most information regarding centrality distributions lies within the first few centralities, cf. Table II. Therefore, we build another group of features for each centrality measure: We calculate the subsequent difference in centrality values of the four most important nodes in descending order for both graphs separately regarding a single measure. This yields three positive valued differences for each graph resulting in 24 features over all four centrality measures.

TABLE II. EXAMPLE FOR CENTRALITY VALUES

DegreeCentrality		EigenvectorCentrality	
T. Fontane: Effi Briest	C. Brontë: Jane Eyre	T. Fontane: Effi Briest	C. Brontë: Jane Eyre
0.366	0.174	0.222	0.206
0.140	0.153	0.208	0.145
0.087	0.111	0.154	0.103
0.076	0.097	0.115	0.095
0.070	0.090	0.077	0.094
0.058	0.083	0.066	0.090
0.058	0.076	0.053	0.078
0.052	0.076	0.049	0.068

iii) Finally, in order to characterize importance across the ten most important characters and interactions, we apply curve fitting to the ten highest centrality values for each centrality measure and each graph which are roughly power law distributed (see Table II). We fit a power law curve $f(x) = a \cdot x^b$ and extract the two parameters a and b of the fitted curve. Thus, we have two parameters for each curve, two graphs and four centrality measures, resulting in a total of 16 features.

IV. Evaluation

In the following, we describe the data sets used in our experiments, the different classifiers and the results for the different feature sets.

Data sets. In this paper, we use a corpus consisting of 1682 German novels freely available at TextGrid[2], DTA[3] and Gutenberg[4]. The list of titles used in this work will be published online[5]. Domain experts identified 11 of them as prototypical social and 21 as prototypical educational novels. This forms our first labeled subset, called *prototype*, and represents 32 novels which have very accurate labels. Another disjoint set of 100 novels was labeled by domain experts with the same classes yet not necessarily representing prototypical examples of either category. Of these 100 novels, 66 belong to class social and 34 to class educational. The overall 132 labeled novels form the second data set, called *labeled.*

All novels were written in or translated into German and date of origin ranges from the 16th to the 20th century. Most authors are male including for example Charles Dickens, Theodor Fontane, Karl May, Sir Walter Scott or Émile Zola. Text lengths range from 4000 to over one million words, the average word count being 100000. In contrast articles in the New York Times typically run from 400 to 1200 words[6]. To the best of our knowledge this is the first corpus containing genre-labeled German novels.

Classifiers. There exists a wide array of different classification algorithms. We will evaluate k-Nearest Neighbour (kNN),

[1] https://github.com/MarkusKrug/NERDetection

[2] http://www.textgridrep.de/

[3] http://www.deutschestextarchiv.de/

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

[5] http://dmir.org/genre-data

[6] http://www.nytimes.com/content/help/site/editorial/op-ed/op-ed.html

Naive Bayes (NB), Fuzzy Rule Learning (Rule), C4.5 pruned and unpruned (Tree), Multilayer Neural Network (NN) and linear Support Vector Machine (SVM), each implemented in KNIME[7] with standard parameters. As baseline we use a majority vote classifier (MV) which yields an accuracy score of 0.66 for data set *prototype* and 0.58 for *labeled*.

Feature sets. We use different subsets of the features introduced in Section III for classification, namely stylometric (st), topics (t), social (so), stylometric and topics (st+t), stylometric and social (st+so), topics and social (t+so) and all. We determine classification accuracy for each subset and each classifier. To account for the small data sets, we use 100 iterations of 10-fold cross validation. The depicted result are the average over these 1000 accuracy values.

Results. Tables IV and V show accuracy results for the *prototype* and *labeled* data set respectively. Those classifier-feature combinations which did significantly better than the majority vote baseline (MV) are marked bold in the tables. Statistical significance was tested using a t-test at $\alpha = 0.01$. Additionally, the best result for each classifier is underlined.

Generally, several classifier-feature tuples yield significantly (below: sign.) better results than the baseline. This indicates that we actually defined features which tend to capture the difference between the two genres, social and educational. Overall, results are better on the *prototype* data. Since the corresponding novels are prototypical for each genre, this strengthens the assumption that our features capture the actual genre characteristics. At the same time, while the accuracy values are slightly smaller, more classifier-feature tuples deliver sign. better accuracy for the larger *labeled* data set. The drop in accuracy is expected since genre characteristics are weaker in this data set. Yet, the larger number of sign. better accuracy values indicates that our features also work on larger data sets with varying strengths in genre characteristics.

Regarding different classifiers, basic classifiers like kNN, NB and Rule perform better on the small *prototype* data and worse on the *labeled* data when comparing against the SVM which might be due to the different class distributions. Fuzzy Rule Learning performs sign. worse than the baseline for every feature set on the labeled data whereas Naive Bayes yields better results over all feature sets and performs comparably well especially on the *prototype* data. For the small but securely labeled sample *protopye* pruning generally helps sign. to enhance decision trees, probably due to avoiding overfitting. Overall, SVM yields sign. the best results on every feature set apart from social features for labeled data, indicating that it may be the best choice for further applications.

Even though literature suggests the orthogonality of genres and topics [8] and mentions social features to be characteristic for certain genres [13], overall topic features score sign. best and social features worst among all feature sets. Adding social or stylometric features to the topic set results in a sign. better performance in only two cases. Hence, topic features alone are the best discriminative factors for educational and social novels. This indicates that despite orthogonality of topic and genre, topics may still be useful for genre classifiction. The bad performance of social features on the other hand may lie in the error prone named entity recognition or in the particular feature generation process we use. Overall, the best accuracy is achieved when using a Support Vector Machine in conjunction with topic features: 0.83 and 0.81 respectively.

TABLE III. TWO TOPICS AND THEIR MOST LIKELY WORDS

topic	most likely words
1	frau herr paris madame franken liebe mann frauen
2	liebe leben selbst herz mutter seele vater welt augen

TABLE IV. AVERAGE ACCURACY FOR CLASSIFICATION ON 32 GERMAN NOVELS USING 100 ITERATIONS AND 10-FOLD CROSS VALIDATION

features	MV	kNN	NB	Rule	Tree	pTree	NN	SVM
all	0.66	0.67	**0.72**	**0.69**	0.53	0.65	**0.72**	**0.72**
st	0.66	0.62	**0.73**	0.54	0.66	**0.69**	0.61	0.61
t	0.66	**0.78**	**0.71**	0.63	**0.71**	**0.76**	**0.69**	**0.83**
so	0.66	0.63	0.58	0.45	0.47	0.46	0.57	0.62
st + t	0.66	**0.70**	**0.71**	0.60	0.56	0.65	**0.74**	0.64
st + so	0.66	0.69	**0.72**	0.53	0.63	0.68	0.63	0.64
t + so	0.66	**0.71**	**0.71**	0.63	0.64	**0.75**	**0.71**	**0.71**

TABLE V. AVERAGE ACCURACY FOR CLASSIFICATION ON 132 GERMAN NOVELS USING 100 ITERATIONS AND 10-FOLD CROSS VALIDATION

features	MV	kNN	NB	Rule	Tree	pTree	NN	SVM
all	0.58	**0.64**	**0.70**	0.51	**0.65**	**0.65**	**0.70**	**0.73**
st	0.58	**0.64**	**0.64**	0.54	**0.61**	**0.60**	**0.64**	**0.69**
t	0.58	**0.67**	**0.69**	0.49	**0.67**	**0.69**	**0.69**	**0.81**
so	0.58	0.58	**0.61**	0.42	0.52	0.56	0.59	**0.59**
st + t	0.58	**0.67**	**0.69**	0.50	**0.67**	**0.65**	**0.71**	**0.74**
st + so	0.58	**0.64**	**0.63**	0.48	**0.61**	**0.60**	**0.64**	**0.71**
t + so	0.58	0.57	**0.69**	0.52	**0.65**	**0.67**	**0.69**	**0.78**

As topic features yield the best results, we take a closer look at the specific topics that are used in the classification process. Among the tested classifiers decision trees are best suited for interpretable results and the scores of pruned trees are above the baseline. In the following, we take a look at the first decision of the pruned trees using only topic features. For the *prototype* data, the pruned decision trees test the topic characterized by titles (Mrs., Mr., madame) as well as persons (woman and man) which can be denoted by the same words in German (Frau, Herr) first, see Topic 1 in Table III. If this topic is present, novels are more likely to be labeled as social by the decision tree. This is in line with the fact that social novels talk a lot about persons in a formal or descriptive way. For the *labeled* data another topic is used as the first decision indicator: It includes references to emotions (love, heart, soul) as well as to family members (mother, father), see Topic 2 in Table III. If this topic is present, novels are more likely to be labeled with the genre educational by the decision tree. This is in line with the fact that educational novels focus on feelings, family and experiencing life.

V. DISCUSSION

In this work, we have introduced two new feature types for genre classification in the context of novels and conducted experiments showing that topic based features perform well. In this section we discuss potential limitations of the approach and outline future work to be addressed in this context.

Limited data set and choice of genres. Our data set consists of almost 1700 novels with only a small subset being labeled as educational or social. We had 132 labeled instances containing 32 labeled as genre protoypes. While we believe that this is a

[7] https://www.knime.org/

good starting point for initial evaluation of feature performance and usefulness, which was the goal of this article, we also acknowledge that a more extensive study on different data sets and more genres classes needs to be conducted in order to further deepen the understanding of how the proposed features interact with genres in a more general way. Feature performance may vary given different environments.

Joint topic models. Since topic based features have been performing best, we believe that further research in this direction is justified. Building joint genre-topic models in the same way as author-topic models [22] is a promising line of future work. [22] also suggest to incorporate stylometric features to further improve their model which matches the common application of such features for genre classification [3].

Advanced NER and social evolution. In our current study we use conservative rules for NER i.e., we avoid currently error prone features like cross referencing. This may be one reason for the observed performance levels below baseline. Also, there exist more advanced methods to build social networks as described in [23]. After generating the graphs there is another large array of measures and methods to derive features which are then utilized for classification. One reason to further go into this direction is that different types of interactions are arguably part of the characteristics of different genres [13]. In particular, character development projects directly into the evolution of social and interaction networks throughout the novel, which we would therefore like to inspect further.

VI. Conclusion

In this paper, we have addressed the issue of genre classification in the context of novels. To this end, we applied different classification algorithms and evaluated a diverse set of features. Besides stylometric features common to genre classification we introduced two types of features which, to the best of our knowledge, have not been applied to genre classification before. That is, topic based features derived from statistical topics automatically generated using Latent Dirichlet Analysis and features based on social networks extracted from the novels. We evaluated how these features affected classification performance and found that the new features based on topics in combination with Support Vector Machine classification works best. This is especially interesting since genres and topics are considered to be orthogonal concepts. Overall, we successfully apply new feature types for genre classification in the context of novels and give directions for further research in this area.

In future work, our study can be extended by using larger data sets and different sets of genre types. Additionally, since topic based features work well, further research in this area is promising. In particular, joint genre-topic models in line with author-topic models are an interesting direction. Furthermore, even if social network and interaction based features have not been yielding the best results, advanced NER tools as well as considering current work in extracting static and dynamic networks may improve the performance of this type of features.

References

[1] R. Rosenberg, “Kanon,” *Reallexikon der deutschen Literaturwissenschaft. Bd. II*, pp. 224–227, 2000.

[2] C.-H. Joerdens, *Lexikon deutscher Dichter und Prosaisten.-Leipzig, Weidmann 1806-11*. Weidmann, 1810, vol. 5.

[3] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, “Text genre detection using common word frequencies,” in *Proc. 18th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 2000, pp. 808–814.

[4] J. Karlgren and D. Cutting, “Recognizing text genres with simple metrics using discriminant analysis,” in *Proc. 15th conference on Computational linguistics-Volume 2*, 1994, pp. 1071–1075.

[5] B. Kessler, G. Numberg, and H. Schütze, “Automatic detection of text genre,” in *Proc. 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 1997, pp. 32–38.

[6] L. Jappe, O. Krämer, and F. Lampart, *Figurenwissen: Funktionen von Wissen bei der narrativen Figurendarstellung*. Walter de Gruyter, 2012, vol. 8.

[7] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, “Text classification using machine learning techniques.” *WSEAS Transactions on Computers*, vol. 4, no. 8, pp. 966–974, 2005.

[8] S. M. Zu Eissen and B. Stein, “Genre classification of web pages,” in *KI 2004: Advances in artificial intelligence*. Springer, 2004, pp. 256–269.

[9] D. Biber, “The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings,” *Computers and the Humanities*, vol. 26, no. 5/6, pp. pp. 331–345, 1992.

[10] A. Finn and N. Kushmerick, “Learning to classify documents according to genre,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1506–1518, 2006.

[11] A. J. Devitt, “Generalizing about genre: New conceptions of an old concept,” *College composition and Communication*, pp. 573–586, 1993.

[12] D. Y. Lee, “Genres, registers, text types, domain, and styles: Clarifying the concepts and navigating a path through the bnc jungle.” *Language Learning & Technology*, vol. 5, no. 3, pp. 37–72, 2001.

[13] M. Hirsch, “From great expectations to lost illusions: The novel of formation as genre,” *Genre*, vol. XII, no. 3, p. 299, 1979.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[15] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.

[16] Z. Rasheed and M. Shah, “Movie genre classification by exploiting audio-visual features of previews,” in *Proc. 16th International Conference on Pattern Recognition*, vol. 2. IEEE, 2002, pp. 1086–1089.

[17] Y.-B. Lee and S. H. Myaeng, “Text genre classification with genre-revealing and subject-revealing features,” in *Proc. 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 145–150.

[18] W. N. Francis and H. Kucera, “Brown corpus manual,” *Brown University*, 1979.

[19] J. F. Burrows, “Word-patterns and story-shapes: The statistical analysis of narrative style,” *Literary and linguistic Computing*, vol. 2, no. 2, pp. 61–70, 1987.

[20] M. L. Jockers, *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[21] T. Underwood, M. L. Black, L. Auvil, and B. Capitanu, “Mapping mutable genres in structurally complex volumes,” in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 95–103.

[22] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, “The author-topic model for authors and documents,” in *Proc. 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 487–494.

[23] D. K. Elson, N. Dames, and K. R. McKeown, “Extracting social networks from literary fiction,” in *Proc. 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 138–147.

[24] F. Jannidis, M. Krug, I. Reger, M. Toepfer, L. Weimer, and F. Puppe, “Automatische Erkennung von Figuren in deutschsprachigen Romanen,” *DHd 2015*, 2015.

[25] A. Noori, “On the relation between centrality measures and consensus algorithms,” in *High Performance Computing and Simulation (HPCS), 2011 International Conference on*. IEEE, 2011, pp. 225–232.