Computational Linguistics in the Netherlands Journal 11 (2021) 85-104 Submitted 10/2021; Published 12/2021

Aspect-based Sentiment Analysis for German: Analyzing "Talk of Literature" Surrounding Literary Prizes on Social Media

Lore De Greve* LORE.DEGREVE@UGENT.BE
Pranaydeep Singh** PRANAYDEEP.SINGH@UGENT.BE
Cynthia Van Hee** CYNTHIA.VANHEE@UGENT.BE
Els Lefever** ELS.LEFEVER@UGENT.BE
Gunther Martens* GUNTHER.MARTENS@UGENT.BE

*Department of Literary Studies, Faculty of Arts and Philosophy, Ghent University, Belgium

**LT^3 Language and Translation Technology Team, Department of Translation, Interpreting and Multilingual Communication, Faculty of Arts and Philosophy, Ghent University, Belgium

Abstract

Since the rise of social media, the authority of traditional professional literary critics has been supplemented – or undermined, depending on the point of view – by technological developments and the emergence of community-driven online layperson literary criticism. So far, relatively little research (Allington 2016, Kellermann et al. 2016, Kellermann and Mehling 2017, Bogaert 2017, Pianzola et al. 2020) has examined this layperson user-generated evaluative "talk of literature" instead of addressing traditional forms of consecration. In this paper, we examine the layperson literary criticism pertaining to a prominent German-language literary award: the Ingeborg-Bachmann-Preis, awarded during the Tage der deutschsprachigen Literatur (TDDL).

We propose an aspect-based sentiment analysis (ABSA) approach to discern the evaluative criteria used to differentiate between 'good' and 'bad' literature. To this end, we collected a corpus of German social media reviews, retrieved from Twitter, and enriched it with manual ABSA annotations: *aspects* and aspect categories (e.g. the motifs or themes in a text, the jury discussions and evaluations, ...), *sentiment expressions* and *named entities*. In a next step, the manual annotations are used as training data for our ABSA pipeline including 1) aspect term category prediction and 2) aspect term polarity classification. Each pipeline component is developed using state-of-the-art pre-trained BERT models.

Two sets of experiments were conducted for the aspect polarity detection: one where only the aspect embeddings were used and another where an additional context window of five adjoining words in either direction of the aspect was considered. We present the classification results for the aspect category and aspect sentiment prediction subtasks for the Twitter corpus. These preliminary experimental results show a good performance for the aspect category classification, with a macro and a weighted F1-score of 69% and 83% for the coarse-grained and 54% and 73% for the fine-grained task, as well as for the aspect sentiment classification subtask, using an additional context window, with a macro and a weighted F1-score of 70% and 71%, respectively.

1. Introduction

In recent times, the knowledge of a limited number of professional 'pundits' is being challenged by technological developments and the 'wisdom of the crowds'. Ample research has been devoted to how technology causes shifts in the consecration of literary texts, affecting gatekeepers such as literary prizes (English 2009, Sapiro 2016), or professional critics' position of authority (Dorleijn et al. 2009, Löffler 2017, Thomalla 2018, Schneider 2018, Kempke et al. 2019, Chong 2020). Nevertheless, comparatively little research (Allington 2016, Pianzola et al. 2020, e.g.) has actually attempted to directly mine the content of user-generated online literary criticism by means of natural language

processing. Text mining could help to examine the role of peer-to-peer recommendation systems and layperson critics as new literary gatekeepers and cultural transmitters. It is an important case in point to study the differences between professional critics and this 'wisdom of the crowd', especially since traditional gatekeepers of the literary field (e.g. publishers, reviewers) are increasingly trying to tap the potential of online reading communities.

We aim to present the language technologies used in the context of the FWO-funded research project entitled "Evaluation of literature by professional and layperson critics. A digital and literary sociological analysis of evaluative talk of literature through the prism of literary prizes (2007-2017)".[1] This project aims to compare, analyse and mine the evaluative 'talk of literature' of both professional and layperson critics surrounding six prominent literary prizes in three different languages. In this paper, we will present our annotated corpus and suggest a sentiment analysis-based methodology to examine the professional and layperson literary criticism pertaining to the German-language *Ingeborg-Bachmann-Preis* and the *Tage der deutschsprachigen Literatur (TDDL)*.[2] During this event, all nominated contenders read an unpublished narrative text in front of a jury and a live (television) audience. This literary prize is unique because the contributions are discussed by the professional jury in the presence of the author and the live audience, but also, and even increasingly so, by an online audience. A devoted following of ca. 1000 Twitter followers reacts, by using the #tddl-hashtag, both to the literary text and its discussion by the official jury. For several years now, the organisers have been encouraging the online audience to use the formerly inofficial, originally community-driven #tddl-hashtag when tweeting or posting about the literary prize and competition.

Our ultimate goal is to gain insight into the evaluative criteria used by both professional and layperson critics to discern 'good' from 'bad' literature, as well as to engage with the differences in evaluative practices across platforms and media.[3] In order to do this, we aim at performing fine-grained aspect-based sentiment analysis (ABSA) on an annotated corpus consisting of comments and reader reviews on social media platforms, such as Twitter, Instagram and Goodreads. In the future, this system should make it possible to detect which sentiment is being expressed about a certain *aspect* or *feature expression* (e.g. contender, nominated book, jury, etc.) and *named entities* mentioned in such comments, and by whom. Consequently, we search to construct literary value through evaluative diction by using ABSA. In this paper, we will use a corpus of German tweets about the *TDDL* that have been enriched with manually added ABSA annotations on three levels: firstly the *aspects* and aspect categories, secondly, the *sentiment expressions*, and thirdly, *named entities*.

It should be clarified that there is a certain overlap between *feature expressions* and *named entities*. What we call *feature expressions* or *aspects* consist of a target word or phrase that is labelled or identified as an entity type or *main aspect category* and an additional attribute or *aspect subcategory*. Several of the *aspects*, however, such as mentions of the names of the jury members or contenders, are at the same time a *named entity* as well. Take the example sentence 'Ich liebe Birgit Birnbacher!' (translation: 'I love Birgit Birnbacher!'). On the one hand, the target words 'Birgit Birnbacher' are an *aspect*, which would be labelled as *main aspect category* "Contender" and *aspect subcategory* 'general'. This would result in the annotation label 'CONTENDER_General'. The entity type consists of uppercase letters and is separated from the attribute label by an underscore. If the attribute label consists of multiple words (cf. included examples of aspect categories), these are also

1. For more information: https://research.flw.ugent.be/en/projects/evaluation-literature-professional-and-layperson-critics-digital-and-literary-sociological or https://www.talklitmining.ugent.be/.
2. Translation: "Ingeborg-Bachmann-Prize" and "Days of German-Language Literature"
3. On the one hand, the texts that are presented during the TDDL might be considered examples of 'good' literature by virtue of being nominated. As a consequence, it could be argued that they therefore do not represent 'bad literature'. On the other hand, however, it is important to keep in mind that the judges each invite or nominate two authors to write a text instead of nominating an already existing and published work. The other judges, and indeed the audience as well, do not necessarily agree on the status of these texts as 'good literature', for instance the heated jury discussion and controversy surrounding Martin Beyer's text *Und ich war da* in 2019.

separated by an underscore. On the other hand, 'Birgit Birnbacher' is of course also a *named entity*, namely the personal name of one of the competitors.

These manual annotations are then used as training data for our ABSA pipeline including both aspect term category prediction and aspect term polarity classification. Our focus will mainly be on both the advantages and technical challenges raised by the nature of the corpus and the annotation system. Furthermore, we aim to describe the preliminary conditions and results for arriving at a model that will allow us to perform the aspect term category prediction and aspect term polarity classification on our corpus. Each pipeline component is developed using state-of-the-art pre-trained BERT models and we conducted two sets of experiments for the aspect term polarity classification. In the first experiment, only the aspect embeddings were used, whereas in the second one an additional context window of five adjoining words in either direction of the *aspect* was taken into consideration. We thus present the classification results for the aspect category and aspect sentiment prediction subtasks for our training corpus.

2. Related Research

Sentiment Analysis concerns the automatic identification and analysis of "people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu 2012, p. 1). The polarity is then the value or direction of opinion expressed (usually positive, negative and neutral). The study of sentiment analysis is possible at three levels: the document, sentence, and aspect level (Hu and Liu 2004). The latter approach is the task of Aspect-Based Sentiment Analysis (ABSA), which has mainly focused on customer reviews from websites and e-commerce platforms such as Amazon, Yelp, etc. ABSA then identifies the polarity of target entities and their aspect categories (Pontiki et al. 2016).

While ABSA is thus a common task (De Clercq et al. 2017, Balahur et al. 2019) with regard to domains such as restaurants, consumer technology and, to a lesser extent, movies, there have been few attempts to apply ABSA to domains that express sentiment in less lexicalized or straightforward ways. Jurafsky has done similar work in his article "linguistic markers of status in food culture" (Jurafsky et al. 2016). In an article on Australian book reviews, Stinson argued that "[c]omputational sentiment analysis—at least the kind enabled by off-the-shelf software tools—does not yet present an adequate means for determining polarity of book reviews"(Stinson 2016, p. 114). Stinson also argues for the necessity of going beyond sentence-level sentiment mining and doing ABSA in view of a recurring trait of corpora containing literary criticism, namely their tendency to voice criticism by means of the "compliment sandwich"(Stinson 2016, p. 108), in other words, by making use of elements of the epideictic discourse strategy of praising and blaming for purposes of nuance and comparison.

BERT-based models have been successful on a wide range of NLP tasks, including sentiment analysis (Xu et al. 2020). Yet, most approaches to sentiment analysis in German are still operating lexicon-based (Fehle et al. 2021) or at sentence-level (Jacobs et al. 2020). Aspect detection in (literary) book reviews is still in its infant stages (Villaneau et al. 2018). While the interest in amateur reviews of fictional literature is growing, neither case-based (Peplow et al. 2015) nor corpus-based research into online reading groups (Moser and Dürr 2021) actually engages in distant reading. There are ongoing efforts to take into account implicit sentiment (e.g. figurative speech) in online amateur reviews (Berenike and Messerli 2019). Cultural analytics approaches (Walsh and Antoniak 2021) target amateur criticism in a similar large-scale way, though without ABSA. The approach of Hofmann and colleagues (2021) is the most sustained effort to date to enrich sentiment/emotion analysis with world knowledge through semantic role and event identification attuned to specific domains. A related angle of attack is to be found in the Ghent-based research on exploring implicit sentiment evoked by events rather than by explicit evaluative diction (Van Hee et al. 2021). Efforts are under way to mine literary texts by means of sentiment analysis, though mostly concentrated

on drama texts that feature explicit speaker identification (Schmidt et al. 2021). To the best of our knowledge, this is the first study investigating fine-grained sentiment analysis on German social media posts about literary prizes.

3. Corpus Construction and Annotation

While the overarching project's entire corpus comprises comments on literary prizes in three different languages (i.e. English, Dutch and German), we focus on the specific challenges raised by performing sentiment mining on a German-language subsection of our corpus, namely on the tweets about the Ingeborg-Bachmann-Preis. The time frame for the collected data ranges from 2007, when Twitter was founded and the very first tweets about the Bachmann-Preis were created, up until 2019. The Bachmann-Preis has its own official Twitter account, @tddlit, and encourages the online audience to use #tddl as the hashtag when tweeting about the TDDL and the Bachmann-Preis.[4] In addition to the official #tddl-hashtag, we scraped similar relevant hashtags, for example by adding the year to the official hashtag, such as #tddl16 and #tddl2016, or by looking for other terms that might refer to the prize, e.g. #bachmannpreis or #bachmannwettbewerb. This led to a definitive list of 46 hashtags used between 2007 and 2019 (see Figure 1). We then collected all tweets containing these hashtags and removed those created outside of the examined time frame. In our paper, we

All scraped TDDL-Hashtags on Twitter (2007-2019)	
• bachmannbewerb	• tddl08
• bachmannpreis	• tddl09
• bachmannpreis2010	• tddl10
• bachmannpreis2013	• tddl11
• bachmannpreis2014	• tddl12
• bachmannpreis2015	• tddl13
• bachmannpreis2016	• tddl14
• bachmannpreis2017	• tddl15
• bachmannpreis2018	• tddl16
• bachmannpreis2019	• tddl17
• bachmannpreisträger	• tddl18
• bachmannpreisträgerin	• tddl19
• bachmannpreisträgerinnen	• tddl2009
• bachmannpreiswettbewerb	• tddl2011
• bachmannwettbewerb	• tddl2012
• bachmannwettbewerb2018	• tddl2013
• ingeborgbachmannpreis	• tddl2014
• ingeborgbachmannpreis2018	• tddl2015
• ingeborgbachmannpreisträgerin	• tddl2016
• tagederdeutschsprachigenliteratur	• tddl2017
• tagederdeutschsprachigenliteratur2018	• tddl2018
• tddl	• tddl2019
• tddl07	• tddlkanon

Figure 1: Overview of the scraped TDDL-related hashtags.

will present the annotation procedure and the subsequent steps towards automatising annotation of this corpus using a semi-supervised learning system. We annotated our training data, which will be used to set up said semi-supervised learning system, using INCEpTION (Klie et al. 2018). In INCEpTION, we created three layers, one for the *aspects* or *feature expressions*, one for the *named*

4. To safeguard the personal and privacy rights, tweets will be cited by mentioning only the tweet-ID, name of the website, date and last access of the collected tweet, e.g. 867326032038199297. *Twitter*, 24 May 2017. Accessed 14 September 2020.

entities and one for the *polarity triggers* or *sentiment expressions*. In order to categorise the *aspects* that are mentioned and evaluated in the tweets, we identify the relevant target words and label these, using a layered labelling system consisting of 7 main categories that are relevant in the context of literary prizes, namely "Text", "Reading", "Contender", "Jury", "Onsite Audience", "Meta" and "Allo-References". These categories group all *aspects* referring respectively to the nominated and competing texts, the live author-readings of said texts, the competing authors, the official jury of the prize, the audience present in the Bachmann-Preis studio, the meta-aspects of the prize and the references or comparisons made by the jury or the Twitter-users to other authors, literary works, musicians etc. Each main category is then divided into smaller and more specific subcategories, as illustrated by Figure 2, to gain a more detailed understanding of the content of the literary discourse. For the Text-category, for example, there are specific subcategories for those *feature expressions* that concern the characters, the flow, rhythm or punctuation of the text, its form, the text in general, the general content or plot, the language or style, motifs or themes, the point of view or narration, quotes from the text and the title. One of our aims is to discover how fine-grained the automatised identification and labelling of such *aspects* or *feature expressions* may be. We also annotated the *named entities*, the current focus however, is on the *aspects* and polarity. The list of main and subcategories was specifically designed based on a close reading of around 10,000 of the tweets surrounding the Bachmann-Preis and was created as to be suitable for other literary prizes and other corpora, e.g. Instagram posts, Goodreads reviews, reviews in newspapers, etc. Contrary to *named entities*, which are by definition explicit, *aspects* can sometimes be implicit. If there is an implicit reference to a certain aspect, we mark the word, the phrase or, if necessary, even the whole sentence as an *aspect* and label it with a tag that refers to the corresponding implied aspect. These references can be personal pronouns or articles, for example in "Das ist nur deshalb nicht schlechter als Stella, weil es nicht ganz so lang ist. #tddl.[5] In this tweet "Das" refers to the nominated text in general and would therefore be tagged as "TEXT_General" and is evaluated negatively. The implied aspect may also consist of other words or phrases; in the tweet "Yes! Gratulation an Birgit Birnbacher! #tddl", the explicit *aspect* is the contender "Birgit Birnbacher", who is being congratulated and is thus evaluated positively, which is expressed by the *sentiment expressions* "yes" with exclamation mark and the word "Gratulation".[6] Additionally, the word "Gratulation", however, also implies another aspect, namely the award ceremony, which is also evaluated positively. This word would therefore be tagged as an *aspect* with the label "META_Winner_Award-Ceremony".

We employ a tripartite polarity, using the labels 'positive', 'neutral' and 'negative' to label the *sentiment expressions* or *polarity triggers* in our corpus. These are then linked to the *aspects* or *feature expressions* as well as the *named entities* occurring in the tweets. In our corpus, positive and negative sentiment is most frequently expressed explicitly by sentiment bearing words or explicit evaluative diction in the form of adjectives and adverbs (e.g. "gut", "begeistert", "innovativ", "großartig", "peinlich", "schlecht", "furchtbar", "leider", "zu", ...), verbs (e.g. "gefallen", "lieben", "befürchten", gratulieren", "nerven", "sich freuen", "hassen", ...), nouns (e.g. "Lieblingstext", "Kitsch", "FavoritIn", "Kolportage", "Problem", "SprachkünstlerIn", "Lob", ...) as well as combinations of these, sometimes in addition to negations. However, during annotation we also take punctuation, most often exclamation marks, (e.g. "Es geht wieder los!") as well as capitalisation or alternative spelling of words (e.g. "DANKE", "Es geht looooos", "WAAAAAS?!", ...) into account as a possible indication or amplifier of sentiment. Nevertheless, *sentiment expressions*, similar to *aspects*, can be (more) implicit as well. These implied *polarity triggers* may consist of figurative speech or descriptions of actions that indicate (dis)approval (e.g. "fulminanten Internet-Applaus" or "ich schwenke Pompoms"), of comparisons to other works or authors - a comparison to Ingeborg Bachmann or Franz Kafka will imply a positive evaluation, whereas comparisons to works like

5. 1144937827244920834. *Twitter*, 29 June 2019. Accessed 18 January 2022.
Translation: "The only reason this is not worse than Stella is that it is not quite as long. #tddl".

6. 1145260003303022593. *Twitter*, 30 June 2019. Accessed 18 January 2022.
Translation: "Yes! Congratulations to Birgit Birnbacher! #tddl".

Würger's *Stella* will be negative - but also of interjections or exclamations (e.g. "o Gott", "meine Güte", "na ja", "puuh", "uups", "boah", ...) et cetera. In the case of implicit sentiment, we simply tag the corresponding word or phrase. In the case of the following tweet, "Meine Vermutung für die Shortlist 2019: Birnbacher, Fischer, Jost, Schultens, Wipauer, Othmann, Federer. #TDDL", a positive sentiment and evaluation is implied by the Twitter-user.[7] Because of the Twitter-user's guess or assumption that Birnbacher, Fischer, Jost, Schultens, Wipauer, Othmann and Federer will be selected for the shortlist, they suggest that these authors are, in their eyes, the best and most worthy, even though there are no explicit sentiment bearing words. For this example, we would label "Meine Vermutung für die Shortlist 2019" as a positive *sentiment expression* which refers to "Birnbacher", "Fischer", "Jost", "Schultens", "Wipauer", "Othmann" and "Federer", which are both *aspects* (CONTENDER_General) and *named entities* (PERSON_Contender). Neutral sentiment, on the other hand, is in itself a lack of sentiment bearing words or other elements that may suggest positive or negative sentiment. As a consequence, it is not possible to tag *sentiment expressions* for neutral sentiment in the same manner as for positive or negative sentiment. Depending on the context, we use two approaches. If the entire tweet contains no negative or positive *sentiment expressions*, the §-symbol at the beginning of the tweet is tagged as a *sentiment expression* and labelled as 'Neutral' in order to signal the tweet's lack of sentiment bearing words etc. In other cases, the tweets may contain both *aspects* or *named entities* that are mentioned in a neutral context as well as that are evaluated and can be linked to *polarity triggers*. In this case, The §-symbol at the beginning of the tweet is not tagged, the available *polarity triggers* are tagged and linked to the *aspects* or *named entities* to which they refer and the *aspects* or *named entities* about which no sentiment is expressed are not linked to any *polarity trigger*. When the annotated files are exported and processed into CSV-files, these 'neutral' *aspects* or *named entities* receive a 'NA'-label to indicate the lack of 'Positive'- or 'Negative'-label, we then interpret this as a stand-in for 'Neutral'. For the experiments presented in this article, however, we solely work at identifying positive and negative sentiment at present.

As a rule, we only annotate *aspects*, *named entities* and *sentiment expressions* that are relevant concerning our research focus. Should a Twitter-user also discuss and evaluate topics that are not related to literary prizes or literature, e.g. if they also express a sentiment regarding their work/pet/..., these are consequently not tagged and labelled, but ignored. The span length and complexity of the *feature* and *sentiment expressions* tend to vary, depending on whether the expression is implicit or explicit, as longer phrases may have to be labelled in case of implicit descriptions or figurative speech, but also on the *aspect* type. One of the annotated aspect subcategories are quotes, either from the nominated text, the jury or the contenders. These tend to be expressions with a longer and more complex span, frequently consisting of entire sentences, in "'Das hat was von einem Stephen-King-Setting', meint Winkels. Ich glaube, er hat noch nichts von King gelesen. #tddl", for example, "'Das hat was von einem Stephen-King-Setting'" (quotation marks included) would be tagged and labelled as a jury quote ('JURY_Quote').[8] Similarly, prepositional phrases related to location are regularly tagged in the context of the meta-aspect location-subcategory as well. As a consequence the span of longer, complicated expressions may inter alia include prepositional phrases and relative clauses.

7. 1145011751232135168; *Twitter*, 29 June 2019. Accessed 17 January 2022.
Translation: "My guess for the 2019 shortlist: Birnbacher, Fischer, Jost, Schultens, Wipauer, Othmann, Federer. #TDDL".

8. 1144230613685362689. *Twitter*, 27 June 2019. Accessed 18 January 2022.
Translation: "'It has something of a Stephen King setting,' Winkels says. I believe he has never read anything by King. #tddl".

Feature Expressions: Categories	
Main Category	**Subcategories**
Text	• Characters • Flow/ Rhythm/ Punctuation • Form • General • General Content/ Plot • Language/ Style • Motifs/ Themes • Point of View /Narration • Quote • Title
Reading	• Flow/ Rhythm/ Punctuation • General • Pronunciation/ Intonation/ Understandability
Contender	• Age • Appearance/ Clothing • Gender • General • Quote • Voice/ Language Use
Jury	• Age • Appearance/ Clothing • Behaviour • Discussion/ Valuation • General • Quote • Voice/ Language Use
Onsite Audience	• Age • Appearance/ Clothing • Behaviour • General
Meta	• Literature/ Literary Prizes • Location • Longlist • Main Event • Montage • Music • Online Assessment • Opening/ Opening Speech • Shortlist • Side Event • Technology/ Social Media • Videoportrait • Voting • Weather • Winner/ Award ceremony

Allo-References	• General
	• Music
	  o Musician
	  o Music
	• Other Person
	• Screen
	  o Director/Actor
	  o Film/Tv
	• Text
	  o Other Author
	  o Other Text

Figure 2: Overview of all aspect or feature expression (FE) categories

Figure 3 illustrates the annotation approach, using one of the tweets from the training corpus as an example.[9] "Kastberger" refers to Klaus Kastberger, one of the members of the professional jury, this word is therefore both a *feature expression* and a *named entity*. As a consequence, it is both tagged as the *named entity* "PERSON_Jury_Moderator" (yellow) and as the *aspect* "JURY_General" (orange), since it mentions Kastberger in general. The tweet also contains a second feature expression, namely "Textauswahl", which in turn concerns the text in general ("TEXT_General"). Both of these *aspects* and the *named entity* are evaluated positively, indicated by the *sentiment expression* "wirklich fantastisch" (purple). The *sentiment expression* expresses a non-ironic positive polarity and thus receives the "-Irony|Positive" label.

9. 1144220726494486529. *Twitter*, 27 June 2019. Accessed 8 September 2020.
Translation: "Kastberger's text selection is really fantastic #tddl".

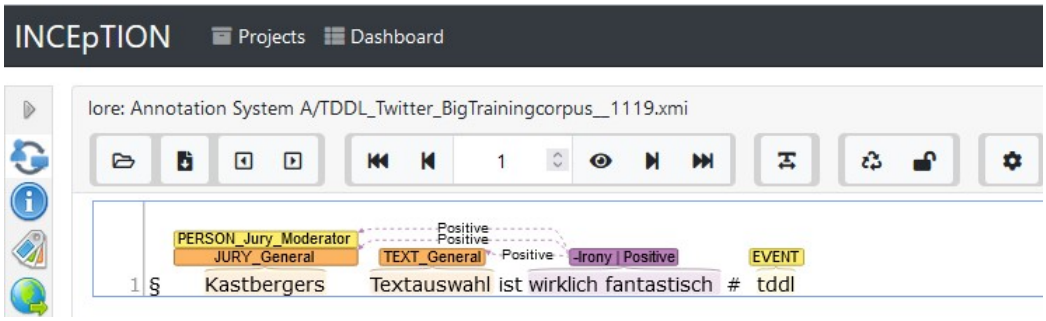

Figure 3: Example of an annotation in INCEpTION.

The sentiment is then linked to the *aspect* and *named entity* it refers to and the polarity is once again specified in the link. Thus, both Kastberger, as a *named entity* and as a *feature expression*, as well as the text *aspect* are evaluated positively. Finally, the tweet contains one more *named entity*, namely "tddl", within the hashtag and is tagged as an event. We do not tag it as a *feature expression* here, since it is not in itself one of the topics being discussed - although, implicitly, of course the TDDL are being discussed in this tweet - but simply acts as a hashtag. In other contexts, however, the hashtag may be tagged as a *feature expression* as well if it fulfills a double function.[10] Please note that the annotation system discussed here distinguishes itself from the one presented in "Wertung von Literatur 2.0: Eine digitale und literatursoziologische Analyse der Online-Twitter-Diskussion zu den Tagen der deutschsprachigen Literatur #tddl" (De Greve, Lore and Martens, Gunther 2021).[11]

Moving on to the corpus composition, we have manually annotated the majority of the 2019 run of the literary prize's online back-channel. To ensure the relevance of the selected training corpus, we decided to use those tweets created during the TDDL (26th-30th June) and containing the query or official hashtag "tddl", consisting of a total of 4364 tweets.[12] After we ran our first experiment using 400 tweets (for more information, see section 4.2.1), we decided to supplement our training corpus with an additional corpus consisting of 43 and 114 tweets from 2018 containing queries referring to the live, onsite audience and the author readings (see "Onsite Audience" and "Reading" in Figure 2), respectively, to deal with the data skew of these two main categories.[13] The annotation effort of our training corpus thus resulted in 4521 manually annotated tweets. Table 1 and Table 2 give an overview of the distribution of the aspect categories and polarity labels, which will be used to train and test the systems described in the next section (Section 4).

10. e.g. in the tweet "Egal. Wichtig sind sowieso nur #tddl", in which "tddl" is both a *named entity* referring to the event and the *feature expression* "META_Main-Event".
1143902091745865730. *Twitter*, 26 June 2019. Accessed 8 October 2021.
Translation: "Whatever. Only the #tddl are important anyway".
11. Firstly, the aspect categories that are employed here are somewhat different and slightly more fine-grained. Secondly, in "Wertung von Literatur 2.0", we looked at the *aspects* on a tweet level, meaning that multiple mentions of the same aspect category with the same polarity would only be taken into account once per tweet. Thirdly, the content of the jury discussion and quotes was not annotated, but simply received the "JURY_Discussion_Valuation" and/or "JURY_Quote" tag, whereas here *feature expressions*, *named entities* and *sentiment expressions* within citations etc. are tagged as well. Lastly, in the annotation for "Wertung von Literatur 2.0" only the authors who were nominated that year (2019) were tagged as contenders, here, however, previous contenders were tagged as such as well.
12. The reasoning behind this decision is that we assume that tweets that are both posted during the event and use the official hashtag are more likely to actually pertain to the *TDDL*.
13. Because the semi-supervised learning system will be used to automatically annotate the *TDDL*-related tweets from 2007 until 2017, we were able to collect the additional data from the tweets created in 2018 without having to resort to using a part of this corpus as a training dataset.

Aspect Category	**Total Nr**	**Neutral Labels**	**Positive Labels**	**Positive With Irony Marker**	**Negative Labels**	**Negative With Irony Marker**
'TEXT_Title'	11	5	2	1	4	0
'TEXT_Quote'	188	68	65	14	55	0
'TEXT_PoV_Narration'	57	3	24	0	30	0
'TEXT_Motifs_Themes'	72	26	20	0	26	0
'TEXT_Language_Style'	293	36	112	6	145	0
'TEXT_General-Content_Plot'	376	63	124	22	189	0
'TEXT_General'	1026	141	413	19	472	1
'TEXT_Form'	39	16	9	1	14	0
'TEXT_Flow_Rhythm_Punctuation'	25	2	11	0	12	0
'TEXT_Characters'	62	13	16	1	33	0
'READING_Pronunciation_Intonation_Understandability'	24	1	6	1	17	0
'READING_General'	316	172	88	5	56	0
'READING_Flow_Rhythm_Punctuation'	20	1	5	1	14	0
'ONSITE-AUDIENCE_General'	33	15	13	0	5	0
'ONSITE-AUDIENCE_Behaviour'	77	31	20	4	26	0
'ONSITE-AUDIENCE_Appearance_Clothing'	14	6	3	0	5	0
'ONSITE-AUDIENCE_Age'	4	0	1	1	3	0
'META_Winner_Award-Ceremony'	210	93	82	5	35	0
'META_Weather'	75	10	15	6	50	0
'META_Voting'	112	64	23	1	25	0
'META_Videoportrait'	84	6	40	2	38	0
'META_Technology_Social-Media'	311	131	91	8	89	0
'META_Side-Event'	19	9	6	0	4	0
'META_Shortlist'	52	26	11	1	15	0
'META_Opening-Speech'	125	57	51	3	17	0
'META_Online-Assessment'	217	162	42	3	13	0
'META_Music'	51	8	25	6	18	0
'META_Montage'	30	11	9	1	10	0
'META_Main-Event'	825	313	234	33	278	2
'META_Longlist'	4	3	0	0	1	0
'META_Location'	121	77	30	0	14	0
'META_Literature_Literary-Prizes'	116	61	26	4	29	0
'JURY_Voice_Language-Use'	41	4	14	6	23	0
'JURY_Quote'	165	75	42	11	48	0
'JURY_General'	276	82	117	9	77	2
'JURY_Discussion_Valuation'	1145	316	317	57	512	1
'JURY_Behaviour'	38	9	12	7	17	1
'JURY_Appearance_Clothing'	37	6	17	1	14	0
'JURY_Age'	3	0	0	0	3	0
'CONTENDER_Voice_Language-Use'	14	2	7	0	5	0
'CONTENDER_Quote'	15	11	2	0	2	0
'CONTENDER_General'	991	437	362	18	192	0
'CONTENDER_Gender'	47	19	5	2	23	0
'CONTENDER_Appearance_Clothing'	34	9	17	1	8	0
'CONTENDER_Age'	4	1	1	0	2	0
'ALLO-REFERENCES_TEXT_Other-Text'	111	67	26	1	18	1
'ALLO-REFERENCES_TEXT_Other-Author'	189	116	36	3	37	4
'ALLO-REFERENCES_SCREEN_Film_Tv'	49	22	19	4	8	0
'ALLO-REFERENCES_SCREEN_Director_Actor'	16	13	2	0	1	0
'ALLO-REFERENCES_OTHER_Person'	8	4	1	0	3	0
'ALLO-REFERENCES_MUSIC_Musician'	16	5	8	0	3	0
'ALLO-REFERENCES_MUSIC_Music'	3	1	1	0	1	0
'ALLO-REFERENCES_General'	73	41	14	4	18	1
Total	8264	2870	2637	273	2757	13

Table 1: This table shows the total number of times a specific aspect subcategory occurs, the number of times this subcategory is evaluated positively or negatively, as well as the number of positive and negative polarity labels that were tagged with an irony marker.

Aspect Category	Total Nr	Neutral Labels	Positive Labels	Positive With Irony Marker	Negative Labels	Negative With Irony Marker
'TEXT'	2149	373	796	64	980	1
'READING'	360	174	99	7	87	0
'ONSITE-AUDIENCE'	128	52	37	5	39	0
'META'	2352	1031	685	73	636	2
'JURY'	1705	492	519	91	694	4
'CONTENDER'	1105	479	394	21	232	0
'ALLO-REFERENCES'	465	269	107	12	89	6
'Total'	8264	2870	2637	273	2757	13

Table 2: This table shows the total number of times a specific main aspect category occurs, the number of times this main aspect category is evaluated positively or negatively, as well as the number of positive and negative polarity labels that were tagged with an irony marker.

4. Experiments

4.1 Experimental Setup

Mining for sentiment is feasible because of the somewhat ritualised and formulaic nature of the communication involved. However, there is a fair deal of ambiguity in the actual rhetoric of praise and criticism. In a small-scale experiment, we performed sentence-level sentiment analysis by means of the standard bert-base-uncased model from the Transformers repository by HuggingFace.[14] The models in question are trained for the objective of optimising MLM (Masked Language Modelling). The MLM objective gives the models an unprecedented grasp of the syntax and vocabulary while also adding contextualisation to the generated embeddings. The advanced language modelling makes these transformers very efficient at being trained for downstream tasks like part-of-speech tagging, named-entity recognition, or in this case, sentiment analysis.

In the presented research, we choose to focus primarily on the second and third task of ABSA, i.e. Category and Polarity Classification for the *aspects*. Consequently, we do not extract the *aspects* automatically, but we use the gold standard *aspects* for all experiments. We fine-tuned the pre-trained German BERT for these two different downstream tasks, namely (1) Aspect Category Classification, and (2) Aspect Sentiment Analysis, each time performing training on 80% of the data and evaluation on the remaining 20%. The first set of experiments was done on an original set of 400 tweets containing the hashtag "tddl", after which the additional 157 tweets about the reading and onsite audience category were added, whereas the second set of experiments was performed on the entire corpus of 4364 tddl-tweets from 2019 in addition to the 157 tweets from 2018. For the polarity detection task, two different approaches were applied: one model was trained using the target word embeddings only, whereas a second model was trained incorporating the averaged embeddings for an additional context window of 5 adjoining words (both preceding and following the target word).

4.2 Results with Original Dataset

4.2.1 Aspect Category classification

Firstly, we used the fine-tuned pre-trained German BERT for the coarse-grained Aspect Category Classification on a small training corpus consisting of 400 tddl-tweets from 2019. Our goal of running the experiment on a reduced corpus when taking only the main aspect categories into account was to get an initial insight into the performance of the learning system and the corpus itself before moving

14. https://github.com/huggingface/transformers

on to a larger or more complicated task. Table 3 displays the results of the first Aspect Category Classificication Experiment. As demonstrated by the table below, the results vary greatly depending on the main aspect category. For some categories, namely the "Jury" and "Onsite Audience", the results were already adequate, achieving an F1-score of 0.70. Others with larger support (i.e. more training instances) perform even better with an F1-score of 0.83 for "Text", 0.86 for "Contender" and 0.87 for "Meta". The F1-score for the "Allo-References"- and "Reading"-categories, however, is very low in comparison, namely 0.40. An examination of the results and the number of examples of each category in this original dataset revealed a data skew regarding the categories for "Allo-References", "Reading" and "Onsite Audience". A closer inspection of the complete dataset of 4364 tweets from 2019 divulged that this skew would automatically be resolved for the "Allo-References"-category, once we run the experiment on the larger dataset. This was not the case with regards to the "Reading" and "Onsite Audience", of which there were comparatively few examples in the entire training dataset. Therefore, we decided to add the additional corpus of 157 tweets from 2018 that mentioned these two *aspects* to try and improve the results. Nevertheless, the attained Macro-F1 score was already at 0.68 while the weighted F1 was at 0.79. Table 4 shows the results for the second run of the experiment with the additional 2018 tweets. The score increase is quite noticeable as the Macro-F1 score increases by 8%. The F1-scores improved for several categories, from 0.83 to 0.88 for "Text", 0.87 to 0.88 for "Meta", 0.40 to 0.67 for "Allo-References", and 0.40 to 0.81 for "Reading".[15] Strangely enough, however, there was a decrease in the F1-score of both the "Onsite Audience"-(0.70 to 0.67) and "Contender"-category (0.86 to 0.69). As these were not caused by a data skew and due to the improved overall accuracy, we continued with the second set of experiments.

Category	**Precision**	**Recall**	**F1-score**	**Support**
Accuracy: 0.80				
'TEXT'	0.82	0.84	0.83	32
'JURY'	0.63	0.79	0.70	24
'META'	0.85	0.88	0.87	66
'ALLO-REFERENCES'	0.50	0.33	0.40	3
'ONSITE-AUDIENCE'	0.88	0.58	0.70	12
'READING'	0.67	0.29	0.40	7
'CONTENDER'	0.86	0.86	0.86	21
Macro avg	0.74	0.65	0.68	165
Weighted avg	0.80	0.80	0.79	165

Table 3: Coarse-grained category classification results for the original small experimental corpus.

Category	**Precision**	**Recall**	**F1-score**	**Support**
Accuracy: 0.81				
'TEXT'	0.85	0.91	0.88	43
'JURY'	0.63	0.79	0.70	24
'META'	0.86	0.91	0.88	76
'ALLO-REFERENCES'	1	0.50	0.67	4
'ONSITE-AUDIENCE'	0.73	0.62	0.67	13
'READING'	0.89	0.75	0.81	32
'CONTENDER'	0.68	0.70	0.69	33
Macro avg	0.81	0.73	0.76	236
Weighted avg	0.81	0.81	0.80	236

Table 4: Coarse-grained category classification results for the extended small experimental corpus.

15. This increase can be explained by the fact that the additional 157 tweets may also include *aspects* other than the "Reading"- or "Onsite Audience"-categories, thus expanding the support for those other aspect categories as well, resulting in better F1-scores.

4.2.2 Polarity Classification

The second downstream task we performed was the Aspect Polarity Classification or Polarity Detection, to test how accurately the system was already able to predict whether an *aspect* was evaluated positively or negatively. For this initial run of the experiment we did not take irony into account and worked with the original small extended dataset of 557 tweets. Firstly, we ran this experiment using only the *aspect* or target word embeddings as training data (see Table 5). The experiment achieved an overall accuracy of 63%. The F1-score was higher for the predicted negative polarity (0.67) than for the positive polarity, which only attained an F1-score of 0.59. There was also a clear bias for negative sentiment, which in itself was not wholly surprising, as the large extended dataset contains more negative than positive sentiment (see Table 1). Secondly, we trained another model (Table 6) for the Aspect Polarity Classification. Instead of only taking the embeddings of the aspect tokens into account, we incorporated an additional context window of 5 adjoining words in either direction of the aspect. This model performed much better than the previous one. The overall accuracy went up by 9%, reaching 72%. The F1-score of predicted positive polarity became 0.68 and the predicted negative polarity resulted in an F1-score or 0.75. The bias towards negative sentiment remained, but the results of the polarity detection by this model were nonetheless substantially better than those of the previous model.

Category	**Precision**	**Recall**	**F1-score**	**Support**
Accuracy: 0.63				
positive	0.59	0.59	0.59	51
negative	0.67	0.67	0.67	64
Macro avg	0.63	0.63	0.63	115
Weighted avg	0.63	0.63	0.63	115

Table 5: Polarity classification results for the small extended experimental corpus when only using the embeddings of the target word.

Category	**Precision**	**Recall**	**F1-score**	**Support**
Accuracy: 0.72				
positive	0.69	0.67	0.68	51
negative	0.74	0.77	0.75	64
Macro avg	0.72	0.72	0.72	115
Weighted avg	0.72	0.72	0.72	115

Table 6: Polarity classification results for the small extended experimental corpus when also incorporating the average of the embeddings of a context window of five words preceding and following the target word.

4.3 Results with Extended Dataset

4.3.1 Aspect Category classification

After running the experiments on the original dataset, consisting of 557 tweets, we proceeded with the extended dataset of all 4364 tweets from 2019 as well as the 157 supplementary tweets for the underrepresented "Reading"- and "Onside Audience"-categories. Table 7 shows the results for the new run of the coarse-grained Aspect classification task, obtaining an accuracy score of 83%, an improvement of 2% compared to the result of the previous run with a smaller dataset. Adding more data thus improved the overall accuracy. For some categories, however, the F1-score slightly decreased for the extended dataset. Although we do not offer a conclusive reason for this trend,

one possible explanation may be the frequent implicitness of the *aspects* in the tweets. These were annotated to the extent possible, but an increase in data is also accompanied by an increase of these more implicit *aspects*, which may influence the Aspect Classification. The F1-score for the "Text"-category went from 0.88 to 0.85, for "Meta" from 0.88 to 0.86, and for the "Reading"-category from 0.81 to 0.73.[16] However, there was an increase in the F1-scores for all other aspect categories. A smaller one for the "Onsite Audience"-category (0.67 to 0.70) and the "Allo-References" (0.67 to 0.71), and a substantial improvement regarding the "Jury", which rose from 0.70 to 0.81, and the "Contender", which now attained an F1-score of 0.86 instead of 0.69.

Category	**Precision**	**Recall**	**F1-score**	**Support**
Accuracy: 0.83				
'TEXT'	0.87	0.83	0.85	408
'JURY'	0.75	0.88	0.81	311
'META'	0.87	0.86	0.86	415
'ALLO-REFERENCES'	0.74	0.68	0.71	76
'ONSITE-AUDIENCE'	0.73	0.67	0.70	24
'READING'	0.78	0.69	0.73	42
'CONTENDER'	0.92	0.82	0.86	198
Macro avg	0.71	0.68	0.69	1475
Weighted avg	0.84	0.83	0.83	1475

Table 7: Coarse-grained category classification results for the extended dataset.

Due to the high accuracy score the coarse-grained aspect category classification task obtained, we decided to also run a fine-grained classification task, taking all subcategories shown in Figure 2 into consideration. Table 8 lists the experimental results for this fine-grained category classification, with an overall accuracy score of 73%.

The F1-score varies greatly between the subcategories, ranging from 0 to 0.92, depending, on the one hand, on the number of examples or the size of the subcategory. Some subcategories are mentioned very infrequently, demonstrating that these categories are less relevant for the current corpus. However, this does not mean they will be irrelevant for media platforms other than Twitter or for different prizes, which is why they are still included in the list of aspect categories. They therefore do not signal a lack of success of our system, but indicate the relevance of a certain *aspect* regarding the annotated corpus. On the other hand, a low F1-score may also result from the specific difficulties of recognising a specific subcategory, either because the *aspect* is often only implicitly mentioned, or because the target word(s) referring to this category are more diverse. Some subcategories are often expressed by the same target word(s). As an example, we can refer to the "general" subcategories, which mostly consist of the same recurring word(s), such as the "CONTENDER_General" and "JURY_General"-subcategories, which mainly consist of the names of the contenders or jury members, "READING_General" with frequent mentions of "Lesung" or "liest", the "META_Main-Event" which are mostly variations of "tddl", "Bachmannpreis" etc, and so on. Other subcategories are not as easily recognisable, because there are no "set" references to rely on, e.g. quotes or the mentions of text's content or plot are generally far less formulaic and tend to diverge more.

16. The latter may be related to the overlap between the main aspect categories of "Reading" and "Contender".

Category	Precision	Recall	F1-score	Support
Accuracy: 0.73				
'TEXT_Title'	0.00	0.00	0.00	1
'TEXT_Quote'	0.50	0.63	0.56	35
'TEXT_PoV_Narration'	0.64	0.47	0.54	15
'TEXT_Motifs_Themes'	0.91	0.62	0.74	16
'TEXT_Language_Style'	0.65	0.66	0.65	50
'TEXT_General-Content_Plot'	0.49	0.57	0.53	77
'TEXT_General'	0.85	0.85	0.85	188
'TEXT_Form'	0.60	0.50	0.55	6
'TEXT_Flow_Rhythm_Punctuation'	0.00	0.00	0.00	4
'TEXT_Characters'	0.62	0.50	0.56	10
'READING_Pronunciation_Intonation_Understandability'	0.50	0.25	0.33	4
'READING_General'	0.82	0.79	0.80	42
'READING_Flow_Rhythm_Punctuation'	1.00	0.40	0.57	5
'ONSITE-AUDIENCE_General'	0.86	0.67	0.75	9
'ONSITE-AUDIENCE_Behaviour'	0.80	0.62	0.70	13
'ONSITE-AUDIENCE_Appearance_Clothing'	0.50	0.33	0.40	3
'ONSITE-AUDIENCE_Age'	0.00	0.00	0.00	2
'META_Winner_Award-Ceremony'	0.68	0.72	0.70	32
'META_Weather'	1.00	0.57	0.73	14
'META_Voting'	0.79	0.55	0.65	20
'META_Videoportrait'	0.89	0.80	0.84	10
'META_Technology_Social-Media'	0.80	0.69	0.74	54
'META_Side-Event'	1.00	0.20	0.33	5
'META_Shortlist'	1.00	0.64	0.78	14
'META_Opening-Speech'	1.00	0.75	0.86	24
'META_Online-Assessment'	0.93	0.90	0.92	42
'META_Music'	0.43	0.60	0.50	5
'META_Montage'	0.67	0.67	0.67	6
'META_Main-Event'	0.88	0.92	0.90	150
'META_Longlist'	0.00	0.00	0.00	2
'META_Location'	0.83	0.62	0.71	24
'META_Literature_Literary-Prizes'	0.82	0.56	0.67	25
'JURY_Voice_Language-Use'	0.00	0.00	0.00	6
'JURY_Quote'	0.37	0.36	0.37	36
'JURY_General'	0.78	0.77	0.77	47
'JURY_Discussion_Valuation'	0.59	0.83	0.69	198
'JURY_Behaviour'	0.00	0.00	0.00	7
'JURY_Appearance_Clothing'	0.56	0.56	0.56	9
'CONTENDER_Quote'	1.00	0.33	0.50	3
'CONTENDER_General'	0.86	0.88	0.87	181
'CONTENDER_Gender'	0.60	0.60	0.60	5
'CONTENDER_Appearance_Clothing'	1.00	0.33	0.50	6
'ALLO-REFERENCES_TEXT_Other-Text'	0.44	0.29	0.35	14
'ALLO-REFERENCES_TEXT_Other-Author'	0.70	0.61	0.66	31
'ALLO-REFERENCES_SCREEN_Film_Tv'	0.67	0.40	0.50	5
'ALLO-REFERENCES_SCREEN_Director_Actor'	1.00	0.67	0.80	3
'ALLO-REFERENCES_MUSIC_Musician'	1.00	0.40	0.57	5
'ALLO-REFERENCES_General'	0.67	0.44	0.53	9
Macro avg	0.64	0.50	0.54	1475
Weighted avg	0.74	0.73	0.73	1475

Table 8: Fine-grained category classification results for the extended dataset.

4.3.2 Polarity classification

This section reports on the classification of the binary polarity prediction task ('negative', 'positive') for the extended dataset. This time, we wished to examine the influence of irony as well, which was ignored in the polarity detection task for the small dataset. Although the system had not been trained to recognise irony, we decided to verify its performance if irony was to be taken into account. Once again, we ran the experiment twice: Table 9 shows the classification results for the model only relying on embedding information for the target word, whereas Table 10 gives an overview of the results when also including the averaged embeddings of the context window including five words preceding and following the target word(s). The polarity prediction that relied solely on the aspect embeddings obtained an accuracy score of 68%, 5% higher than previously for the original small dataset, and an F1-score of 0.66 instead of 0.59 for positive polarity and 0.70 instead of 0.67 for negative polarity. As a consequence, the increase of data has clearly improved the performance of this model. The overall accuracy is one percent lower (67%), however, if the ironic instances are taken into account, and the F1-score decrease for the prediction of both positive (0.64) and negative polarity (0.69).

Category	**Precision**	**Recall**	**F1-score**	**Support**
with ironic instances				
Accuracy: 0.67				
positive	0.69	0.60	0.64	498
negative	0.65	0.73	0.69	506
Macro avg	0.67	0.67	0.66	1004
Weighted avg	0.67	0.67	0.66	1004
without ironic instances				
Accuracy: 0.68				
positive	0.65	0.67	0.66	434
negative	0.71	0.69	0.70	516
Macro avg	0.68	0.68	0.68	950
Weighted avg	0.68	0.68	0.68	950

Table 9: Polarity classification results for the extended dataset when only using the embeddings of the target word.

Category	**Precision**	**Recall**	**F1-score**	**Support**
with ironic instances				
Accuracy: 0.68				
positive	0.66	0.64	0.65	467
negative	0.69	0.72	0.70	537
Macro avg	0.68	0.68	0.68	1004
Weighted avg	0.68	0.68	0.68	1004
without ironic instances				
Accuracy: 0.71				
positive	0.71	0.60	0.65	424
negative	0.71	0.81	0.76	526
Macro avg	0.71	0.70	0.70	950
Weighted avg	0.71	0.71	0.71	950

Table 10: Polarity classification results for the extended dataset when also incorporating the average of the embeddings of a context window of five words preceding and following the target word.

Once more, the accuracy is higher for the second model, which includes the averaged embeddings of the context window consisting of the five words that precede and follow the aspect. This model obtains an accuracy score of 71%, slightly lower than the previous run on the original dataset, despite the increase in data. The prediction of positive polarity obtains an F1-score of 0.65 (cf. 0.68 in Table 6), the prediction of negative polarity has an F1-score of 0.76 (cf. 0.75 in Table 6). If the model has to take irony into account as well, the accuracy drops to 68%, with an F1-score of 0.65 for positive and of 0.70 for negative polarity.

As neither model was specifically trained to identify irony, it is to be expected that the accuracy and F1-scores decrease when ironic instances are taken into consideration. This indicates that the models would need additional training for the recognition and correct prediction of irony to improve the results, were we to proceed with irony prediction in addition to the basic polarity detection. We can also conclude that the second model appears to have reached its maximum performance, as the accuracy no longer improved after additional data was added. In order to further improve the performance, further steps would be needed, such as pre-training specifically for the context of literary criticism.

In order to provide some insight into the limitations and difficulties related to the automatised Polarity Classification, we have included some examples of instances in Table 11 where the predicted sentiment differs from the annotated sentiment. The left column contains the tweets.[17] The *aspects* about which the sentiment in question is expressed are printed bold, the corresponding *sentiment expressions* are either teal, indicating they received the 'Positive'-label during annotation, or red, if they were labelled as 'Negative'. The second and third column contain the annotated and predicted sentiment, respectively. In the first example, the system relied on the positive adverb "grandios" for its prediction, but it did not take into account that it is used here to amplify the negative sentiment expressed by "verhauende". In the second example, the system focused on the word "Fucking", which is here used as an amplifier instead of as a curse, instead of the positive sentiment bearing adjective "Tolle". There is quite a distance span between the *aspect* and the positive *sentiment expression* "freue mich" in the third tweet. Because of the span, the system no longer recognizes the link between the *aspect* and *sentiment expression* and instead predicted negative sentiment. Concerning the fourth tweet, the cause for the difference in prediction is not immediately discernable, however, it may be connected to the distance between the *aspect* and the core of the negative *sentiment*

17. Citations and translations per example:

1. 1144159541745131521. Twitter, 27 June 2019. Accessed 19 January 2019. Translation: "There are some grandiosely bungled sentences in the text. Didn't anyone proofread this beforehand? #tddl".
2. 1144160247025754112. Twitter, 27 June 2019. Accessed 19 January 2019. Translation: "Captivity in the (fictional) imagination, i.e. Fucking Kayfabe and its power of seduction! Great speech by Clemens J. Setz at the tddl opening. (But now continuing with online abstinence)! https://www.derstandard.de/story/2000105488380/bachmannpreis-fucking-kayfabe ...".
3. 1144133352334774273. Twitter, 27 June 2019. Accessed 19 January 2019. Translation: "Today at 10 a.m. I'll be watching the #TDDL for the first time in my life. Katharina Schultens is reading a text of which I was privileged to witness the genesis and I am delighted that the whole world will now hear it. https://bachmannpreis.orf.at/stories/2978635/ ...".
4. 1144160422364352512. Twitter, 27 June 2019. Accessed 19 January 2019. Translation: "I'll start with food first, this is really too irritating for me right now after 2 hours of sleep #tddl".
5. 1144160468786982912. Twitter, 27 June 2019. Accessed 19 January 2019. Translation: "#tddl First Reading. Future world - radically envisioned. Brutal and poetic at the same time. pic.twitter.com/f3BFdmWgyE".
6. 1144168331727032320. Twitter, 27 June 2019. Accessed 19 January 2019. Translation: "Atmospherically, I also thought of 'Never let me go' by Kazuo Ishiguro. #tddl".
7. 1144173077724573696. Twitter, 27 June 2019. Accessed 19 January 2019. Translation: "Progress: A video portrait which I don't find shitty. #tddl".
8. 1144176638961377280. Twitter, 27 June 2019. Accessed 19 January 2019. Translation (quote included): "'There are omens not only for death, but also for becoming a ghost.' <3 #tddl #wipauer".

	Tweet	Annotated Sentiment	Predicted Sentiment
1	Da sind grandios verhauende **Sätze im Text**. Hat das denn vorher keiner lektoriert? #tddl	Negative	Positive
2	Die Gefangenschaft in der (fiktiven) Vorstellung, also Fucking Kayfabe und dessen Verführungs-Kraft! Tolle **Rede von Clemens J. Setz zur tddl-Eröffnung**. (Nun aber weiter mit der Online-Abstinenz)! https://www.derstandard.de/story/2000105488380/bachmannpreis-fucking-kayfabe ...	Positive	Negative
3	Heute um 10 Uhr gucke ich das erste Mal in meinem Leben bei den #TDDL rein. Katharina Schultens liest einen **Text**, dessen Genese ich beiwohnen durfte und ich freue mich, dass ihn nun die ganze Welt hören wird. https://bachmannpreis.orf.at/stories/2978635/ ...	Positive	Negative
4	Ich fang erst mal mit Essen an, **das** ist mir gerade echt zu nervig nach 2 Std Schlaf #tddl	Negative	Positive
5a	#tddl Erste Lesung. **Künftige Welt - radikal gedacht**. Brutal und poetisch zugleich. pic.twitter.com/f3BFdmWgyE	Positive	Negative
5b	#tddl Erste Lesung. Künftige Welt - radikal gedacht. **Brutal und poetisch** zugleich. pic.twitter.com/f3BFdmWgyE	Positive	Negative
6	**Atmosphärisch** dachte ich auch an "Never let me go" von Kazuo Ishiguro. #tddl	Positive	Negative
7	Fortschritt: Ein **Videoporträt**, das ich nicht scheisse finde. #tddl	Positive	Negative
8	**"Nicht nur für den Tod gibt es Vorzeichen, auch für die Gespenstwerdung."** <3 #tddl #wipauer	Positive	Negative

Table 11: Examples of differences between annotated and predicted sentiment regarding the Polarity Classification.

expression ("nervig"), as well as the somewhat informal language use. The sentiment was predicted incorrectly twice for similar reasons in the following tweet. In both cases, the adverb "radikal" and the adjective "brutal" are interpreted as expressing negative sentiment by the system, whereas they are actually expressing positive sentiment in the context of literary criticism. The fifth tweet contains an implicit *sentiment expression* which is conveyed by the comparison to a celebrated novel and author. However, this is not "noticed" by the system due to a lack of explicit evaluative diction. In the next example, the system does not take the litotes, the negation of the negative adjective "scheisse", into account. The final tweet illustrates the difficulties caused by pictorial language, which is frequently used in social media contributions. The system does not recognize the heart-emoticon, which communicates a positive evaluation.

5. Conclusion and Future Work

In this research, we investigated an aspect-based sentiment analysis (ABSA) approach to discern the evaluative criteria used to differentiate between 'good' and 'bad' literature. To this end, we collected a corpus of German social media reviews from Twitter and manually labelled *aspects*, aspect categories, *sentiment expressions* and *named entities*. Next, machine learning experiments were performed to automatically predict aspect term categories and polarities. Supervised classification systems were trained using pre-trained BERT models in two different flavours: one where only the aspect embeddings were used, and a second one where an additional context window of five adjoining words in either direction of the *aspect* was considered. We show promising results, with an overall weighted F1-score of 83% for the coarse-grained and 73% for the fine-grained aspect category classification task, and 71% for the aspect sentiment prediction task.

In future work, we intend to pursue different directions to improve sentiment-based learning of the model. Firstly, we hypothesize that German BERT, which is trained on the German Wikipedia and other similar corpora with formal language, may lack inherent insight into social media language and its nuances. To this end, we propose to re-train German BERT with additional Twitter data. Secondly, as mentioned above, sentiment can be a multi-layered construct in the context of art reviewing. We therefore also hypothesise that simply a fraction of our corpus may not be sufficient to instil sufficient knowledge of the complications with polarities and aspect detection. To solve this issue, we propose to pre-train our version of German BERT with related English sentiment datasets, by translating them to German. We believe these two additional learning signals might be sufficient to outperform previous approaches to ABSA for literary reviews. In addition, we also plan to investigate a model to not simply classify the aspect categories, but to actively and automatically extract the aspect terms from literary reviews (Tulkens and van Cranenburgh 2020). This task will come with its own set of challenges, especially regarding the subcategories that obtained a lower F1-score in Section 4.3.1. On the one hand, the challenges will be related to the diversity of the target words that refer to a certain aspect. On the other hand, the task may also be complicated by the length of the strings, as for instance quotes or descriptions of a text's plot or content can sometimes be extensive.

Acknowledgements

This research was supported by the Flemish Research Fund FWO-Flanders under project code G087119N.

References

Allington, Daniel (2016), 'Power to the Reader' or 'Degradation of Literary Taste'? Professional Critics and Amazon Customers as Reviewers of 'The Inheritance of Loss', *Language and Literature* **25** (3), pp. 254–278. https://doi.org/10.1177/0963947016652789.

Balahur, Alexandra, Roman Klinger, Veronique Hoste, Carlo Strapparava, and Orphée De Clercq (2019), *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics.

Berenike, Herrmann and Thomas C. Messerli (2019), Metaphors we read by: Finding metaphorical conceptualizations of reading in web 2.0 book reviews, *DH2020 Conference*.

Bogaert, Xiana (2017), *'ICH WÜRDE AM LIEBSTEN MIT DER JURY DISKUTIEREN! #TDDL' - Der Ingeborg-Bachmann-Preis: Ein Vergleich zwischen der Jury- und Laienkritik auf Twitter.*, Master's thesis, Ghent University.

Chong, Phillipa K. (2020), *Inside the Critics' Circle: Book Reviewing in Uncertain Times*, Princeton University Press. http://www.jstor.org/stable/j.ctvkwnph1.

De Clercq, Orphée, Els Lefever, Gilles Jacobs, Tijl Carpels, and Véronique Hoste (2017), Hello ABSA! Collaboration between LT3 and Hello Customer to develop an aspect-based sentiment analysis pipeline, *ATILA17 Conference (Tilburg University)*.

De Greve, Lore and Martens, Gunther (2021), Wertung von Literatur 2.0 : eine digitale und literatursoziologische Analyse der Online-Twitter-Diskussion zu den Tagen der deutschsprachigen Literatur #tddl, *in* Ruf, Oliver and Winter, Christoph H., editor, *Small critics: transmediale Konzepte feuilletonistischer Schreibweisen der Gegenwart*, Vol. 3 of *Mikrographien / Mikrokosmen. Kultur-, literatur- und medienwissenschaftliche Studien; Band 3*, Königshausen & Neumann.

Dorleijn, Gillis J., Dirk De Geest, and Koen Rymenants (2009), *Kritiek in crisistijd: Literaire kritiek in Nederland en Vlaanderen tijdens de jaren dertig*, Nijmegen: Vantilt.

English, James F. (2009), *The Economy of Prestige: Prizes, Awards, and the Circulation of Cultural Value*, Harvard University Press. https://books.google.be/books?id=vY3UOFDA2sAC.

Fehle, Jakob, Thomas Schmidt, and Christian Wolff (2021), Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques, *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, KONVENS 2021 Organizers, Düsseldorf, Germany, pp. 86–103. https://aclanthology.org/2021.konvens-1.8.

Hofmann, Jan, Enrica Troiano, and Roman Klinger (2021), Emotion-Aware, Emotion-Agnostic, or Automatic: Corpus Creation Strategies to Obtain Cognitive Event Appraisal Annotations, *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Hu, Minqing and Bing Liu (2004), Mining and summarizing customer reviews, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Acm, New York, NY, pp. 168–177.

Jacobs, Arthur M., Berenike Herrmann, Gerhard Lauer, Jana Lüdtke, and Sascha Schroeder (2020), Sentiment Analysis of Children and Youth Literature: Is There a Pollyanna Effect?, *Frontiers in Psychology* **11**, pp. 2310. https://www.frontiersin.org/article/10.3389/fpsyg.2020.574746.

Jurafsky, Dan, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith (2016), Linguistic Markers of Status in Food Culture: Bourdieu's Distinction in a Menu Corpus, *Cultural Analytics*.

Kellermann, Holger and Gabriele Mehling (2017), *Laienrezensionen auf amazon.de im Spannungsfeld zwischen Alltagskommunikation und professioneller Literaturkritik.*, Würzburg: Königshausen und Neumann, pp. 173–202.

Kellermann, Holger, Gabriele Mehling, and Martin Rehfeldt (2016), *Wie bewerten Laienrezensenten? Ausgewählte Ergebnisse einer inhaltsanalytischen Studie.*, Würzburg: Königshausen und Neumann, pp. 229–238.

Kempke, Kevin, Lena Vöcklinghuis, and Miriam Zeh (2019), *Institutsprosa: Literaturwissenschaftliche Perspektiven auf akademischen Schreibschulen*, Leipzig : Spector Books.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych (2018), The inception platform: Machine-assisted and knowledge-oriented interactive annotation, *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, pp. 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/.

Liu, Bing (2012), Sentiment analysis and opinion mining, *Synthesis lectures on human language technologies* **5** (1), pp. 1–167, Morgan & Claypool Publishers.

Löffler, Sigrid (2017), Danke, kein Bedarf? Wie die totgesagte Literaturkritik ihr Ableben überleben könnte, *Stimmen der Zeit <Freiburg, Breisgau>* **235** (12), pp. 850–814.

Moser, Doris and Claudia Dürr (2021), *Über Bücher reden: Literaturrezeption in Lesegemeinschaften*, V&R unipress, Göttingen. Google-Books-ID: cok7EAAAQBAJ.

Peplow, David, Joan Swann, Paola Trimarco, and Sara Whiteley, editors (2015), *The Discourse of Reading Groups: Integrating Cognitive and Sociocultural Perspectives*, Routledge, London. Google-Books-ID: GJz4CgAAQBAJ.

Pianzola, Federico, Simone Rebora, and Gerhard Lauer (2020), Wattpad as a Resource for Literary Studies. Quantitative and Qualitative Examples of the Importance of Digital Social Reading and Readers' Comments in the Margins, *PLOS ONE* **15** (1), pp. 1–46, Public Library of Science. https://doi.org/10.1371/journal.pone.0226708.

Pontiki, Maria, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit (2016), SemEval-2016 task 5: Aspect based sentiment analysis, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, Association for Computational Linguistics, San Diego, California, pp. 19–30. https://aclanthology.org/S16-1002.

Sapiro, Gisèle (2016), The Metamorphosis of Modes of Consecration in the Literary Field: Academies, Literary Prizes, Festivals, *Poetics* **59**, pp. 5–19. https://www.sciencedirect.com/science/article/pii/S0304422X16000103.

Schmidt, Thomas, Katrin Dennerlein, and Christian Wolff (2021), Towards a Corpus of Historical German Plays with Emotion Annotations, *in* Gromann, Dagmar, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch, editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, Vol. 93 of *Open Access Series in Informatics (OASIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, pp. 9:1–9:11. https://drops.dagstuhl.de/opus/volltexte/2021/14545.

Schneider, Ute (2018), *Bücher zeigen und Leseatmospären inszenieren. Vom Habitus enthusiastischer Leserinnen und Leser.*, München: Edition Text+Kritik, pp. 113–123.

Stinson, Emmett (2016), How Nice is too Nice? Australian Book Reviews and the 'Compliment Sandwich', *Australian Humanities Review* **60**, pp. 108–126.

Thomalla, Erika (2018), *Bücheremphase: Populäre Literaturkritik und Social Reading im Netz.*, München: Edition Text+Kritik, pp. 124–136.

Tulkens, Stéphan and Andreas van Cranenburgh (2020), Embarrassingly simple unsupervised aspect extraction, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 3182–3187.

Van Hee, Cynthia, Orphee De Clercq, and Veronique Hoste (2021), Exploring Implicit Sentiment Evoked by Fine-grained News Events, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, Online, pp. 138–148. https://aclanthology.org/2021.wassa-1.15.

Villaneau, J., Stefania Pecore, and Farida Saïd (2018), Aspect Detection in Book Reviews: Experimentations, *NL4AI@AI*IA*.

Walsh, Melanie and Maria Antoniak (2021), The Goodreads "Classics": A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism, *Journal of Cultural Analytics* **4**, pp. 243–287.

Xu, Hu, Lei Shu, Philip Yu, and Bing Liu (2020), Understanding Pre-trained BERT for Aspect-based Sentiment Analysis, *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 244–250. https://aclanthology.org/2020.coling-main.21.