Exploring aspect-based sentiment analysis methodologies for literary-historical research purposes

Tess Dejaeghere[1,4], Pranaydeep Singh[4], Els Lefever[1,4], Julie M. Birkholz[1,2,3]
[1]GhentCDH, Ghent Center for Digital Humanities, Ghent University
[2]Ghent University Department of History
[3]KBR, Royal Library of Belgium, Brussels, Belgium
[4]LT3 - Ghent University Department of Translation, Interpreting and Communication
{tess.dejaeghere, pranaydeep.singh, els.lefever, julie.birkholz}@ugent.be

Abstract

This study explores and compares aspect-based sentiment analysis (ABSA) methodologies for literary-historical research, aiming to overcome the limitations of traditional sentiment analysis in capturing the nuanced aspects of literature. Through the analysis of an English corpus of 19th and 20th-century travelogues, the study develops annotation guidelines and evaluates three ABSA toolchains: a rule-based system, a machine learning-based approach based on both BERT and MacBERTh embeddings, and a prompt-based workflow using the open-source generative large language model Mixtral 8x7B. Findings reveal insights into the challenges and potentials of ABSA methodologies for literary-historical analysis, highlighting the need for context-aware annotation strategies, required technical skills and time investment. The research contributes to the following: (1) the curation of a multilingual corpus comprising 3078 travelogues sourced from online repositories in German, English, French, and Dutch; (2) the publication of an annotated multilingual literary-historical dataset of travelogues for aspect-based sentiment analysis, focusing specifically on environment-related aspects and their associated sentiment scores; (3) creation of openly available and adaptable Jupyter Notebooks with the Python code developed for each modelling approach; (4) publication of pilot experiments for ABSA on literary-historical texts using the English subset of the dataset; and (5) formulation of future endeavors aimed at advancing ABSA methodologies within the realm of literary-historical research.

Keywords: aspect-based sentiment analysis, travelogues, methodology

1. Introduction

The influx of Natural Language Processing (NLP) methodologies in literary-historical research settings remains limited to date (Blevins and Robichaud, 2011; Kuhn, 2019; Kuhn and Reiter, 2015; McGillivray et al., 2020; Suissa et al., 2022). Sentiment analysis (SA) in particular, a popular text mining approach to automatically categorize textual entities as positive, neutral or negative, is critically regarded in literary studies, and often deemed inept to cater to the meticulous research needs of humanist researchers (Kim and Klinger, 2018b; Schmidt and Burghardt, 2018). The reasons for this critique stem largely from the fact that literary analysis can hardly be fit to the inflexible polarity scheme ("positive", "neutral" and "negative") employed by contemporary SA-tools (Buechel et al., 2016; Kim and Klinger, 2018a,b; Kim, 2022; Schmidt and Burghardt, 2018). As a consequence, the application of sentiment analysis in (digital) humanities remains under-explored, and historians and literary scholars are eventually nudged back to a familiar praxis of close reading and manual analysis (Kim and Klinger, 2018b,a; Kuhn, 2019).

1.1. Aspect-based sentiment analysis

To account for the rigid nature of SA-tools, aspect-based sentiment analysis (ABSA) has steadily gained traction. Rather than procuring a polarity label on the level of the document, paragraph or sentence, ABSA systems operate on the aspect-level by combining multiple information extraction subtasks to extract 1) aspect terms 2) aspect categories, 3) opinion terms and 4) sentiment polarities (Birjali et al., 2021; Zhang et al., 2022).

While ABSA is an up-and-coming area of research in NLP, and opening up promising avenues and levels of granularity for sentiment mining, its application currently remains limited to commercial domains such as customer reviews (Zhang et al., 2022). To the knowledge of the authors, the application of ABSA has thus far not been explored for literary-historical textual material.

Unsurprisingly so, perhaps, given that affective patterns in literature often deliberately transcend conventional linguistic structures to translate the enigmatic realm of the intimate human experience (Rebora, 2023). Furthermore, NLP tools are known for introducing their very own implicit (sentiment) theories and biases, contributing an additional stratum of opacity to their application. The rapid advancement of NLP-tools from explainable rule-based systems to models which try to capture abstractions of

LT4HALA 2024@LREC-COLING-2024, pages 129–143

human reasoning further complicates its use in literary analysis contexts, where a demarcation of perspective is paramount. Consequently, a cross-pollination of practices between the two fields is further dwindling, requiring an increasingly intricate set of computational skills and knowledge to build methodological bridges and foster mutual understanding (McGillivray et al., 2020; Rebora, 2023).

The current divide raises the question of whether ABSA as a technique could be a way to circumvent the rigidity of conventional SA-models - granting a more fine-grained and explainable perspective on aspect representation and sentiment expression in literary text. Answering to the calls for exploratory research and evaluation of NLP-based methods, this study presents a pilot endeavor to test a number of ABSA methodologies for literary-historical research contexts (Rebora, 2023).

2. Related work

2.1. Aspect-based sentiment analysis in computational literary studies

In contemporary settings, ABSA is often used in the context of e-commerce to achieve a better understanding of public opinion towards specific aspects of their offered services and products, or to analyze opinions expressed on social media platforms (Mowlaei et al., 2020; D'Aniello et al., 2022; Troya et al., 2022). While sentiment categories are usually constrained to a five- or three-point scale – previous work explored fine-grained emotion categories tied to an aspect to improve customer relation management (De Geyndt et al., 2022).

Literature on ABSA is characterized by its scattered nature, and the scientific terminology employed to delineate this task lacks uniformity. While "aspect-based sentiment analysis" is largely accepted as the standard nomenclature – the task has been referred to as ACOD (aspect-category-opinion-sentiment quadruple extraction), TOWE (target-oriented opinion word extraction) (Xu et al., 2020), ELSA (entity-level sentiment analysis) (Rønningstad et al., 2023), TSA (targeted sentiment analysis) (Zhang et al., 2016), ASAP (aspect category sentiment analysis and rating prediction) (Bu et al., 2021) among a myriad of other denominations. Indeed, "the terminologies of ABSA studies are often used interchangeably, but sometimes they have different meanings according to the context [...] This may cause unnecessary confusion and often makes the literature review incomplete (Zhang et al., 2022).". Next to the mere terminological nature of this debate – what is defined as an aspect and a sentiment differs across applications and "must be treated using completely different approaches as they lead to different kind of results" (D'Aniello et al., 2022). While this fuzzy use of terminology is likely not the primary impediment to the adoption of this technique in DH settings – it may further obscure the definition of the methodology itself and the necessary distinct subtasks involved, posing an additional hurdle for scholars less familiar with NLP jargon when attempting to integrate this technique or assess its application range.

Depending on the desired output, different learning strategies are combined for the aspect recognition and sentiment analysis subtasks respectively – ranging from unsupervised (e.g.: frequency, statistics, heuristics, dependency parsing, rule-based approaches, zero-shot classification or topic modelling, etc.), semi-supervised (e.g.: lexicons and lexicon generation, dependency trees or knowledge graphs, etc.), and supervised (e.g.: machine learning, decision trees, neural networks, etc.) strategies (Birjali et al., 2021; D'Aniello et al., 2022; Keshavarz and Abadeh, 2017; Pattakos, 2021; Xu et al., 2021; Zhang et al., 2022). In more recent work, the power of generative language models for zero-shot and few-shot classification were also explored (Hosseini-Asl et al., 2022; Pangrazzi, 2022; Vector Institute, 2023).

Considering the traction gained by tasks such as Named entity recognition (NER), relation extraction (REX) and sentiment analysis (SA) in humanist research (Al-Razgan et al., 2021; Arnoult et al., 2021; Gamallo and Garcia, 2019; Jänicke et al., 2017; Li, 2022; Neudecker, 2016; Pineda et al., 2020; Todorov and Colavizza, 2020; Won et al., 2018) – it is but a small step to envision the potential of ABSA, which amalgamates the capabilities of these individual techniques. Apart from recent work which compares the application of ChatGPT to an in-house fine-tuned BERT architecture applied to a set of literary reviews (Martens et al., 2023) – ABSA has not yet been applied within the domain of computational literary studies.

While positing ABSA as a panacea would be a gross exaggeration, trying new methodologies to assess the applicability of NLP in DH practice is paramount. Rather than presenting a full-fledged solution, this study aims to answer to the calls for an exploratory approach in NLP-infused literary analysis methodologies, guided by the principle that "a criticism of the tools and methods currently adopted in sentiment analysis is as necessary as a free exploration of its potential (Rebora, 2023)".

2.2. Annotation and evaluation

While the annotation process is widely considered essential for the development and evaluation of information extraction tasks, literary texts are known to be extraordinarily tedious and difficult to annotate due to their subjective nature and stylistic properties (Kleymann and Stange, 2021; Ivanova et al., 2022; Ehrmann et al., 2021). Figurative language such as metaphors, personification and metonymy; stylistic and language-specific peculiarities across authors' works and the specific research needs of literary scholars and historians hamper a standardisation of annotation practices across the entire literary domain (Bamman et al., 2019). Additionally, the historical variety space in which a text resides further obfuscates its interpretation and, therefore, the annotation process for targeted information extraction tasks (Plank, 2022).

Despite previous attempts at the creation of annotated datasets and annotation frameworks for NER within the domain of English literature by for example LitBank (Bamman et al., 2019) and the calls for targeted approaches and "agreed-upon annotation guidelines to be used for the annotation of literary novels (Ivanova et al., 2022)" - the highly individual text analysis needs of literary scholars and historians require a more flexible approach (D'Aniello et al., 2022; Jacobs, 2019; McGillivray et al., 2020).

Regarding evaluation, utilizing or merging existing datasets to serve as a benchmark representative of the "literary data" domain has not yielded fruitful results. Because of the wide variety of annotation practices and the diverse characteristics featured across these test sets, using different partitions of the gold standard annotations may lead to vastly different evaluation outputs (Ivanova et al., 2022). Additionally, NLP-native evaluation metrics such as accuracy and F1 scores often do not cater to the meticulous evaluation practices in the humanities - thus making annotation and evaluation "[...] all the more challenging as the scope of needs and applications in humanities research is much broader than the one usually addressed in modern NLP (Ehrmann et al., 2021)" (Klinger et al., 2020; Rebora, 2023).

3. Methodology

3.1. Travelogues as data

As a use-case to test these methodologies, attention is geared towards the application of ABSA to a textual corpus of travelogues from the 19th and 20th centuries.

Travelogues are an extraordinarily interesting source in this respect - as they constitute an idiosyncratic lens on the author's travel experiences - thus granting readers an intimate glimpse into the writer's identity and views on their surroundings (Colletta et al., 2015; José and Joseph Parathara, 2018; Sprugnoli, 2018). Leveraging this unique characteristic, the study zooms in on a set of aspects related to the environment as perceived and documented in the travelogue. Not only standard aspects such as people, locations, organizations are annotated, but we further enriched the data with aspect annotations related to weather phenomena, natural landforms, human landforms, biomes, fauna, and flora. While beyond the current study's scope, the resulting open-source dataset could serve as a catalyst to foster a more profound understanding of the historical value attributed to nature through literary analysis, or as a benchmark dataset for future ABSA methodologies in the literary-historical domain (Virdis, 2023; Correia et al., 2021; Langer et al., 2021; van Erp et al., 2018).

1. **Dataset collection**: as a first step, the collection of a multilingual corpus comprising of 3078 non-fictional travelogues from the 19th to the 20th century in English, French, Dutch and German from a range of online repositories is described.

2. **The development of annotation guidelines** tailored to the annotation of aspects and sentiments in travelogues is explained, as well as the selection of annotators. As a proof of concept, a subset of the corpus consisting of 58 texts across languages is subjected to annotation according to these guidelines by three trained student annotators.

3.2. ABSA pipeline development

The development and evaluation of three ABSA-pipelines, one supervised system and two unsupervised systems, is further detailed.

1. **A rule-based system** is developed for 1) aspect extraction based on spaCy's noun extraction module, 2) opinion word identification using spaCy's POS-tagger to extract adjectives, adverbs and auxiliary constructions and 3) sentiment analysis based on the extracted opinion words using the SenticNet package. In the case of negated sentiment words, NLTK's synset module was used to fetch the word's antonym and generate a score (Loper and Bird, 2002; Cambria et al., 2020; Montani et al., 2023).

2. **A machine learning-based** pipeline is developed in two steps. The aspect extraction task is tackled by training two Flair-based

sequence taggers on the annotations. One of the sequence taggers is based on BERT embeddings, while the other is trained using MacBERTh embeddings. Their performances are evaluated on the gold standard aspects using 5-fold cross-validation, and compared. For the sentiment analysis task, BERT and MacBERTh models were fine-tuned on the gold standard aspects. These embeddings subsequently serve as input for diverse machine learning classification architectures, including SVM, AdaBoost, Random Forest, and MLP classifiers (Devlin et al., 2019; Manjavacas Arevalo and Fonteyn, 2021; Greve et al., 2021).

3. **A prompt-based zero-shot workflow** using the multilingual generative Large Language Model Mixtral-8x7B-Instruct-v0.1 is developed. Experiments with prompts, parameter settings and output parsing steps are discussed for the aspect and sentiment extraction tasks respectively (Jiang et al., 2024).

Our developed methodologies are compared in terms of time investment, required expertise, and the level of transparency and usability for humanist research purposes. The final evaluation is conducted from a methodological point of view, and not geared towards the improvement or comparison of model performances. Furthermore, we evaluate the suitability of the ABSA approaches for the literary-historical domain and propose directions for future research.

4. Results and discussion

4.1. Data gathering

The travelogues feature diverse genres such as nature writing, travel memoirs, journals, and poetry. It must also be acknowledged that a non-fictional nature of these texts cannot be fully assumed – as these stories are often, though not always, a concoction of fact and fiction. The documents were sourced from various online repositories as outlined below, and resulted in a dataset of 3,320 texts across the languages English, French, Dutch and German as shown in Table 1. Opposite to the other collections, the texts gathered from the Biodiversity Heritage Library as well as those fetched from the Travelogues project included OCR-related mistakes. Using the garbageness score as a quality filter, the most extreme cases were filtered out (Ryan, 2015).

1. Travel-related texts from the Biodiversity Heritage Library[1] were scraped via API using travel-related terms and primarily feature non-fictional travel reports by biologists and naturalists .

2. The subcollection sourced from DBNL (Digitale Bibliotheek voor Nederlandse Letteren) [2] consists mainly of Dutch stories and reports on colonial explorations by Dutch-speaking settlers.

3. Italian travel reports comprise narratives about Italy written by English authors in the 1930s (Sprugnoli, 2017).

4. The Arctic Travellers dataset was manually collected from the Internet Archive[3].

5. Non-fictional travel reports were gathered from Project Gutenberg[4].

6. A set of German travelogues from the Travelogues project, available for download on their GitHub repository, were automatically compiled by domain experts (Rörden et al., 2020)[5].

Language	18thC	19thC	20thC	Total
English	41	782	668	**1,491**
French	5	145	50	**200**
Dutch	25	92	242	**359**
German	972	218	80	**1,270**
Total	**1,043**	**1,163**	**897**	**3,320**

Table 1: Overview of languages contained in the travelogues corpus (approx. 5,000 tokens/text)

Finally, 58 texts were annotated across all the languages present in the corpus (English, French, Dutch and German) using the platform INCEPTION (Klie et al., 2018). As a proof of concept, this work focuses on the English subset of this gold standard data. This is a subset of 22 texts. After training the students to use the annotation platform and the annotation guidelines, 14 texts of approximately 500 tokens each were annotated by all annotators to calculate the inter-annotator agreement (Fleiss' kappa score) for the aspect categories and the sentiment annotation on both aspect and sentence levels as shown in Table 2. Interestingly, these results indicate that students found it more difficult to annotate sentiment on the level of the sentence than on the level of the aspect. This may be because it is simply harder to assess the sentimental value of an entire sentence. While the Kappa score for the aspect categories PERSON, LOCATION,

[1]https://www.biodiversitylibrary.org/

[2]https://www.dbnl.org/

[3]https://www.archive.org/

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

[5]https://www.travelogues-project.info/

ORGANISATION, FAUNA, FLORA, BIOME, HUMAN_LANDFORM, NATURAL_LANDFORM, NATURAL_PHENOMENON, WEATHER, MYTH and BIOME was quite high, categorization of these aspects is not the focus of this work.

Annotation	Kappa
Aspect category	0.88
Sentiment (aspect)	0.64
Sentiment (sentence)	0.24

Table 2: Overview of the inter-annotator agreement Fleiss' Kappa scores across sentiment and aspect annotations for English

4.2. Annotation process

Student annotators were chosen based on their language proficiency across the languages featured in the corpus. The students were working on studies in history or multilingual communication. At all times, with the exception of the annotations used to calculate the IAA, the students were allowed to engage in discussions with one another to foster an exchange of historical and linguistic expertise. The texts' metadata was released to the students and included information on release dates, full titles and authors, allowing them to look up more contextual information if needed. Discussions regarding recurring ambiguous aspect categories often spontaneously took on a rather philosophical nature (e.g.: should we indicate "God" as a PERSON or MYTH aspect?), and decisions were gradually adjusted depending on the cases encountered. Metaphors also regularly surfaced (e.g.: "Eternal City" as a denomination for "Rome") and annotated.

Because we attempt to model the readers' evaluative response to the text rather than the intended sentiment value of the author, the students were asked to annotate sentiment based on their own affective evaluation of the text. Given the unpredictable shape of literary text, the only rule implemented to distinguish between the extreme categories 1 and 5 was the presence of intensifiers in the chunk (adverbs such as "very" or "extremely"). It quickly became evident during our discussions that the five-point scale for sentiment introduced too much ambiguity, and during the modelling phase it was decided to compress the categories to a three-point scale and compare performances. Examples of ambiguous cases are legion and their thorough discussion could easily be the subject of separate research efforts. One example is shown in Figure 1, where a colonial traveller discusses an encounter with the indigenous Indian population and refers to them in his travelogue by describing them as "civilised". Our annotator deemed this a positive expression connected to the aspect "Indians", but given the colonial context in which this text was written, the need of the author to explicitly mention the "civilised" nature of these people expresses a level of surprise, harbouring a condescending and thus negative depiction of the aspect indians through a contemporary reader's lens.

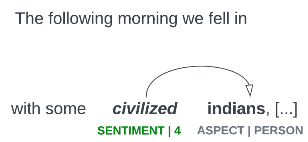

Figure 1: Example of ambiguous sentiment annotation.

Another example depicted in Figure 2 showcases the layered sentimental expression often present in literary sentences, and underlines the usefulness of ABSA as a fine-grained methodology.

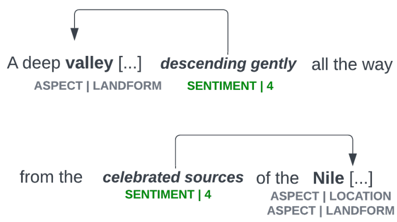

Figure 2: Example of the annotation of layered sentimental expression in a single sentence.

4.3. Aspect extraction

Our annotations were converted to a BIO-format, and the output of our aspect extraction models was evaluated using the nervaluate package[6] and a strict macro F1 approach and shown in Table 3.

4.3.1. Unsupervised approaches

Our rule-based system constituted a simple approach which follows the notion of noun chunks as optimal aspect candidates, while adjectives and adverbs serve as potential opinion words - as suggested by previous work (Anwar et al., 2023; Nandhini et al., 2018; Mai and Zhang, 2020; Nandhini et al., 2018; Anwar et al., 2023). SpaCy was used to extract nouns and proper nouns from the noun chunks which were then converted to BIO-format and evaluated against the annotations. The discrepancy of our annotations and this rather one-dimensional approach is reflected in the low strict

[6]https://pypi.org/project/nervaluate/

F1 scores (0.20). A manual analysis of the errors showed that the rule-based system's mistakes are logically mostly due to extraction of irrelevant nouns as aspects (e.g.: "Sunday", "a brief visit", "lower end", "ugliness", "unusual distance") which are not part of the categories under consideration. Conversely, in some cases, the approach revealed entities that were missed by the annotators.

The other unsupervised system using the generative Mixtral-8x7B-Instruct-v0.1 model, was implemented through the LangChain development framework as a zero-shot approach (Harrison, 2022). Being a recently developed technology at the time of writing, pitfalls and strengths of these generative LLMs across domains are yet to be discovered. The biggest challenge for a digital humanist to overcome here is not necessarily producing the code itself, but finding the correct way of constructing a prompt of which the output can be consistently parsed while retaining awareness of the model's inherent bias and tendencies to hallucinate. Using a development set of annotated samples as input texts, we experimented with the temperature setting, which was eventually set to the low value of 0.01 as this intuitively renders the least convoluted results. To make the output easy to parse, we designed a JSON output schema as the example shown in Figure 5 and asked the model to generate the output accordingly. Without this structural element, the model's output was unstructured and consequently impossible to parse consistently.

Categories were added as context information as a string object, and included a short definition for each category between brackets. The input sentence was indicated in the prompt using designated symbols to ensure the model relies solely on the input sentence to construct an answer. The finalized prompt is shown in Figure 3. A clear task description ("Extract the relevant named entities from the given sentence") was used as input. Upon experimentation, it became clear that the model produced better results when asked to extract "named entities" as compared to "aspects". It was noted that in both cases, the model extracted common names, personal pronouns ("he", "her") as well as proper names, which may need to be tweaked through the prompt depending on the use-case. Interestingly, the concept of "location" was quite literally interpreted by the model, and snippets such as "convenient places", "over there" and "the latter place" were also extracted. We experimented with adding a personality to the model (e.g.: "You are a historian and literary scholar with expertise on historical travel literature"). Interestingly, adding this feature sometimes caused the model to add an unrequested lengthy explanation about its reasoning, a feature which could be useful for humanists to adjust their prompting techniques and decide on

```
question: "Extract the relevant named entities from the given sentence."
template: """Your task is to identify the named entities in a sentence.
Named entities include {categories}.
Structure the answer according to {schema_entity}.
The sentence is indicated by <<<>>>.
Question: {question}
Sentence: <<<{sentence}>>>


Answer: """
```

Figure 3: Prompt for aspect extraction

which contextual information and examples to add in a few-shot setting, allowing for an adequate and intuitive human-in-the-loop setting.

While the results of this approach on a held-out test set are not high (0.34 F1), the output was still impressive considering the limited contextual information that was given in the prompt, and warrants further research in this domain.

4.3.2. Supervised approach

Our machine-learning based aspect extraction model was made using Flair's SequenceTagger module, and evaluated through 5-fold cross-validation on equal splits of the data. BERT- and MacBERth-embeddings were used respectively to train two different taggers. Surprisingly, the BERT embeddings in this case rendered a better macro F1 score (0.62) and trumped the MacBERTh embeddings made for historical English (0.59). The code for this operation was easy to retrieve and adapt through Flair's documentation, but does require a basic understanding of embeddings and parameter settings.

Unsupervised models	**F1**
Rule-based system	0.20
Mixtral 8x7b	0.34
Supervised models	
SQT Flair BERT	**0.62**
SQT Flair MacBERTh	0.59

Table 3: Overview of scores for the English aspect extraction models on a test set

4.4. Sentiment analysis

4.4.1. Unsupervised approaches

For the rule-based system, we wanted to evaluate the system on the text snippets that were labelled with a sentiment and connected to an aspect in the gold standard data. Thus, we had to look for language-specific tools that were able to output a sentiment score based on a given text chunk. For English, luckily, quite a few lexicon-based tools are available for sentiment analysis. Eventually, the tool Senticnet was applied and evaluated on the opinion words in the annotated data (Cambria et al., 2020). This tool was chosen for its ease of use and transparency in terms of the used emotion ontology and polarity scoring principles. Using a sigmoid function, the resulting float scores returned by Senticnet $\in [-1:1]$ were normalized into a [0:1] float range for each sentiment-bearing word. The final "sentiment score" is the mean of the scores for each word. After that, a threshold was determined and linked to a respective sentiment label (if the mean score is equal to or less than 0.20, the sentiment label is 1; if the score is equal to or similar to 0.40, the sentiment label is 2 and so forth) to match the range of the annotations. Negations occurring in the noun phrases (e.g.: "*not* beautiful") were addressed by finding the antonym of the negated word using NLTK's synsets module - and then applying SenticNet to the fetched antonym (Loper and Bird, 2002). Intensifiers, given that these were explicitly mentioned in the annotation guidelines and are thus expected to influence the annotations, were considered by checking whether an adverb is present in the noun phrase, and pushing the mean score into category 1 if it's below or equal to the 0.50 threshold, or 5 if it's above. As shown in 4, the system consistently performed better when compressing the scoring system in the 1-3 range. The packages used, while multilingual, are not tailored to historical language, which was not a serious shortcoming for English, but undoubtedly would be in the case of lesser-resourced (historical) languages, which makes this approach less advisable in most cases.

The prompt for the Mixtral 8x7b was constructed much in the same way as that of the aspect extraction. A clear indication of the sentence under consideration and an expected output structure as shown in Figure 5 was confirmed to be really important. Here, too, the personality addition ("you are a historian") made for a more convoluted output and produced a string of reasoning, which made the output unpredictable and difficult to parse, but was interesting to further scrutinize. In one example, the aspect "officers" in the sentence "[...] he, accordingly to a plan long since proposed , formed the Indians into Companies and by degrees taught them to feel the convenience of having officers set apart to each , which they were soon not only reconciled to but highly pleased with , by which means he gave some degree of method and form to the most Independent race of the Indians [...]", was positively evaluated, because, according to the model: "The sentence expresses that the officers were able to teach the Indians to feel the convenience of having officers set apart to each, which they were soon not only reconciled to but highly pleased with. This implies that the officers were able to positively influence the Indians and make them feel more organized and structured.", echoing a contextless and historically unnuanced assessment of the text material which may be considered dangerously biased in research contexts.

```
question: "Is this aspect very positive, positive, neutral,
negative or very negative in the sentence?"

template: """Your task is to identify the sentiment of an aspect
in the categories "very positive", "positive", "neutral", "negative"
or "very negative".
Sentences are only very positive or very negative if an
intensifier such as "very" or "extremely" is present.
The sentence is indicated by <<<>>>;
The aspect you have to evaluate is indicated by <<>>.
Structure the answer according to {schema}.

Question: {question}
Sentence: <<<{sentence}>>>
Aspect: <<{aspect}>>
Answer: """
```

Figure 4: Prompt for sentiment analysis using Mixtral 8x7b

```
schema = {
    "properties": {
        "sentiment": ["response"],
        },
    }
```

Figure 5: Output JSON schema for sentiment analysis using Mixtral 8x7b

Unsupervised models	F1
Rule-based system (1-5)	0.32
Rule-based system (1-3)	**0.37**
Mixtral 8x7b (1-5)	0.33
Mixtral 8x7b (1-3)	**0.42**

Table 4: Overview of unsupervised sentiment model scores for English

4.4.2. Supervised approaches

Our approach was adapted from previous work by Greve et al. (2021), which trained embeddings using BERT and used them as features in machine learning models to differentiate between positive and negative literary reviews. BERT and MacBERTh embeddings were trained for sentiment labels on a 1-5 point scale and labels on a 1-3 point scale respectively, and used as input for a variety of ML-models (SVM, MLP, RF and AdaBoost). The MLP classifier, a Multi-Layer Perceptron classifier, consistently outperformed the other networks as shown in Table 5.

Embeddings	Model	F1
BERT (1-5)	SVM	0.53
	MLP	**0.56**
	RF	0.49
	AdaBoost	0.42
MacBERTh (1-5)	SVM	0.55
	MLP	**0.57**
	RF	0.49
	AdaBoost	0.43
BERT (1-3)	SVM	0.60
	MLP	**0.61**
	RF	0.50
	AdaBoost	0.50
MacBERTh (1-3)	SVM	0.57
	MLP	**0.62**
	RF	0.51
	AdaBoost	0.49

Table 5: Overview of supervised sentiment model scores for English

4.5. Qualitative comparison of methodologies

Designing the rule-based model was a time-consuming process, and requires not only thorough knowledge of the content of the data, but also of the linguistic manifestation of sought-after information. While most corpora for literary-historical use-cases are indeed limited in size, nouns phrases are, as expected, unfit to uncover complicated literary vehicles such as metaphors and simile, which may skew results. Additionally, sentiment lexica and tools for historical vernaculars were hard to find for the English language domain, let alone for other lesser-resourced languages, which would be a considerable impediment for developing a rule-based system in most DH research settings. However, the transparency of this white-box approach does grant the user a sense of control over the output, and does not require a thorough knowledge of modelling practices. Summarized, this approach seems advisable in the case of small corpora and cases where the grammatical structure of the aspects to be extracted is known and relevant to the use-case, or where sentiments are expressed using predictable words and formulae.

In the case of the generative model Mixtral, it was noted that the model sometimes had a tendency to hallucinate aspect categories that were not given in the prompt. Depending on how the prompt is formulated, the output included unrequested information beyond defined aspects or sentiment categories, and was sometimes unpredictable in shape and thus difficult to parse. Additionally, how a sentiment value is calculated exactly based on the input sentence is not clear, and one must keep into account that even this output may be no more than a model's best guess. From a technical point of view, this approach is quickly gaining traction at the time of writing, as many new open- and closed-source models and prompting techniques are being developed. This oversupply could make it challenging for the humanist researcher to find a fitting and well-documented generative open-source approach for a specific use-case. Multiple existing frameworks and models are currently behind a pay-wall, which raises questions regarding the privacy of research output and impedes widespread use. The open-source models through HuggingFace have installed a limit on server requests, which should be taken into account when planning to apply this methodology to large datasets. Indeed, it is possible to use these models to produce output quite easily, even with a basic understanding of the inner workings of generative LLM and programming, which makes them an attractive option for information extraction, but a thorough evaluation of its output is advised.

Machine learning and deep learning approaches have been favoured in computational literary-historical settings in the last couple of years. Logically, fine-tuning these systems creates output which remains more faithful to the annotations than the rule-based or generative model-based approaches, making it a reliable methodology in this context. Adapting existing code or creating new systems requires at least basic background understanding of embeddings, Tensor operations and a meta-understanding of neural networks and modelling using the HuggingFace platform. For digital humanists without this knowledge of NLP-practices, adapting and implementing this code may be too time-consuming. Additionally, machine learning models are data-hungry, and the effort required to produce annotations and enable training may be disproportionate if its application will be limited to a small case-study.

4.6. Contributions

This study makes the following contributions that includes the sharing of annotated data to knowledge on these practices:

1. A novel multilingual dataset of 3,320 travelogues ranging from the 19Th until the 20Th centuries is gathered from a range of online sources and made public on our GitHub repository [7].

2. Insights are formulated on the creation of annotation guidelines and their application to the literary-historical domain.

3. The annotated subset of this dataset for aspect-based sentiment analysis in English, Dutch,

[7] https://github.com/TessDejaeghere/Travelogues

German and French as well as the annotation guidelines used are made open-source, encouraging reuse for further research endeavours in the domain of aspect-based sentiment analysis and literary-historical research on travelogues.

4. We introduce pioneering work on the assessment of aspect-based sentiment analysis methodologies for the domain of computational literary studies, ranging from white box rule-based techniques to state-of-the-art black box techniques using a generative LLM. The code developed for this research is made open-source in the form of annotated Jupyter Notebooks to facilitate adaptability and reuse by computational linguists and (digital) humanists alike.

5. Conclusion and future research

The research explored three methodologies for ABSA in literary-historical research contexts. First and foremost, it must be noted that annotating biodiversity in travelogues is a fully-fledged research project an sich. Annotating literary-historical texts for research purposes is exceptionally challenging, and as opposed to the breadth-oriented approach in contemporary NLP settings, Digital Humanists can hardly escape the depth-oriented strategy to cater to their meticulous needs. Rather than adopting an exploratory lens, it is advised to use these information extraction techniques for well-defined research ends, and the availability of sufficient data and time should warrant its development (Chun and Elkins, 2023). The methodologies assessed come with their unique set of advantages and disadvantages: in the absence of sufficient annotated material, a large corpus as a use-case or knowledge of NLP practices, machine learning approaches may oftentimes not merit the effort. Rule-based systems do not require this knowledge of NLP-techniques and may work well in settings where the aspect and the sentiment expressions follow strict and formulaic patterns, but are often time-consuming to create. Unlike the level of expertise required for ML approaches, the methodology involving prompting the generative LLM Mixtral 8x7b is fairly straightforward. However, one must tread with great care when applying this methodology for literary-historical research applications, as our experiments confirmed the tendency of these models to hallucinate unrequested information. Additionally, specifically in the case of sentiment analysis, it is unclear how the engine makes its assessment. At the time of writing, a myriad of new models are created on a daily basis, which makes choosing an adequate model rather challenging. Researchers should also be aware of privacy concerns when using closed models versus open-source models on their dataset. However, generative LLMs could present an exciting new way to answer the call for a grey-box human-in-the-loop approach, but further research is needed to explore pitfalls and possible evaluation schemes:

- Future research may delve into the implementation of ABSA within a case-study framework, juxtaposed with a manual methodology for comparison.
- Using the novel multilingual travelogues dataset annotated for ABSA presented in our research, we aim to gear our future efforts towards methodological research expansion in sentiment analysis, NER and ABSA across diverse linguistic and literary-historical contexts. Future research endeavors might be directed towards the development of novel evaluation methodologies that transcend the conventional metrics employed in NLP. Such inquiries could contemplate whether outputs divergent from gold standard data necessarily constitute inaccuracies, or if they offer alternative perspectives that could augment human assessment.
- Further exploration into generative models across varied contexts presents an intriguing avenue. This includes investigating the impact of bias and model hallucinations on information extraction tasks like ABSA, as well as experimenting with different prompting techniques, incorporating contextual information, and even diverse modalities. Such endeavors could establish the groundwork for a human-in-the-loop grey-box evaluation methodology, wherein researchers engage in dialogue with the corpus, assess output samples, and adapt prompts accordingly.

6. Bibliographical References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Muna Al-Razgan, Asma Alrowily, Rawan N. Al-Matham, Khulood M. Alghamdi, Maha Shaabi, and Lama Alssum. 2021. Using diffusion of innovation theory and sentiment analysis to analyze attitudes toward driving adoption by Saudi women. *Technology in Society*, 65:101558.

Muchamad Anwar, Dedy Trisanto, Ahmad Juniar, and Fitra Sase. 2023. Aspect-based Sentiment Analysis on Car Reviews Using SpaCy Dependency Parsing and VADER. *Advance Sustainable Science Engineering and Technology*, 5:0230109.

Sophie I. Arnoult, Lodewijk Petram, and Piek Vossen. 2021. Batavia asked for advice. Pretrained language models for Named Entity Recognition in historical texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 21–30, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Lidia Bocanegra Barbecho. 2023. Review: BERT for Humanists. *Reviews in Digital Humanities*, IV(4).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*, 1st ed edition. O'Reilly, Beijing ; Cambridge [Mass.]. OCLC: ocn301885973.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Cameron Blevins and Andrew Robichaud. 2011. 2: A Brief History » Tooling Up for Digital Humanities.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A Chinese Review Dataset Towards Aspect Category Sentiment Analysis and Rating Prediction. ArXiv:2103.06605 [cs].

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016. Feelings from the Past—Adapting affective lexicons for historical emotion analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 54–61, Osaka, Japan. The COLING 2016 Organizing Committee.

Lawrence Buell, Ursula K. Heise, and Karen Thornber. 2011. Literature and Environment. *Annual Review of Environment and Resources*, 36(1):417–440.

Jon Chun and Katherine Elkins. 2023. eXplainable AI with GPT4 for story analysis and generation: A novel framework for diachronic sentiment analysis. *International Journal of Digital Humanities*, 5(2):507–532.

L. Colletta, J. Buzard, C. Chard, C.E. Hornsby, L. Olcelli, S. Russell, N. Stanley-Price, J. Suh, and A. Thompson. 2015. *The Legacy of the Grand Tour: New Essays on Travel, Literature, and Culture*. Fairleigh Dickinson University Press.

Ricardo A. Correia, Richard Ladle, Ivan Jarić, Ana C. M. Malhado, John C. Mittermeier, Uri Roll, Andrea Soriano-Redondo, Diogo Veríssimo, Christoph Fink, Anna Hausmann, Jhonatan Guedes-Santos, Reut Vardi, and Enrico Di Minin. 2021. Digital data sources and methods for conservation culturomics. *Conservation Biology*, 35(2):398–411.

Michael Cronin. 2022. *Eco-Travel: Journeying in the Age of the Anthropocene*. Elements in Travel Writing. Cambridge University Press.

Ellen De Geyndt, Orphee De Clercq, Cynthia Van Hee, Els Lefever, Pranaydeep Singh, Olivier Parent, and Veronique Hoste. 2022. SentEMO: A multilingual adaptive platform for aspect-based sentiment and emotion analysis. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 51–61, Dublin, Ireland. Association for Computational Linguistics.

Giuseppe D'Aniello, Matteo Gaeta, and Ilaria La Rocca. 2022. KnowMIS-ABSA: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. *Artificial Intelligence Review*, 55(7):5543–5574.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named Entity Recognition and Classification on Historical Documents: A Survey. *arXiv:2109.11406 [cs]*. ArXiv: 2109.11406.

Martin Paul Eve. 2022. Distance and History. In Martin Paul Eve, editor, *The Digital Humanities and Literary Studies*, page 0. Oxford University Press.

Justyna Fruzińska. 2022. *Nineteenth-century visions of race: British travel writing about America*. Routledge studies in nineteenth century literature. Routledge, Taylor & Francis Group, New York.

Pablo Gamallo and Marcos Garcia. 2019. Editorial for the Special Issue on "Natural Language Processing and Text Mining". *Information*, 10(9):279.

Lore De Greve, Pranaydeep Singh, Cynthia Van Hee, Els Lefever, and Gunther Martens. 2021. Aspect-based Sentiment Analysis for German: Analyzing "Talk of Literature" Surrounding Literary Prizes on Social Media. *Computational Linguistics in the Netherlands Journal*, 11:85–104.

Jo Guldi. 2023. Why Textual Data from the Past Is Dangerous. In *The Dangerous Art of Text Mining: A Methodology for Digital History*, pages 25–56. Cambridge University Press, Cambridge.

Andrew J. Hansen, Ruth S. DeFries, and Woody Turner. 2012. Land Use Change and Biodiversity: A Synthesis of Rates and Consequences during the Period of Satellite Imagery. In Freek D. Van Der Meer, Garik Gutman, Anthony C. Janetos, Christopher O. Justice, Emilio F. Moran, John F. Mustard, Ronald R. Rindfuss, David Skole, Billy Lee Turner, and Mark A. Cochrane, editors, *Land Change Science*, volume 6, pages 277–299. Springer Netherlands, Dordrecht. Series Title: Remote Sensing and Digital Image Processing.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis. ArXiv:2204.05356 [cs].

Rositsa Ivanova, Marieke van Erp, and Sabrina Kirrane. 2022. Comparing Annotated Datasets for Named Entity Recognition in English Literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3788–3797, Marseille, France. European Language Resources Association.

Arthur M. Jacobs. 2019. Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. *Frontiers in Robotics and AI*, 6.

George José and Thomas Joseph Parathara. 2018. Travelogues: Concept and The Text. *International Journal of Creative Research Thoughts*, 6(2).

Dan Jurafsky and James H. Martin. 2023. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 3rd ed edition. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, N.J. OCLC: 213375806.

S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. 2017. Visual Text Analysis in Digital Humanities. *Computer Graphics Forum*, 36(6):226–250.

Selin Kesebir and Pelin Kesebir. 2017. A Growing Disconnection From Nature Is Evident in Cultural Products. *Perspectives on Psychological Science*, 12(2):258–269.

Hamidreza Keshavarz and Mohammad Saniee Abadeh. 2017. ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowledge-Based Systems*, 122:1–16.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744.

Evgeny Kim and Roman Klinger. 2018a. A survey on sentiment and emotion analysis for computational literary studies. *CoRR*, abs/1808.03137.

Evgeny Kim and Roman Klinger. 2018b. Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hoyeol Kim. 2022. Sentiment Analysis: Limits and Progress of the Syuzhet Package and Its Lexicons. *DHQ: Digital Humanities Quarterly:*, 16(2).

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021.

Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 2611–2624. Curran Associates, Inc.

Rabea Kleymann and Jan-Erik Stange. 2021. Towards Hermeneutic Visualization in Digital Literary Studies. *DHQ: Digital Humanities Quarterly*, 2(15).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Roman Klinger, Evgeny Kim, and Sebastian Padó. 2020. Emotion Analysis for Literary Studies: Corpus Creation and Computational Modelling. In *Emotion Analysis for Literary Studies: Corpus Creation and Computational Modelling*, pages 237–268. De Gruyter.

Jonas Kuhn. 2019. Computational text analysis within the Humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation*, 53(4):565–602.

Jonas Kuhn and Nils Reiter. 2015. *A Plea for a Method-Driven Agenda in the Digital Humanities* . Sydney, Australia.

Lars Langer, Manuel Burghardt, Roland Borgards, Katrin Böhning-Gaese, Ralf Seppelt, and Christian Wirth. 2021. The rise and fall of biodiversity in literature: A comprehensive quantification of historical changes in the use of vernacular labels for biological taxa in Western creative literature. *People and Nature*, 3(5):1093–1109.

Nicolas Le Guillarme and Wilfried Thuiller. 2022. TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3):625–641.

Jingxia Li. 2022. Emotion Expression in Modern Literary Appreciation: An Emotion-Based Analysis. *Frontiers in Psychology*, 13.

Churnjeet Mahn. 2016. Travel writing and sexuality : queering the genre. In Churnjeet Mahn, editor, *The Routledge companion to travel writing*, Routledge companions, pages 46–56. Routlege, Taylor & Francis Group, London ; New York.

Deon Mai and Wei Emma Zhang. 2020. Aspect Extraction Using Coreference Resolution and Unsupervised Filtering. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 124–129, Suzhou, China. Association for Computational Linguistics.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

Stephan Marche. 2012. Literature Is not Data: Against Digital Humanities. Section: culture.

Gunther Martens, Lore De Greve, and Pranaydeep Singh. 2023. A comparison of ChatGPT and fine-tuned BERT with regard to ABSA in the domain of literary criticism - Aspect-Based Sentiment Analysis for Literary Criticism: Experts vs. Social Critics on Literary Prizes.

Claudio Masolo, Emilio Sanfilippo, Marion Lamé, and Perrine Pittet. 2019. Modeling Concept Drift for Historical Research in the Digital Humanities. In *1st International Workshop on Ontologies for Digital Humanities and their Social Analysis (WODHSA)*, Graz, Austria.

Barbara McGillivray, Thierry Poibeau, and Pablo Ruiz Fabo. 2020. Digital Humanities and Natural Language Processing: Je t'aime... Moi non plus. *Digital Humanities Quarterly*, 014(2).

Julie Mennes, Ted Pedersen, and Els Lefever. 2019. Approaching terminological ambiguity in cross-disciplinary communication as a word sense induction task: a pilot study. *Language Resources and Evaluation*, 53(4):889–917.

MIT Global Shakespeare. 2023. Generative AI and the Digital Humanities: Pedagogical Implications.

Franco Moretti. 2014. "Operationalizing": or, the Function of Measurement in Modern Literary Theory. *The Journal of English Language and Literature*, 60(1):3–19.

Mohammad Erfan Mowlaei, Mohammad Saniee Abadeh, and Hamidreza Keshavarz. 2020. Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Systems with Applications*, 148:113234.

Trevor Muñoz and Katie Rawson. 2019. Against cleaning. In *Debates in the Digital Humanities 2019*, pages 279–292. University of Minnesota Press.

M. Devi Sri Nandhini, Pradeep Gurunathan, and Yuvaraj D. 2018. Exploring Aspect-Based Sentiment Analysis – A Survey.

Clemens Neudecker. 2016. An Open Corpus for Named Entity Recognition in Historic Newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4348–4352, Portorož, Slovenia. European Language Resources Association (ELRA).

Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. How We Do Things With Words: Analyzing Text as Social and Cultural Data. *Frontiers in Artificial Intelligence*, 3:62.

Fabian Offert and Peter Bell. 2020. Generative digital humanities. In *Workshop on Computational Humanities Research*.

Michele Pangrazzi. 2022. Building an aspect-based sentiment analysis pipeline using GPT-3.

Aris Pattakos. 2021. Aspect-Based Sentiment Analysis Using Spacy & TextBlob | by Aris Pattakos | Towards Data Science.

Murray G. Phillips and Gary Osmond. 2015. Australia's Women Surfers: History, Methodology and the Digital Humanities. *Australian Historical Studies*, 46(2):285–303.

Luis A. Pineda, Noé Hernández, Iván Torres, Gibrán Fuentes, and Nydia Pineda De Avila. 2020. Practical non-monotonic knowledge-base system for un-regimented domains: A Case-study in digital humanities. *Information Processing & Management*, 57(3):102214.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *CoRR*, abs/1608.07836.

Barbara Plank. 2022. The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. Publisher: arXiv Version Number: 1.

Gabrina Pounds. 2021. The values of trees and woodland: a discourse-based cross-disciplinary perspective on integrating 'revealed' evaluations of nature into environmental agendas. *Critical Discourse Studies*, 18(4):461–480. Publisher: Routledge _eprint: https://doi.org/10.1080/17405904.2020.1752757.

Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 017(2).

Sharon Richardson. 2022. Exposing the many biases in machine learning. *Business Information Review*, 39(3):82–89.

Aynat Rubinstein. 2019. Historical corpora meet the digital humanities: the Jerusalem Corpus of Emergent Modern Hebrew. *Language Resources and Evaluation*, 53(4):807–835.

Baumann Ryan. 2015. Automatic evaluation of OCR quality.

Jan Rörden, Doris Gruber, Martin Krickl, and Bernhard Haslhofer. 2020. Identifying Historical Travelogues in Large Text Corpora Using Machine Learning.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2023. Entity-Level Sentiment Analysis (ELSA): An exploratory task survey. Publisher: arXiv Version Number: 1.

Satvika, Vikas Thada, and Jaswinder Singh. 2021. A Contemporary Ensemble Aspect-based Opinion Mining Approach for Twitter Data. *International Journal of Advanced Computer Science and Applications*, 12(5).

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

Thomas Schmidt, M. Burghardt, and Christian Wolff. 2019a. Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti.

Thomas Schmidt and Manuel Burghardt. 2018. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. pages 139–149, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thomas Schmidt, Manuel Burghardt, and Christian Wolff. 2019b. Toward multimodal sentiment analysis of historic plays: A case study with text and audio for lessing's emilia galotti. In *Digital Humanities in the Nordic Countries Conference*.

Amina Siba, Rédda Aboura, Réda Kechairi, Mustapha Maatouk, and Bouthaina Sebbah. 2022. Diachronic study (2000-2019) of bioclimate and land use in Tlemcen region, Northwest Algeria. *International Journal of Environmental Studies*, 0(0):1–14.

Thomas Smits and Melvin Wevers. 2023. A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. *Digital Scholarship in the Humanities*, 38(3):1267–1280.

Rachele Sprugnoli. 2017. "Two days we have passed with the ancients...": a Digital Resource of Historical Travel Writings on Italy.

Sprugnoli, Rachele. 2018. *Arretium or Arezzo? A Neural Approach to the Identification of Place Names in Historical Texts*.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, Villains, and Victims, and GPT-3: Automated Extraction of Character Roles Without Training Data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. 2022. Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology*, 73(2):268–287.

Konstantin Todorov and Giovanni Colavizza. 2020. Transfer Learning for Historical Corpora: An Assessment on Post-OCR Correction and Named Entity Recognition. In *CHR*, CEUR Workshop Proceedings, pages 310–339, Amsterdam. Aachen: CEUR-WS.

Anina Troya, Reshmi Gopalakrishna Pillai, Dr. Cristian Rodriguez Rivero, Dr. Zulkuf Genc, Dr. Subhradeep Kayal, and Dogu Araci. 2022. Aspect-Based Sentiment Analysis of Social Media Data With Pre-Trained Language Models. In *Proceedings of the 2021 5th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '21, pages 8–17, New York, NY, USA. Association for Computing Machinery.

Marieke van Erp, Jesse de Does, Katrien Depuydt, Rob Lenders, and Thomas van Goethem. 2018. Slicing and Dicing a Newspaper Corpus for Historical Ecology Research. In Catherine Faron Zucker, Chiara Ghidini, Amedeo Napoli, and Yannick Toussaint, editors, *Knowledge Engineering and Knowledge Management*, volume 11313, pages 470–484. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Vector Institute. 2023. Aspect-Based Sentiment Analysis Using Prompt Engineering | Vector Applied Intern Project Talks.

Daniela Francesca Virdis. 2023. Ecostylistics: texts, methodologies and approaches. *Journal of World Languages*. Publisher: De Gruyter Mouton.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5:2.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning Span-Level Interactions for Aspect Sentiment Triplet Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4755–4766, Online. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-Aware Tagging for Aspect Sentiment Triplet Extraction. Publisher: arXiv Version Number: 3.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, pages 3087–3093.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. ArXiv:2203.01054 [cs].

7. Language Resource References

Language Resources

Bamman, David and Popat, Sejal and Shen, Sheng. 2019. *An annotated dataset of literary entities*. Association for Computational Linguistics. PID https://github.com/dbamman/litbank.

Cambria, Erik and Li, Yang and Xing, Frank Z. and Poria, Soujanya and Kwok, Kenneth. 2020. *SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis*. Association for Computing Machinery, CIKM '20. PID https://sentic.net/.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. PID https://huggingface.co/google-bert/bert-base-uncased.

Harrison, Chase. 2022. *Langchain*. PID https://python.langchain.com.

Albert Q. Jiang and Alexandre Sablayrolles and Antoine Roux and Arthur Mensch and Blanche

Savary and Chris Bamford and Devendra Singh Chaplot and Diego de las Casas and Emma Bou Hanna and Florian Bressand and Gianna Lengyel and Guillaume Bour and Guillaume Lample and Lélio Renard Lavaud and Lucile Saulnier and Marie-Anne Lachaux and Pierre Stock and Sandeep Subramanian and Sophia Yang and Szymon Antoniak and Teven Le Scao and Théophile Gervet and Thibaut Lavril and Thomas Wang and Timothée Lacroix and William El Sayed. 2024. *Mixtral of Experts*. PID https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Loper, Edward and Bird, Steven. 2002. *NLTK: The Natural Language Toolkit*. arXiv. PID https://www.nltk.org/.

Manjavacas Arevalo, Enrique and Fonteyn, Lauren. 2021. *MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)*. NLP Association of India (NLPAI). PID https://huggingface.co/emanjavacas/MacBERTh.

Ines Montani and Matthew Honnibal and Matthew Honnibal and Adriane Boyd and Sofie Van Landeghem and Henning Peters. 2023. *explosion/spaCy: v3.7.2: Fixes for APIs and requirements*. Zenodo. PID https://spacy.io/.

Rörden, Jan and Gruber, Doris and Krickl, Martin and Haslhofer, Bernhard. 2020. *Identifying Historical Travelogues in Large Text Corpora Using Machine Learning*. arXiv. PID https://github.com/travelogues/travelogues-corpus. ArXiv:2001.01673 [cs].