From Sentiment to Interpretation: Teaching NLP for Literary Understanding Across Educational Contexts

Karl-Emil Kjær Bilstrup[1,3], Kirstine Nielsen Degn[2], Morten Schultz[1], Alexander Conroy[2] Jens Bjerring-Hansen[2], Daniel Hershcovich[1],

[1]Computer Science, University of Copenhagen
[2]Nordic Studies and Linguistics, University of Copenhagen
[3]keb@di.ku.dk

Abstract

We developed Litteraturmaskinen, a graphical annotation and exploration interface that enables students to collaborate on labeling sentiment in literary passages, comparing their decisions with model predictions, and justifying their interpretations. We deployed the system in two educational settings: A university module on computational literary studies and regular teaching by two first-language high school teachers. Based on observations, collected teaching plans, and interviews, we find that tensions between epistemic and academic traditions are both a barrier for integration and a productive entry point for literary reflection and argumentation. We conclude with recommendations for integrating NLP into literature and first-language curricula.

1 Introduction

The rapid diffusion of language technologies based on large language models (LLMs) has transformed how students encounter text, interpretation, and analysis across disciplines. While natural language processing (NLP) has traditionally been taught within computer science curricula, recent years have seen growing interest in introducing NLP concepts to students in the humanities (Jockers, 2014), language studies (McEnery and Hardie, 2012), and secondary education (Berendt et al., 2020). This shift raises a central pedagogical challenge for NLP education: How to teach computational models of language in ways that respect and leverage the interpretive practices of humanistic disciplines, rather than reducing them to technical abstractions.

In this paper, we argue that sentiment analysis of literary texts offers a particularly effective entry point for teaching NLP to literature students. Although sentiment classification is often presented as a simple supervised learning task, applying it to literary language immediately exposes tensions around ambiguity, narrative perspective, historical context, and cultural specificity. These tensions are not obstacles to learning; rather, they provide fertile ground for developing computational thinking alongside critical reflection on what language models can and cannot capture.

Our work is situated in the context of Danish literary education, where students engage with historically and culturally embedded texts that differ markedly from the English-dominated, contemporary datasets typically used in NLP instruction. Literary texts resist stable labels, foreground interpretive plurality, and challenge assumptions about linguistic meaning as context-independent. When students are asked to annotate sentiment in such texts, they naturally begin to question whose sentiment is being modeled, how context is truncated, and why language models appear to "know" sentiment at all.

To support this pedagogical approach, we developed Litteraturmaskinen, a graphical user interface for text annotation and exploration that allows students to label sentiment in literary passages, compare their interpretations with model predictions, and generate visualizations of sentiment in larger text pieces and corpora. We deployed this system in two educational settings: a university-level module in computational literary studies and two first-language (L1) high school teachers' regular teaching.

Analyzing interviews with students and teachers, teaching plans and observation notes, we find that sentiment annotation in literature promoted both literary discussions and reflections on the opportunities of NLP technologies within literature analysis and teaching. The disagreement between human and model interpretations functioned as a productive teaching moment, encouraging students to articulate assumptions about language, meaning, and modeling. Last, the findings show how NLP methods conflict with epistemic traditions and academic identities among teachers and students, both

Proceedings of the Seventh Workshop on Teaching Natural Language Processing (TeachNLP 2026), pages 118–128

of whom questioned whether NLP methods were developing or diluting their academic practice. We conclude with practical recommendations for educators seeking to integrate NLP into literature and language curricula.

2 Related Work

Our contribution aligns with and extends prior work on teaching NLP through socially meaningful tasks, such as hate speech detection and societal empowerment (Cignarella et al., 2024; Yang et al., 2025). However, our focus differs in two important respects. First, rather than addressing contemporary social media language, we center on literary texts, where ambiguity and contested interpretation are intrinsic rather than exceptional. Second, rather than aiming primarily at awareness-raising or ethical positioning, we emphasize how literary analysis can serve as a theoretical and pedagogical resource for understanding NLP models themselves. Below, we account for existing research on teaching NLP in Upper-secondary education (high school) and higher education.

2.1 NLP in Upper-Secondary Education

Most applications of Natural Language Processing (NLP) in secondary education are centered on generative AI and focus on improving language proficiency and assessment in foreign and second language learning contexts. Recent studies predominantly focus on enhancing writing and reading comprehension (Xiao, 2025), facilitating vocabulary acquisition (Alsakaker, 2025), generating or tailoring student feedback (Zhang et al., 2025), and teachers' and students' attitudes towards the use of generative AI in education (DeVito et al., 2025). While establishing AI and NLP as powerful instruments for individualized instruction and skill development, this body of work has rarely been extended to students' interpretive or analytical engagement with literary texts *per se*, and least of all in L1 education.

Reviews on how AI is integrated in K-12 (kindergarten -to 12th grade) education show diverse examples of how learners can grasp core AI concepts through hands-on, collaborative, and project-based activities, but that there is a need for more curriculum-integrated designs (Martins and Gresse Von Wangenheim, 2022) and pedagogical frameworks to integrate AI literacy effectively (Ng et al., 2023). In a Danish context (same as this study), Bundgaard and Kalsgaard Møller (2024) show that students extensively use generative AI in homework and assignments alongside limited pedagogical guidance and calls for more research exploring different approaches to teaching and learning about the technologies as well as building the necessary literacies. Allred et al. (2024) explores a "social" approach to text annotation, in which students comment and respond to shared texts that support dialogic teaching in secondary English classes. The results showed that while structured teacher guidance was necessary, it promoted rich interactions, interpretive reasoning, and collaborative text engagement. Similarly, Connelly et al. (2025) design a web-application and paper-based activities, where students build small physical language models, to make NLP teaching more accessible and graspable in high school education. Bilstrup et al. (2025) co-design NLP activities with L1 teachers and demonstrate how NLP and L1 competencies interplayed with each other when NLP technologies and methods were sufficiently integrated into the subject.

2.2 NLP in Higher Education

Most instances of NLP in higher education similarly focus on improving academic writing, comprehension, and assessment. A substantial body of work addresses automated evaluation and feedback (Perelman, 2014; Zhang and Hyland, 2022). Other studies focus on examining AI tools as tutors and guidance mechanisms in self-regulated learning in higher education (Kasneci et al., 2023; Lee et al., 2024). While this research demonstrates the potential of NLP tools for scalable feedback and learner support, it typically frames students as end-users of AI systems – particularly LLMs – rather than as critical investigators of how such technologies function and how they shape knowledge production. In this regard, Southworth et al. (2023) argue that AI literacy initiatives at the university level often prioritize tool use over conceptual understanding, ethical reflection, and methodological critique. Moreover, these initiatives frequently remain confined to STEM disciplines, marginalizing perspectives from the social sciences and especially the humanities.

Within Digital Humanities (DH), increasing attention has been paid to the pedagogical implications of applying computational methods in humanities education. In general, DH pedagogy has conceptualized the digital both as a set of tools

and methods that support humanities research and teaching, and as an object of critical inquiry in its own right (Georgopoulou et al., 2025). However, several scholars note that DH teaching in higher education often emphasizes computational methods primarily as analytical tools, at the cost of critical engagement with NLP models themselves. This has prompted calls for a greater focus on Critical DH and Critical AI pedagogies, which foreground methodological reflexivity and engage students with NLP tools not as neutral instruments, but as cultural artifacts embedded in specific historical, social, and political contexts – and thus shaped by particular biases (Georgopoulou et al., 2025; Schneider and Oliveira, 2025; Roe et al., 2025).

Responding to these calls, we explore how both secondary and higher education can integrate data annotation activities and sentiment analysis into humanities teaching. We propose an approach that positions NLP methods and models not only as analytical tools but also as objects of critical inquiry that support interpretive literacy, methodological awareness, and reflective engagement with language technologies.

3 Theoretical Framing & Design Rationale

We provide a theoretical background and two design rationales for our educational tool and activities:

1. Sentiment Analysis as a Literary Task: Over the past two decades, sentiment analysis (SA) has become one of the most widely adopted methods in applied NLP, developing into an established research area with dedicated conference tracks and workshops (Kim and Klinger, 2021). Researchers have used SA to investigate a broad set of questions traditionally of interest to literary scholars, including character analysis (Vishnubhotla et al., 2024), narrative structure (Reagan et al., 2016), aesthetic positioning (Elkins, 2022), literary quality (Bizzoni et al., 2023), and canonicity (Degn et al., 2025). Literary meaning is shaped by different elements, such as focalization, narrator stance, tone, figurative language, and cultural context, all of which complicate the relationship between textual features and sentiment (Piper et al., 2021).

Engaging with SA requires students to draw on narrative theory, to examine how emotions are conveyed through subtle linguistic cues, and to recognize the culturally situated nature of sentiment in literary texts. For example, annotating sentiment compels students to consider narrative voice, focalization, and degrees of narrator reliability. It also prompts reflection on how genre conventions shape the expression and interpretation of emotion. By having students use their literary knowledge to annotate sentences from relevant texts, they experience how this simple annotation task is not trivial, but requires foundational literary competencies. They encounter the interpretive dimensions through the coarse-grained classifications used in sentiment analysis that foreground literature's complexity through their very inadequacy. The friction between computational categories and literary nuance becomes an instructive moment. In this way, sentiment classification renders visible the challenges of modeling interpretive judgment and emotional subtlety – competencies that lie at the core of both applied NLP and literary analysis.

2. How LMs Learn Sentiment from Text: Modern sentiment classifiers build on representations learned through language modeling, where meaning is acquired via distributional semantics and co-occurrence patterns: words and expressions that frequently appear in similar contexts are assigned similar representations (Mikolov et al., 2013). Because large language models are trained on corpora such as Wikipedia, books, and web text, they implicitly learn associations between lexical patterns and evaluative language, enabling downstream sentiment prediction (Devlin et al., 2019). There are also limitations of this approach. Language models operate on fixed context windows and lack explicit access to broader narrative structure (Levy et al., 2024), cultural background (Hershcovich et al., 2022), or historical conventions (Al-Laith et al., 2024), which is particularly consequential for literary texts.

Students encounter this through participating in fine-tuning a classifier to a specific literary corpus, using it to analyze the corpus, and comparing the results with generic classifiers. Through this activity, they experience why sentiment classification works in practice and why it systematically fails in cases involving irony, perspective shifts, or culturally specific expression.

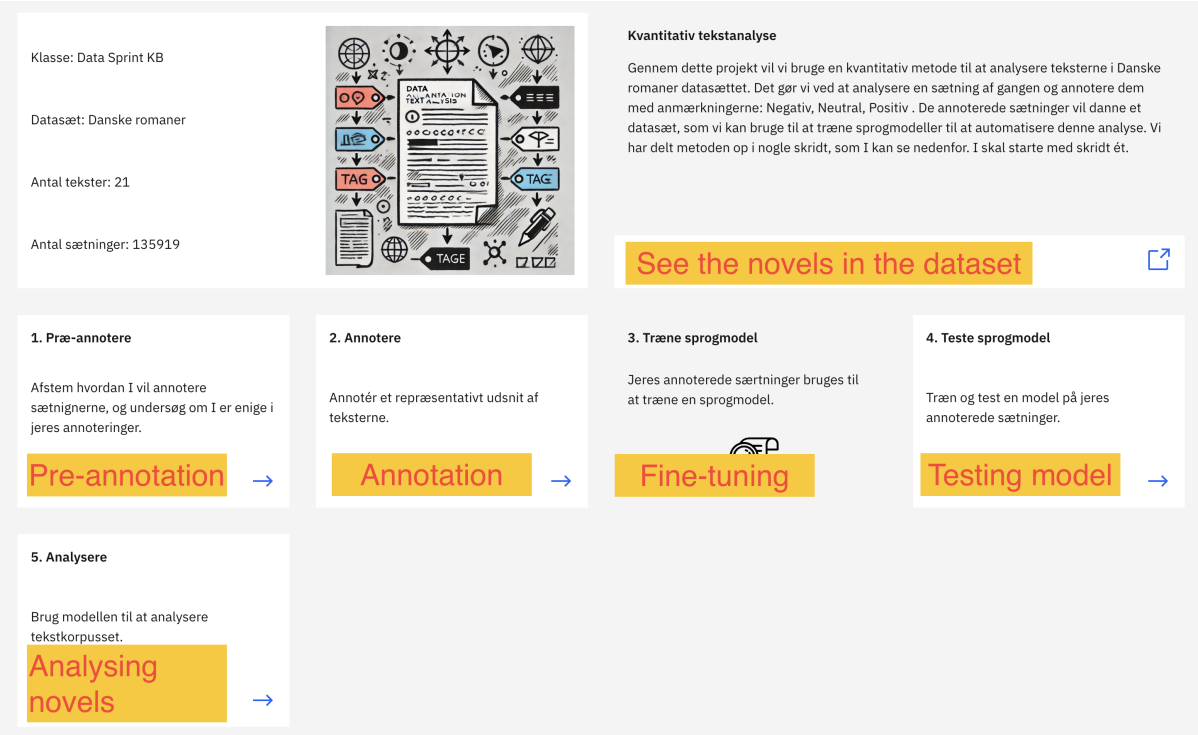

Figure 1: When students log into Litteraturmaskinen, they see a front page organized into five steps that mirror a simplified language-model fine-tuning process.

4 The Litteraturmaskinen Interface

Litteraturmaskinen ('The Literature Machine' in English) is a web application for orchestrating text annotation activities in classroom contexts. We designed it to support integration of digital/NLP methods into literary analysis activities in humanities classes. It gives a low-barrier-of-entry to structured annotation and quantitative analysis of literature, thus making such activities more accessible for non-technical students and teachers. Further, it aims to bridge the tensions in the epistemic traditions by connecting close reading of literary texts and knowledge of literary periods with structured annotation tasks and quantitative argumentation.

The tool provides a graphical user interface that enables students to annotate sentences, compare annotations, test fine-tuned models, and generate visualizations of how fine-tuned models label novels and text corpora. The classroom works on a shared project and provides annotations to a shared dataset. Further, the web-application has a teacher page, where the teacher can create new annotation projects, give students access to different steps in the annotation process, inspect student annotations, and submit datasets to fine-tune new models. Lastly, the system has a backend that stores the text corpora and annotations from different classrooms in a database, fine-tunes Bert models on student-annotated data when a new dataset is submitted, and deploys new endpoints that make the models available in the web application.

4.1 Interface Features

When students log into the system, they see a front page (see Figure 1) organized into five steps that mirror a simplified language model fine-tuning process: pre-annotation, annotation, fine-tuning, testing model, and analyzing novels. Each tile (squared button) links to a sub-page which guides students through the step, except for step three (fine-tuning) which is controlled by the teacher who submits the annotated sentences to the backend as explained above. The teachers can control which steps the student can access through a teacher page; e.g., the teacher can give access to the annotation step when the classroom have aligned their annotation strategy through the pre-annotation step. Figure 2 shows two sub-pages of the interface, where students annotate sentences and can inspect the distribution of annotations for each sentence. Figure 3 shows two types of visualizations from the 'analyzing novels', based on the predictions of the fine-tuned models.

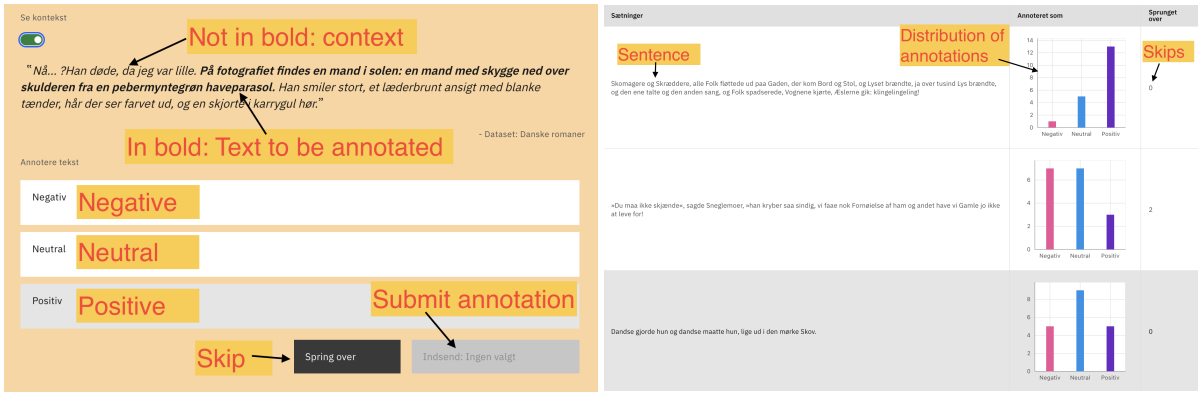

Figure 2: Two sub-pages of the Litteraturemaskinen interface: Left) The interface for annotating a sentences. Right) Interface for inspecting the distribution of annotations for each sentence.

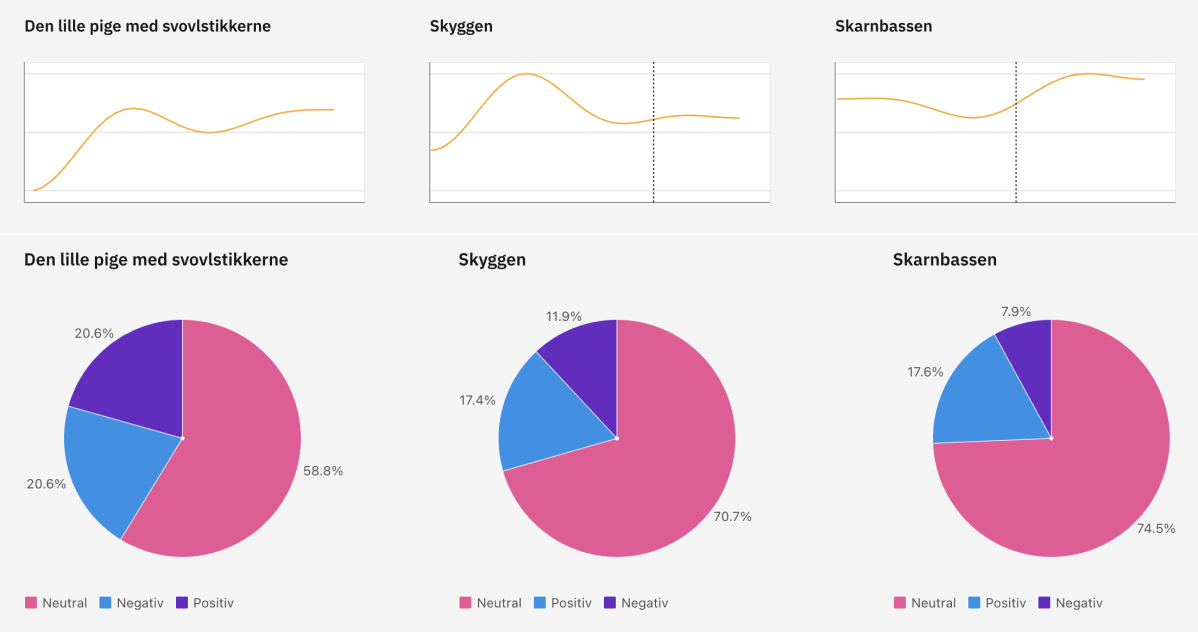

Figure 3: Two types of visualizations of sentiment in the Litteraturmaskinen interface. The visualizations are based on predictions from classifiers fine-tuned on student-annotated data. Here sentences from H.C. Andersen fairy tales. Top) Emotional arcs of fairy tales. Bottom) The fairy tales' distributions of sentiment.

5 Evaluation

We evaluated our educational design of Litteraturmaskinen in two different contexts and from two different perspectives. 1) Two high school L1 teachers adopted Litteraturmaskinen for standard class instruction to teach about H.C. Andersen's fairy tales. 2) We conducted a workshop with undergraduate humanities students on comparing how sentiment is used in both prestige fiction and crime fiction. In both contexts, students annotated the sentence-level sentiment polarity (negative, neutral, or positive). Teachers and workshop participants were assured of anonymity and informed of the research purpose, data storage and usage. The student-annotated sentences were only collected for educational purposes and will not be reused for other purposes.

5.1 High School Setting

We instructed two high school L1 teachers in using Litteraturmaskinen in a three-hour workshop, where we also introduced them to literary SA, basic NLP concepts, and a corpus with H.C. Andersen's fairy tales. Both teachers integrated activities with Litteraturmaskinen into their regular Danish teaching. Teacher (T1) taught in a first year classroom and teacher (T2) taught in a third year classroom. They both taught two modules of 1.5 hours each with Litteraturmaskinen where the students annotated sentences from the fairy tales in the first mod-

ule and used the resulting classifier in the second module to analyze and compare fairy tales (through the types of visualizations shown in Figure 3).

In total, the high school students annotated 1131 sentences with Krippendorff's $\alpha = 0.68$.

5.2 University Setting

We also conducted a workshop within a university-level teaching context. This three-hour extracurricular session was open to humanities students, both undergraduate and graduate, at a Danish university. Eighteen participants were present at the outset of the workshop, of whom ten participated for its full duration (many had to leave to attend a lecture). Most participants were students from the Danish Program and therefore had a direct academic interest in the workshop's thematic focus on Danish literature, and the application of NLP to Danish-language materials.

Against this academic background, instructors and workshop participants jointly prefaced the annotation process by formulating theoretically grounded hypotheses at both the overall corpus level and the individual novel/plot level, which could then be tested by running the fine-tuned model. At the corpus level, the hypothesis was that crime fiction would display stronger sentiment switching and polarization than prestige fiction, reflecting its plot-driven dynamics, while at the level of individual novels the resolution of the crime would be reflected in the sentiment distribution.

In total, the participants annotated 249 sentences with Krippendorff's $\alpha = 0.72$.

6 Data Collection & Analysis

We collected qualitative data through observations, lesson plans, and interviews. In the high school context, we asked both of the teachers to make detailed lesson plans (describing what they did in the classrooms and why) and interviewed each of them after they had taught their last module with Litteraturmaskinen. In the higher education context, four researchers (authors) made observations during the workshop, and we interviewed four participants (Danish university students) after the workshop. Both teacher and student interviews were conducted with a semi-structured interview guide that asked them about their initial thoughts about the activities, how the activities gave new insights about literary texts, how the activities gave new insights about AI technologies, and how/if they could use similar NLP methods in their future practice. We conducted an inductive analysis of the data, which resulted in the four themes presented in the next section.

7 Findings

We present insights from the case study in the form of four themes that emerged from the data analysis. In the following 'teachers' (T) will refer to high school teachers and 'students' (S) will refer to university students if nothing else is specified. We interviewed two teachers and four students.

7.1 Tension Between NLP Methods and Academic Identities

Students and teachers were critical of the value of computational methods in literary analysis and questioned whether they were developing or diluting academic practice. Both their attitudes and subsequent activities with the tool evidenced this.

The two teachers negotiated their sense of professional expertise in the activities differently. T1 articulated a pronounced tension between a self-identified position as a humanist and the perceived technical demands of working with computational or AI-based tools: "*It suddenly dawned on me that I had to teach something that reaches well beyond my field of expertise.*" [T1]. T2, by contrast, acknowledged moments of uncertainty but framed them more pragmatically, presenting the activity as a joint exploration with students: "*Let's see what we can get out of this.*" [T2] Whereas T1's reflections foregrounded disciplinary boundaries, T2 focused on adaptive framing and shared inquiry as strategies for managing uncertainty.

The students had a general skepticism towards AI, e.g., '*As a Danish student [humanist], you have a tendency to think that AI is your worst enemy*" [S1]. However, they all indicated that they participated in the workshop because they hoped to be challenged in their skepticism. S1 and S2 found it difficult to accept the premise of the activity, e.g.: "*All the small nuances that can be embedded in irony, double meanings, and metaphors. They simply disappear.*" [S2]. Both S1 and S2 acknowledged that it could be a way "*to start with something extremely simplified*" [S1]. S2 also admitted that she stopped listening when we explained how the annotated data was used to fine-tune a language model, because she did not perceive herself as having the background knowledge to understand the

technical setup. S3 found it interesting to compare the language of prestige literature and crime fiction, but she was also worried about our intentions with the workshop. Did she "*contribute to automate her own function as a literary scholar?*" [S3].

The findings demonstrate how students and teachers were cautious towards the NLP methods but also curious about how they could be integrated into their existing practices.

7.2 Epistemic Tensions Promoted Literary Reflections

Disagreements between human and model interpretations of the literary texts promoted reflections on how text is interpreted in different contexts and worked as a productive teaching moment.

Both teachers reported high student engagement during the activities. T1 observed "*heated discussions across groups*" [T1] as students debated how to assign sentiment to isolated sentences. She thought that these disagreements prompted deeper disciplinary conversations about context, genre, and authorship. T2 highlighted that Litteraturmaskinen made linguistic features visible and tangible, while T1 emphasized that the interpretive ambiguities inherent in sentiment analysis led to valuable L1 discussions. Both teachers thus found that the tools could catalyze interpretive dialogue, though through slightly different modalities. T2 emphasized playful experimentation and T1 emphasized analytical tension.

In general, the students indicated that the workshop had taught them new perspectives on literature: how taste in literature and literary interpretations are embedded in cultural and social contexts. E.g., S4 reflected on how they interpreted 'death' as negative in their sentence analysis but that it could be interpreted differently in other cultures. It made her wonder if she could use NLP to investigate how different cultures and religions understand death or other concepts. Similarly, S2 had become more aware of how complex language is and how "*We can read things extremely differently depending on the perspective we bring to them.*" [S2].

We also observed this during the workshop. The annotation task prompted a discussion among the participants about the narratological level at which the annotation should be performed – for example, whether annotations should target individual characters, the narrator, or the text's global evaluative stance. In parallel, the students discussed the amount of contextual information included in annotation and training, and whether the model can distinguish, on the basis of context, between, for example, heroes and villains, irony, and humor, as well as the role of linguistic features – such as syntax and sociolect – as potential sources for emotion and sentiment annotation.

While polarity classification is often introduced as a simple supervised task, the findings demonstrate how — in a literary context — it immediately raises questions about interpretation, ambiguity, perspective, and cultural context, which prompted reflections and insights into how literacy is embedded in cultural and social contexts.

7.3 Reflection on the Capabilities and Opportunities of NLP Technologies

Both students and teachers saw new opportunities in NLP technologies after having used Litteraturmaskinen but they had also become more aware of the technology's limitations.

Both teachers openly framed the classroom activities as 'experimental', i.e., the use of Litteraturmaskinen was to be seen as a new tool in literary exploration. T1 foregrounded disciplinary boundaries and potential epistemic tensions, being concerned with maintaining an L1 focus amid perceived technological complexity. T2's account emphasized pedagogical adaptability and student motivation, describing the activities around Litteraturmaskinen, as "*quite successful as it gave rise to relevant discussions in class on the importance of context*," and how being given only individual sentences had the students focus much more intently on "*the tone of the sentence and its constituent words*". T2 also suggested ways to further adapt the tool for L1, such as allowing teachers to upload their own texts.

The activity had started reflections among the interviewed students on how language models are trained. S1 had become more aware of what it requires to train a model for a specific task; how biased chatbots are because of the human judgments that go into the models: "*[H]ow it is socially conditioned [how you annotate the sentences] by where you come from culturally, where in the country you are from, and your background. And how much influence we actually have on language models.*" [S1]. She further wondered how the model would have performed if it was trained on a smaller corpus of text. S4, similarily, wondered how it would work if the model was trained on children's literature which can be more explicit and unambigious in its emotions. We also observed these

types of questions being raised during the workshop. One student, for instance, asked whether the training of the classification model started from scratch using only the workshop's training data, or whether it built upon a pre-trained model that was further fine-tuned on the dataset. Another student raised the question of why next-word prediction is a central training objective, and how this objective enables a model to acquire linguistic competence.

The interviewed students were critical of the validity of the results. S1 would have liked to have time to iterate more on the model to improve the issues they identified when testing the model. S3 were intrigued by how a language model can be trained to navigate in different contexts through fine-tuning. She imagined how a model could be trained to understand the emotional intents of people from different professions or cultures: "*If there is a teacher facing a very multicultural group of students who use terms that she—or he—may not be used to using.*" [S3] During the workshop, we observed the students identify and discuss the limitations of NLP technologies. For instance, they highlighted potential sources of error in annotation, including subjectivity and the cultural embeddedness of emotional perception, as well as ethical reflections on annotation bias and the implications of training language models on biased data.

The findings show how the activities with Litteraturmaskinen gave a deeper understanding of how language models are trained and prompted reflections on the capabilities and opportunities of NLP technologies.

7.4 Reflections on Practical NLP-Integration

Both teachers and students reflected on how NLP technologies and methods can be integrated into their practices.

T1 raised explicit concerns about whether Litteraturmaskinen were "*L1-content relevant enough, or whether instead it would be more at home in a separate subject, such as technology comprehension.*" [T1]. She also reflected critically on the epistemic dimension of the activities, asking: "*Once having completed these activities, is it us, then, or the machine, who come out the wiser?*" [T1]. In contrast, T2 perceived Litteraturmaskinen as meaningfully connected to the L1 subject, albeit in different ways. The Litteraturmaskinen was, for her "*a student-friendly way of fostering a linguistic focus,*" [T2]. Where T1 questioned whether computational methods belonged within the disciplinary frame of L1, T2 argued for their contribution to students' general formation and digital literacy. T2 also saw a clear pedagogical transfer: After the initial activity, students successfully applied sentiment annotation to a new text, albeit working in an analogue fashion, demonstrating heightened awareness of how sentiment is conveyed linguistically.

The workshop spurred reflections on how AI tools could be used in literary studies. The students indicated that they had experienced the workshop as a way to be equipped to use AI tools in more reflexive ways: "*It is positive to use AI not to generate text and replace one's own thinking, but to use it actively as a tool.*" [S4]. They also highlighted how they could gain more control of how AI should influence their subject through understanding and working with the technology: "*It is important that, as a student of Danish [literature], you gain knowledge about this [NLP methods] and learn how to use it, so that it doesn't suddenly overtake you from the inside. That way, we can have influence over how we end up using it.*" [S1]. S2 did, however, also indicate that she saw it more as a necessity to learn about than something she found interesting.

The findings show how the activities provided a starting point for reflections on how NLP methods and technologies can be tighter integrated into practices in literature research and teaching.

8 Discussion

In the context of literary studies, sentiment analysis is an act of reduction as it steers the interpretative attention away from the text as whole and towards individual sentences. Treating all sentences and texts equally, rather than highlighting a selected few as in conventional close reading practices, situates the students within a quantitative mode of literary analysis, often referred to as 'distant reading' (Moretti, 2013) or 'macroanalysis' (Jockers, 2013). Both university students and high school teachers took a curious but skeptical approach to this reductive method and emphasized the ambiguity and cultural context when conducting the sentiment analysis. This skeptical stance mirrors broader debates in literary studies, where distant reading and other quantitative approaches have often been met with concern about reduction and loss of interpretive nuance (Da, 2019). It is important to note that the participating university students took part in the workshop as an extracurricular activity, suggesting a degree of self-selection and curios-

ity. When integrating NLP tools into curricular teaching, it may be valuable to begin from students' own literary questions and interests, and to invite reflection on what kinds of literary inquiry they themselves would like to pursue through computational and quantitative approaches. Thus, this skepticism can refine the tool's role as a catalyst for, and not a replacement of, interpretation.

The pedagogical value of classification tasks such as sentiment analysis lies primarily in making interpretive disagreements visible (Plank, 2022). From a didactic perspective, and as our findings suggest, these moments of disagreement can be understood as productive rather than problematic. They create opportunities for students to reflect on interpretive plurality, categorical reduction, and the human assumptions embedded in computational models. Altogether, the findings indicate that NLP-based literary activities can be meaningfully integrated into literary education across educational levels when framed as exploratory, critical, and dialogic.

While sentiment classification provides an effective entry point for introducing NLP concepts, future work could move from predictive tasks to generative models that can articulate explanations and arguments (Wiegreffe et al., 2022). Rather than asking models to assign labels, students can engage with systems that generate justifications for a given interpretation, compare alternative readings, or articulate counter-arguments grounded in different narrative perspectives (Sui et al., 2025). Such uses shift the focus from correctness to reasoning (Wei et al., 2022), aligning more closely with literary pedagogy while also exposing students to central challenges in NLP, e.g., faithfulness, uncertainty, and cultural grounding (Bender et al., 2021). Framing generative models as tools for argumentation rather than authoritative interpreters reinforces critical engagement and supports a transition from classification to interpretive reasoning.

9 Recommendations

Based on the presented work with Litteraturmaskinen, we propose four key recommendations for integrating NLP into literature and 1L curricula.

Leverage model–human disagreement. Disagreements between student annotations and model predictions should be foregrounded as pedagogical opportunities rather than treated as errors. Such mismatches naturally prompt discussion of narrative perspective, ambiguity, and contextual limits, turning error analysis into interpretive inquiry.

Articulate interpretive reasoning in annotation. Annotation tasks should require brief justifications alongside sentiment labels. This encourages reflective annotation, aligns with literary pedagogy with interpretations anchored in text-specific arguments, and provides insight into how students reason about sentiment, perspective, and context.

Introduce distributional semantics early. A concise explanation of how language models learn from co-occurrence patterns helps students understand why sentiment prediction is possible at all. Framing this learning process as statistical rather than semantic establishes realistic expectations and supports critical evaluation of model outputs.

Support interdisciplinary course design. Effective integration of NLP into literature curricula benefits from the collaboration between literature and computer science educators. This collaboration helps to ensure that computational methods are introduced responsibly while remaining anchored in established interpretive practices.

10 Conclusion

This paper contributes to integrating NLP methods into literature and L1 education. We argue that sentiment analysis provides a productive bridge between NLP and literary interpretation. To support this approach, we developed Litteraturmaskinen as an interface that supports reflective annotation and reasoning and evaluated it with university students and high school teachers. We found that the epistemic tensions between NLP methods and close-reading promoted literary reflections and engaged students and teachers in understanding NLP technologies. Based on these experiences, we provide four recommendations for integrating NLP into literature and L1 curricula.

References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.

Johnny B. Allred, Sean P. Connors, and Christian Z. Goering. 2024. Social annotation and dialogic teaching and learning in english language arts. *Journal of Adolescent; Adult Literacy*, 68(5):515–525.

Saleh Mohammad Alsakaker. 2025. Investigating efl learners' perceptions of using ai to enhance english vocabulary acquisition based on the technology acceptance model. *Forum for Linguistic Studies*, 7(2).

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bettina Berendt, Allison Littlejohn, and Mike Blakemore. 2020. Ai in education: learner choice and fundamental rights. *Learning, Media and Technology*, 45(3):312–324.

Karl-Emil Kjær Bilstrup, Luke Connelly, Line Have Musaeus, Magnus Høholt Kaspersen, and Marianne Graves Petersen. 2025. From automation to integration: Designing opportunities for students and teachers to act skillfully around ai in existing k-12 subjects. In *Proceedings of the 24th Interaction Design and Children*, IDC '25, page 221–235, New York, NY, USA. Association for Computing Machinery.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. Sentimental matters - predicting literary quality by sentiment analysis and stylometric features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Association for Computational Linguistics.

Kristine Bundgaard and Anders Kalsgaard Møller. 2024. Use of ai-powered technologies in upper secondary language learning. *Learning Tech*, (14):14–35.

Alessandra Teresa Cignarella, Elisa Chierchiello, Chiara Ferrando, Simona Frenda, Soda Marem Lo, and Andrea Marra. 2024. From hate speech to societal empowerment: A pedagogical journey through computational thinking and NLP for high school students. In *Proceedings of the Sixth Workshop on Teaching NLP*, pages 54–65, Bangkok, Thailand. Association for Computational Linguistics.

Luke Connelly, Karl-Emil Kjær Bilstrup, and Marianne Graves Petersen. 2025. Beyond llms as black boxes: Activities and an educational tool supporting unplugged and digital ai learning activities for k-12 classrooms. In *Adjunct Proceedings of the Sixth Decennial Aarhus Conference: Computing X Crisis*, AAR Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Nan Z Da. 2019. The computational case against computational literary studies. *Critical inquiry*, 45(3):601–639.

Kirstine Nielsen Degn, Jens Bjerring-Hansen, Ali Al-Laith, and Daniel Hershcovich. 2025. Unhappy texts?: A gendered and computational rereading of the modern breakthrough. *Scandinavian Studies*, 97(1):1–24.

Paulina DeVito, Akhil Vallala, Sean Mcmahon, Yaroslav Hinda, Benjamin Thaw, Hanqi Zhuang, and Hari Kalva. 2025. Unpacking generative ai in education: Computational modeling of teacher and student perspectives in social media discourse.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*.

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Elements in Digital Literary Studies, ahead of print.

Maria Sofia Georgopoulou, Christos Troussas, Evangelia Triperina, and Cleo Sgouropoulou. 2025. Approaches to digital humanities pedagogy: A systematic literature review within educational practice. *Digital Scholarship in the Humanities*, 40(1):121–137.

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.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana, IL.

Matthew L. Jockers. 2014. *Text Analysis with R for Students of Literature*. Springer.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, and et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Evgeny Kim and Roman Klinger. 2021. A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift für digitale Geisteswissenschaften*.

Hsin-Yu Lee, Pei-Hua Chen, Wei-Sheng Wang, Yueh-Min Huang, and Ting-Ting Wu. 2024. Empowering chatgpt with guidance mechanism in blended learning: Effect of self-regulated learning, higher-order thinking skills, and knowledge construction. *International Journal of Educational Technology in Higher Education*, 21(1):16.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *Annual Meeting of the Association for Computational Linguistics*.

Ramon Mayor Martins and Christiane Gresse Von Wangenheim. 2022. Findings on teaching machine learning in high school: A ten - year systematic literature review. *Informatics in Education*.

Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations*.

Franco Moretti. 2013. *Distant Reading*. Verso, London.

Davy Tsz Kit Ng, Jiahong Su, Jac Ka Lok Leung, and Samuel Kai Wah Chu. 2023. Artificial intelligence (ai) literacy education in secondary schools: a review. *Interactive Learning Environments*, 32(10):6204–6224.

Les Perelman. 2014. When 'the state of the art' is counting words. *Assessing Writing*, 21(7):104–111.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Jasper Roe, Mike Perkins, and Leon Furze. 2025. Reflecting reality, amplifying bias? using metaphors to teach critical ai literacy. *Journal of Interactive Media in Education*, 2025(1).

Britta Schneider and Milene Oliveira. 2025. Developing critical ai language literacy—prompting experiments on raciolinguistic bias to understand large language models as cultural artefacts. *AI & Society*. Advance online publication, November 6.

Jane Southworth, Kati Migliaccio, Joe Glover, and et al. 2023. Developing a model for ai across the curriculum: Transforming the higher education landscape via innovation in ai literacy. *Computers and Education: Artificial Intelligence*, 4(1):100127.

Peiqi Sui, Juan Diego Rodriguez, Philippe Laban, J. Dean Murphy, Joseph P. Dexter, Richard Jean So, Samuel Baker, and Pramit Chaudhuri. 2025. KRISTEVA: Close reading as a novel task for benchmarking interpretive reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32829–32849, Vienna, Austria. Association for Computational Linguistics.

Krishnapriya Vishnubhotla, Adam Hammond, Graeme Hirst, and Saif Mohammad. 2024. The emotion dynamics of literary novels. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Yanling Xiao. 2025. The impact of ai-driven speech recognition on efl listening comprehension, flow experience, and anxiety: a randomized controlled trial. *Humanities and Social Sciences Communications*, 12(1).

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2025. Socially aware language technologies: Perspectives and practices. *Computational Linguistics*, 51:689–703.

Zhe (Victor) Zhang and Ken Hyland. 2022. Fostering student engagement with feedback: An integrated approach. *Assessing Writing*, 51:100586.

Zhihui Zhang, Scott Aubrey, Xiaomeng Huang, and Thomas K. F. Chiu. 2025. The role of generative ai and hybrid feedback in improving l2 writing skills: a comparative study. *Innovation in Language Learning and Teaching*, page 1–19.