African Journal of Inter/Multidisciplinary Studies 2024 | *6(S3): 1-13* | *DOI:* https://doi.org/10.51415/ajims.v6i1.1558

RESEARCH ARTICLE:

Using Natural Language Processing Models to Automate Text Labelling: Categorising Semantic Density in Preservice Teachers' Lesson Observation Reports

Thato Senoamadi[1], Lee Rusznyak[2] and Ritesh Ajoodha[3]

Received: 07 July 2024 | **Revised:** 30 October 2024 | **Published:** 09 December 2024

Reviewing Editor: Dr. Anisa Vahed, Xi'an Jiaotong-Liverpool University

Abstract

Education researchers have long had to choose between studies that provide rich insight into teaching and learning in a particular context and insight into broad patterns revealed from large-scale studies. The advances in natural language processing models potentially generate research that offers detailed analysis of specific cases and reveals broader patterns in a much larger dataset. This paper reports on the findings of a study that tested the accuracy of advanced natural language processing models to assign labels to a qualitative dataset. The dataset for this analysis comes from lesson observation reports written by a cohort of preservice teachers pursuing a Postgraduate Certificate in Education (PGCE). Their responses were manually analysed using Legitimation Code Theory (LCT) and graded from simple descriptive observations to complex ones that suggested an interpretation of teachers' pedagogic actions. The Bidirectional Encoder Representations from Transformers (BERT) and its derivatives, namely DistilBERT and RoBERTa, were trained to recognise coding decisions made by researchers on a subset of empirical data. This study compares the efficacy of BERT models in assigning appropriate labels to sections of the dataset by comparing its assigned labels to those allocated manually by the research team. Built upon a dataset consisting of 2167 manually annotated sections, the natural language processing models were trained, refined, and tested in labelling the dataset. A comparative analysis of BERT, DistilBERT, and RoBERTa offers insights into their strengths, efficiencies, and adaptability, achieving an accuracy rate between 72% and 78%. The metrics reveal the current efficacy of these models in coding semantic density in lesson observation reports and create possibilities for analysing massive datasets of similar text. The challenges experienced also reveal the potential limitations of this approach.

Keywords*: natural language processing; Bidirectional Encoder Representations from Transformers; Legitimation Code Theory; semantic density; teacher education*

Introduction

Manual coding of qualitative data can be labour-intensive and time-consuming, yet essential for research in the social sciences. In fields like teacher education, monitoring changes in preservice teachers' understandings of classroom practices is crucial for designing curricula and assessment tasks that support their professional learning. The development of their professional learning can be investigated by analysing practice-based texts written by preservice teachers during their preparation (e.g., Kumm and Graven, 2024; Rusznyak and Walton, 2014). Adler (2002: 10) argues that qualitative research into the complexities of teaching should ideally "enable description and comparison across a range of teachers and classrooms, and with sufficient teachers within the range for patterns of practice to be identified." A challenge for researchers is the limits of qualitative and quantitative data and the methods used to analyse them. Adler (2002: 10) argues that "capturing the complexity of teaching, and indeed the ways this is shaped over time, requires in-depth, qualitative research approaches…that enable rich descriptions of learning by a particular teacher in a particular context". These studies typically collect detailed qualitative data through in-depth interviews, focus group discussions and practice-based artefacts. For example, studies on preservice teachers' development have used essay responses on their own academic and literacy practices (e.g.,

[1]University of the Witwatersrand, thatosenoamadi007@gmail.com | https://orcid.org/0000-0002-1750-3793
[2]University of the Witwatersrand, Lee.Rusznyak@wits.ac.za | https://orcid.org/0000-0002-6835-821
[3]University of the Witwatersrand, Ritesh.Ajoodha@wits.ac.za | https://orcid.org/0000-0002-6443-8592

Dison *et al.,* 2019; Mendelowitz *et al.,* 2023); observations and interpretations of observed lessons (James, 2024; Langsford and Rusznyak, 2024), and written reflections on their own teaching (e.g., Pennefather, 2023; Rusznyak, 2022). Because many variables affect how practices are enacted in contextually appropriate ways, findings from small-scale studies into the development of practice-based expertise may not always be transferable across contexts. However, teacher education research also needs to "investigate practices that extend across diverse contexts and conditions" (Adler, 2002: 10). Typically, such larger-scale studies of teachers' perceptions and practices tend to generate quantitative datasets using questionnaires and surveys (e.g., Matoti and Odora, 2013; Moodley *et al*., 2019). The broad analyses of these kinds of large datasets may sacrifice the nuanced understanding that smaller, more focused studies provide. Researchers undertaking practice-based studies in education often make trade-offs between richly textured insights from in-depth studies of practices within specific contexts, and identifying broad trends obtained from analysing the practices of large cohorts of participants in diverse contexts. While in-depth and intricate analysis is manageable in small-scale case studies, the nature of data and analysis poses significant challenges when researchers attempt to identify patterns of change and continuity in massive datasets.

A substantial challenge in practice-based research is twofold: generating comparable datasets across a diverse range of participants and contexts and finding suitable analytic tools that reveal broad patterns with rich descriptions of learning in massive quantitative datasets. This challenge raises the question this article addresses: Can advanced natural language processing models be trained on manually coded qualitative datasets to accurately detect broad patterns in far larger datasets? This article reports on a study that used a manually coded dataset from a massive qualitative study of preservice teachers' written lesson observations to train machine learning models to recognise categories of coding labels and then tested their accuracy in coding other subsets of the data. This study addresses the following research questions:

- How accurately do Bidirectional Encoder Representations from Transformers (BERT) and its adaptations, DistilBERT and RoBERTa, automate the labelling of complexity in lesson observation reports written by preservice teachers?
- How might such technological innovations enhance our understanding of the affordances and limitations of systems integration software to reveal patterns in massive qualitative datasets?

The findings of this study suggest that these learning models can be used to reveal patterns in larger uncoded datasets, achieving a reliability rate between 72% and 78%. However, a challenge encountered in training the language processing models is that before the analysis could be run on a massive dataset, the extended writing needed to be deconstructed into meaningful units of analysis. Researchers still needed to perform this task manually.

Investigating Practice-Based Learning of Preservice Teachers

Policy governing teacher preparation in South Africa requires that preservice teachers undertake practical learning that includes "aspects of *learning from practice* (e.g., observing and reflecting on lessons taught by others), as well as *learning in practice* (e.g., preparing, teaching and reflecting on lessons presented by oneself)" (Department of Higher Education and Training 2015: 10). Practice-based studies of students' professional development have tended towards small-scale studies in particular contexts. In studies adopting a *learning-in-practice* approach, Nkambule and Mukeredzi (2017) and Pennefather (2023) investigate the pedagogical learning of small groups of preservice teachers from urban universities doing their respective teaching practicals in rural schools in the KwaZulu-Natal and Mpumalanga provinces of South Africa. Similarly, school-based learning in special schools has been the subject of studies by Coates *et al.* (2020), Walton and Rusznyak (2013), and Kang and Martin (2018) in the United Kingdom, South Africa, and Korea, respectively. While offering textured accounts of preservice learning in classroom contexts, the generalisability of these studies is somewhat limited. Not only do preservice teachers bring different sets of dispositions into teacher education programmes (Dison *et al.,* 2019), but there is enormous variation between different teacher preparation programmes offered by South African universities (Council for Higher Education, 2010; Deacon, 2016) and their professional knowledge, skills, and capacity for reasoning at different stages of their professional preparation (James, 2024).

The potential for a massive, comparative study into the pedagogic learning of South African preservice teachers arose during the COVID-19 pandemic during a national lockdown. As schools were closed for an extended period, preservice teachers could not complete all requirements to fulfil their school-based learning requirements. As part

of a governmental capacity-building initiative, a group of teacher educators from different South African universities collaborated to develop an online module designed to supplement and enrich school-based learning (Bertram and Rusznyak, 2024). In contrast to the other studies mentioned, the module adopts a *learning-from-practice* approach and prepares preservice teachers to observe and analyse lessons in scholarly ways. In its first five years, more than 65 000 preservice teachers from 24 South African universities completed the module as a part of their work-integrated learning requirements. The module consists of six units, which guide students through the pedagogic decisions teachers make and how their practices are informed by the priorities and possibilities within the context, the nature of content to be taught, considerations about learner identities that inform the choice and use of learning support materials. Using a wide selection of recorded lessons, preservice teachers learn to observe, analyse and interpret the diverse classroom practices of real teachers across a range of subjects, grades, and school contexts (Rusznyak and Bertram, 2021).

A large-scale research project attached to the module investigated how pre-service teachers interpret the classroom practices they observe at different stages of their teacher preparation. The project obtained permissions from the Department of Higher Education and Training and ethical clearance from various university-based ethics committees. After completing all module requirements, participants were invited to make their responses available for research into preservice teachers' developing teaching practices. The data included lesson observation reports written during the module and their responses to other practice-based tasks. Therefore, the project has generated a massive set of qualitative data where students from different programmes, institutions, and years of study responded to the same recorded lessons and practice-based tasks. The data used in this study was collected from preservice teachers who participated in this module. One significant advantage of the dataset collected during the Teacher Choices in Action research project is its potential to analyse how diverse preservice teachers observe and interpret the same recorded lessons across different institutions, contexts, and stages of their preparation. Langsford and Rusznyak (2024) compared the complexity of lesson observation reports written by a cohort of 83 Postgraduate Certificate in Education (PGCE) preservice teachers at the start and the end of the Teacher Choices in Action module. Their coded dataset provides the basis for the present study. Participants were asked to observe a Grade 7 lesson on peer pressure recorded by Read to Learn SA and used in the study with permission. Preservice teachers were asked to describe the lesson, explain what the teacher does and suggest reasons for why.

To analyse participants' interpretation of the classroom practices they observed, the concept of *semantic density* (SD) from Legitimation Code Theory (LCT) was used to distinguish between simpler and more complex accounts of why the teacher taught as she did. Semantic density conceptualises complexity in terms of the relations of meanings within practices (Maton, 2014; Maton and Doran, 2017). The more meanings are related to or within an expression of practice, the stronger its semantic density. In the case of this study, the strengths of semantic density reflect the extent to which preservice teachers provide networked and theorised interpretations of the teaching they observe in a lesson. Table 1 presents a translation device that shows how different categories of semantic density can be recognised in the empirical data of this study. The table provides the indicators used to assign each unit of data to categories of semantic density and gives illustrative examples from the dataset.

The researchers divided each lesson observation report into units, where each teacher's action constituted a unit of analysis. This yielded 2 119 units of text in the dataset. Using the translation device shown in Table 1, pre-service teachers' interpretations of why the teacher did what she did were analysed and categorised. Researchers checked each other's data labelling and discussed discrepancies until inter-rater reliability was achieved. The categorised data used in the Langsford and Rusznyak (2024) study were a sub-sample from the Teacher Choices in Action project, focusing on one cohort of preservice teachers from one participating university. The project has access to another 37 612 lesson observation reports on the same lesson written by highly diverse cohorts of preservice teachers from different teacher preparation programmes and at different stages of their studies. Although a large-scale comparative study of how first-year and final-year preservice teachers interpret teaching practices could generate invaluable insights into the development of preservice teachers, the dataset is too large to do a complete, in-depth analysis. This factor raises the pivotal question for this study: To what extent can advances in artificial intelligence be used to learn and automate the labelling of categories of semantic density to accurately and efficiently code the remaining uncoded dataset?

Table 1: Translation device showing indicators for different categories of semantic density used to analyse the complexity of preservice teachers' interpretations of classroom practices and examples of how they manifest empirically (adapted by Maton and Rusznyak from Langsford and Rusznyak, 2024)

Concept	Categories	Indicators	Examples
stronger semantic density ↑	SD++	Students describe patterns in the teacher's actions and suggest overarching reasons for those patterns. Typically, students explicitly mention technical terms from educational theories and approaches.	"The teacher used a content-based instruction approach, using content about peer pressure to teach the language skills of reading, identifying keywords, rearranging and putting sentences back in order, the spelling of the vocabulary words, and writing simple sentences." "Throughout the lesson, the teacher asks questions to prompt thinking. The teacher does not answer her own questions. She gives learners the opportunity to reply whether their answers are correct or not. Furthermore, she allows them sufficient time to answer. This gives the learners opportunities to think about their responses."
	SD+	Students describe the teacher's actions and suggest a specific reason for each individual action. Typically, students do not explicitly mention technical terms from educational theories and approaches.	"The teacher asks questions to see if the learners understand." "The teacher walked around to assist those in need and offer encouragement and guidance."
↓ weaker semantic density	SD–	Students describe the teacher's actions without suggesting reasons for those actions.	"The teacher checks on each group." "The teacher started the lesson by asking the learners to define the terms peer and peer pressure."
	SD– –	Students describe classroom practice (e.g., lesson structure, activities, etc.) without mentioning the teacher's actions or their potential reasons.	"The lesson is well-structured and is broken down into sections." "The learners followed the lesson well and were motivated to participate."

Natural Language Processing Models and Text Classification

This section explores the evolving landscape of Natural Language Processing (NLP), specifically focusing on pivotal developments that advanced machine-learning strategies to automate text classification. While the power of these models has been demonstrated across various domains, their application to educational texts—such as lesson reports generated by student teachers—remains relatively unexplored. Here, we investigate how advanced deep learning models like BERT, DistilBERT, and RoBERTa can be applied to classify and label educational text, offering new tools for researchers to analyse large qualitative datasets in fields like teacher education.

The explosive growth of data-driven research and digital communication has recently resulted in an abundance of unlabelled text corpora from the web, social media and digital archives. These diverse datasets, rich in semantic meanings, offer possibilities for developing various advanced computational models for analysis. However, the complexities of these vast texts have also raised challenges, such as the need for AI coding models to handle noise, ambiguity, the subtleties of human language, and nuanced meanings. The strides in language modelling offer the potential for computers to analyse human language patterns, find information and perform tasks such as answering questions, semantic similarity assessments, language translation, and text classification (Radford and Narasimhan, 2018). The text classification process involves two main approaches: rule-based methods and machine-learning techniques. Both methods start with inputting a document or text section into the model, followed by either manual or automatic feature extraction. The final classification step assigns probability distributions over possible classes. This three-step process of text classification is illustrated in Figure 1 below.

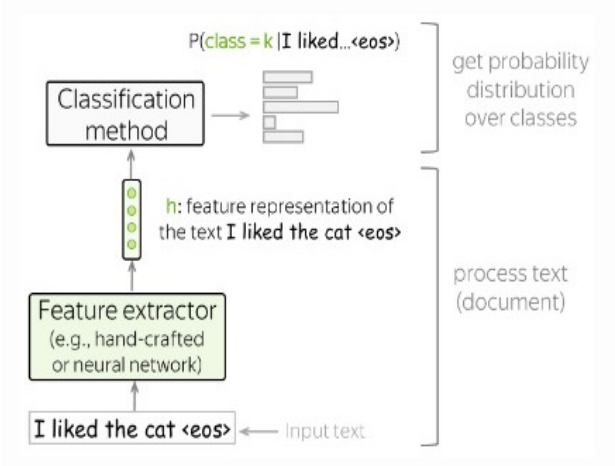

Figure 1: Diagrammatic representation of the text classification process
Source: Voita, 2023.

Neural Approaches to Text Classification

Neural network methodologies gained prominence for managing high-dimensional data and resilience against noisy or inconsistent inputs (Baharudin *et al.,* 2010). Despite being criticised for their black-box nature, they revolutionised natural language programming by introducing word embeddings. Embeddings are mathematical representations of words that capture their meanings based on their usage in context. They convert words into dense vectors of numbers that represent the semantic relationships between words. In the 2010s, advances in neural networks, notably recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, allowed a more nuanced understanding and generation of language. The most foundational neural network models are feed-forward neural networks that use word embeddings

Recurrent Neural Networks process text as a sequence of words, capturing interdependencies and overall structure. Traditional RNNs face challenges with vanishing gradients, where the influence of information decreases exponentially over time, making it difficult for the model to learn and retain the influence of earlier input on later outputs. This issue makes it challenging for the network to handle long sentences or documents where context from the beginning is crucial for understanding content at the end. Long Short-Term Memory networks (LSTMs) are a type of RNN explicitly designed to overcome this problem. LSTMs incorporate special mechanisms called memory cells and gates to maintain information over arbitrary time intervals, preserving long-term dependencies (Tai *et al.* 2015). Both RNNs and LSTMs are adept at handling structured input by recursively applying a set of weights, thus allowing them to generate structured outputs or vector representations, particularly useful for variable-sized inputs, as depicted in Figure 2.

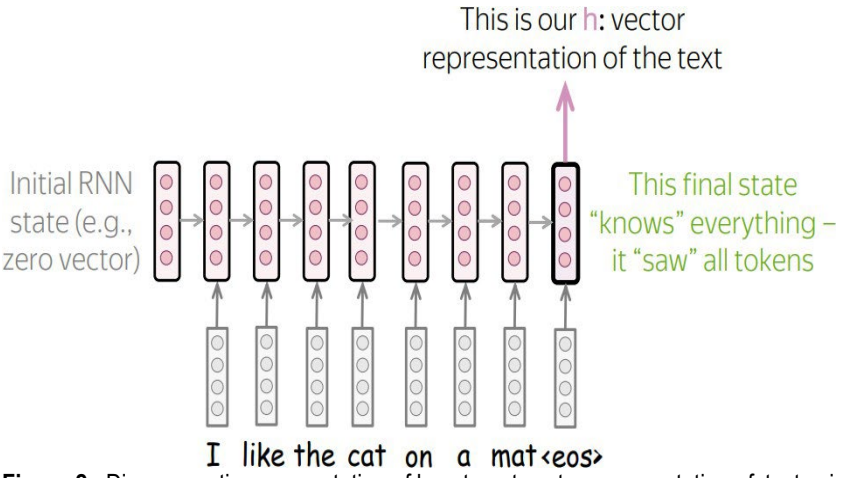

Figure 2: Diagrammatic representation of how to get vector representation of text using the RNN model

Source: Voita, 2023.

Convolutional Neural Networks (CNNs) proved proficient in pattern recognition within data, identifying local, position-invariant patterns like phrases (Johnson and Zhang, 2017). They have been applied to text classification using word embeddings and convolutional layers to capture relationships between words and phrases (Kim, 2014; Conneau *et al.,* 2018).

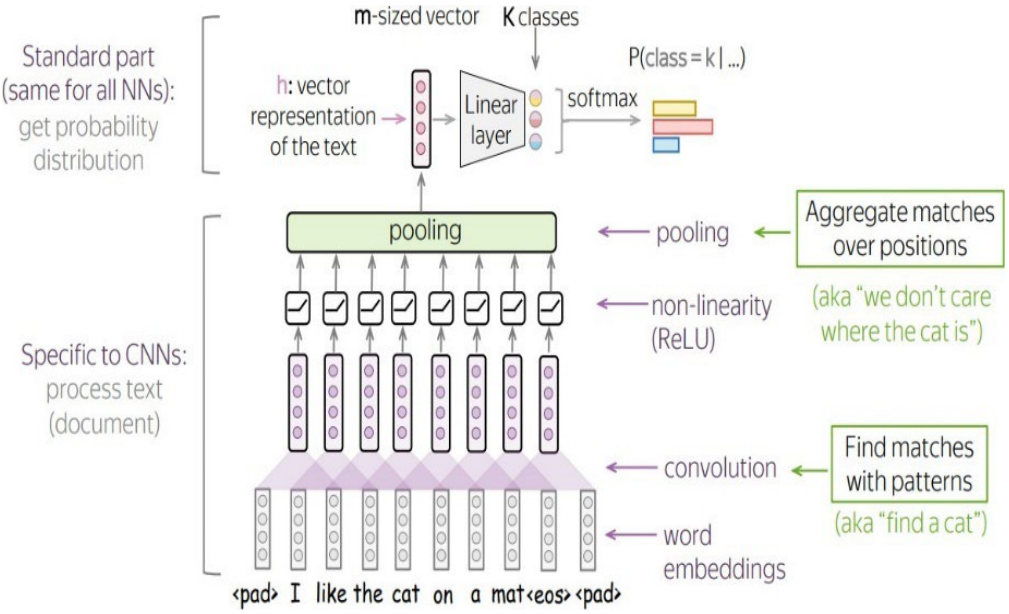

Figure 3: Diagrammatic representation of text classification using Convolutional Neural Networks
Source: Voita, 2023

Attention Networks prioritise different input segments, improving decision-making. These networks identify significant words or phrases within the text. Introducing attention mechanisms, particularly in sequence-to-sequence models, has significantly enhanced context understanding and nuance (Bahdanau *et al.,* 2014; Liu *et al.,* 2016; Yang *et al.,* 2016). Although they advanced text classification, both recurrent neural networks and convolutional neural networks have limitations in handling lengthy texts and their potential for scalability (Vaswani *et al.* 2017).

Transformers' Revolutionary Approach to Text Classification

The advent of transformer-based architectures, particularly BERT and its successors, has led to significant advancements in how computers process language. These technologies are especially good at simultaneously processing many pieces of information and understanding the context of words in a sentence. Transformers use self-attention mechanisms dynamically informed by surrounding words. They predict missing words in a text corpus, endowing them with foundational linguistic comprehension honed through task-specific fine-tuning (Vaswani *et al.,* 2017). BERT models are initially trained on a vast corpus and then fine-tuned for specialised tasks, such as answering questions, categorising text, or filling in missing information in sentences. This versatility makes them highly effective for complex language tasks, leveraging their initial broad training and fine-tuning it to specific needs.

Since 2018, natural language processing has seen significant advances through two main types of pre-trained language models (PLMs). Autoregressive pre-trained language models, exemplified by OpenGPT, anticipate subsequent words in a sequence based on preceding content. In contrast, autoencoding pre-trained language models, such as BERT, reconstruct hidden or masked words within a sentence, benefitting from bidirectional context to enhance understanding. The development of the transformer model by Vaswani *et al.* (2017) led to the creation of Bidirectional Encoder Representations from Transformers (BERT) and its derivatives, DistilBERT and RoBERTa (Devlin *et al.* 2018). These models can capture the context of words in a way not previously possible and have led to unprecedented performance in a wide variety of tasks requiring natural language programming. The ability of BERT to understand context allows it to discern varying semantic implications of words. For instance, BERT can determine the varying semantic implications of the word "bank" in different contexts, overcoming the

constraints of traditional word embeddings like Word2Vec and GloVe.

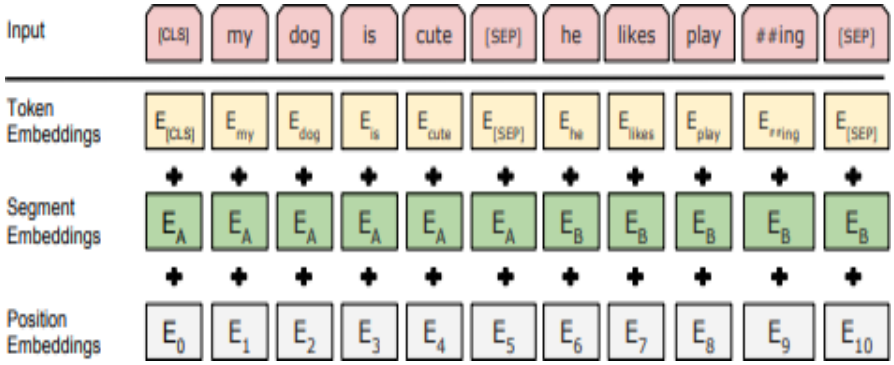

Figure 4: The word vector representations in BERT are more context-dependent because they consider the token embedding, the segmentation embeddings, and the position embeddings together
Source: Voita, (2023)

Refinements of BERT, like RoBERTa (Liu *et al.* 2019) and DistillBERT (Sanh *et al.* 2019), have further optimised transformer models regarding speed and size, making them suitable choices for training to classify texts from extensive datasets. Despite this, its computational identification remains nascent, and few empirical studies are available.

In one empirical study, Bozanta *et al. (*2021) used various classification models to label 100,000 stock movements of five stocks as either "bullish" or "bearish". By working with just two labels and numerical data in a quantitative study, their study showed that BERT and its derivatives could be labelled with an accuracy of between 80% and 85%, as shown in Table 2.

Table 2: Performance of classification models on different stock datasets

Dataset	Algorithm	Accuracy	Precision	Recall	F1 - score
Apple	BERT	0.84	0.84	0.84	0.84
	DistilBERT	0.82	0.81	0.83	0.82
	RoBERTa	0.85	0.84	0.85	0.84
Amazon	BERT	0.86	0.86	0.86	0.86
	DistilBERT	0.85	0.85	0.84	0.85
	RoBERTa	0.86	0.88	0.83	0.85
Boeing	BERT	0.84	0.84	0.84	0.84
	DistilBERT	0.80	0.80	0.81	0.81
	RoBERTa	0.83	0.84	0.83	0.83
Walt Disney	BERT	0.85	0.85	0.85	0.85
	DistilBERT	0.85	0.85	0.85	0.85
	RoBERTa	0.87	0.86	0.87	0.87
SPY	BERT	0.87	0.87	0.87	0.87
	DistilBERT	0.85	0.86	0.85	0.85
	RoBERTa	0.88	0.87	0.88	0.88

Source: Bozanta *et al.* (2021)

Working with four labels and using extended written text in a qualitative study within an educational setting is yet to be attempted, which is the methodological and analytic contribution this article seeks to make.

Methodology

Our study sought to evaluate the capacity of the suite of BERT models to be trained to label a set of qualitative data from pre-service teachers' written lesson observation reports. The dataset contained lesson observation reports written by preservice teachers participating in the "Teacher Choice in Action" module. Building upon the foundational rule-based methods where categories of semantic density were manually labelled (Langsford and

Rusznyak, 2024), this study investigated how machine learning can be leveraged to scale up these classifications. The research adapted BERT and similar models to the task of labelling categories of semantic density on a qualitative dataset. This dataset was appropriate for training BERT derivative models and testing its ability to label qualitative text for several reasons. First, the indicators of categories of semantic density (Table 1) allowed us to use clear, unambiguous categories to train and test natural language processing techniques. The dataset contained authentic text that was linguistically rich, contextually varied and generated. The dataset thus presents an authentic practice-based challenge for algorithmic interpretation. Second, the availability of a set of pre-labelled data within the massive dataset by Langsford and Rusznyak (2024) facilitates the training phase of the model and provides a benchmark for its evaluative accuracy. Third, the complete dataset is massive, with more than 30,000 contributions by thousands of preservice teachers from varied backgrounds and different teacher education institutions at different stages of their professional development. This diversity ensures that any model trained on this data must achieve a robust generalisability level reflecting real-world educational environments' complexities.

In keeping with the ethical undertakings of the research project attached to the 'Teacher Choices in Action' module, an anonymised dataset was provided with participant identifiers generated by an *MS Access* database in which the raw empirical data is stored and managed. The dataset utilised in this study thus protects the identity and privacy of participants.

The models were trained to recognise four categories of semantic density in the dataset. This involved gathering a large set of coded lesson reports and randomly dividing them into three subsets. Data preprocessing initiates the automated classification of the datasets by removing unwanted symbols and deconstructing the text into smaller, manageable units, a process known as tokenisation. Tokenisation involves splitting the text into individual words, phrases, or other meaningful segments the model can analyse. Subsequently, the BERT model underwent pre-training on a large dataset comprising subsections of the lesson reports. This pre-training captured the preservice teachers' general language patterns and representations to describe a lesson. The model was then fine-tuned on a smaller labelled dataset, enabling it to adapt specifically to the text classification task. BERT allocated each bit of data one of four labels, ranging from SD++ to SD– –. The BERT models went through cycles of refinement by providing feedback on where its labelling differs from that done manually by the researchers. The third dataset evaluated BERT's accuracy by comparing their labelling of categories of semantic density with another subset of the data coded by researchers.

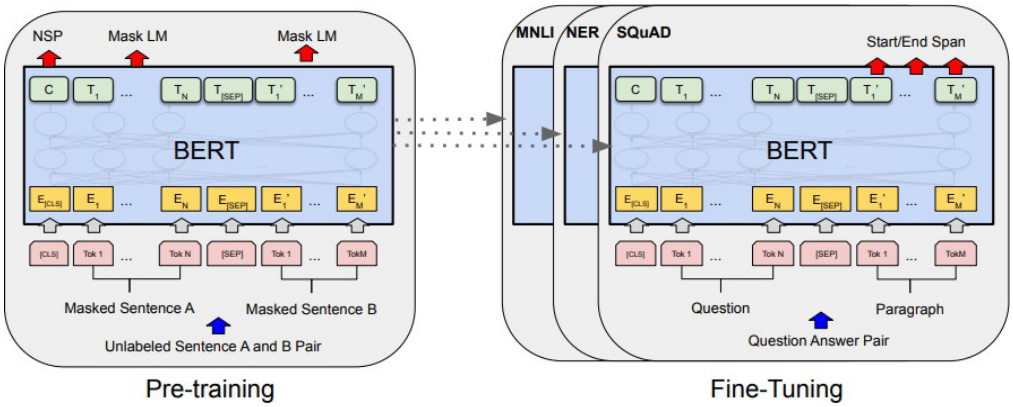

Figure 5: Schematic overview of the BERT pre-training and fine-tuning process
Source: Devlin *et al.* (2018)

The classification prowess of the three BERT models can be assessed through accuracy, precision, recall, and F1-score. Accuracy represents the proportion of correctly classified instances out of the total cases, indicating overall correctness. Precision measures the accuracy of the positive predictions, defined as the proportion of true positive predictions out of all positive predictions made by the model. Recall assesses the model's ability to identify all relevant instances, defined as the proportion of true positive predictions out of all actual positive instances. The F1-score is the harmonic mean of precision and recall, balancing the two metrics.

There were three principal variables in this study. The input variable comprised lesson observation reports by student teachers, which underwent tokenisation and subsequent conversion into contextual embeddings. The processing variables were the model's operation, which involved distilling features essential for accurate semantic density

labelling. The outcome variables were the labels for semantic density (SD), ranging from stronger (SD++) to weaker (SD– –), as shown in Table 1. These reflected preservice teachers' ability to observe and interpret teaching practices with varying levels of complexity. The variables interacted in various ways. The BERT models had to consider the interplay between the pedagogical content of the lesson reports and the educational discourse and linguistic subtleties used by preservice teachers. For more precise labelling of categories of semantic density, it is crucial to fine-tune the model's hyperparameters—settings that control how the model learns from the data, such as the learning rate, batch size, and number of training iterations. Optimising these hyperparameters helps improve the model's performance by ensuring it learns effectively from the data without failing to generalise to new samples. The envisaged output is an array of semantic density labels attained through fine-tuning pre-trained BERT models to capture the depth of the textual context.

For reproducibility, the hyperparameters for BERT, DistilBERT, and RoBERTa were selected to balance model performance and computational efficiency. We used the AdamW optimiser with a learning rate 1e-05, batch size 16, and fine-tuning for four epochs. The AdamW optimiser helps regulate weight updates, reducing overfitting by correcting for bias in weight decay. A learning rate 1e-05 ensures gradual updates to the model's weights for stable training. A batch size 16 was selected to balance computational efficiency and training stability while fine-tuning for four epochs provided sufficient exposure to the data without overtraining the model. These parameters were chosen based on a grid search that tested different combinations to prevent overfitting during model training. Additionally, we addressed potential class imbalances by employing a stratified sampling technique so that each semantic density category was adequately represented during training. This mitigated the risk of the model becoming biased towards more prevalent classes, thereby improving the generalisability of our findings.

We also conducted an error analysis to understand the instances where the models made incorrect predictions. We identified common patterns in these errors, particularly in the SD++ and SD– – categories, which often involve more nuanced and context-dependent interpretations. Errors typically occurred when the models encountered ambiguous or contextually complex sentences, indicating a need for improved contextual understanding.

Findings

We achieved notable outcomes by applying BERT and its variants to semantic density labelling within our dataset. Table 2 details the hyperparameters used for fine-tuning three transformer-based models (BERT, DistilBERT, and RoBERTa) for semantic density labelling. All models used the AdamW optimiser, which combines adaptive learning rates with weight decay regularisation to prevent overfitting. They were trained for four epochs, balancing the need for sufficient training with the risk of overfitting. A small learning rate 1e-05 was chosen to ensure gradual and stable updates to the model parameters, allowing for precise fine-tuning of these pre-trained models for the specific task. Table 3 summarises the performance metrics of the BERT, DistilBERT, and RoBERTa models on the lesson reports dataset, including accuracy, precision, recall, and F1-score.

Table 3: Performance metrics for different algorithms used on the lesson reports dataset

Dataset	Algorithm	Accuracy	Precision	Recall	F1 - score
Lesson Observation Reports	BERT	0.740	0.744	0.740	0.739
	DistilBERT	0.726	0.727	0.726	0.726
	RoBERTa	0.780	0.780	0.780	0.780

In this study, RoBERTa achieved the highest accuracy (0.78), precision (0.78), recall (0.78), and F1-score (0.78), indicating its superior performance in correctly classifying the semantic density of lesson reports compared to BERT and DistilBERT, which also performed well but with slightly lower metrics. The analysis of the results suggests that the RoBERTa model performed better in classifying the lesson reports into the defined semantic density categories. RoBERTa demonstrated a particularly robust capability in distinguishing between stronger and weaker categories of semantic density, reflecting its proficiency in recognising varying levels of complexity within the reports.

The efficacy of using the BERT models extends the findings of other leading-edge studies in labelling and text classification. The results of our study did not achieve an accuracy as high as that of the study by Bozanta *et al.,* (2021) of 100 000 stock movements. The slightly better performance in their case was likely due to having only two labels. In contrast, our research has four, using a much larger dataset and the effect of using techniques such as 5-fold cross-validation. Furthermore, their study used a more straightforward dataset, whereas ours was based

on extended text analysis. Future research has the opportunity to explore a lot more techniques that could improve the accuracy of these models, such as pre-training on the unlabelled available dataset or using Hybrid models, such as the combination of BERT and Bayesian Networks as Liu *et al.* (2019) described in their paper. The observed high levels of accuracy validate our proposition that transformer-based architectures like BERT are aptly suitable for fine-tuning to execute complex tasks such as labelling categories of semantic density. It is crucial to acknowledge our study's intrinsic constraints. These include the dataset's particularity, the selected hyperparameters, and the inherent opacity of deep learning models, which collectively introduced elements of unpredictability. In pursuing our research objectives, we encountered several challenges that impacted the efficacy and applicability of the BERT models. These challenges provided valuable insights and directions for future research to enhance the robustness and applicability of advanced natural language processing models and their use for analysing qualitative data in educational contexts.

The intricate structure of BERT models required significant computational power, which posed a constraint given our available resources. However, we effectively mitigated this issue by optimising the fine-tuning process. Ensuring the dataset captured the full spectrum of semantic density in educational materials was also challenging. We addressed this by incorporating labelled data using lesson observations written at different points in time, with meticulous data quality oversight. In addition, we faced the risk of overfitting, whereby models might align too closely with the training data and fail to generalise to unseen data. We carefully tuned hyperparameters and implemented regularisation techniques such as dropout to combat this. An unexpected challenge was the need to deconstruct the dataset into 'episodes of pedagogic reasoning' before the trained models could reveal patterns within the massive unlabelled dataset. The next phase of this study will use the trained RoBERTa model on the massive uncoded dataset. Before this occurs, we would need to determine if large language models can be used to segment the lesson observation reports into units describing the teachers' actions and interpretations.

Conclusion

While BERT, DistilBERT, and RoBERTa performed well overall, their robustness in handling edge cases or ambiguous sections of lesson reports warrants further scrutiny. These models might struggle with atypical examples or unclear contexts within the lesson reports, potentially leading to misclassifications. Ensuring the models can handle these challenging scenarios is critical for their reliable application in diverse educational contexts. In addition, the training dataset used in this study might contain inherent biases that could inadvertently influence the models' decision-making processes. Such biases might arise from how lesson reports were collected and annotated or from imbalances in the dataset, such as overrepresenting certain types of observations. If the models learn these biases, it might lead to skewed results, affecting the fairness and accuracy of semantic density labelling. Despite achieving high accuracy, the complex nature of BERT models poses challenges to the interpretability of their decision-making processes. This aspect is particularly relevant in our study as educators and researchers need to understand why a model made a specific prediction about the semantic density of a lesson report. Enhancing interpretability is necessary for validating and trusting the models' outputs, thereby providing deeper insights into pedagogical practices.

Our study investigated the versatility, promise, and limits of transformer-based models like BERT in natural language processing for labelling categories within a qualitative dataset. This method potentially enables the labelling process of massive unlabelled datasets and reveals the expected level of precision and consistency. Bassey (2001) distinguishes between scientific generalisation and generalisability in the context of social sciences; he argues that scientific generalisations can seldom be made in educational research or the social sciences due to the many variables present. The scholar suggests that, at best, "fuzzy generalisations" using qualifiers in statements of the form 'x in y circumstances may result in z', can still provide valuable and informed insight into social phenomena (Bassey 2001: 10). This study contributes to a potential way in which researchers in teacher education work between the trade-offs Adler identified between the rich insights offered by in-depth studies of specific exemplars, and insights from broad generalised patterns obtained from massive datasets. The study reveals the degree of confidence that could be claimed from the automated labelling of qualitative text data using natural language processing models.

As more complex models are introduced with further technological advancements, natural language comprehension and processing frontiers are headed for significant expansion. Subsequent research initiatives in this regard will likely be significantly informed by the groundwork established by studies such as ours and those that came before.

References

Adams, J. D. and Mabusela, M. S. 2015. Pre-service Educators' Attitude Towards Inclusive Education: A Case Study. *Journal of Social Sciences*, 43(1): 81-90.

Adler, J. 2002. Global and Local Challenges of Teacher Development. In: Adler, J. and Reed, R. eds. *Challenges of Teacher Development: An Investigation of Take-Up in South Africa*. Pretoria: Van Schaik, 1-16.

Baharudin, B., Lee, L. H. and Khan, K. 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1(1): 4-19.

Bahdanau, D., Cho, K. and Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. Available: https://formacion.actuarios.org/wp-content/uploads/2024/05/1409.0473-Neural-Machine-Translation-By-Jointly-Learning-To-Align-And-Translate.pdf (Accessed 12 January 2024).

Bassey, M. 2001. A Solution to the Problem of Generalisation in Educational Research: Fuzzy Prediction. *Oxford Review of Education*, 27(1): 5-22.

Bertram, C. and Rusznyak, L. 2024. Navigating Tensions in Designing a Curriculum that Prepares Preservice Teachers for School-based Learning. *Education as Change,* 28(1): 1-23.

Bozanta, A., Angco, S., Cevik, M. and Basar, A. 2021. Sentiment Analysis of Stocktwits Using Transformer Models. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9680251 (Accessed 18 January 2024).

Coates, J. K., Harris, J. and Waring, M. 2020. The Effectiveness of a Special School Experience for Improving Preservice Teachers' Efficacy to Teach Children with Special Educational Needs and Disabilities. *British Educational Research Journal*, 46(5): 909-928.

Conneau, A., Schwenk, H., Barrault, L. and Lecun, Y. 2016. Very Deep Convolutional Networks for Text Classification. Available: https://arxiv.org/abs/1606.01781 (Accessed 13 June 2024).

Council of Higher Education. 2010. Report on the National Review of Academic and Professional Programmes in Education. Available: https://www.che.ac.za/sites/default/files/publications/Higher_Education_Monitor_11.pdf (Accessed 03 June 2024).

Deacon, R. 2016. The Initial Teacher Education Research Project: Final Report. Available: https://www.jet.org.za/resources/deacon-iterp-final-composite-report.pdf (Accessed 27 May 2024).

Department of Higher Education and Training. 2015. National Qualifications Framework Act (67/2008): Revised Policy on the Minimum Requirements for Teacher Education Qualifications. Available: https://www.dhet.gov.za/Teacher%20Education/National%20Qualifications%20Framework%20Act%2067_2008%20Revised%20Policy%20for%20Teacher%20Education%20Quilifications.pdf (Accessed 12 May 2024).

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available: https://arxiv.org/abs/1810.04805?amp=1 (Accessed 22 August 2023).

Dison, L., Shalem, Y. and Langsford, D. 2019. Resourcefulness Matters: Student Patterns for Coping with Structural and Academic Challenges. *South African Journal of Higher Education*, 33(4): 76-93.

James, L. M. 2024. Observing Literacy Pedagogies: A Comparison of How First and Fourth-Year Preservice Teachers Analyse Practice. Master's dissertation, University of the Witwatersrand.

Johnson, R. and Zhang, T. 2017. Deep Pyramid Convolutional Neural Networks for Text Categorization. Available: https://aclanthology.org/P17-1052.pdf (Accessed 22 July 2023).

Kang, D. Y. and Martin, S. N. 2018. Improving Learning Opportunities for Special Education Needs (SEN) Students by Engaging Pre-Service Science Teachers in an Informal Experiential Learning Course. *ASIA Pacific Journal of Education*, 38(3): 319-347.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. Available: https://arxiv.org/pdf/1408.5882 (Accessed 25 October 2023).

Kumm, M. and Graven, M. H. 2024. Exploring Pre-Service Teachers' Knowledge of Efficient Calculation Strategies. *South African Journal of Childhood Education*, 14(1): 1-13.

Langsford, D. and Rusznyak, L. 2024. Observing Complexity in Teachers' Choices: The Impact of Preparing Preservice Teachers for Work-Integrated Learning. *Education as Change,* 28(1): 1-22.

Liu, S., Tao, H. and Feng, S. 2019. Text Classification Research Based on the Bert Model and Bayesian Network. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8996183 (Accessed 18 May 2023).

Liu, Y., Sun, C., Lin, L. and Wang, X. 2016. Learning Natural Language Inference Using Bidirectional LSTM Model and Inner Attention. Available: https://arxiv.org/pdf/1605.09090 (Accessed 18 July 2023).

Maton, K. 2014. *Knowledge and Knowers: Towards a Realist Sociology of Education*. Abington: Routledge.

Maton, K. and Doran, Y. 2017. Semantic Density: A Translation Device for Revealing Complexity of Knowledge Practices in Discourse. Available: https://pensamientoeducativo.uc.cl/index.php/onom/article/view/30395/41835 (Accessed 24 June 2023).

Matoti, S. N. and Odora, R. J. 2013. Student Teachers' Perceptions of Their Experiences of Teaching Practice. *South African Journal of Higher Education*, 27(1): 126-143.

Mendelowitz, B., Ferreira, A. and Dixon, K. 2023. *Language Narratives and Shifting Multilingual Pedagogy*. London/New York: Bloomsbury.

Moodley, T., Sadeck, M. and Luckay, M. 2018. Developing Student Teachers' Professional Knowledge (Including Teaching Practice) in the Further Education and Training Phase. In: Sayed, Y., Carrim, N., Badroodien, A., McDonald, Z. and Singh, M. eds. *Learning to Teach in a Post-Apartheid South Africa*. Cape Town: Sun Press, 131-148.

Nkambule, T. and Mukeredzi, T. G. 2017. Pre-Service Teachers' Professional Learning Experiences during Rural Teaching Practice in Acornhoek, Mpumalanga Province. *South African Journal of Education*, 37(3): 1-9.

Pennefather, J. 2023. Student Teacher Learning in Rural Contexts: Challenges and Opportunities. *Journal of Education*, 90: 87-108.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. 2018. Improving Language Understanding by Generative Pre-Training. Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (Accessed 18 August 2023).

Rusznyak, L. and Walton, E. 2014. Using Metaphors to Gain Insight into South African Student Teachers' Initial and Developing Conceptions of 'Being a Teacher'. *Education as Change*, 18(2): 335-355.

Rusznyak, L. and Bertram, C. 2021. Conceptualising Work-Integrated Learning to Support Pre-Service Teachers' Pedagogic Reasoning. *Journal of Education*, 83: 34-52. http://dx.doi.org/10.17159/2520-9868/i83a02

Rusznyak, L. 2022. Using Semantic Pathways to Reveal the 'Depth' of Pre-Service Teachers' Reflections. *Education as Change*, 26(1): 1-24. http://dx.doi.org/10.25159/1947-9417/10013

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper, and Lighter. Available: https://arxiv.org/pdf/1910.01108 (Accessed 19 October 2023).

Shum, H. Y., He, X. D. and Li, D. 2018. From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots. *Frontiers of Information Technology and Electronic Engineering*, 19: 10-26.

Tai, K. S., Socher, R. and Manning, C. D. 2015. Improved Semantic Representations from Tree-Structured Long Short-Term Memory. Available: https://arxiv.org/pdf/1503.00075 (Accessed 18 July 2023).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. 2017. Attention Is All You Need. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (Accessed 19 June 2023).

Voita, L. 2023. Natural Language Processing Course. Available: https://lena-voita.github.io/nlp_course.html (Accessed 07 July 2024).

Walton, E. and Rusznyak, L. 2013. Pre-service Teachers' Pedagogical Learning during Practicum Placements in Special Schools. *Teaching and Teacher Education*, 36: 112-120. https://doi.org/10.1016/j.tate.2013.07.011

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. Available: https://aclanthology.org/N16-1174.pdf (Accessed 28 July 2023).