technologies

MDPI

Article

Automated Information Extraction from Safety and Material Data Sheets—A Domain-Specific NLP Pipeline for Structured Material Data Management in Battery Cell Production

Simon Otte *, Felix Bayer, Sebastian Schabel and Jürgen Fleischer *

wbk Institute of Production Science, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
* Correspondence: simon.otte@kit.edu (S.O.); juergen.fleischer@kit.edu (J.F.)

Abstract

The performance of lithium-ion batteries is strongly determined by material properties, which are provided in technical data sheets but often in inconsistent formats and terminology. Automated extraction of these parameters could enable downstream applications such as process optimization, traceability, and hazard assessment. However, current approaches are unsuitable for industrial use. This work presents a prototype NLP-based extraction pipeline for material and safety data sheets. Using fine-tuned SpaCy models, F1-scores above 0.7 are achieved for key parameters such as CAS number, molecular mass, and density. The resulting structured material database provides a foundation for data-driven applications in battery cell production. The feasibility of domain-specific NLP for automated material information extraction is demonstrated and potential pathways for integration with process control and optimization workflows are discussed.

Keywords: battery cell production; material data sheet; natural language processing; process optimization

1. Motivation and Objective

The performance of a battery cell depends on a variety of factors, such as the used materials [1,2]. Beyond major material choices, such as anode type (graphite versus silicon–graphite) or cathode chemistry (LFP versus NMC), variations within the same material category affect downstream processes and cell performance [3,4]. Changes in particle morphology, particle size distribution, and material purity alter mixing behavior and ultimately determine production yield [5].

Material suppliers already provide relevant material information through safety data sheets (SDSs) and material data sheets (MDSs). However, extracting and utilizing this data systematically remains challenging. The MDSs can have a highly customized format that varies between suppliers and even within a company. In addition to standardized terms, such as CAS number or density, the core problem is nomenclature heterogeneity. Identical material properties are referred to by different terms across supplier documents. For example, the residual moisture content of active material may appear as residual moisture, remaining moisture, remaining water content, relative humidity, or similar terms. Furthermore, particle size may be specified as d50, percentile ranges, or size distribution descriptions. In addition to the variety of names, the placement of information also varies, whether in continuous text, as a list, or in a table.

Received: 23 March 2026
Revised: 24 April 2026
Accepted: 7 May 2026
Published: 9 May 2026

To solve this problem, three solution pathways exist. Cloud-based language models such as ChatGPT, Microsoft Copilot and DeepSeek theoretically could handle terminology variability, but are unsuitable for use in industrial production. Refs. [6,7] identified trustworthiness concerns with large language models, while [8] highlighted specific data privacy risks in SDS processing. Commercial on-premises extraction tools perform well on structured documents such as invoices but fail on documents with heterogeneous terminology and variable formats. Refs. [9,10] demonstrated that commercial systems produce incomplete results on less structured documents, and MDSs are substantially more heterogeneous than the documents these tools were designed to process.

A more viable option is to develop company-specific models based on established architectures. Named entity recognition (NER), a fundamental natural language processing (NLP) task that identifies and classifies specific information categories in text, is such an alternative. The ChemDataExtractor achieved an 87 percent F1-score extracting chemical properties from scientific publications. Ref. [11] applied BERT-based NER to SDSs and achieved 93 percent precision for hazard extraction. Ref. [12] demonstrated text mining approaches for synthesis parameter extraction. These successes suggest NER could address MDS processing. However, existing approaches operate on domains with relatively standardized terminology. Prior work has not systematically addressed the nomenclature heterogeneity characteristic of uncontrolled material supplier documents to consolidate paired SDS and MDS documents with their different information structures and complementary content.

This paper addresses this gap by developing a prototype NLP-based extraction pipeline for material and SDSs. The primary objective is to demonstrate that fine-tuned NER models can effectively extract material information from heterogeneous supplier documents despite terminology variability. The secondary objective is to develop a complete end-to-end pipeline including PDF text extraction, entity recognition, conflict resolution, and intelligent consolidation of paired documents. This work is explicitly scoped as a technical feasibility study and does not include validation of downstream process optimization applications, integration with live production systems, or quantification of industrial impact.

2. Background on Natural Language Processing

NLP is an interdisciplinary field combining computational linguistics and artificial intelligence to enable machines to process human language. The discipline has evolved from early rule-based approaches to statistical, machine learning-based and, more recently, neural methods [13]. Modern NLP systems now predominantly employ statistical and machine learning methods to handle language's inherent ambiguity and variability [14].

NER represents a fundamental NLP task that identifies and classifies textual references to real-world entities. NER enhances text understanding by identifying entities such as persons, locations, or organizations. Essentially, any noun with a proper name constitutes a named entity. NER aims to determine the token span of such entities and assign them appropriate labels. Common NER tags include: PER (person), LOC (location), ORG (organization), and GPE (geopolitical entity). NER can also be extended to expressions like dates, times, or numerical data (e.g., prices) [15].

This work uses these features to identify melting points, product names, Chemical Abstracts Service (CAS) numbers and more by training a model for their detection. NER generally serves as a crucial first step for many NLP applications (e.g., sentiment analysis). Unlike part-of-speech-tagging (POS-tagging), which identifies word classes, NER adds semantic information not derivable from a word's grammatical function alone. This leads to ambiguous cases (e.g., "JFK" could refer to a person, airport, or school), whereas POS-tags would remain unambiguous due to grammatical rules [14]. While language-specific models

exist (e.g., for German [16]), their tags remain limited to persons, locations, and organizations, necessitating fine-tuning for the additional entities required to be recognized for this body of work. Alternatives like BiLSTM-CNN hybrids are another viable alternative.

2.1. SpaCy's NER Model

SpaCy [17] is a popular open-source Python library for NLP, offering pretrained deep neural network (DNN) models for NER, POS-tagging, and text classification. Its efficiency and robustness make it a standard for practical NLP pipelines. SpaCy v2 furthermore employs subword tokenization (e.g., splitting impossible into im + possible) and Bloom embeddings, which reduce memory usage via hash-based vector mapping [18]. SpaCy models use the BILUO tagging scheme [19], relevant for training data preparation.

2.2. BERT NER Models

BERT's effectiveness for NER comes from its pretraining fine-tuning paradigm and bidirectional architecture. The model employs self-supervised learning during initial pretraining, where it processes large unlabeled text corpora (∼3.3 billion words) through two key tasks [20]:

- Masked language modeling (MLM): Random tokens (15% of input) are masked and predicted using contextual information from both directions, forcing the model to learn deep linguistic representations [20,21].
- Next sentence prediction (NSP): The model determines whether two text segments appear consecutively in the original corpus, enhancing discourse understanding [20].

Our work uses German BERT (trained on German National Library texts [16]) and XLM-RoBERTa (pretrained on 2.5 TB of multilingual CommonCrawl data [22]), which share BERT's core architecture. Both models share BERT's core architecture but differ in pretraining corpora and optimization strategies, affecting their NER performance on specialized domains. Notably, the XLM-RoBERTa model uses 24 instead of 12 layers of transformer blocks that each hold 16 instead of 12 multihead attention layers. The hidden layer is also bigger with a size of 1024 instead of 768 as is the case for the standard BERT model. This results in a larger model with 550 instead of 100 million parameters [14,22].

3. State of the Art

Building upon the foundational concepts, this chapter shows current NLP applications in materials science, focusing on information extraction from technical documents. The rise of pretrained language models like BERT has significantly advanced domain-specific entity recognition capabilities through fine-tuning approaches [20].

3.1. Information Extraction from Scientific Literature

Scientific publications contain valuable unstructured data that is costly to process manually. Ref. [23] developed ChemDataExtractor, an automated pipeline combining NLP techniques to identify chemical entities and their properties from PDF documents. The modular system achieves an F1-score of 87.11% for chemical NER, demonstrating the feasibility of automated data extraction despite challenges like information-dense text and structural formula interpretation. Similar tools like OpenChemIE specialize in reaction data extraction from documents, achieving 64.3% precision for a large variety of chemical literature.

3.2. NLP in Materials Science

Materials science comes with unique NLP challenges due to data heterogeneity and domain-specific terminology. Olivetti et al. [24] identify key obstacles: (1) decentralized data production lacking standardized formats, (2) multiscale material descriptions (nano-

to-macroscopic), and (3) technical jargon requiring specialized models. Kononova et al. [25] further highlight nomenclature inconsistencies and publication bias favoring positive results, limiting training data quality. Despite these challenges, successful applications exist, such as Mahbub et al.'s [12] text mining approach for solid electrolyte synthesis parameter extraction.

3.3. Safety Data Sheet Processing

While no systems currently process MDS, several approaches exist for chemical SDS. The SDSParser [26] uses regular expressions for limited manufacturer-specific extraction. Suman et al. [27] developed a more advanced hybrid system combining BERT-based NER (precision: 0.931) with computer vision for table extraction from SDS documents. Their multi-stage pipeline demonstrates the effectiveness of task-specific model architectures. This review reveals critical research gaps: (1) insufficient labeled datasets in materials science, (2) no existing solutions for MSDS processing, and (3) the need for specialized NER models handling technical nomenclature. Our work addresses these challenges by developing a fine-tuned NER system for automated data extraction from both SDS and MSDS documents.

4. Methodology

This chapter presents our approach for extracting material property data from SDS and MDS using fine-tuned NER models. The pipeline comprises four key stages: text extraction/preprocessing, model architecture/training, parameter optimization, and post-processing (Figure 1). The individual process stages are explained below.

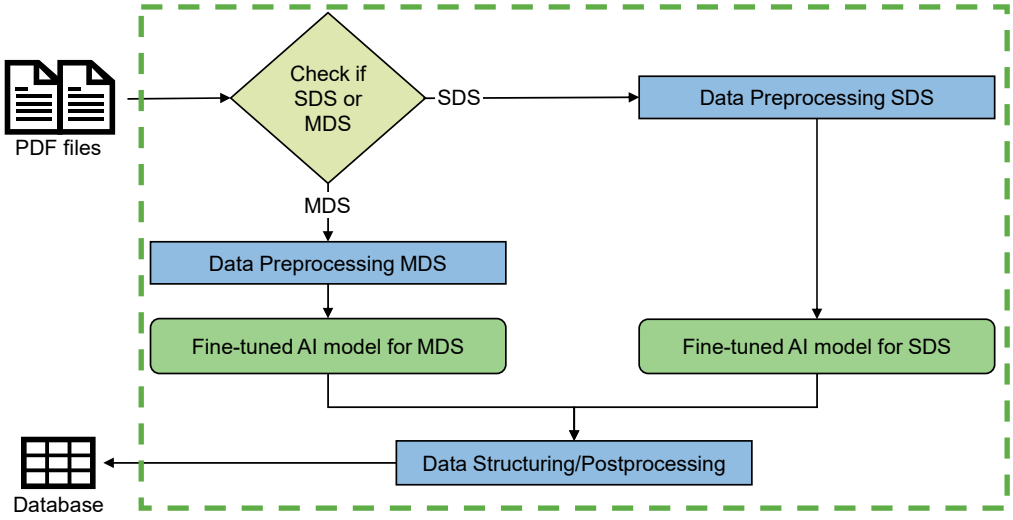

Figure 1. Structure for NLP-based extraction of material information from technical data sheets.

4.1. Structure and Training

Two model architectures were employed: SpaCy's transition-based NER system and BERT-based transformers. Both models are initialized with German language pretrained weights (de_core_news_sm for SpaCy, bert-base-german-cased and xlm-roberta-base for transformers). The original architecture is used, and only fine-tuning is applied to the existing models. The selection of German language models reflects the fact that material and SDSs in the European battery manufacturing supply chain are predominantly provided in German by suppliers in Germany, Austria, and Switzerland. Additionally, the majority of documents in our training dataset (96 of 160 documents) were in German. English documents (48 documents) were automatically translated to German during preprocessing to maintain consistent terminology. While this approach is optimized for German-language documents,

the methodology itself is language-agnostic. The pipeline can be adapted to other languages by substituting the language-specific pretrained models with appropriate alternatives.

The models are separately fine-tuned for SDS and MDS processing due to structural differences between document types. SDSs follow standardized regulatory templates with defined sections. Material data sheets exhibit highly variable formats and layouts across different suppliers. Additionally, SDS documents require extensive preprocessing to reduce text volume by 80 percent (from 26,000 to 3000 characters), whereas MDS documents in single- or double-page format require minimal preprocessing. These structural and preprocessing differences motivate separate model training. Training uses a 90/10 data split, with 144 documents for training and 16 documents for validation. Adam optimization and dropout regularization are applied. Early stopping prevents overfitting on the dataset of 59 MDS and 101 SDS files.

4.2. Text Extraction and Preprocessing

PDF documents underwent domain-specific preprocessing, which consists of various steps (Figure 2). Text was first extracted via pdfplumber with layout parameters set to an x_tolerance of 2 and a y_tolerance of 4. These parameters control the spatial tolerance for grouping text elements into lines and words. The values were selected through preliminary testing to optimize text reconstruction from the digitally created PDF documents in this study.

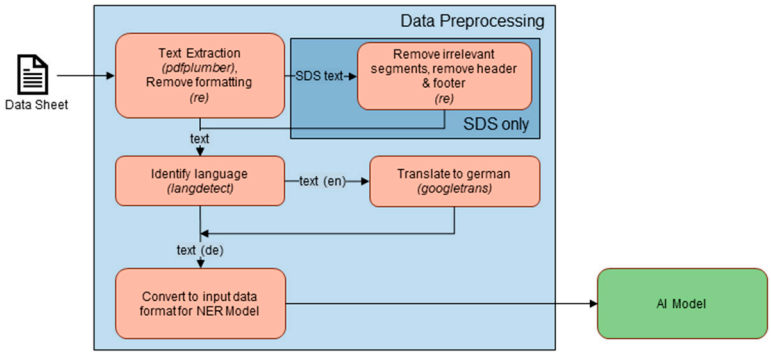

Figure 2. Process structure of text extraction and preprocessing.

For SDS documents, headers and footers were removed and relevant sections extracted using regex patterns. This reduces SDS text volume by 80% (from 26,000 to 3000 characters) while preserving key information. MDS documents do not require shortening due to their single- or double-page format. All documents in the dataset were digitally created PDFs. Optical character recognition was not required.

Language identification was performed for each data sheet to ensure uniform input terminology. The langdetect library was used to detect document language, achieving 100 percent accuracy on both MDS and MDS documents. English language documents (48 MDS and 48 SDS) were automatically translated to German using the googletrans library. This translation approach was selected because alternative transformer-based models (such as Helsinki-NLP, T5, and mBART-50) are limited to 512-token input sequences, which is insufficient for full SDS documents. Additionally, these models exhibit information loss in special entity categories such as CAS numbers and hazard classifications. Googletrans can process texts of arbitrary length with acceptable precision for material nomenclature. Translation was performed before text cleaning to preserve context and minimize errors. Some translation errors occur when document layout information is lost; however, these errors are unavoidable given the constraints of automatic translation and the absence of sufficient English-language training data to train separate English models.

4.3. Model Training with Parameter Optimization

The NER models are trained to recognize 10 entity types. These were selected based on their relevance to electrode slurry mixing process control and regulatory compliance requirements. The entity types are product name (PROD_NAME), manufacturer name (MANU_NAME), CAS number (CAS), hazard classifications (HAZ), melting point (MELT_POINT), density (DENSITY), moisture content (MOISTURE), pH value (PH), molecular weight (MOL_WEIGHT), and particle size (PARTICLE_SIZE). The available data was manually labeled by extracting property values directly from each document and entering them into corresponding cells in an Excel table. This manual approach was selected because single-annotator labeling with spreadsheet tools is efficient for this dataset size, whereas formal annotation tools such as Doccano or BRAT are more cost-effective for multi-annotator projects. All document text was searched for each property value. In cases where properties appeared multiple times or were not present, missing values were flagged as negative one (negative 1) to enable proper training data creation.

Hyperparameter optimization was performed using Optuna version 4.0.0 with a probabilistic model of the objective function. The optimization direction was set to maximize F1-score. The hyperparameter search spaces were defined as follows. The number of epochs was set to range from 10 to 100. Batch size was fixed at 8 due to the limited size of the training dataset. The learning rate was optimized in the range from 1 times 10 to the negative 5 power to 8 times 10 to the negative 5 power on a logarithmic scale, which is standard practice for BERT models. the dropout rate was optimized in the range from 0.05 to 0.35. Optuna performed 20 trials with pruning enabled to terminate unpromising trials early and conserve computational resources. The training process applied Adam optimization for SpaCy models and the standard Huggingface Trainer optimization for BERT models.

4.4. Postprocessing

The data sheet processing pipeline concludes with three postprocessing steps to transform raw model predictions into structured material databases. As the NER model processes entire text sequences, it often makes multiple predictions for the same property across different sections of a document. To resolve these conflicts, a frequency-based selection strategy is employed, retaining the entity value predicted most frequently for each material property. This approach is particularly important for CAS numbers, which may appear multiple times or resemble similar numerical formats. However, an exception is made for hazard statements, where all identified instances are consolidated into a single cell separated by spaces. This reflects the principle that comprehensive hazard documentation is essential for chemical safety.

The document paring of SDS and MDS is achieved by matching filenames, whereby an MDS is paired with an SDS if the filenames are identical except the ‘MDS’ suffix and case are ignored. Paired documents are merged using a hierarchical consolidation strategy, whereby only empty cells are filled in with values from the complementary document. If a cell already contains an extracted value, that entry is retained as authoritative, and any conflicting information from the paired document is discarded. This ensures data integrity and reduces the risk of information corruption. Once entity resolution and consolidation are complete, the processed data is structured and exported to a defined format like Excel, with each row representing a complete material record. This end-to-end architecture enables the rapid transformation of heterogeneous PDF folders into unified material property databases. This provides a standardized foundation for downstream analytical and process control applications, eliminating the need for manual data entry.

5. Results and Discussion

The NER models are trained to minimize training loss as the primary objective. During training, both SpaCy models demonstrate convergence behavior, where loss decreases rapidly in the first 10 epochs and then stabilizes. Figure 3 shows that the F1-score increases over the training epochs and reaches a plateau around epoch 30 for the SDS model and epoch 60 for the MDS model. The convergence of both loss and F1-score occurs concurrently, indicating successful model learning.

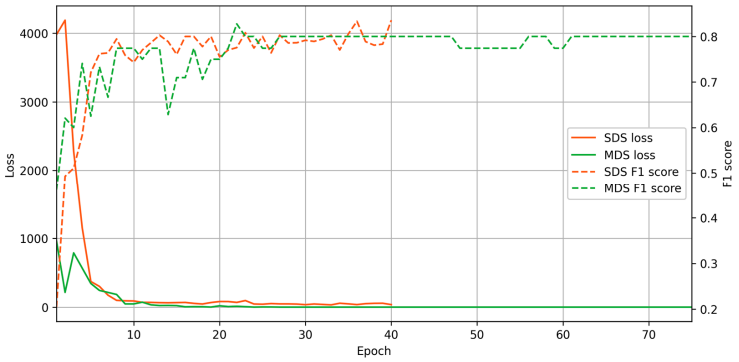

Figure 3. Loss and F1-score of the best SpaCy SDS and MDS models.

When training with small datasets, there is a general risk of overfitting. However, the absence of data leaks in our methodology mitigates this concern. The validation dataset is strictly separated from the training set, with no overlap. The F1-score improvements (see Figure 3) reflect true generalization to previously unknown validation data rather than the memorization of training examples. The parallel convergence of training loss and validation F1-score further supports that the models learn transferable patterns rather than exploiting dataset-specific artifacts.

The F1-scores of the different models are plotted over the trained epochs in Figure 4, and their corresponding parameter configurations are shown in Table 1. SpaCy models performed significantly better than BERT and RoBERTa models on the available training and evaluation dataset, reaching F1-scores of 0.836 and 0.8 (Figure 5). The poor performance of the BERT and RoBERTa models is likely due to the small dataset, which limits effective training and leads to suboptimal learning compared to the SpaCy models. Additionally, the task is highly domain-specific, involving specialized, non-standardized terminology and complex information structures commonly found in material and SDSs. These characteristics pose challenges for BERT-based models, which are pretrained on general language corpora and optimized primarily for contextual sentence understanding. SpaCy's NER architecture, on the other hand, is better suited for extracting information from semi-structured documents. It demonstrates higher data efficiency, enabling more robust learning from smaller datasets.

Table 1. Hyperparameters and F1-score of resulting models.

Model	Epochs Trained	Learning Rate	Dropout	F1-Score
SpaCy SDS model	40	1.955×10^{-6}	0.041	0.836
SpaCy MDs model	75	1.724×10^{-2}	0.055	0.800
BERT SDS model	66	2.157×10^{-5}	0.050	0.182
BERT MDS model	63	4.451×10^{-5}	0.063	0.242
RoBERTa SDS/MDS model [1]	-	-	-	0

[1] None of the models achieved an F1-score greater than zero.

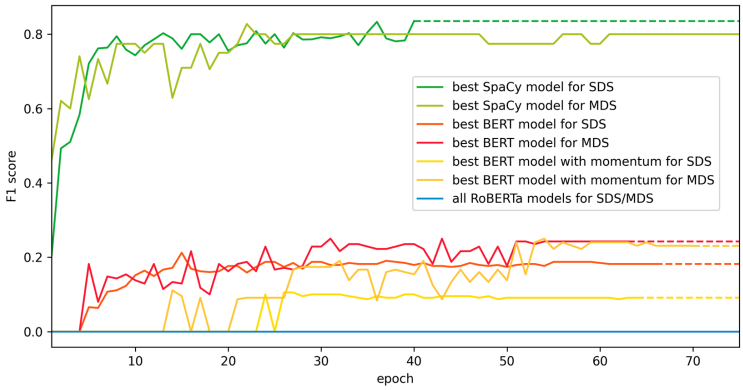

Figure 4. F1-scores of all models over training time.

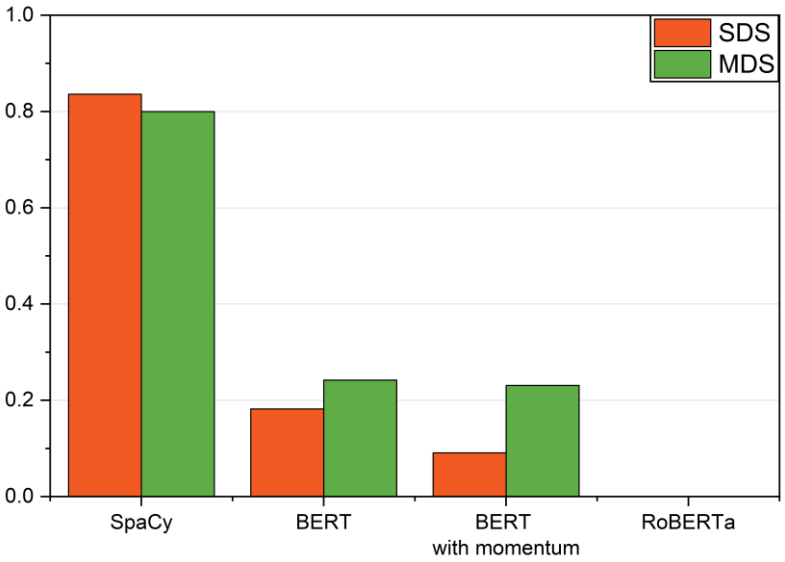

Figure 5. A comparison of F1-scores for the various NLP models, with SpaCy achieving the highest F1-score for both SDS and MDS.

The SpaCy NER models achieved a good F1-score, considering that in the comparison by Vychegzhanin and Kotelnikov [28], the evaluated models for the standard labels achieved a maximum F1-score of 0.887. In particular, the SpaCy NER model was tested as part of this comparison and for the labels PER, LOC, and ORG, the labels together resulted in an F1-score of 0.597. The stagnation of the F1-score can be explained by the reduction in new information that can be learned from the training data.

For the entire pipeline, the evaluation metrics for the resulting table of data sheet information are defined as follows:

- True positive (TP): cells in the actual output that contain the same (correct) value as the corresponding cells in the desired output;
- False positive (FP): cells that are empty in the desired output but contain a value in the actual output as well as cells of the actual output table that contain the wrong value;
- False negative (FN): cells that contain a value in the correct output but are empty in the actual output;
- True negative (TN): cells that are empty in both the desired output and the actual output.

Using BERT and RoBERTa models in the final extraction pipeline did not yield an F1-score greater than zero. The complete system using the SpaCy NER model for SDS and MDS results in an F1-score of 0.693 (0.844 cells are considered correct, which vary in format, and each hazard classification is considered a separate cell).

There were differences in how well different entities were recognized (Table 2 and Figure 6). The system performed best on extracting CAS numbers, presumably due to their frequency in the training data and strict format. The molecular weight and density were also extracted reliably, with an F1-score of 0.833 and 0.8. While CAS numbers follow their unique two-dash format, molecular weight and density both have unambiguous units that are used nowhere else in the document. If entities did not follow a unique or even a uniform format, extraction was less reliable, having the worst F1-scores for melting point, moisture content and particle size. In addition to the format of the entities of these classes varying, there were only a few entries for these entities in the training data. This poses a significant challenge. A very large dataset with thousands of material data sheets is expected to improve the model's performance. However, due to limited access to material data sheets from a small number of suppliers, the training data set was relatively small. Nevertheless, it was demonstrated that a pipeline for information extraction could be developed for material data sheets with a very heterogeneous structure and that it achieved good performance for parameters with clearly defined formats and/or unique identification features.

Table 2. System performance using the SpaCy NER model.

	TP	FP	FN	TN	Frequency	Precision	Recall	F1-Score
PROD_NAME	8	2	4	0	14	0.8	0.667	0.727
MANU_NAME	7	3	3	1	13	0.7	0.7	0.7
CAS	13	1	0	0	14	0.929	1	0.963
HAZ	3	6	0	5	10	0.33	1	0.5
MELT_POINT	0	0	9	5	10	0	0	0
DENSITY	6	3	0	5	9	0.667	1	0.8
MOISTURE	0	0	3	11	4	0	0	0
PARTICLE_SIZE	0	0	1	13	2	0	0	0
PH	2	0	2	10	5	1	0.5	0.667
MOL_WEIGHT	5	0	2	7	8	1	0.714	0.833

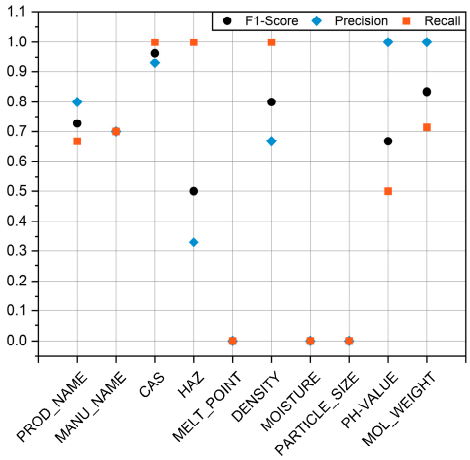

Figure 6. System performance using the SpaCy NER model for different entity types.

The domain-specific, NLP-based extraction pipeline shows that it is possible to extract information from unstructured data sheets. However, there are still limitations that require further research.

Several important limitations of the current work should be noted. The training dataset is small, comprising only 160 documents from a limited number of suppliers. This is primarily due to NDAs and intellectual property protections forced by material suppliers, resulting in a small dataset of publicly available data sheets. The size constraint directly impacts extraction performance for entities with low frequency in the training data, like melting point, moisture

content, and particle size. These entities appear infrequently in the available documents and exhibit high terminology variability across suppliers. Addressing this limitation would require access to thousands of material data sheets from diverse suppliers. Such datasets are rarely available due to the proprietary nature of supplier documentation.

The extraction performance for SDS documents is preprocessing-dependent. The text extraction and section-filtering process (reducing SDS documents by 80 percent) must be executed without error for the NER model to function effectively. If preprocessing fails to preserve relevant information or incorrectly removes important sections, the downstream entity recognition will suffer. Different SDS formats from different suppliers require customization of the preprocessing regex patterns. This makes the system less universally applicable than a format-agnostic approach would be.

Integration with existing manufacturing systems remains entirely future work. Extracted data must flow seamlessly into material specification databases, production planning systems, and quality management systems. This integration requires close collaboration between data science teams and manufacturing engineering teams. The technical infrastructure for such integration does not currently exist in most battery manufacturing facilities.

A hybrid human-in-the-loop approach is recommended for initial deployment rather than fully automatic decision-making. Operators would review and verify extracted values before those values influence production decisions. This staged approach reduces risk during the period when the system's reliability is still being established in real production environments. As the system demonstrates consistent performance over time, the verification burden could be gradually reduced.

Most critically, validation studies in actual manufacturing environments remain essential. This work demonstrates that automated extraction is technically feasible but does not demonstrate that extracted material data actually improves production outcomes. Pilot studies are needed to confirm that better material data translates to measurable benefits such as reduced scrap rates, improved cycle time, or improved electrode quality. Without such validation, the business case for deployment cannot be established.

6. Potential Future Industrial Applications of NLP-Based Information Extraction

The extraction pipeline developed in this work provides a structured data foundation for several industrial applications in battery manufacturing. It is important to note that this study focuses on demonstrating the technical feasibility of automated extraction. The downstream integration with production systems and validation of actual process improvements remain. This section outlines promising application directions for future work based on the extraction performance achieved in this work.

The greatest potential is seen here in process control and optimization, battery passport and traceability, and root-cause analysis, as these actively contribute to either regulatory requirements being met (battery passport) or cost savings and process optimization being achieved. These factors should not be underestimated, particularly during the start-up phase of a production line. High scrap rates, long ramp-up times, financing bottlenecks and cost overruns can all pose a threat to the economic success of a battery cell factory. This has been the case for some European companies [29].

6.1. Supplier Quality Management and Hazard Documentation

A primary opportunity lies in automating supplier quality management. The system reliably extracts CAS numbers with 92.9 percent precision and 100 percent recall (F1-score 0.963, Table 2). This high reliability enables the automated compilation of supplier material catalogs

without manual verification. Similarly, density extraction achieved an F1-score of 0.8 (Table 2), providing reliable capture of this critical parameter for batch specification comparison.

Extracted material parameters can quantify deviations from internal specifications and historical baseline values. Manufacturers could compare incoming material batches against specification ranges to detect inconsistencies between suppliers or between shipments from the same supplier. For example, particle size distribution could be automatically extracted and compared against tolerance ranges to flag non-compliant material before production. This enables data-driven decisions regarding material suitability without manual inspection of each data sheet. Quality indicators become machine-readable and directly comparable across documents.

Hazard classification extraction achieved 100 percent recall, meaning all hazard statements present in documents were identified. In a future implementation, this capability could automatically aggregate hazard information from multiple supplier documents for consolidated safety reviews. The system could flag material batches with critical hazards before processing starts. However, the current 33% precision indicates that false positives must be addressed before production deployment. This would require either improving model performance or implementing human-in-the-loop verification for hazard classifications.

6.2. Process Control and Batch-Specific Optimization

Automated extraction of material parameters enables future process simulation applications. The system extracted density and molecular weight reliably (F1-scores 0.8 and 0.833, respectively). These parameters could replace nominal supplier values in physics-based mixing simulations or hybrid data-driven models. Batch-specific simulation would enable operators to anticipate process deviations before they occur. However, several conditions must be met before real-world deployment. Integration pathways with manufacturing execution systems must be developed. Cost–benefit analysis should quantify whether automation reduces manual material specification effort sufficiently to justify implementation. Most importantly, validation studies with actual production data are required to confirm that extracted parameters actually improve mixing behavior or reduce scrap rates.

6.3. Regulatory Compliance and Battery Passport

Starting in February 2027, the European Battery Regulation will mandate digital battery passports documenting material origin and properties. Structured material data from automated extraction could provide the foundation for these regulatory documents. Our system successfully extracts material property values in a consistent format, eliminating the current manual compilation of information scattered across heterogeneous supplier documents. These structured material descriptions can then be integrated into digital process models, enabling automatically generated material profiles and allowing for the determination of the required values for the battery passport at the cell, module, and pack levels.

6.4. Root-Cause Analysis and Traceability

The extraction pipeline also provides a foundation for root-cause analysis by enabling the systematic linkage of supplier material parameters with production outcomes. For quality management, extracted parameters could enable automated comparison of incoming material batches against specification ranges. Manufacturers can identify how specific variations affect mixing behavior, slurry stability, coating uniformity, or electrode adhesion. Common defects such as poor wetting or delamination can be traced back to deviations in material properties documented in supplier data sheets. For knowledge management, structured material data enables rapid searching and comparison of materials with specific characteristics.

7. Conclusions

Material properties play an important role in battery cell production, as variations in characteristics such as particle morphology, size distribution, or purity can significantly affect downstream processes and cell quality. Extracting this information from supplier documentation remains a significant challenge. The core problem is nomenclature heterogeneity. Identical material properties are referred to by different terms across documents, making manual compilation of material specifications time-consuming and error-prone. Standard extraction approaches such as cloud-based language models and commercial tools are unsuitable for industrial battery cell production. This is because there are concerns regarding security, privacy, and data management. These methods are also not specialized enough to address the specific challenges at hand.

This paper addresses this gap by presenting a prototype NLP-based extraction pipeline specifically designed for heterogeneous supplier material and SDSs. This work is explicitly scoped as a technical feasibility study. The pipeline uses fine-tuned SpaCy models to transform unstructured PDF documents into structured material databases. Key parameters such as CAS numbers, molecular mass, and density are extracted with F1-scores above 0.7, demonstrating technical feasibility. The system achieves 92.9% precision for CAS numbers and 100% recall for hazard statements. These results show that domain-specific NER models trained on small datasets can effectively handle the nomenclature variability characteristic of supplier documentation.

The pipeline demonstrates several important findings. First, SpaCy outperforms larger transformer-based models on this domain-specific task with small training datasets. Second, structured entity extraction is achievable for properties with consistent formats and terminology. Third, parameters with variable formats or low training data frequency require substantially larger datasets to achieve reliable performance. Fourth, the complete end-to-end system provides a practical foundation for downstream applications despite these performance variations.

The structured material database created by this pipeline provides the data foundation necessary for data-driven approaches to battery manufacturing. By eliminating manual compilation of supplier information and providing machine-readable material properties, the system addresses a fundamental bottleneck in current production systems. While significant work remains to be done before full production deployment, this study demonstrates that automated extraction of material information from heterogeneous documents is both technically feasible and industrially valuable. The approach offers manufacturers a pathway toward data-driven decision-making based on supplier material documentation, with the potential to improve process stability, reduce scrap rates, enhance regulatory compliance, and enable more efficient resource use in battery cell production.

Advancing the model application toward production deployment requires addressing current limitations. Low-frequency entities require expanded training datasets. Preprocessing dependencies must be resolved through format-agnostic approaches. Integration with manufacturing execution systems must be developed. Human-in-the-loop verification should be implemented initially. Most importantly, pilot studies in real manufacturing environments must validate that extracted material data translates to measurable production improvements.

Author Contributions: Conceptualization, S.O.; methodology, S.O.; software, F.B. and S.O.; validation, F.B. and S.O.; formal analysis, S.O. and F.B.; investigation, S.O. and F.B.; writing—original draft preparation, S.O. and F.B.; writing—review and editing, S.O., F.B., S.S. and J.F.; visualization, F.B. and S.O.; supervision, J.F.; funding acquisition, J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by German Federal Ministry of Research, Technology and Space (BMFTR) for supporting the project "IntelliPast" (funding code: 03XP0343A).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code for the NLP-based extraction pipeline can be found at: https://github.com/OtteSimon/Data_Extraction_NLP/ (accessed on 6 May 2026). The datasets used in the study for the training and validation models are publicly available on figshare under https://doi.org/10.6084/m9.figshare.31841773.

Acknowledgments: We would like to thank Elsa Olivetti for hosting Simon Otte's research stay in her group at the Department of Materials Science and Engineering at the Massachusetts Institute of Technology (MIT) and for creating such a welcoming environment. The infrastructure, expertise and knowledge exchange were valuable to the work for this publication. Special thanks also go to Kevin Huang for his valuable support and technical input and for the enriching discussions during Simon's time there. We are also grateful for his ongoing assistance in the conception and revision of this manuscript. The research stay of Simon Otte at the Massachusetts Institute of Technology (MIT) was partly supported by the Karlsruhe House of Young Scientists (KHYS). This work contributes to the research performed at KIT-BATEC (KIT Battery Technology Center) and at CELEST (Center for Electrochemical Energy Storage Ulm Karlsruhe). During the preparation of this manuscript/study, the author(s) used DeepL Write (version 25.48, free of charge) for the purposes of checking spelling, grammar, and punctuation. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Li, J.; Fleetwood, J.; Hawley, W.B.; Kays, W. From Materials to Cell: State-of-the-Art and Prospective Technologies for Lithium-Ion Battery Electrode Processing. *Chem. Rev.* **2022**, *122*, 903–956. [CrossRef] [PubMed]
2. Kwade, A.; Haselrieder, W.; Leithoff, R.; Modlinger, A.; Dietrich, F.; Droeder, K. Current status and challenges for automotive battery production technologies. *Nat. Energy* **2018**, *3*, 290–300. [CrossRef]
3. Kim, M.; Yang, Z.; Son, S.-B.; Trask, S.E.; Jansen, A.; Bloom, I. Effect of cathode on crosstalk in Si-based lithium-ion cells. *J. Mater. Chem. A* **2021**, *9*, 26904–26916. [CrossRef]
4. Zhang, J.; Qiao, J.; Sun, K.; Wang, Z. Balancing particle properties for practical lithium-ion batteries. *Particuology* **2022**, *61*, 18–29. [CrossRef]
5. Liu, H.; Cheng, X.; Chong, Y.; Yuan, H.; Huang, J.-Q.; Zhang, Q. Advanced electrode processing of lithium ion batteries: A review of powder technology in battery fabrication. *Particuology* **2021**, *57*, 56–71. [CrossRef]
6. Feretzakis, G.; Verykios, V.S. Trustworthy AI: Securing Sensitive Data in Large Language Models. *AI* **2024**, *5*, 2773–2800. [CrossRef]
7. Zhao, G.; Song, E. Privacy-Preserving Large Language Models: Mechanisms, Applications, and Future Directions. *arXiv* **2024**, arXiv:2412.06113. [CrossRef]
8. Pekel, F.; Kalkman, G.; Lemcke, E.; van Stokkum, R.; Pronk, A.; Godderis, L.; Goossens, J.; de Raeve, H.; Coene, E.; Kuijpers, E. Implementing generative pretrained transformer models for text recognition tasks in safety data sheets. *Ann. Work. Expo. Health* **2026**, *70*, wxaf081. [CrossRef] [PubMed]
9. Balasubramanian, J.B.; Adams, D.; Roxanis, I.; de Gonzalez, A.B.; Coulson, P.; Almeida, J.S.; García-Closas, M. Leveraging large language models for structured information extraction from pathology reports. *J. Pathol. Inform.* **2025**, *19*, 100521. [CrossRef] [PubMed]
10. Yashwant, S.; Dubey, A.; Paikray, P.; Thulsiram, G. Invoice Information Extraction: Methods and Performance Evaluation. *arXiv* **2025**, arXiv:2510.15727. [CrossRef]
11. Kumar, A.; Starly, B.; Lynch, C. *ManuBERT: A Pretrained Manufacturing Science Language Representation Model*; Elsevier: Amsterdam, The Netherlands, 2023.
12. Mahbub, R.; Huang, K.; Jensen, Z.; Hood, Z.D.; Rupp, J.L.M.; Olivetti, E.A. Text mining for processing conditions of solid-state battery electrolytes. *Electrochem. Commun.* **2020**, *121*, 106860. [CrossRef]

13. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [CrossRef] [PubMed]
14. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2009.
15. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2026.
16. Labusch, K.; Neudecker, C.; Zellhöfer, D. BERT for Named Entity Recognition in Contemporary and Historical German. In Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), Erlangen, Germany, 9–11 October 2019.
17. Honnibal, M.; Montani, I. spaCy · Industrial-Strength Natural Language Processing in Python. 2024. Available online: https://spacy.io/ (accessed on 10 July 2024).
18. Honnibal, M.; Boyd, A.; Warmerdam, V.D. Compact Word Vectors with Bloom Embeddings · Explosion. 2022. Available online: https://explosion.ai/blog/bloom-embeddings (accessed on 21 August 2024).
19. Ammar, W.; Peters, M.E.; Bhagavatula, C.; Power, R. The AI2 system at SemEval-2017 Task 10 (ScienceIE): Semi-supervised end-to-end entity and relation extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*; Bethard, S., Carpuat, M., Apidianaki, M., Mohammad, S.M., Cer, D., Jurgens, D., Eds.; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 592–596.
20. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
21. Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://openai.com/index/language-unsupervised/ (accessed on 29 May 2024).
22. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
23. Swain, M.C.; Cole, J.M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894–1904. [CrossRef]
24. Olivetti, E.A.; Cole, J.M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T.Y.-J.; Hiszpanski, A.M. Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **2020**, *7*, 041317. [CrossRef]
25. Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E.A.; Ceder, G. Opportunities and challenges of text mining in materials research. *iScience* **2021**, *24*, 102155. [CrossRef] [PubMed]
26. Stepe, A. astepe/sds_parser. 2024. Available online: https://github.com/astepe/sds_parser (accessed on 20 May 2024).
27. Suman, A.; Khan, M.; Talreja, V.; Penfield, J.; Crowell, S. A machine learning driven automated system for safety data sheet indexing. *Sci. Rep.* **2024**, *14*, 4415. [CrossRef] [PubMed]
28. Vychegzhanin, S.; Kotelnikov, E. Comparison of Named Entity Recognition Tools Applied to News Articles. In Proceedings of the 2019 Ivannikov Ispras Open Conference (ISPRAS), Moscow, Russia, 5–6 December 2019; pp. 72–77. [CrossRef]
29. Murray, C. 'Over 50% Scrap Is Not Sustainable': European Battery Industry Looks to Learn from Mistakes of the Past. Energy Storage News. 22 May 2025. Available online: https://www.energy-storage.news/over-50-scrap-is-not-sustainable-european-battery-industry-looks-to-learn-from-mistakes-of-the-past/ (accessed on 20 April 2026).