Available online at www.sciencedirect.com

Chemical Engineering Research and Design

journal homepage: www.elsevier.com/locate/cherd

A text mining framework for screening catalysts and critical process parameters from scientific literature - A study on Hydrogen production from alcohol

Avan Kumar[a], *Swathi Ganesh*[b], *Divyanshi Gupta*[a], *Hariprasad Kodamana*[a,c,*]

[a] *Department of Chemical Engineering, Indian Institute of Technology Delhi, 110016, India*
[b] *Department of Chemical Engineering, Indian Institute of Technology Madras, 600036, India*
[c] *Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, 110016, India*

ARTICLE INFO

Article history:
Received 1 March 2022
Received in revised form 3 May 2022
Accepted 16 May 2022
Available online 25 May 2022

Keywords:
Catalyst
Process parameter
LDA
Hydrogen
Alcohol
NLP
SciBERT
Classification
NER
Ex-SciBERT

ABSTRACT

Hydrogen production is an active area of research with a vast amount of available scientific literature. However, this data is unstructured and scattered, making its utilization difficult from an academic and industrial point of view. This work aims to develop a recommendation system to identify optimal process conditions and catalyst information using Natural Language Processing (NLP) tools. To this end, full-text articles were extracted using the Elsevier API key followed by a custom XML parser. Latent Dirichlet allocation (LDA) was applied on this dataset to form clusters of topics. The experimental section of each article is annotated using state-of-the-art sentiment analysis techniques and divided into four categories based on the presence of catalyst and process information. This dataset is used to develop a dedicated NLP model, "Ex-SciBERT" by performing transfer learning on the "Sci-BERT" model. This model performs classification followed by Named Entity Recognition (NER) to extract catalyst and process parameters. Ex-SciBERT model produces an accuracy score of 0.915 (train dataset) and 0.890 (test dataset) for the classification of sentences task and an excellent accuracy score of 0.998 (train dataset) and 0.997 (test dataset) for the NER task. Deployment of this model will automate and accelerate the screening of relevant information from literature by reducing manual efforts.

1. Introduction

The world's energy requirement is increasing tremendously with the increase in population and energy consumption (Wang et al., 2016). More than 80% of world energy requirements are currently produced through fossil fuels (Nguyen et al., 2020). Excessive use of these non-renewable sources has led to depletion and exhaustion of these sources and substantial environmental pollution. One of the cleaner alternatives to traditional energy resources is the transformation to a hydrogen energy-based ecosystem (Searmsirimongkol et al., 2011; Khor et al., 2022; Torkian et al., 2022). Hydrogen is a clean fuel and energy resource with high energy conversion, large storage capacity, zero-emission and reliability. It is also environmentally friendly, renewable, and widely available. Hydrogen is considered one

[*] Corresponding author at: Department of Chemical Engineering, Indian Institute of Technology Delhi, 110016, India.
E-mail address: kodamana@iitd.ac.in (H. Kodamana).

https://doi.org/10.1016/j.cherd.2022.05.018

of the most promising energy sources of the future (Susanti et al., 2014; Zhang et al., 2021; Lorenzut et al., 2011; Dosado et al., 2015). To this extent, the identification of novel reaction routes and catalysis pathways is the key to the ongoing transition in energy from the use of fossil fuel (Lee et al., 2021; Akhoondi et al., 2021). Researchers are actively working in this area to identify optimal operating conditions, catalyst composition, support materials, etc. (Hojjati-Najafabadi et al., 2022; Mansoorianfar et al., 2022; Hojjati-Najafabadi and Salmanpour, 2021a), that give maximum yield with minimal energy investment. A quick check on Google Scholar with the keyword “Hydrogen production from alcohol” results in scientific articles around 3.5 million in number and numerous patents, which will only increase over time.

As a potential driver of future energy security, it is crucial that the salient research in this direction is abstracted and a recommendation system built for quick reference. The critical information regarding optimal synthesis routes, catalysts, and support characteristics is available as text. This salient summary of these scientific research articles will aid industries and the research community. This information is scattered in various formats such as books, journals, and other articles. A large part of this information is also in unstructured data such as images, tables and plots (Venugopal et al., 2021; Kaur and Chopra, 2016; Hojjati-Najafabadi et al., 2021b; Vaucher et al., 2020). Also, it is cumbersome to manually go through all the available resources and find the relevant information. It also requires expertise in the field to make sense of the information distributed in various sources.

In this context, recent advancements in machine learning (ML), particularly natural language processing (NLP), pave the way to automate this process, derive useful information from unstructured data, and yield decision enabling information. Recent advancements in ML and NLP provide a way to comprehend text data and carry out specific downstream tasks on the data. These tasks include text comprehension, querying, and knowledge extraction from given documents (Balyan et al., 2017; Zhang et al., 2018). In particular, the functionalities such as (i) topic modelling and clustering, (ii) text classification, and (iii) named entity recognition (NER), a task to locate and classify named entities mentioned in unstructured text into predefined categories, are beneficial in this context (Vo et al., 2022).

Topic modelling of text corpus is a class of unsupervised NLP techniques used to categorise a text document into topics. It is used to uncover the semantic structures and themes of the papers and understand the underlying information in the text (Jacobi et al., 2016). Latent Dirichlet allocation (LDA) is one of the most popular topic modelling and discovery models. It is a generative probabilistic model relating to the distribution of topics and words in text corpora (Blei et al., 2003; Nikolenko et al., 2017). This three-level hierarchical Bayesian framework models each collection item as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities (Blei et al., 2003; Jelodar et al., 2019).

As it stands, we can clearly see that, vast amount of information is available in the form of nonstructural data such as text, plots, etc. NLP tools could help us to extract valuable information from the related literature databases. To this end, there have been various attempts to develop deep neural network models for NLP applications and several of them suffers from various drawbacks. The Bidirectional RNN (Bi-RNN) is a computationally slower NLP model (Liew (2021)). One of the primary drawbacks of text classification in CNN is that it is impossible to know whether a text classifier’s class prediction is true and human intervention may be required (Jogin et al., 2018). Another model, the Bi-LSTM, has two LSTM cells, so it is expensive to utilise on a large scale (Bhuvaneshwari et al., 2022). Various researchers have reported that training of the conditional random field (CRF) is challenging and it is hard to get a good prediction (Inatani et al., 2014). Hence, many researchers have utilized the hybrid version of these models, such as Bi-LSTM+CRF, CNN+Bi-RNN, BERT+Bi-LSTM+CRF and many more (Hua et al., 2018; An et al., 2022; Ma et al., 2022; Li et al., 2022). The introduction of Bi-directional encoder representation of transformers (BERT) by Google in 2018 (Devlin et al., 2018) has largely transformed the landscape of NLP problems.

BERT is trained on large text corpora such as Wikipedia, books, and journals and exhibited higher performance on question-answer and classification tasks, etc. (Zhou et al., 2016). There are two classes of the BERT model based on the size of decoder-encoder (i) $BERT_{large}$ (total parameters = 340 M) and (ii) $BERT_{base}$ (total parameters = 110 M) (Devlin et al., 2018). However, it has been shown that, even a well pre-trained model such as $BERT_{large}$ could not perform well on scientific literature due to the specific vocabulary and sentence semantics pertaining to the scientific literature. Hence, BERT is trained on scientific literature specifically to learn its semantics yielding to the ‘SciBERT’ model (Beltagy et al., 2019)). Now there are various domain-specific BERTs are available to learn and process domain-specific vocabularies and terminologies. For instances, BioBERT (Lee et al., 2020), a biomedical-specific language model, clinicalBERT (Alsentzer et al., 2019) trained on 2 million clinical notes in the MIMIC-III v1.4 database ; (Johnson et al., 2016; Afzal et al., 2022) mBERT (Libovicky` et al., 2019) for multilingual machine translations tasks, patentBERT (Lee and Hsiang 2019) for patent classification, and FinBERT for financial NLP tasks (Araci 2019).

It is interesting to note some of the related works employing NLP tools with applications to process development as well. Villarreal et al. have been shown to extract the synthesis parameter for NiMO sulfide catalyst by using the “ChemDataExtractor” tool (Villarreal and Villarreal 2019). Koripelly et al. (2020) presented a promising parser-based approach for extracting relations from scientific papers based on polymer synthesis (Koripelly et al., 2020). Court and Cole (2018) built the database containing chemical entities by utilising the “ChemDataExtractor” on 68,078 chemistry and physics articles (Court and Cole 2018). Bass et al. have shown the classification of environment regulation sentences with the BERT neural network (Bass et al., 2019). Copara et al., (2020) have worked on NER on chemical patents with BERT, CRF, CNN, ChEMUBERT and ensemble models (Feng et al., 2021; Copara et al., 2020) modelled classification based on hazardous levels from HAZOP data by the BERT model (Feng et al., 2021). Some recent works introduced an attention-based deep learning technique for NER task for material science literature (Yang and Hsu, 2021; Trewartha et al., 2022). Kuniyoshi et al. (2020) reported a synthesis procedure for all-solid-state batteries with a literature corpus made of 243 articles (Kuniyoshi et al., 2020).

From the critical analysis, it has been clear, to the best of the authors' knowledge, that hardly any attempt has been made to apply NLP tools to extract decision enabling information from the large corpora of scientific literature in the area of chemical engineering. Chemical engineering is similar to the other engineering fields in terms of domain-specific terminology and the non-uniform format of the literature. These similarities reveal the potential of NLP in chemical engineering and how a domain-specific NLP model can be used to screen details of critical process parameters and catalyst materials by exploring the literature. In this regard, in the current study, we try to examine LDA and SciBERT for the topic modelling and specific tasks (classification and NER) for scientific literature data based on the keyword 'Hydrogen production from alcohol'. In particular, we employ transfer learning of SciBERT on text data comprising a corpus of 5901 scientific literature, specifically focusing on catalyst and process conditions text data, and develop a new NLP model termed Extended SciBERT (Ex-SciBERT). We believe that this framework would eventually aid in accelerated frameworks for screening the catalysts.

The remainder of the article is structured as follows. A detailed discussion of data collection and pre-processing is given in Section 2. Section 3 discusses the text mining tools utilised, and Section 4 talks about the proposed model "Ex-SciBERT". Section 5 presents results and discussion, and section 6 provides conclusive remarks.

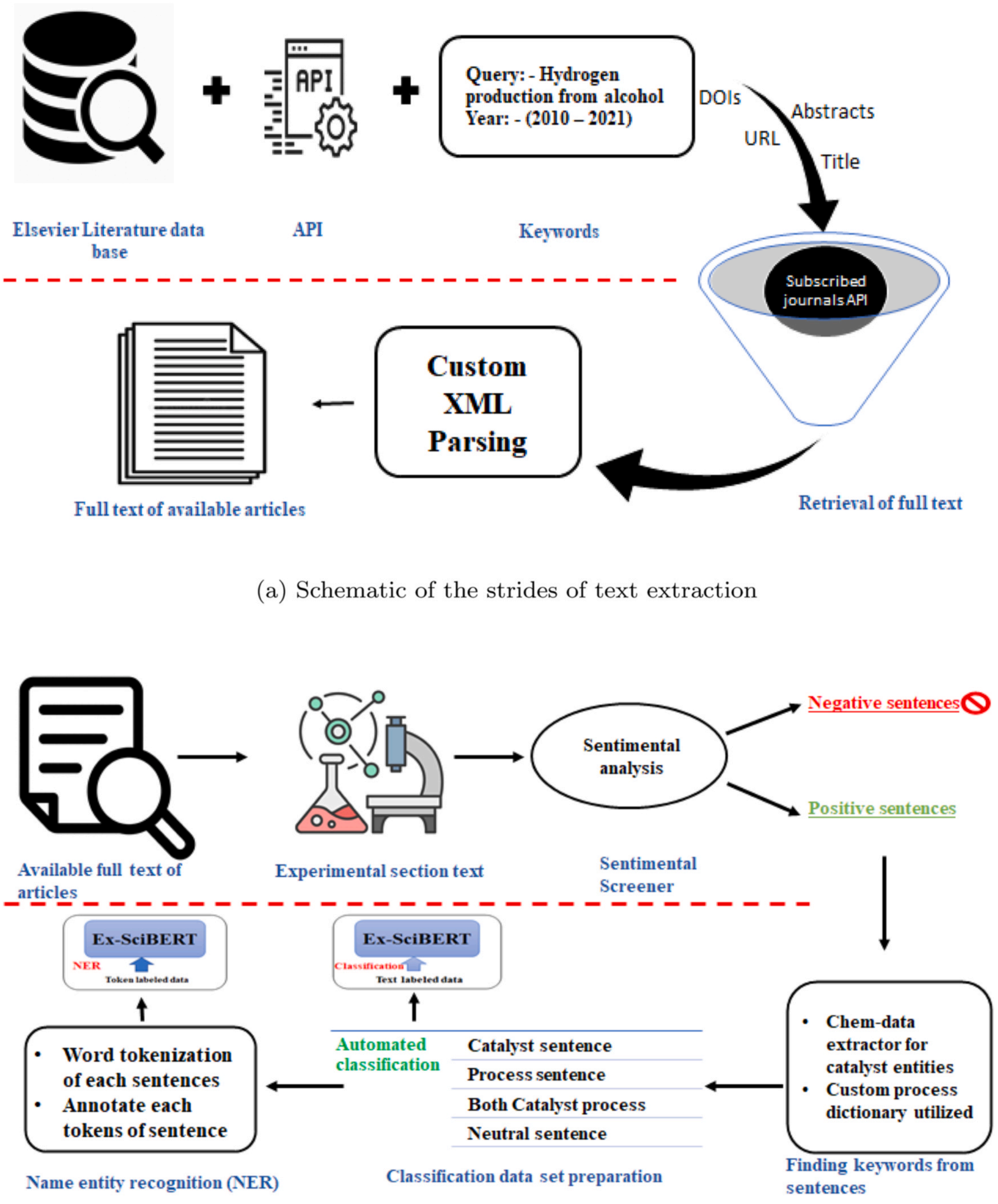
2. Data collection and pre-processing

2.1. *Data extraction*

The first requirement for developing an NLP model is the representative text data corpus. We have used the "Elsevier Science Direct" application programming interface (API) to extract the same. This API searches the entire Elsevier database for literature related to the keyword. We used the keyword 'Hydrogen production from alcohol' and mined scientific literature for ten years from 2010 to 2021. This query returned a list of 5945 DOIs, around 5901 abstracts, and the titles and URL links. Available DOIs are utilised to extract the full text of the articles using the "Elsevier Science Direct" API associated with the institute token linked to IIT Delhi. After passing the above list, we could retrieve 5901 full texts of scientific literature. The various data extraction steps are shown with a pipeline in Fig. 1 (a). At the end of this operation, we have retrieved extensible markup language (XML) formatted output. We have utilized the custom XML parser as we need to preferentially extract certain sections of the scientific literature. We have developed the XML parser using commands from ElementTree XML API module by Python. The parser iterates through the text identifying relevant sections of data including author, publication details, abstract, introduction, experimental/ methodology and result sections separately and stores them in CSV format for easy access. This data is further used for preprocessing and vectorization. For the development of Ex-SciBERT, we have focused mainly on experimental section of scientific literature.

We require annotated datasets to train the SciBERT for specific tasks such as sentence classification and NER. We have entirely utilised the retrieved text files to develop these datasets. We have extracted the "Experimental section" from the full text of the articles as it contains the details of the catalyst, chemicals, and process conditions. In cases where the experimental section of the article is clearly not defined by a heading or is split into multiple sub-sections, the XML parser identifies all parts after "Introduction" and before "Results and Conclusion" as relevant and stores them. Then, an NLP library, namely, 'NLTK', is utilised to perform sentence tokenisation of paragraph. These sentences are further processed using "FLAIR", state of the art pretrained NLP framework to predict the positive and negative sentiment of the sentences based on the keyword "Catalyst" (Visser and Dunaiski, 2022; Akbik et al., 2019). The sentences with positive sense are processed further, and sentences with negative sense are discarded as they are not helpful for annotation. A sentence with positive sentiment here means that any sentence that talks about catalyst, process or process parameters positively in context of process synthesis. It is expected that they yield more information than a negative sentiment sentence. We have employed Flair sentimental analysis toolbox to perform and segregate positive and negative sentiment sentences. Further, every sentence is analysed by the "ChemDataExtractor" library (Swain and Cole 2016), it extracts all chemical entities from every sentence, and these entities are stored. The corpus of chemical entities is further subjected to a manual screening, where we have excluded all non-catalyst entities. Various steps for annotation are schematically presented in Fig. 1(b). For representative purposes, the frequently occurring chemical name of catalysts are shown in Fig. 2(a) as Zipf's plot, where the y-axis depicts the frequency, and the x-axis shows the rank of the catalyst. In Fig. 2(b), the histogram plot, the y-axis is represented by the log of frequency of catalyst. Further, a unique catalyst material word cloud is shown in Fig. 2(c) to illustrate words that frequently appear in the source text graphically.

The verified unique catalyst material list is utilised to label the positive sentences. If any catalyst entity is present in the sentence, it is numerically labelled as '1′ and categorised as a catalyst sentence. Similarly, if any positive sentence with the process parameter content is labelled as a process sentence, '0′ is assigned. This is done with the help of a custom made dictionary of process parameters such as pressure, temperature, etc. After that, we screen out all sentences containing catalyst and critical process parameters in a single sentence. This category is labelled as 2. To make the dataset representative, we have included some sentences with neither catalyst materials nor process parameters in the corpus and labelled them as 3. The total number of annotated sentences is 19,083; where the respective category-wise details are captured in the Table 2, corresponding to the train and test set. The dataset thus prepared are manually cross-checked to verify the sanity. Further, we feed the dataset to the Ex-SciBERT model for


Elsevier Literature data base
API
Query: - Hydrogen production from alcohol
Year: - (2010 – 2021)
Keywords
DOIs
URL
Abstracts
Title
Subscribed journals API
Custom XML Parsing
Full text of available articles
Retrieval of full text
Available full text of articles
Experimental section text
Sentimental analysis
Negative sentences
Positive sentences
Sentimental Screener
Ex-SciBERT
NER
Token labeled data
Ex-SciBERT
Classification
Text labeled data
Word tokenization of each sentences
Annotate each tokens of sentence
Automated classification
Catalyst sentence
Process sentence
Both Catalyst process
Neutral sentence
Chem-data extractor for catalyst entities
Custom process dictionary utilized
Name entity recognition (NER)
Classification data set preparation
Finding keywords from sentences

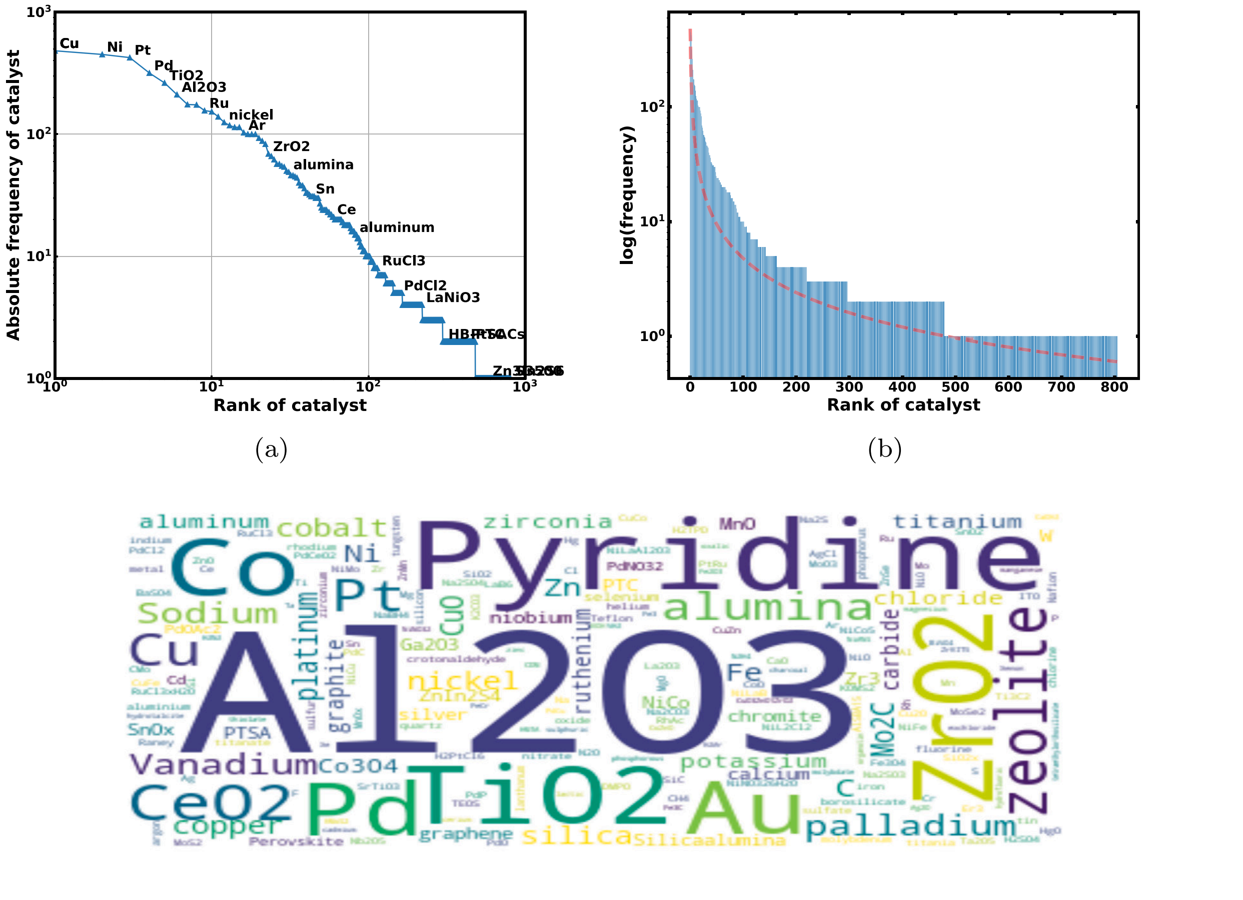
(a) Schematic of the strides of text extraction

(b) Data annotation pipeline

Fig. 1 – Data annotation for classification and NER.

classification and training in the transfer learning framework.

After annotating datasets for sentence classification, we proceed to annotate datasets for NER. After the first level of annotation, we have three buckets of sentences containing only catalyst, only process conditions, and both catalyst and process conditions information. All verified categorical datasets are further processed for token labelling. Each sentence is tokenised by the word tokeniser of the 'NLTK' library. We have developed a python script to differentiate between non-catalyst, catalyst and process tokens (words), labelled as '0', 'catalyst' and 'process', respectively. This dataset is utilised to perform transfer learning of the Ex-SciBERT model for the NER task. This dataset for the NER task has three unique labels as '0', 'catalyst', and 'process' and their category-wise details are collated in the Table 2, where it is mentioned with train and test set.



(a)

(b)

(c)

Fig. 2 – The plots represent the catalyst frequency that exists in the data by employing (a) Zipf plot, (b) histogram plot, and (c) a catalyst cloud image of all chemical entities.

Table 1 – This table collated all necessary information of all nine topics extracted by LDA.

Topic Number	Number of Documents	Top 10 most dominant words
0	31	coal, compound, sodium, wine, clay, biofilm, sugar, fruit, food, resistant
1	623	hydrogen, reaction, temperature, gas, high, ethanol, yield, product, increase, co
2	527	pva, film, alcohol, property, show, polymer, increase, result, bond, study
3	301	alcohol, activity, induce, cell, enzyme, level, increase, ethanol, effect, study
4	1588	catalyst, reaction, hydrogenation, catalytic, high, alcohol, co, support, selectivity, activity
5	941	hydrogen, tio, photocatalytic, production, high, evolution, activity, photocatalyst, performance, composite
6	786	production, hydrogen, yield, concentration, high, fermentation, rate, produce, substrate, study
7	246	membrane, alcohol, model, water, temperature, solvent, system, parameter, mixture, study
8	858	hydrogen, fuel, production, energy, process, technology, review, system, high, produce

3. Text mining tools

3.1. LDA

LDA is a topic modelling technique with the underlying basic idea that documents are represented as random mixtures over latent topics, where a distribution over words characterises each topic. The topics are represented in an n-dimensional simplex (segment for two topics, triangle for three, tetrahedron for four and so on, etc.) with α and β as corpus level parameters of Dirichlet distributions, assumed to be sampled once in the process of generating a corpus.

Given a corpus D consisting of M documents, K number of topics, with N_d words in a document d. First, multinomial distributions ϕ (probability of word w occurring in topic k) for topic and θ (probability of topic k occurring in document d) for documents are chosen from dirichlet distributions for topic-words (with parameter β) and document-topics (with parameter α). Then, for each word positions in document d, topic z_n from θ and word w_n from ϕ are selected. The variables θ_d are document level variables, sampled once per document. z_{dn} and w_{dn} are word-level variables sampled once for each

Table 2 – The summary of annotated data points utilisation in train and test datasets for Ex-SciBERT transfer learning.

Type of data	Name of label	Counts
Classification		
Train	Process (0)	10,595
	Catalyst (1)	1,846
	Both catalyst and process (2)	493
	Neutral (3)	2,332
Total	15,266	
Test	Process (0)	2,677
	Catalyst (1)	455
	Both catalyst and process (2)	123
	Neutral (3)	562
Total	3,817	
NER		
Train	0	461,242
	Catalyst	3,466
	Process	17,142
Total	481,850	
Test	0	114,833
	Catalyst	903
	Process	4,216
Total	119,952	

word in each document. These words are combined to obtain a document, and the process is repeated multiple times to create a corpus. The corpus thus obtained is compared to the original one to find an arrangement of points in distributions that maximise the likelihood of similarity between LDA generated corpus and the original set of documents, as shown by eq. (1).

$$P(:\quad, =, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P\left(W_{j,t}|\varphi_{Z_{j,t}}\right) \tag{1}$$

3.2. SciBERT

Annotated unstructured text datasets are fed as input to the BERT neural network (Devlin et al., 2018). BERT is a class of recurrent neural networks (RNN) that can handle sequential input data. However, a key difference with traditional RNN is that this network can process text data in any order and is not necessarily fully sequential due to the embedded attention mechanism. As a result, BERT neural nets can identify the contextual meaning of each word in the sentence. The BERT model is well trained on a more extensive corpus of the vocabulary of books, Wikipedia, etc. but not on any scientific literature corpora. Hence, researchers retrained the BERT on scientific literature from scratch to yield SciBERT, which stands for scientific BERT. Here, the vocabulary differs from BERT's vocabulary and contains words from computer science and biological literature. However, the literature corpora on which SciBERT is trained do not include catalyst and process condition data. As a result SciBERT, as an NLP model, still is not helpful in our context. To address this lacuna, we retrain SciBERT to yield a new NLP model termed "Ex-SciBERT" in a transfer learning framework to learn the text data relevant to catalyst and process conditions obtained by curating dataset obtained with the keyword "Hydrogen production from alcohol" as prescribed in Section 2 (see Fig. 3(a)).

4. The proposed model: Ex-SciBERT

The "Ex-SciBERT" model is developed based on the $BERT_{Base}$ architecture, as shown in Fig. 3(b). It has 12 layers of encoders, and each element of the encoder is also shown in Fig. 3(b). The first layer of the encoder can be fed with a maximum of 512 tokens. We have retrained the weights of the last layer in the SciBERT model to obtain the Ex-SciBERT model, using the pre-processed data as given in Section 2. For the following tasks: (a) Classification of sentences (b) NER.

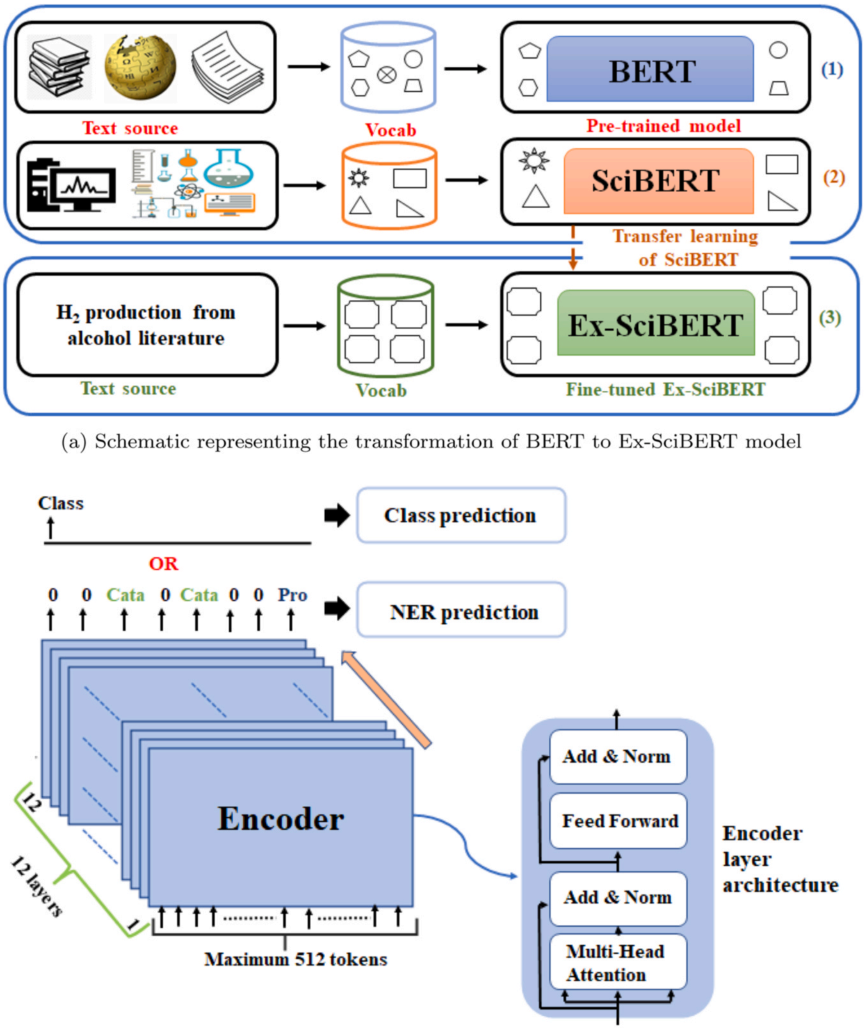
4.1. *Transfer learning framework*

We have two significant datasets after the data collection and data pre-processing. First for the classification task and second for the NER task. The summary of datasets is presented in the Table 2. Training the SciBERT from scratch requires enormous computation power and massive data. Hence, the SciBERT model is fine-tuned for specific tasks with the unique data source prescribed above. While training on a new dataset, it optimises the last layer's weights without altering the weights of other layers (Agrawal et al., 2022). Some crucial parameters to be specified during transfer learning are maximum sequence length, learning rate, validation fraction, and the number of training epochs. The training is performed on a compute cluster having 32 GB GPU in the "PyTorch" machine learning framework in the Python environment. For the best performance we have optimized the following hyperparameters: maximum sequence length = 256, learning rate = 0.0003, attention dropout = 0.1, hidden activation= "gelu", hidden dropout = 0.1, hidden size= 768, maximum position embedding 512, number attention heads = 12, and vocabulary size = 28,996 and loss function = "Cross-entropy loss". Similarly, for the NER task, we have employed the same set of hyper parameter values except the learning rate = 0.00003 and maximum sequence length = 178. While training the model, 10% of data is randomly utilised per epoch as validation data to avoid overfitting. Also, to avoid over-fitting early stopping criteria are applied. The overall algorithmic pipeline employed in the study is presented in the Fig. 4.

5. Result and discussion

5.1. *Topic modelling using LDA*

For LDA analysis, 5901 abstracts obtained from Elsevier search API are cleaned to remove punctuation, unnecessary spaces and stop-words. The text corpora are further lemmatised to remove inflectional endings and get the base form of words or lemma. The sentences are then converted to lowercase and tokenised into a list of words. These tokens are then checked for stopwords and lemmatisation and made into trigrams. Before modelling with LDA, we



(a) Schematic representing the transformation of BERT to Ex-SciBERT model

(b) The architecture of $Ex-SciBERT$ based on the $BERT_{base}$ architecture with 12 layers of encoders, a description of encoder

Fig. 3 – SciBERT and Ex-SciBERT.

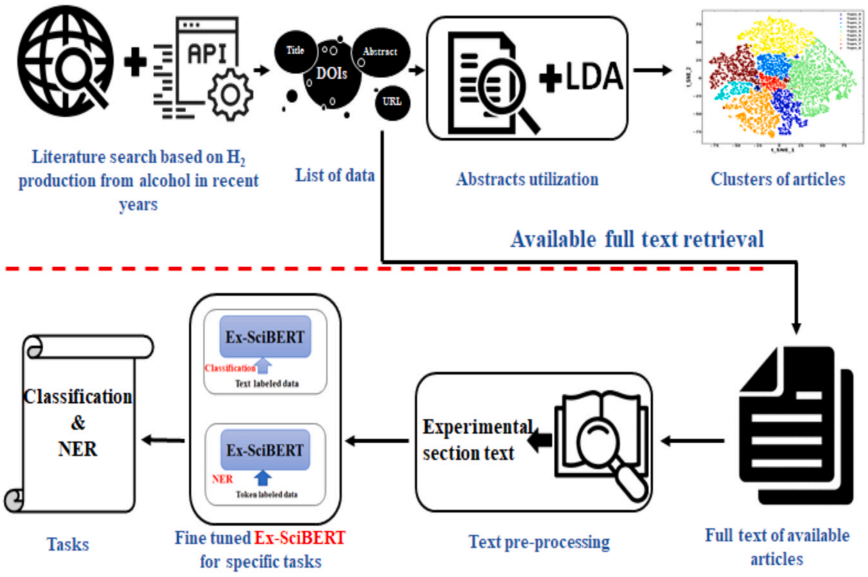
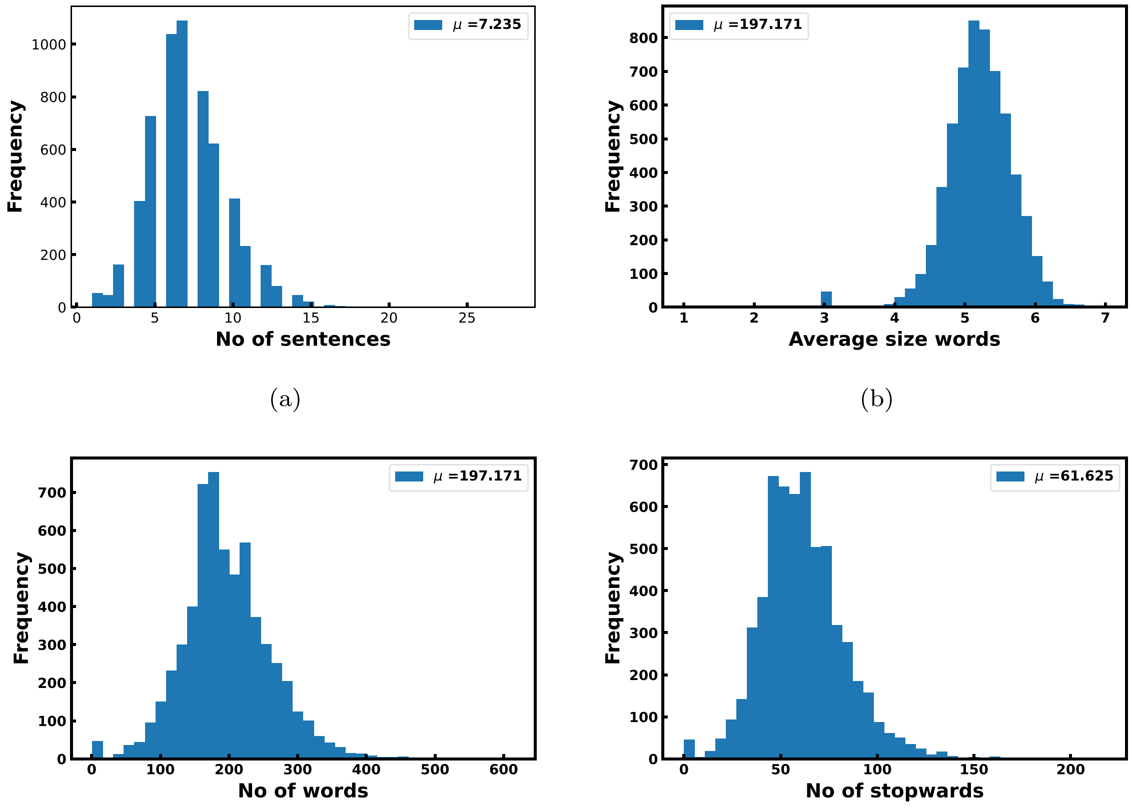

API
Title
DOIs
Abstract
URL
+LDA
Literature search based on H_2 production from alcohol in recent years
List of data
Abstracts utilization
Clusters of articles
Available full text retrieval
Classification & NER
Ex-SciBERT
Classification
Text labeled data
Ex-SciBERT
NER
Token labeled data
Experimental section text
Tasks
Fine tuned Ex-SciBERT for specific tasks
Text pre-processing
Full text of available articles

Fig. 4 – Schematic of the entire pipeline followed in the study.



(a)

(b)

(c)

(d)

Fig. 5 – The histogram plots and statistical analysis of some key parameters are shown, (a) Number of sentences, (b) Average length of words, (c) Number of words and (d) Count of stop words present in each abstract with respective mean of the distribution.

investigated the number of words present in each abstract, the average length of words, how many sentences are used in it, and at last, counted the number of stop-words. The histogram plots depict these key parameters, as shown in the Fig. 5. Gensim library in the python framework is used to create two main inputs of LDA, the corpus and the dictionary. Gensim creates unique identifiers for each word in the document giving a corpus that is a mapping of word identifier to its frequency. Gensim is again called to build the base model for LDA, which is further used to optimise the following hyperparameters: maximum coherence score (Röder et al., 2015), the number of topics (K), and the Dirichlet distribution parameters, α and β (Minka, 2000; Chang et al., 2009). The optimized parameters were found to be number of topics, $K = 9$ (topic numbers from 0 to 8), $\alpha = 0.61$, $\beta = 0.31$ with a Coherence score of 0.489087. The plot of coherence score vs the various number of topics can be seen in Fig. 6(b). Further, the maximum number of documents was represented by topics 4 and 5, as seen in Fig. 6(a). These are described in the form of a word cloud in Fig. 6(d), and visualisation of topics in 2D space using "t-Distributed Stochastic Neighbor Embedding" (t-SNE) is shown in Fig. 6(c), where the t-SNE plot indicates which cluster belongs to which topic explicitly. The green cluster is the biggest and stands for topic 4, as topic 4 covers 1588 abstracts. These are followed by topics 5, 8 and 6, covering 941, 858 and 786 abstracts, respectively. Topic 0 has only 31 abstracts and is the topic with the least number of abstracts, as shown in Fig. 6(a) and Fig. 6(c) with navy blue color.

The Table 1 shows the topic distribution, documents topic number, the number of documents per topic, and the top 10 most dominant words in the topic. Here, it can be seen that the most suitable topic 4 has catalytic hydrogenation of alcohol-related abstracts as the most dominating words extracted. Topic 5 categorises all documents belonging to photo-catalysis. Topic 1 organises all documents belonging to the production of hydrogen from ethanol. Topic 3 is about enzyme catalysis as it has keywords like 'enzyme', 'ethanol', 'cell', etc. Topic 7 talks about membrane methodology to produce hydrogen. All topics highlight results, yield, concentration, and other essential parameters related to selectivity and performance.

5.2. *Sentence classification and NER using Ex-SciBERT*

Firstly, we present the results of transfer learning of Ex-SciBERT. The loss values for train and validation data per epoch are shown in Fig. 7(a) in the case of classification. The loss value is optimised to 0.2495 for the train set and 0.3002 for validation data, and other performance metrics such as precision, recall and F1 score are collated in the table for all four classes in Table 3. The overall accuracy for train test

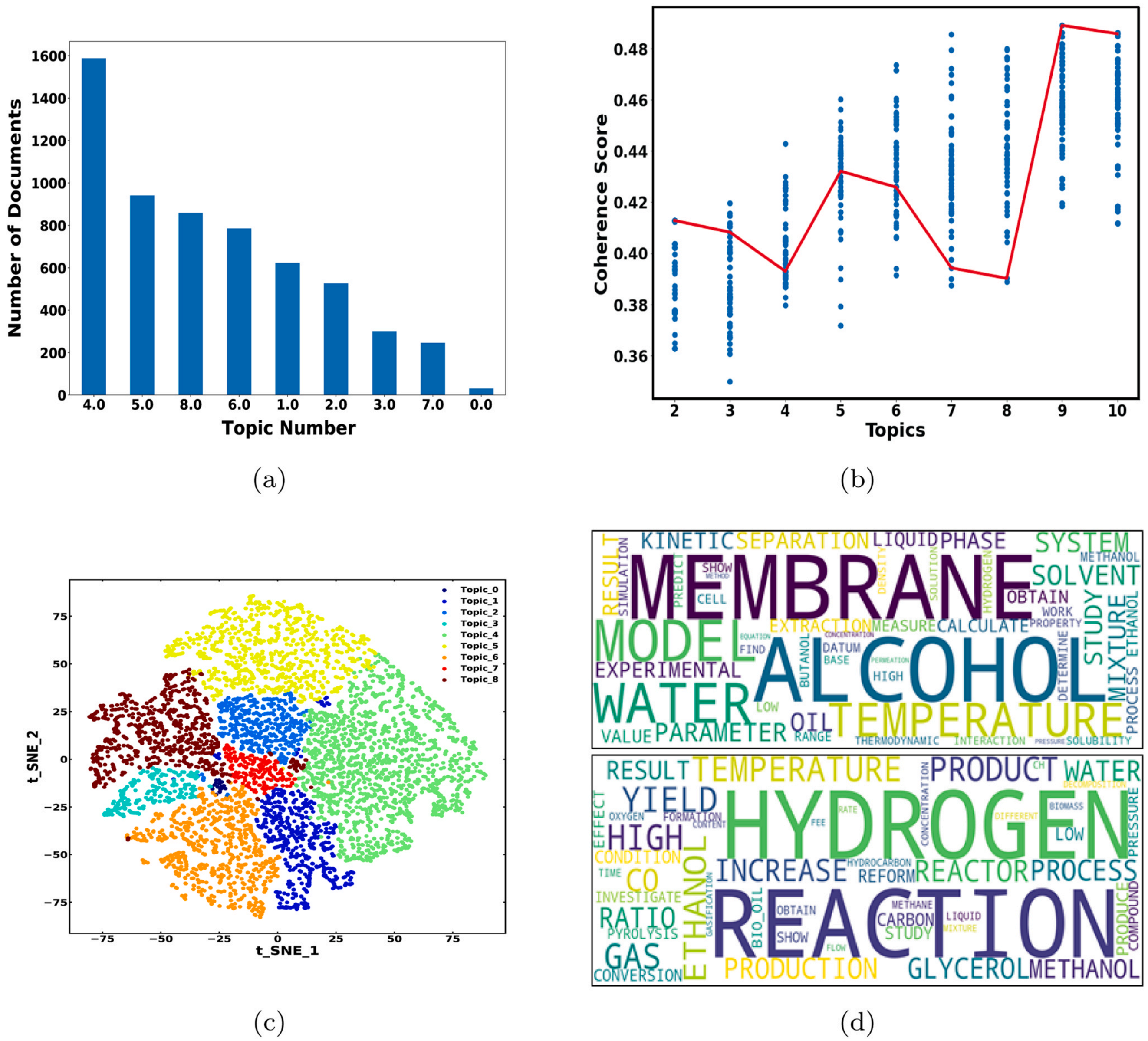


Fig. 6 – The results of LDA are represented as (a) Bar plot of Number of documents vs Topic. (b) figure shows variation of Coherence score with topics for different values of α and β (scatter). The line plots variation for $\alpha = 0.61$ and $\beta = 0.31$ (c) Bokeh plot visualising the cluster of documents in a 2D space (2D from 9D topic simplex used by LDA's Dirichlet distribution) using the t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm. (d) WordCloud of top 50 most dominant words in most represented topics, Topic 4 and Topic 5.

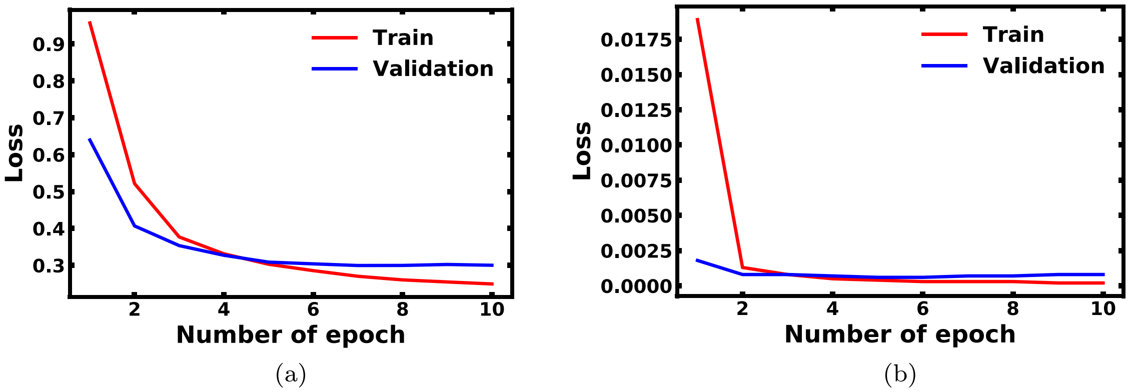


Fig. 7 – The variation of loss with per epoch for tasks, (a) Classification (b) Name entity recognition, while transfer learning the SciBERT model.

Table 3 – Performance of the Ex-SciBERT.

Type of data	Name of label	Precision	Recall	F1
Classification				
Train	Process (0)	0.94	0.98	0.96
	Catalyst (1)	0.78	0.88	0.82
	Both catalyst and process (2)	0.70	0.61	0.65
	Neutral (3)	0.99	0.74	0.84
Accuracy	0.915			
Test	Process (0)	0.92	0.96	0.94
	Catalyst (1)	0.72	0.82	0.77
	Both catalyst and process (2)	0.61	0.57	0.59
	Neutral (3)	0.96	0.70	0.81
Accuracy	0.890			
NER				
	0	1.00	1.00	1.00
	Catalyst	0.98	0.98	0.98
	Process	0.99	0.97	0.98
Accuracy	0.998			
Test	0	1.00	1.00	1.00
	Catalyst	0.95	0.92	0.93
	Process	0.99	0.97	0.98
Accuracy	0.997			

Table 4 – The classification and NER prediction results of Ex-SciBERT. The model first classifies the sentence and then highlights the sentence's catalyst and process parameters. Highlighted bold words stand for catalysts and italics words stand for process contents.

S.No.	Original text with highlighted keywords	predicted class
1	Hydrogen production by photocatalytic alcohol reforming has been studied in the presence of a **Pt** with **TiO2** photocatalyst	catalyst
2	Three **Cu,ZnO,Al2O3** catalysts, with a molar ratio of **copper, zinc** and **aluminum** 45:45:10, were prepared by different methods	catalyst
3	The photocatalytic H2 production activity of a **Cu2O** with **TiO2** photocatalyst was 10 times higher than **Au** with **TiO2** photocatalyst	catalyst
4	Before sealing the vessel it was filled with 3 bar *pressure* H2 and placed in an oven set at 333*K* for 24 h	process
5	The internal *pressure* of reactor was controlled at 3000 *Pa* using dry vacuum pump	process
6	The initial solution *temperature* measured by two thermocouples was 293 *K* which was same with the room temperature	process
7	The percentage concentrations of hydrogen and carbon monoxide increased with the increase of power	Neutral
8	The **TiNTx** photocatalysts were obtained by calcination of **H2Ti3O7** nanotubes at *temperature* between 423 and 1273 *K* for 2 h	catalyst and process
9	Several studies have reported the effective accomplishment of these reactions type in the presence of blue **Pd** alloys catalysts in relative high *pressure*	Catalyst-process but predicted process

datasets is collated in the Table 3. We obtain a reasonable overall accuracy for the training dataset = 0.915 and the test set = 0.890. The accuracy is slightly less for the class for correctly predicting the sentences with both catalyst and process conditions together. This is due to limited availability in the train set as indicated in the Table 2.

For NER, the loss function plots are presented in Fig. 7(b). The final loss values are 0.0002 and 0.0008 for train and validation data. The summary of the NER model performance metrics is shown in the Table 3, where precision, recall and F1 score concerning each class are mentioned. The overall accuracy is also mentioned in the Table 3. The train and test set accuracy is approximately identical, 0.998 and 0.997, respectively, which is impressive. Here, the number of tokens with unique labels is mentioned in the Table 2, where '0′ tags are more. We have taken a hyperparameter to ignore the label '0′ to address this problem to avoid '0′ tokens while training.

Lastly, we have extensively tested our model's performance with random test sentences from some articles based on hydrogen production from alcohol to check classification and NER. The Fig. 8 clearly showcases how a sentence is fed

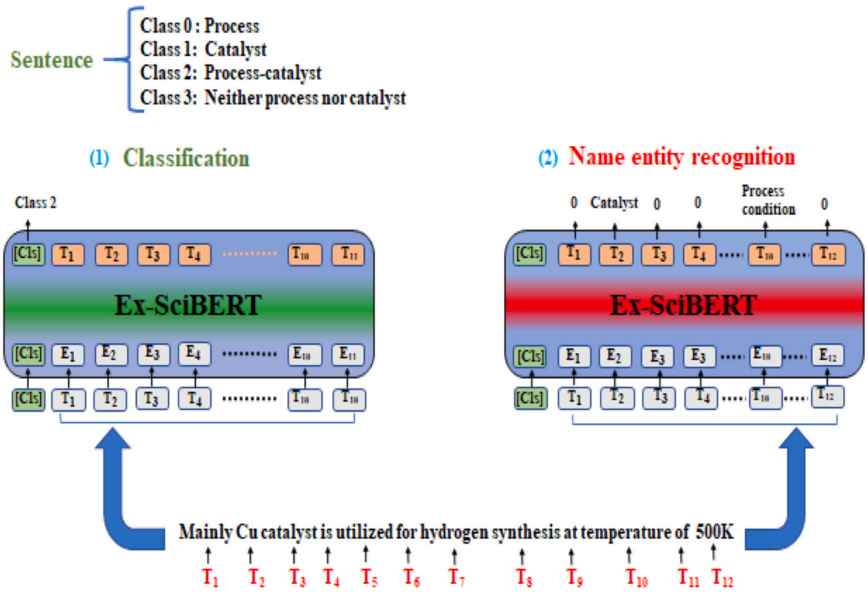

Sentence
Class 0 : Process
Class 1: Catalyst
Class 2: Process-catalyst
Class 3: Neither process nor catalyst
(1) Classification
(2) Name entity recognition
Class 2
[Cls] T1 T2 T3 T4 T10 T11
Ex-SciBERT
[Cls] E1 E2 E3 E4 E10 E11
[Cls] T1 T2 T3 T4 T10 T10
0 Catalyst 0 0 Process condition 0
[Cls] T1 T2 T3 T4 T10 T12
Ex-SciBERT
[Cls] E1 E2 E3 E3 E10 E12
[Cls] T1 T2 T3 T4 T10 T12
Mainly Cu catalyst is utilized for hydrogen synthesis at temperature of 500K
T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12

Fig. 8 – This Schematic represent the working our Ex-SciBERT models (1) Classification (2) Named entity recognition.

to Ex-SciBERT for classification and NER. The highlighted model prediction results of similar test instances are collated and shown in the Table 4. In the Table 4, the first eight examples show the correct prediction of classes and NER, while the last example shows an incorrect prediction of class while correctly predicting the NER.

6. Conclusion

This paper has applied NLP tools to extract decision enabling information from the large corpora of scientific literature in chemical engineering, specifically in the area of Hydrogen production. A vast amount of unstructured and scattered information is available in the literature on hydrogen production from alcohol. The nature of the data and its utility in several industrial processes make summarising this information is crucial to developing a recommendation system. With that in view, we obtained this data using an Elsevier API key and institute token issued to IIT Delhi with “Hydrogen production from alcohol” as the search query. We then parsed the obtained XML files using a custom made XML parser. We used the vital “NLTK” and “Gensim” libraries for text pre-processing and “ChemDataExtractor” to obtain chemical entities in sentences. Using “Hydrogen production from alcohol” as a keyword, we extracted 5901 scientific articles and pre-processed them suitably. We then used LDA for topic modelling and clustered similar articles for hydrogen production into nine optimal topics. Further, using carefully annotated data, we have trained SciBERT, NLP deep learning model to form the extended-SciBERT (Ex-SciBERT). The Ex-SciBERT can perform classification and NER tasks on sentences of corpora. The accuracy for classification on the test dataset is 0.890, and the accuracy of the NER model on the test dataset is 0.997. We are optimistic that Ex-SciBERT will help in reducing the manual effort of screening through scientific literature. The encapsulation will further help uncover new and unimplemented Hydrogen production methodologies available in scientific literature. There is a scope for significant future research in this domain. For instance, we can train Ex-SciBERT from scratch on massive chemical corpora and combine it with the domain knowledge for various physical processes. This will result in more personalised recommendation systems for various industrial applications and open new pathways of connection between active research and industrial utilisation. The complete implementation detail of the present work, including Python codes of the entire study, is available on github.com/avanscholar.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge research grants 51/14/11/2019-BRNS and CRG /2018/001555, from BRNS and SERB India, respectively.

References

M. Afzal, J. Hussain, A. Abbas, H. Maqbool, Multi-class clinical text annotation and classification using bert-based active learning, Available at SSRN 4081033 2022.

Agrawal, A., Tripathi, S., Vardhan, M., Sihag, V., Choudhary, G., Dragoni, N., 2022. Bert-based transfer-learning approach for nested named-entity recognition using joint labeling. Appl. Sci. 12 (3), 976.

A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use framework for state-of-the-art nlp, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, 54–59.

Akhoondi, A., Feleni, U., Bethi, B., Idris, A.O., Hojjati-Najafabadi, A., 2021. Advances in metal-based vanadate compound photocatalysts: synthesis, properties and applications. Synth. Sinter. 1 (3), 151–168.

E. Alsentzer, J. Murphy, W. Boag, W. Weng, D. Jindi, T. Naumann, M. McDermott, Proceedings of the 2nd clinical natural language processing workshop (2019).

An, Y., Xia, X., Chen, X., Wu, F.-X., Wang, J., 2022. Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf. Artif. Intell. Med., 102282.

D. Araci, Finbert: Financial sentiment analysis with pre-trained language models, arXiv preprint arXiv:1908.10063 (2019).

R. Balyan, K.S. McCarthy, D.S. McNamara, Combining machine learning and natural language processing to assess literary text comprehension, Grantee Submission (2017).

Bass, C.R., Benefield, B., Horn, D., Morones, R., 2019. Increasing robustness in long text classifications using background corpus knowledge for token selection. SMU Data Sci. Rev. 2 (3), 10.

I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

Bhuvaneshwari, P., Rao, A.N., Robinson, Y.H., Thippeswamy, M., 2022. Sentiment analysis for user reviews using bi-lstm self-attention based cnn model. Multimed. Tools Appl. 1–15.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M., 2009. Reading tea leaves: how humans interpret topic models. Adv. Neural Inf. Process. Syst. 288–296.

J. Copara, N. Naderi, J. Knafou, P. Ruch, D. Teodoro, Named entity recognition in chemical patents using ensemble of contextual language models, arXiv preprint arXiv:2007.12569 (2020).

Court, C.J., Cole, J.M., 2018. Auto-generated materials database of curie and néel temperatures via semi-supervised relationship extraction. Sci. data 5 (1), 1–12.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

Dosado, A.G., Chen, W.-T., Chan, A., Sun-Waterhouse, D., Waterhouse, G.I., 2015. Novel au/tio2 photocatalysts for hydrogen production in alcohol–water mixtures based on hydrogen titanate nanotube precursors. J. Catal. 330, 238–254.

Feng, X., Dai, Y., Ji, X., Zhou, L., Dang, Y., 2021. Application of natural language processing in hazop reports. Process Saf. Environ. Prot. 155, 41–48.

Hojjati-Najafabadi, A., Salmanpour, S., Sen, F., Asrami, P.N., Mahdavian, M., Khalilzadeh, M.A., 2021. A tramadol drug electrochemical sensor amplified by biosynthesized au nanoparticle using mentha aquatic extract and ionic liquid. Top. Catal. 1–8.

Hojjati-Najafabadi, A., Davar, F., Enteshari, Z., Hosseini-Koupaei, M., 2021. Antibacterial and photocatalytic behaviour of green synthesis of zn0. 95ag0. 05o nanoparticles using herbal medicine extract. Ceram. Int. 47 (22), 31617–31624.

Hojjati-Najafabadi, A., Mansoorianfar, M., Liang, T., Shahin, K., Karimi-Maleh, H., 2022. A review on magnetic sensors for monitoring of hazardous pollutants in water resources. Sci. Total Environ. 824, 153844.

Q. Hua, S. Qundong, J. Dingchao, G. Lei, Z. Yanpeng, L. Pengkang, A character-level method for text classification, in: 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE, 2018, 402–406.

S. Inatani, T. Van Phan, M. Nakagawa, Comparison of mrf and crf for text/non-text classification in japanese ink documents, in: 2014 14th International Conference on Frontiers in Handwriting Recognition, IEEE, 2014, 684–689.

Jacobi, C., Van Atteveldt, W., Welbers, K., 2016. Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit. J. 4 (1), 89–106.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimed. Tools Appl. 78 (11), 15169–15211.

M. Jogin, M. Madhulika, G. Divya, R. Meghana, S. Apoorva, et al., Feature extraction using convolution neural networks (cnn) and deep learning, in: 2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT), IEEE, 2018, 2319–2323.

Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G., 2016. Mimic-iii, a freely accessible critical care database. Sci. Data 3 (1), 1–9.

A. Kaur, D. Chopra, Comparison of text mining tools, in: 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), IEEE, 2016, 186–192.

Khor, S., Jusoh, M., Zakaria, Z., 2022. Hydrogen production from steam and dry reforming of methane-ethane-glycerol: a thermodynamic comparative analysis. Chem. Eng. Res. Des.

A. Koripelly, Z. Hong, K. Chard, Diving for treasure in a sea of scientific literature: Extracting scientific information from free text articles (2020).

F. Kuniyoshi, K. Makino, J. Ozawa, M. Miwa, Annotating and extracting synthesis process of all-solid-state batteries from scientific literature, arXiv preprint arXiv:2002.07339 (2020).

Lee, D., Choi, J., Lee, Y.-W., Lee, J.M., 2021. Design and economic analysis of biodiesel production process of simultaneous supercritical transesterification and partial hydrogenation using soybean oil with pd/al2o3 catalyst. Chem. Eng. Res. Des. 172, 264–279.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. So, J. Kang, Biobert: Pre-trained biomedical language representation model for biomedical text mining. arxiv 2019, arXiv preprint arXiv:1901.08746 (2020).

J.-S. Lee, J. Hsiang, Patentbert: Patent classification with fine-tuning a pre-trained bert model, arXiv preprint arXiv:1906.02124 (2019).

Li, D., Yan, L., Yang, J., Ma, Z., 2022. Dependency syntax guided bert-bilstm-gam-crf for chinese ner. Expert Syst. Appl. 196, 116682.

J. Libovicky`, R. Rosa, A. Fraser, How language-neutral is multilingual bert?, arXiv preprint arXiv:1911.03310 (2019).

C.D. Liew, Survey of machine learning algorithms used in natural language processing and understanding tasks 2021.

Lorenzut, B., Montini, T., De Rogatis, L., Canton, P., Benedetti, A., Fornasiero, P., 2011. Hydrogen production through alcohol steam reforming on cu/zno-based catalysts. Appl. Catal. B: Environ. 101 (3–4), 397–408.

Ma, K., Tan, Y., Xie, Z., Qiu, Q., Chen, S., 2022. Chinese toponym recognition with variant neural structures from social media messages based on bert methods. J. Geogr. Syst. 1–27.

Mansoorianfar, M., Shahin, K., Hojjati-Najafabadi, A., Pei, R., 2022. Mxene–laden bacteriophage: a new antibacterial candidate to control bacterial contamination in water. Chemosphere 290, 133383.

T. Minka, Estimating a dirichlet distribution (2000).

Nguyen, T.A., Nakagawa, K., Duong, H.P., Maeda, Y., Otsuka, K., 2020. Hot-spots and lessons learned from life cycle sustainability assessment of inedible vegetable-oil based biodiesel in northern viet nam. In: Biofuels for a More Sustainable Future. Elsevier, pp. 165–212.

Nikolenko, S.I., Koltcov, S., Koltsova, O., 2017. Topic modelling for qualitative studies. J. Inf. Sci. 43 (1), 88–102.

M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the eighth ACM international conference on Web search and data mining, 2015, 399–408.

Searmsirimongkol, P., Rangsunvigit, P., Leethochawalit, M., Chavadej, S., 2011. Hydrogen production from alcohol distillery wastewater containing high potassium and sulfate using an anaerobic sequencing batch reactor. Int. J. Hydrog. Energy 36 (20), 12810–12821.

Susanti, R.F., Dianningrum, L.W., Yum, T., Kim, Y., Lee, Y.-W., Kim, J., 2014. High-yield hydrogen production by supercritical water gasification of various feedstocks: alcohols, glucose, glycerol and long-chain alkanes. Chem. Eng. Res. Des. 92 (10), 1834–1844.

Swain, M.C., Cole, J.M., 2016. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. J. Chem. Inf. Model. 56 (10), 1894–1904.

Torkian, N., Bahrami, A., Hosseini-Abari, A., Momeni, M.M., Abdolkarimi-Mahabadi, M., Bayat, A., Hajipour, P., Rourani, H.A., Abbasi, M.S., Torkian, S., et al., 2022. Synthesis and characterization of ag-ion-exchanged zeolite/tio2 nanocomposites for antibacterial applications and photocatalytic degradation of antibiotics. Environ. Res. 207, 112157.

Trewartha, A., Walker, N., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K.A., Ceder, G., Jain, A., 2022. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns 3 (4), 100488.

Vaucher, A.C., Zipoli, F., Geluykens, J., Nair, V.H., Schwaller, P., Laino, T., 2020. Automated extraction of chemical synthesis actions from experimental procedures. Nat. Commun. 11 (1), 1–11.

V. Venugopal, S. Sahoo, M. Zaki, M. Agarwal, N.N. Gosvami, N. Krishnan, Looking through glass: Knowledge discovery from materials science literature using natural language processing, arXiv preprint arXiv:2101.01508 (2021).

A. Villarreal, R. Villarreal, Machine learning and natural language processing for the identification of synthesis parameters of nimo sulfide catalysts(2019).

R. Visser, M. Dunaiski, Sentiment and intent classification of in-text citations using bert., Tech. rep., EasyChair (2022).

Vo, N.N., Vu, Q.T., Vu, N.H., Vu, T.A., Mach, B.D., Xu, G., 2022. Domain-specific nlp system to support learning path and curriculum design at tech universities. Comput. Educ.: Artif. Intell. 3, 100042.

Wang, B., Sun, B., Zhu, X., Yan, Z., Liu, Y., Liu, H., Liu, Q., 2016. Hydrogen production from alcohol solution by microwave discharge in liquid. Int. J. Hydrog. Energy 41 (18), 7280–7291.

H. Yang, W. Hsu, 2021. Named entity recognition from synthesis procedural text in materials science domain with attention-based approach., in: SDU@ AAAI, 2021.

Zhang, B., Zhang, S.-X., Yao, R., Wu, Y.-H., Qiu, J.-S., 2021. Progress and prospects of hydrogen production: opportunities and challenges. J. Electron. Sci. Technol., 100080.

Z. Zhang, Y. Wu, Z. Li, S. He, H. Zhao, X. Zhou, X. Zhou, I know what you want: Semantic learning for text comprehension, arXiv preprint arXiv:1809.02794 (2018).

P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, B. Xu, Text classification improved by integrating bidirectional lstm with two-dimensional max pooling, arXiv preprint arXiv:1611.06639 (2016).