Dataset Mention Detection & Intent Classification in Scientific Publications

Robin Spiers
2617736
r.spiers@student.vu.nl

Academic Supervisor: Michael Cochez[1]
Examiners: Al Idrizou[1] and Jan-Christoph Kalo[1]
Internship Supervisor: Hosein Azarbonyad[2]

[1] Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam
[2] Elsevier B.V., Radarweg 29, 1043 NX Amsterdam

Abstract. This study explores the effectiveness of supervised learning models for dataset mention detection and intent classification in scientific publications. In specific, the first task focuses on incorporating a binary classifier as a preliminary filtering step before performing dataset mention detection. The second task is relatively novel, where binary classifiers were trained to distinguish the intention behind mentioning a dataset between background references and signs of own usage. Additionally, experiments for both tasks were designed to gain insights into the differences between traditional models and pre-trained language models. For dataset mention detection, the study explores various combinations of extraction models and classifiers. Results show that a model based on RoBERTa outperforms the traditional models, but also struggles the most with identifying unseen dataset titles. Moreover, incorporating binary classifiers improves both the correctness and efficiency of extraction models. Regarding intent classification, a comparison between SVM and SciBERT reveals the superiority of SciBERT for the task. While the models are able to detect signs of dataset usage outside of the mainly associated sections, the number of false positives is higher as well. Future work suggestions include expanding the range of models and configurations, investigating entity masking and linguistic features, creating a larger annotated dataset and exploring entity linking techniques. In conclusion, addressing the findings and suggestions of this paper advances the understanding and stewardship of dataset usage in scientific literature, enhancing transparency, efficiency, and impact of scholarly discourse.

Keywords: Natural Language Processing · Information Retrieval · Sequence Classification · Transformers · BERT

1 Introduction

In the realm of scientific research, datasets play a crucial role as the bedrock upon which discoveries and advancements are built. These vast repositories of information hold immense potential and offer invaluable insights into various domains, ranging from social sciences to medicine. The effective management and utilisation of datasets are paramount to ensuring the integrity and reproducibility of scientific findings. For the conclusions of a study to stand, the underlying data must be made accessible and open to examination by qualified professionals with the appropriate expertise [5].

Embracing the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) [36] becomes vital for data stewardship, as it empowers researchers to navigate the expanding landscape of information. However, despite the growing recognition of the significance of datasets and the FAIR principles, challenges persist in identifying the mentioning of a dataset and comprehending their presence within scientific texts. Addressing these challenges requires novel approaches and technologies that can automatically detect mentions of datasets [15] and recognise signs of their utilisation [13]. By shedding light on the incorporation of datasets in scientific publications, these advancements hold the potential to enhance the transparency, efficiency, and impact of scholarly discourse.

In this paper we firstly examined how existing dataset mention detection models are affected by the incorporation of binary classification models for preliminary filtering. This was done by taking two of the best performing models from the 2021 Kaggle competition *"Coleridge Initiative - Show US the Data"*[3], which was centered around the task of detecting dataset mentions in the texts of scientific publications [17]. Then, the two detection models were combined with three binary classification models: logistic regression, support vector machine (SVM) and a classifier based on RoBERTa. Each model combination was evaluated by measuring the inter-model differences with respect to the correctness of predictions and overall runtimes. Secondly, we explore the automatic recognition of usage when mentioning a dataset, an uncharted subject in the field of text classification. Due to limited availability of high quality data, experiments were executed in a weakly-supervised learning setup. The main objective was to see how traditional models compare to pre-trained models, namely SVM and SciBERT, and to see how both compare to a simplistic baseline method.

Explicitly, our study is built on the following research questions:

RQ1: How is the performance of existing dataset mention extraction models affected when a binary sequence classifier filters samples beforehand?

1. How do traditional machine learning models perform compared to pre-trained language models for this sequence classification task?
2. How does model correctness differ between detecting mentions of seen datasets and detecting mentions of unseen datasets?

[3] https://www.kaggle.com/competitions/coleridgeinitiative-show-us-the-data

RQ2: To what extent are classification models able to distinguish the intention for mentioning a dataset, between own usage and background references?

1. How do traditional machine learning models and pre-trained language models compare to a baseline of basing intentions off section titles?

The main contributions of this paper are as follows: (1) Preliminary filtering by binary classifiers in a dataset mention detection setup was found to not only reduce the runtime for extracting dataset mentions, but also to improve correctness of the predictions. (2) Although a pre-trained language model like RoBERTa has the potential to outperform traditional models like logistic regression and SVM for said task, experiments showed that it is sensitive for overfitting on known dataset titles. (3) For the novel application of intent classification, experiments showed that SciBERT strictly dominates SVM regarding the correctness of predictions, while performing slightly less than a baseline method in terms of accuracy and precision but outperforming it on recall and F_1. (4) A new dataset is created, containing 979 samples of dataset mentions from scientific publications, manually annotated by their intention between background references and signs of own usage. (5) Suggestions are made for future research directions around dataset mentions, including direct follow-ups, linguistic feature engineering, creating new high quality datasets and shifting focus towards entity linking.

2 Related Work

In this section we discuss previous work that was found to be relevant to the present study. For each research question, we present the relevant preceding work, followed by an identification of the research gap.

2.1 Dataset Mention Detection

Several studies have experimented with the detection of dataset mentions in Natural Language Processing (NLP) as a Named Entity Recognition (NER) task. Early work mostly focused on architectures that combined word representation vectors with Recurrent Neural Networks (RNN), like Long Short-Term Memory networks (LSTM) and Bidirectional LSTMs (BiLSTM) [9,18,32]. More recent studies have experimented with dataset mention detection as a downstream task for pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers), outperforming the previously mentioned models [8,22,39].

In 2021, the Coleridge Initiative launched the *Show US the Data* (SUTD) competition on Kaggle for the task of detecting dataset mentions [17]. The format was to infer the excerpts mentioning a dataset for every publication in the test set. Submissions were evaluated using a Jaccard-based F_β score, with $\beta = 0.5$ to prioritize precision over recall. The winning submission ($F_{0.5} = .576$) used a

masked language modelling ensemble of RoBERTa and SciBERT, combined with several data pre- and post-processing steps. The second best submission ($F_{0.5} = .575$) searched for candidates in the pattern *LONG-FORM (ACRONYM)* and fine-tuned a RoBERTa model on such candidates to recognise dataset titles. The third best submission ($F_{0.5} = .558$) consisted of regex pattern matching, filtering for stopwords and keywords, co-occurrence statistics and referring to a dictionary of known dataset names. The relatively good applicability of simple pattern matching techniques for this task could be explained by naming conventions for datasets in certain domains. For instance, datasets in NLP research often have proper names or distinct acronyms [13].

Binary Classifiers. Recently, Younes and Mathiak used the SUTD dataset to evaluate BERT models for a binary classification task [38]. The objective was to detect whether the section of a publication contained a dataset mention or not. Additional experiments were conducted to measure the impact of data re-sampling methods and a cost sensitive learning setup. The best recall for positive samples was achieved when using section contents, re-sampling techniques, balanced focal loss and the F_β score with $\beta = 3$ as the validation metric. This value for β was selected as recall was more important for the binary sequence classification task, in contrary to the importance of precision for the dataset mention detection task.

Earlier studies on scientific NER have also made use of binary classifiers before applying extraction models [35]. One important argument for this is that such entities are not uniformly distributed over the different sections in scientific articles. Results on relevant tasks have shown that the incorporation of a binary classifier can lead to significant improvements for both precision and recall [23]. Kumar, Ghosal and Ekbal [14] divided dataset mention detection into two tasks, namely sentence classification and dataset mention extraction. The authors combined SciBERT with a multi-layer perceptron (MLP) for the sentence classification task, achieving a macro-averaged F_1 score of 91% on a social sciences dataset.

Research Gap. RQ1 of this thesis bridges a scientific gap by combining the best submissions of the SUTD competition with recent discoveries, mainly those by Younes and Mathiak. Findings from other related work suggest that this combination could have a significant impact on the performance for automated dataset mention detection.

2.2 Intent Classification

To our knowledge, no prior research has been done specifically for classifying the intention behind mentioning a dataset. However, citation intent classification is a similar topic in the field of NLP.

In 2019, Cohan et al. [6] presented the SciCite dataset, containing sentences with citations that are labelled with one of three possible intents: background,

method or results. The paper also presented a framework for the task of citation intent classification, a model including two auxiliary tasks: section title prediction and citation worthiness prediction. Their proposed model was a BiLSTM network with GloVe word representations, ELMo contextualized embeddings and a dot-product attention mechanism. Oesterling et al. [24] showed that the performance of this model was improved when expanded to include hand-generated features and a TF-IDF embedding layer.

Pre-trained Language Models. BERT models achieve competitive results on the SciCite dataset, where SciBERT performs marginally better than variants like regular BERT and RoBERTa [33]. Wright and Augenstein [37] used the SciCite dataset as a downstream task to introduce CiteBERT, a SciBERT model that was fine-tuned for predicting the citation-worthiness of a sentence. The original version of SciBERT achieved an F1 score of 84.83%, while CiteBERT achieved an F1 score of 85.35%. Later, an improvement for SciCite's public leaderboard was submitted by ImpactCite, an XLNet-based model by Mercier et al. [21]. The ImpactCite model achieved a macro-F1 score of 88.93%, outperforming the model by Cohan et al. and BERT-based models. More recently, Lahiri et al. [16] used SciBERT in a prompt-based learning setup. Their submission did improve upon the original BiLSTM with an F1-score of 86.33%, but failed to outperform ImpactCite.

Although traditional machine learning models have also been evaluated for the SciCite dataset [30], they were not able to match the performances of advanced neural network architectures and pre-trained models. However, a recent study by Wahba et al. [34] demonstrated the competitiveness of a linear support vector machine against pre-trained language models BERT, DistilBERT, RoBERTa and XLM on different text classification tasks.

Research Gap. The models and experimental designs from related studies form a good starting point for novel research on intent classification for dataset mention specifically. Although the concepts and structural patterns in natural language are different between citations and dataset mentions, replicating the experiments for this novel task is a relatively straightforward process.

3 Methods

This section presents the approaches to the two research questions, before describing the exact details of the experimental setup.

Within the scope of this paper, we view a dataset mention as any text excerpt that refers to an existing dataset by making a direct reference to its full name, a common alias or an acronym. Ambiguous references like *the dataset* or *the test set* fall outside the scope of this research. Some examples from the SUTD dataset of raw text containing dataset mentions are displayed in Figure 1.

> *"In our application, we use the data obtained from the* ***Alzheimer's Disease Neuroimaging Initiative (ADNI)*** *database, ADNI-1 cohort [17]." ... "This secure data collection system will be implemented beginning with the administration of the longitudinal* ***Baccalaureate and Beyond*** *(B&B) survey."*

Fig. 1: Raw text samples containing dataset mentions.

3.1 Dataset Mention Detection

For the task of dataset mention detection, experiments were conducted by combining the best performing models from the SUTD competition with binary classifiers. The incorporation of binary classifiers draws direct influence from the work of Younes and Mathiak [38]. Their study evaluated several models for the task of labelling article sections as containing a dataset mention or not. Where their research was limited to experimentation with the binary classification models on their own, this study combines them with dataset mention extraction models.

Binary Classifiers. The main goal of RQ1 was to examine how the performance of the dataset mention extraction models is impacted by filtering out texts that are not likely to contain a dataset mention. This was done by training binary classification models for the task of predicting whether a text from a scientific article contains a dataset mention or not.

Since this research design was inspired by the study of Younes and Mathiak [38], the model that performed best for their experiments is the most prominent model of the experiments for this study. The best performance was achieved by the fine-tuned RoBERTa-base model that used balanced focal loss, data resampling and section texts as input features.

To provide a comprehensive evaluation, this study also includes the assessment of logistic regression and support vector machines (SVM) as alternative approaches for comparison. By considering multiple model architectures, this research aims to contribute valuable insights into the effectiveness of different strategies for dataset mention detection and the impact of pre-training on the performance of such models.

Extraction Models. On the private leaderboard of the SUTD competition, there is a gap of 5.5% $F_{0.5}$ points between the third and fourth best submissions. Initially, the aim of this research was to conduct experiments with each of the top three models, but instead experiments are only conducted for the third[4] and second[5] best submissions.

[4] https://github.com/Coleridge-Initiative/rc-kaggle-models/tree/main/3rd%20Mikhail%20Arkhipov

[5] https://www.kaggle.com/competitions/coleridgeinitiative-show-us-the-data/discussion/248296

Winning Kaggle Submission. Before describing the detection models examined in the present study, this paragraph summarises the winning submission of the SUTD competition in more detail to clarify why it was not considered. The winning submission came from participants Khoi Nguyen and Nguyen Quan Anh Minh, who collaborated under the name *Zalo AI*. The main reason why their submission was not considered for the present study is the fact that it used an ensemble approach, which made it difficult to perform reliable and valid follow-up experiments. Classification models based on RoBERTa and SciBERT in the ensemble were fine-tuned on a combination of two different embeddings. The two embeddings were divided into label embeddings, for the dataset mentions themselves, and contextual embeddings, for the surrounding text. The two embeddings were then multiplied into singular sequence embeddings, where an MLP was trained to classify the tokens of a sentence sequence as being part of a dataset mention or not. While the second and third best submissions shared a similar approach, the winning submission was entirely different, putting it out of scope for experiments in the present study.

Kaggle Model 3. The third best submission, or KM3, by Mikhail Arkhipov achieved a score of 55.8% on the private set. Its solution is fully based on pattern matching. For both the training and testing documents, a set of candidate sequences is formed by extracting all capitalized sequences followed by brackets from the text. A second requirement is that all candidates should include data-specific keywords, such as *study*, *survey*, or *dataset*. Candidates that contain English stopwords are excluded. Additionally, candidates must co-occur with the word *data* in the same document for at least 10% of the documents it occurs in. The final candidates are then utilised for straightforward string matching against the raw texts of the test documents. The data flow of KM3 is visualised in Figure 2.

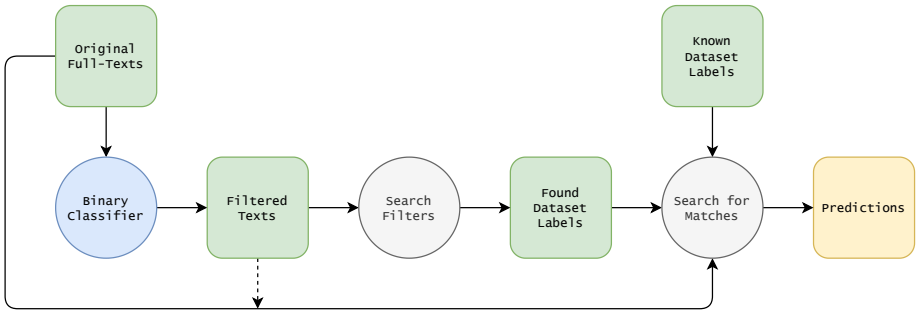


Fig. 2: Kaggle Model 3. Green blocks for text data, yellow blocks for predicted dataset mentions, grey circles for original model steps, blue circles for modifications, solid lines for core data flows, broken lines for optional replacements.

Kaggle Model 2. The second best submission, or KM2, by Chun Ming Lee achieved a score of 57.5% on the private set, 0.1% worse than the winning submission. Its methodology starts by searching through the text of all test documents for strings in the form *LONG-NAME (ACRONYM)* using the Schwartz-Hearst pattern matching algorithm [31]. The long names collected from the raw text are then classified as datasets or not by a RoBERTa-base model. This model was fine-tuned on a set of roughly five thousand manually marked samples, which were all collected from SUTD's training and testing data using the Schwartz-Hearst algorithm. For each document, a matched string is submitted as a prediction if its frequency across the texts of all test documents is higher than a pre-defined threshold or when it has a Regex match with either the word *Study* or *Survey*. Direct matches with labels from the training set are also submitted as predictions. If there is a match with these conditions, and the associated acronym is mentioned in the document, the acronym is also added as a prediction.

In the final submission, the probability threshold for accepting a candidate label as a dataset name was set at 90%, while the minimum frequency threshold to be considered as a highly frequent dataset was set at 50 documents. The data flow of KM2 is visualised in Figure 3.

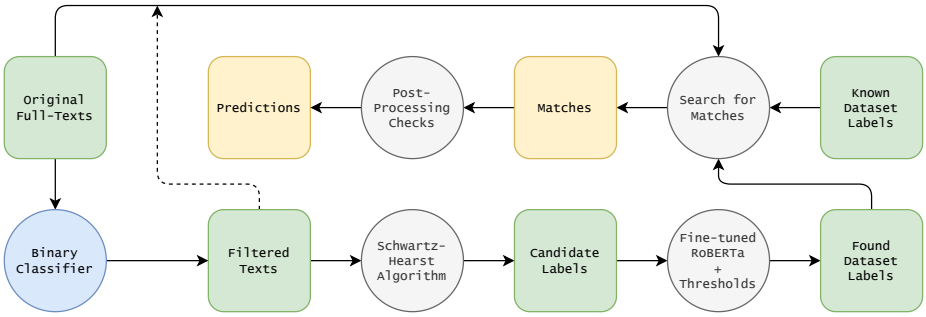


Fig. 3: Kaggle Model 2. Green blocks for text data, yellow blocks for predicted dataset mentions, grey circles for original model steps, blue circles for modifications, solid lines for core data flows, broken lines for optional replacements.

Filtering Usage. For this research, the texts after filtering can be applied at two different steps within the algorithms of the selected detection models. The first step is to filter texts before a pattern matching algorithm scans through them in the search of candidate dataset labels. By filtering at this step, raw texts that are less likely to contain a dataset mention are removed. The hypothesis is that this reduces the number of strings that do match the pattern, but are not dataset mentions. The second possible step is to use the same filtered texts again when going through all test document texts searching for matches with candidates that satisfy all requirements. The hypothesis for filtering at this step

is that texts containing an incorrect candidate label are ignored and that less documents need to be searched through, reducing the computational costs.

3.2 Intent Classification

Previous work extensively explored intent classification for citations in scientific articles, with a focus on identifying the purpose of a citation within the context of the paper. These studies have significantly contributed to the understanding of how authors refer to and utilise external sources to support their research claims. However, a crucial aspect that has received relatively less attention is the classification of dataset mentions within scientific articles.

The present thesis addresses this research gap by focusing on the intent classification of dataset mentions in scientific articles. Kolyada, Potthast and Stein [13] proposed a taxonomy to divide dataset mentions into three distinct categories: reuse, description and reference. However, in order to restrict ambiguity, our experiments are focused on two classes: (0) references to the dataset for background purposes, and (1) own usage of the dataset. To accomplish this task, the experimental setup proposed in this thesis draws inspiration from methodologies employed for citation intent classification.

Baseline. To establish a baseline for our dataset intention recognition task in scientific publications, we adopt a section-based classification approach. In this methodology, we leverage the inherent structure of research papers and assume that the mention of a dataset can be indicative of its intended purpose based on the section it appears in. Specifically, we assign fixed labels to different sections commonly found in scientific articles, such as *"Methodology"*, *"Related Work"*, and others.

The section-based classification baseline allows us to exploit the textual clues and context provided by specific sections in scientific publications. For example, the *"Methodology"* section typically contains details about the experiments, methodologies, and results of the authors' own research, making it likely that mentions of datasets within this section are used in their work. Conversely, the *"Related Work"* section usually focuses on reviewing prior studies and establishes the background knowledge, implying that dataset mentions here are more likely to be background references.

By training a classification model on this section-based labeling scheme, we aim to capture the inherent connection between the content of different sections and the intentions behind dataset mentions. While this baseline may oversimplify the nuanced nature of dataset mentions in scientific literature, it provides a starting point for our investigation and serves as a foundation for comparing the performance of more advanced models.

Models. Two distinct models were evaluated: an SVM model and a SciBERT model. This selection aimed to explore the effectiveness of both traditional machine learning models and pre-trained language models.

The SVM model is a well-established and widely used algorithm in classification tasks. We train the SVM model using a set of features extracted from the dataset mention contexts, including word embeddings and n-grams. By leveraging the SVM's ability to find an optimal hyperplane that maximally separates the two classes, we can assess the discriminative power of the utilised features in distinguishing intentions between own usage and background. Previous work on citation intent classification demonstrated that the SVM model performs better than traditional machine learning techniques such as linear regression, K-nearest neighbors and Naive Bayes [30].

In addition to the SVM model, we employ a SciBERT model, which is a variant of the BERT architecture specifically designed for scientific text understanding [2]. SciBERT is pre-trained on a large corpus of scientific literature from the computer science and biomedical domain, enabling it to capture a contextual understanding of scientific concepts. Previous work on citation intent classification has shown that SciBERT is capable of achieving a competitive performance for a similar task [16,33,37]. We will fine-tune the SciBERT model to learn the nuances of dataset intention recognition. Evaluation will provide insights into the effectiveness of leveraging domain-specific pre-training in capturing the subtle cues necessary to differentiate between background information and own usage intentions.

4 Experiments

In this section we describe the details of our experimental setup. This is done by first describing the research design of experiments, followed by the data that were used and lastly the specific implementation details.

4.1 Research Design

RQ1. The extraction process was evaluated for multiple variants of KM2 and KM3. Among the tested variants were the original version, the combination with each of the three binary classifiers when filtering was only applied at the first described filtering step and when filtering was applied for both the first filtering step and the second filtering step. This implies that seven unique variants were examined for each extraction model.

The various model combinations were also compared to a benchmark method. This algorithm kept a set of all known dataset titles and labels from the train set, and then searched for any matches in the texts of the test documents.

Evaluation. Besides precision and recall, the same $F_{0.5}$ metric for token-based Jaccard scores as described by the SUTD competition was used to evaluate the extraction models. The $F_{0.5}$ measure assigns a bigger weight to the precision of extracted dataset mentions compared to the recall. The metric is relevant for the case of dataset mention extraction, as the correctness of predictions is more important than ensuring that dataset mentions are always detected.

In addition, to measure the impact on the computational costs of the models, the runtime is also used as an evaluation metric. Two phases are relevant in terms of runtime: filtering and extraction. The filtering phase is the set of actions where a list of section texts is filtered by a binary classifier. The extraction phase consists of every step afterwards until a prediction string is assigned to each of the test documents. Runtimes were measured for multiple runs, to tackle stochastic influences, and only for experiments concerning the data that was split by the dataset labels.

RQ2. The experimental setup for intent classification consists of two parts. Firstly is to train and validate the two machine learning models in a weakly supervised learning set-up. Secondly is to evaluate them on the annotated test set. Test set results are compared to the baseline that bases intent classes solely on the corresponding section titles.

Before applying the SVM model, text inputs were vectorized using a *TF-IDF* embedding. Different n-gram ranges were tested beforehand, where a range between unigrams and trigrams came out best. For SciBERT, text inputs were vectorized using a SciBERT-based tokenizer. The training process of SciBERT was fine-tuned on the validation set, after a 80/20 split on the train set. Tuning was performed to improve the number of epochs and the learning rate.

Evaluation. To evaluate model predictions for the intent classification task, we rely on the accuracy, precision, recall and the F1 measure. These four classification metrics are commonly used to evaluate models for NLP tasks.

In addition, a qualitative reflection on the explainability of the models was performed. For SVM, this was done by taking the features with the most positive and most negative model coefficient scores after training. For SciBERT, the most important features were extracted using the *Transformers Interpret* package in Python [27]. Here, attribution scores can be calculated for the outputs of transformer-like models, based on integrated gradients and layer-integrated gradients. All of the attribution scores are calculated using PyTorch's explainability package *Captum* [12].

4.2 Data

RQ1. The dataset from the SUTD competition was used for experiments on dataset mention detection. Annotated dataset mentions from public test set nor the private test set were accessible to the public. Hence, only the training data has been used. Annotated dataset mentions are found in a CSV file, where every row represents the mention of a dataset by a certain publication. The metadata consists of the datasets official title and the label, which is the raw text excerpt containing the dataset mention. For the SUTD competition, predictions were evaluated based on the cleaned versions of the dataset labels. Labels are cleaned by replacing any single character that is not an alphabetical or numerical character with a blank space and lowercasing all alphabetical characters.

In addition, there is a JSON file for every publication containing the full text, separated into sections.

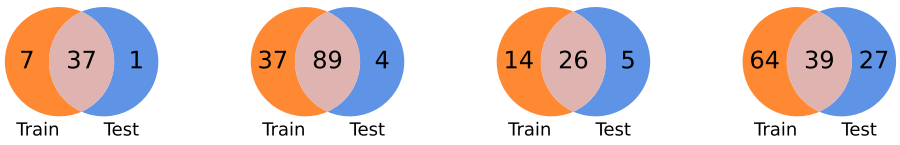
To train the binary classifiers, only the section texts that were at least 20 characters long were used as samples. Samples were labeled as positive (1) when they contained a dataset title or label associated with the publication, otherwise negative (0). Younes and Mathiak [38] demonstrated that this annotation method results in 12% of all samples being labelled as positive.

After pre-processing the documents, two methods of splitting the data into train and test subsets were applied. First, they were split into train and test sets with an 80/20 ratio. Secondly, the publications were split by firstly taking a random sample of 20 dataset labels from the 130 unique labels in total. All publications that were annotated with any of these 20 labels were added to the test set, while the remainder was kept for training. This resulted in 20.16% of the publications being used for testing, roughly similar to the split based on publications.

Informative statistics about the samples after the different split methods can be found in table 1. It can be seen here that one noticeable difference between the two splits is that the relative proportion of positives is smaller for the split on dataset labels. The distribution of dataset titles and labels across documents in the differently split sets is displayed in Figure 4. As a result of splitting at the label-level, there are 27 unseen labels instead of four. Additionally, the label-split test set contains five unseen dataset titles, in contrary to the random split on publications which only has a single unseen dataset title in the test set.

Table 1: Descriptive statistics for the pre-processed SUTD datasets.

Data split	**Set**	**# Publications**	**# Samples**	**# Positives**	**% Positives**
Publications	Train	11,448	186,336	22,451	12%
	Test	2861	46,963	5741	12%
Dataset labels	Train	11,424	188,200	24,270	12.9%
	Test	2885	45,099	3895	8.6%



(a) Distribution of titles for the publication split.

(b) Distribution of labels for the publication split.

(c) Distribution of titles for the label split.

(d) Distribution of labels for the label split.

Fig. 4: Venn diagrams of the overlap between unique dataset titles and labels in the train and test sets of the two different data splits.

RQ2. As this was a supervised learning task, dataset mentions in the text corpus required annotations for the intent target class. However, manual annotation on a full corpus of sufficient size was out of scope for the present thesis. Therefore, we used a large collection of weakly labelled data to train models. The weak labels were generated using the baseline method based on section titles. For example, dataset mentions in related work sections were labelled as background, while mentions in methodology sections were labelled as usage. In contrary, for a reliable evaluation and performance comparison of the selected models, we created a gold-standard test set through manual annotation.

Weakly Labelled Train Set. A novel training set was created by firstly querying the database of ScienceDirect[6], a large collection of scientific publications spanning various disciplines, and secondly extracting text snippets from the matched publications. Publications' XML-structured contents were parsed using the *ElementTree* XML API[7] in Python, providing the ability to extract section-wise contents of articles.

To safeguard reliability of the new dataset, it was formed by only searching for publications that contained an in-text reference to a dataset title or alias from a pre-determined set of aliases. The first source of key dataset aliases came from partners of the Coleridge Initiative. This set contained 165 unique dataset titles corresponding to a total number of 723 unique aliases. After removing non-alphabetical and -numerical characters, lowercasing strings and setting a length threshold of at least 5 characters, 683 aliases remained. The second source of key dataset aliases was the Dataset list [28], a list of popular machine learning datasets from across the web. From the Dataset list, 295 dataset titles were scraped. The final set was formed by taking the union of aliases from the two sources. After manually removing dataset aliases that were too ambiguous on their own, a total of 828 keywords were used for querying.

Querying the ScienceDirect corpus using the key dataset aliases resulted in matches for 125,388 section texts, of which 14,367 were related to the background class and 111,021 to the usage class. Section texts were tokenized into sentences using NLTK's sentence tokenizer [4]. Then, for each section text, only the first sentence that mentioned the matched dataset alias was kept, along with the three sentences before and after it. Furthermore, text snippets were only kept if the corresponding section titles matched with one of the pre-determined section titles. Removing duplicates from the initial 125,388 text samples resulted in 7272 samples for the background class and 41,194 for the usage class. To tackle class imbalance, only 10,000 random samples were selected from the usage class. The 17,272 samples in the final training set were connected to 244 unique dataset aliases. The validation subset, which came from an 80/20 split, contained 175 unique aliases.

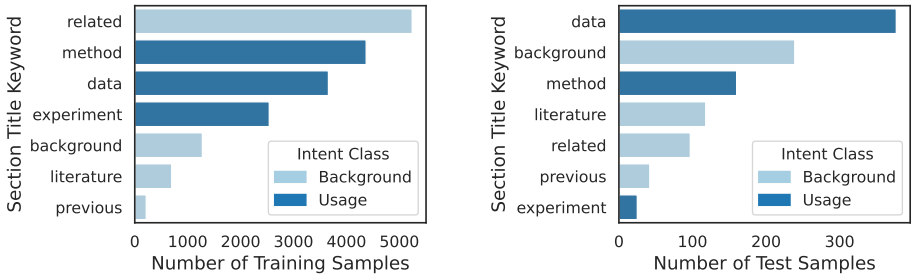
Figure 5a shows how the training samples were distributed across the different categories of section titles.

[6] https://www.sciencedirect.com/
[7] https://docs.python.org/3/library/xml.etree.elementtree.html

Gold-standard Test Set. Samples in the gold-standard set were derived from the SUTD dataset. All dataset titles that were mentioned in the train set of the SUTD competition were also present in the set of key dataset aliases provided by the Coleridge Initiative. Since the SUTD data was considered to be reliable, its samples were utilised as the source for the gold-standard test set. Section texts were selected by looking at those that contained a dataset mention and whose section titles matched with one of the defined intent-related titles. Snippets were then extracted from the section texts by locating the sentences containing dataset mentions and taking said sentence, the three sentences before and the three sentences after it.

Initially, the subset contained 479 samples from background-related sections and 4729 samples from usage-related sections. To avoid influences caused by class imbalance, a subset of 500 samples was taken from the usage-related samples. Among all usage-related samples, certain datasets were more frequently present than others. Hence, 100 of the 500 taken samples came from dataset aliases in the top 10 most popular ones, while the other 400 came from aliases outside it.

Figure 5b displays how the testing samples were distributed with regards to the section title keywords on which they were matched.



(a) Distribution for the training set.

(b) Distribution for the test set.

Fig. 5: Distribution of samples with regards to the section title keywords on which they were matched.

Annotation Procedure. During the annotating process, the 979 samples were randomly divided into 20 batches of 48 to 49 samples. Each sample in a batch contained the text snippet to annotate, as well as the dataset keyword on which the snippet was matched. The annotation process was designed to let each sample annotated by two different annotators. All disagreements between the two initial annotators were later resolved by a third annotator. Alongside one or more batches to annotate, annotators were provided with annotation guidelines. The guidelines contained descriptions of the annotation task and the two classes. Whenever a text sample contained different types of dataset mentions, annotators were instructed to always mark these as the *"usage"* class.

After the first round of annotation, the agreement across all 20 batches was 81.8%, with a standard deviation of 7.6%. In total, there were 178 cases of disagreements between the two initial annotators. Additionally, the inter-annotator agreement was calculated for all batches using Cohen's kappa coefficient [7]. Overall, scores across indicated a substantial agreement ($\mu = .634$, $\sigma = .153$). This annotation process resulted in 474 samples (48.4%) for the background class and 505 samples (51.6%) for the usage class.

4.3 Implementation

RQ1. The LR and SVM models were implemented using the *SGDClassifier* from Scikit-learn [26], set to utilise balanced class weights. Before applying the models, text samples were vectorized using Scikit-learn's *TfIdfVectorizer*, set to extract uni-grams, bi-grams and tri-grams.

The RoBERTa model [19] was implemented using the code provided by Younes and Mathiak [38]. The best model was noted as the variant that utilised Balanced Focal Loss (BFL), re-sampled positive and negative samples to a respective ratio of 4 : 0.55 and used section contents as input data. This model used the *RoBERTa-base* tokenizer from Hugging Face's Transformers library [10] and a 80% : 20% data split for training and validation. Other than that, F_3 was used as the validation metric, while the model was trained on 4 epochs with AdamW optimization [20]. Since BERT models only accept a maximum input of at most 512 tokens, the methodology of Younes and Mathiak was replicated by taking the first 382 and the last 128 tokens of each section text. The exact same values for hyperparameters as published by Younes and Mathiak were used.

For runtime measurements, all experiments were performed on an Amazon Sagemaker notebook instance. An *ml.g4dn.2xlarge* notebook instance type[8] was selected, which provided access to 32 GB instance memory and an NVIDIA Tesla T4 Tensor Core GPU.

RQ2. Similarly to Wahba et al. [34], a linear SVM was implemented using the *TfIdfVectorizer* together with the *LinearSVC* model from the Scikit-learn library [26]. To implement SciBERT, the *SciBERT-base* tokenizer and model from Hugging Face's Transformers library were used. Experiments were again conducted in an Amazon Sagemaker notebook instance. In this case, an *ml.g5.2xlarge* notebook instance type was selected, which provided access to 32 GB instance memory and an NVIDIA A10G Tensor Core GPU.

As SciBERT is an adaptation of the BERT architecture, the maximum length of tokenized input sequences is 512. The distribution of tokenized sequence lengths is displayed in Figure 6. For the weakly labelled data, 0.7% of the samples contained a tokenized input sequence longer than this. Since this proportion of samples was relatively small, it was decided to apply the same token selection to the limit-exceeding samples as the one used for RQ1. Namely, to take the first 382 and the last 128 tokens of tokenized sequences that are too long.

[8] https://aws.amazon.com/sagemaker/pricing/

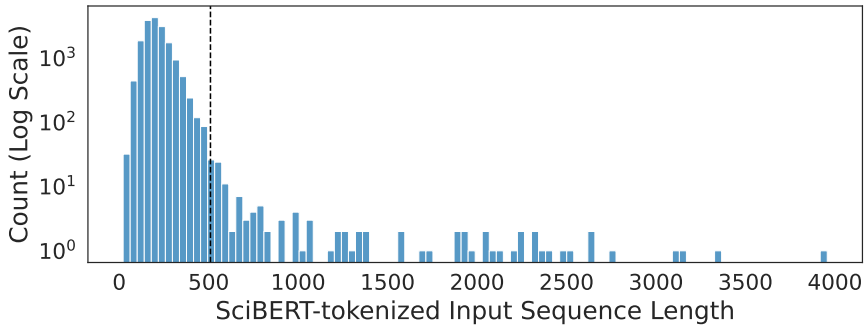


Fig. 6: Length distribution of SciBERT-tokenized input sequences, with counts displayed in log scale. The dashed line marks the maximum limit of 510 tokens.

In terms of hyperparameter tuning, the TF-IDF embedding was set to include unigrams, bigrams and trigrams, while the linear SVM was evaluated using its default settings. For SciBERT, grid search was performed and validated using the validation set for the learning rate, the batch size and the number of epochs. The best F_1 score was achieved by combining a learning rate of 2×10^{-5} with a batch size of 16 and fine-tuning the model for 2 epochs.

5 Results

This section presents the experimental results of this study. First are the results of the binary classifiers that are used for dataset mention detection, followed by results of the detection models, accompanied with further analysis of certain model errors. Last are the results for the intent classification task, where metric results, model bias and explainability of the models is briefly discussed..

5.1 Binary Classifiers for Dataset Mention Detection

Table 2 shows the results after training the binary classifiers and applying them to section samples of the publication-split test set. Here it can be seen that RoBERTa outperforms the other models regarding precision for positive samples and recall for negative samples. However, the three models are relatively within the same margin regarding the precision for negative samples and the recall for positive samples.

Results for the label-based split can be found in table 3. For the new data split, all models demonstrate that precision and recall for the positive label are lower compared to the publication-based split. Out of the three models, it occurs that RoBERTa is most heavily impacted by this change.

Table 2: Binary classifier results for negative and positive samples.

Model	Precision ↑		Recall ↑	
	Negatives	**Positives**	**Negatives**	**Positives**
LR	98%	42%	83%	91%
SVM	**99%**	45%	84%	**92%**
RoBERTa	**99%**	**63%**	**93%**	91%

Table 3: Binary classifier results for the split on dataset labels.

Model	Precision ↑		Recall ↑	
	Negatives	**Positives**	**Negatives**	**Positives**
LR	**96%**	31%	88%	**58%**
SVM	95%	**39%**	92%	52%
RoBERTa	95%	37%	**94%**	41%

5.2 Impact of the Binary Classifiers on the Detection Models

Measurements of the impact of the binary classifiers on the dataset mention detection models are discussed by firstly looking at results for the publication split and secondly for the dataset label split. For both splits, the metric results are followed by brief analyses of certain model errors. On the dataset label split, all detection models and their corresponding binary classifier combinations are also evaluated on runtime.

Split on Publications. The results for experiments concerning KM3 are shown in table 4. Here, it can be seen that the RoBERTa variant with double usage of the filtered texts achieved the best overall score. For RoBERTa, the double usage of filtered texts lead to a 6.1% increase in precision, as well as a 7.9% decline in recall.

The results for experiments concerning KM2 can be found in table 5. For KM2 it becomes clear that a single usage of the filtered texts improves the score for each variant, with RoBERTa achieving the best result with a score of 77.8%, a 5% increase compared to the original version. Double usage of the filtered texts improved precision by 0.5%, but decreased recall with 10.5%, worsening the $F_{0.5}$ score by 0.9%. Double usage did not improve $F_{0.5}$ for the other models either.

One important note is that the benchmark method, which only uses dataset labels from the train CSV file as a knowledge base for finding dataset mentions, resulted in an $F_{0.5}$ score of 80.9%, with a precision of 78.9% and a recall of 90.2%. Neither KM3 or KM2 were able to match the benchmark on $F_{0.5}$ for this data split, due to the precision much higher. However, both Kaggle models did achieve better recall scores compared to the benchmark method.

Table 4: Kaggle Model 3 results for the split on publications.

Model	**Texts**	**Usage**	$F_{0.5}$ ↑	**P** ↑	**R** ↑
Original	100%	-	54.1%	48.9%	**93.1%**
LR	25.3%	single	54.5%	49.4%	**93.1%**
		double	55.5%	51.0%	86.8%
SVM	22.1%	single	54.5%	49.4%	**93.1%**
		double	56.4%	51.7%	88.3%
RoBERTa	15.8%	single	55.0%	49.9%	**93.1%**
		double	**60.1%**	**56.0%**	85.2%

Table 5: Kaggle Model 2 results for the split on publications.

Model	**Texts**	**Usage**	$F_{0.5}$ ↑	**P** ↑	**R** ↑
Original	100%	-	72.8%	68.2%	**99.9%**
LR	25.6%	single	75.6%	71.3%	**99.9%**
		double	74.2%	70.9%	91.5%
SVM	22.5%	single	75.7%	71.4%	**99.9%**
		double	74.9%	71.4%	93.2%
RoBERTa	16.3%	single	**77.8%**	73.8%	**99.9%**
		double	76.9%	**74.3%**	89.4%

False Negatives. As the original version of KM2 achieved a recall score of 99.9%, no experimental combination improved upon it. The false negatives in this case are caused by a single dataset, namely the *RSNA International COVID-19 Open Radiology Database (RICORD)* dataset. It is the only dataset title in the test set that does not occur in the train set.

False Positives. The RoBERTa-single approach resulted in 1409 false positives. Within these false positives, 144 had a Jaccard similarity of at least 50% with a true positive for the same document. Of the 1265 false positives that did not have a high enough Jaccard similarity with a true positive, 209 predicted labels where derived from the annotated train set. Thus, the remaining 1056 false positives were found through the conditional filters of KM2. Besides, 51.1% of all false positives came from acronyms.

Inter-Model Differences. Between RoBERTa-double and the original version of KM2, the RoBERTa variant falsely detected one particular dataset label for five documents, whereas the original version did not. This is the case for *Rural Urban Continuum Codes*. The fault of RoBERTa in this case is that it predicts the full name correctly, like the original version, but it also predicts its acronym *RUCC*. Compared to RoBERTa-single however, there are two documents where

RoBERTa-double predicts the full name but not the acronym. Furthermore, the SVM model falsely detects an alternate variant of the same dataset, one that goes by *RUCCS*, while RoBERTa does not detect it as a dataset name.

Filter Usage. During analysis, some patterns were observed for the differences between models with single and double usage of filtered texts. For instance, there are examples where the model with a single usage predicts both the singular and plural version of the title, while the model after double usage does not submit either. Another type of difference is that the single usage often submits both the full name and the acronym for a certain dataset as separate predictions for the same publication, while models with double usage only predict the full name. There are also occurrences where the model with single usage predicts some unrelated dataset, whereas the double filter combination does not.

Split on Dataset Labels. Experimental results for KM3 and KM2 on the data split based on dataset labels can be found in tables 6 and 7, respectively. Compared to the data split on publications, nearly all scores are lower for both models. KM3's recall is the only exception, which is now slightly higher at 95.5% instead of 93.1%. Of all tested models, the combination of KM2 and RoBERTa with a single filter usage achieved the best $F_{0.5}$ and precision, while recall is 7.6% lower compared to the split on publications.

The benchmark comparison for the original Kaggle models achieved an $F_{0.5}$ score of 71.4% on this data split, with a precision of 92% and a recall of 37.6%. The particularly low recall demonstrates how the benchmark method is not useful for detecting unseen dataset titles, as it only refers to a knowledge base that is never extended with new labels. While the two Kaggle models achieved much better recall scores on this data split compared to the benchmark, neither could match it on $F_{0.5}$ due to the high precision.

False Negatives. From the five unseen dataset titles, the best model combination was able to detect three. *COVID-19 Image Data Collection* and *Our World in Data COVID-19 dataset* are the only unseen dataset titles which the model failed to recognise. The lower recall is mostly caused by dataset titles related to *Our World in Data*, which is related to 75% of the false negatives in the test set.

False Positives. Looking at the false positives by the model combination KM2-RoBERTa-single, it is observed that a large proportion is related the dataset titles *Programme for International Student Assessment* (45%) and *Progress in International Reading Literacy Study* (16%). Furthermore, the false positives contain numerous occurrences where different dataset labels of the same dataset title are predicted. Compared to the original version of KM2, the dataset labels *National Household Education Survey*, *NHES* and *Staff and Schooling Survey* were filtered out by the RoBERTa model. The filtering step resulted in an increase for precision of 7.2%.

Table 6: Kaggle Model 3 results for the split on dataset labels. Averaged runtimes (in seconds) are measured as *Filter Time (FT)* and *Extraction Time (ET)*.

Model	**Texts**	**FT (s)** ↓	**Usage**	$F_{0.5}$ ↑	**P** ↑	**R** ↑	**ET (s)** ↓
Original	100%	-	-	44.8%	39.5%	**95.7%**	783 ± 8
LR	14.5%	99 ± 2.4	single	**46.0%**	40.7%	**95.7%**	758 ± 5.7
			double	44.4%	41.5%	62.1%	722 ± 7.7
SVM	10.3%	91 ± 4.6	single	44.4%	41.5%	62.1%	761 ± 9.8
			double	45.2%	43.1%	56.3%	684 ± 11.9
RoBERTa	10.4%	1565 ± 19	single	45.2%	40.2%	90.5%	744 ± 3.7
			double	45.1%	**43.7%**	51.8%	686 ± 12.7

Table 7: Kaggle Model 2 results for the split on dataset labels. Averaged runtimes (in seconds) are measured as *Filter Time (FT)* and *Extraction Time (ET)*.

Model	**Texts**	**FT (s)** ↓	**Usage**	$F_{0.5}$ ↑	**P** ↑	**R** ↑	**ET (s)** ↓
Original	100%	-	-	60.6%	55.8%	**92.3%**	392 ± 16
LR	14.8%	89 ± 1.7	single	63.3%	58.7%	91.7%	205 ± 3.7
			double	58.2%	57.6%	60.4%	196 ± 2.6
SVM	10.6%	90 ± 6.9	single	65.7%	61.4%	91.7%	170 ± 5.4
			double	59.9%	61.3%	54.6%	160 ± 6.6
RoBERTa	10.7%	1684 ± 31	single	**67.3%**	**63.0%**	92.1%	176 ± 3.1
			double	60.5%	62.4%	54.0%	156 ± 6.5

Runtimes. When considering the runtimes for filtering the initial set of texts, it can be seen that the filtering process for RoBERTa takes much longer than the other two models. Combining an extraction model with a binary classifier leads to a lower extraction time in every considered combination. This is especially the case for KM2, where runtimes are more than halved for SVM and RoBERTa. Furthermore, double usage of filtered texts instead of single usage leads to a lower extraction time in every considered combination.

5.3 Intent Classification

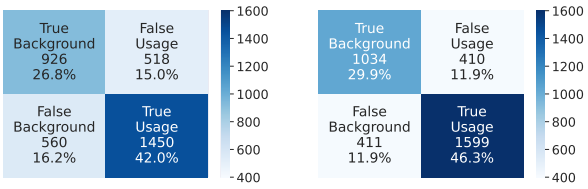
Metric results for the intent classification task can be found in Table 8. On the validation set, the SVM model is outperformed by SciBERT with a margin of at least 6% for every metric. The differences between the two models are smaller for the test set, but SciBERT still outperformed SVM at each metric. Looking at the baseline results, SVM was not able to match the baseline regarding accuracy, precision and F_1, but it did exceed the baseline's recall by 3%. In contrary, SciBERT performed better in terms of both recall and F_1, yet it also was not able to match the baseline in terms of accuracy or precision.

Table 8: Metric results on the validation and test sets. Precision, recall and F_1 are measured for *"usage"*, the positive class.

Model	**Accuracy ↑**		**Precision ↑**		**Recall ↑**		F_1 **↑**	
	Val.	**Test**	**Val.**	**Test**	**Val.**	**Test**	**Val.**	**Test**
Baseline	-	**79%**	-	**78%**	-	79%	-	78%
SVM	69%	74%	74%	72%	72%	82%	73%	76%
SciBERT	**76%**	78%	**80%**	75%	**79%**	**87%**	**79%**	**81%**

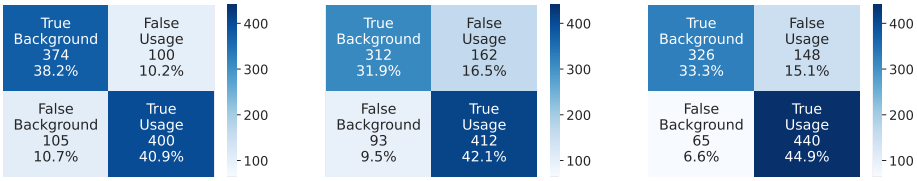
Confusion matrices of the predictions are displayed in Figure 7 and Figure 8 for the validation set and the test set, respectively. On both evaluation sets, SVM and SciBERT are better capable of identifying usage samples compared to the background samples. Notable for the test set, is that SVM and SciBERT ended up with a relatively higher number of false predictions for the usage class compared to the background class. Nonetheless, while the two models correctly recognised more samples of usage, neither could match the number of correct background samples that were identified by the baseline.

Between the predictions of SVM and SciBERT, there was a moderate agreement on the validation set (agreement = 74.8%, Cohen's $\kappa = .485$) and a substantial agreement on the test set (agreement = 82.2%, Cohen's $\kappa = .632$). This indicates that the models share a better understanding of recognising signs of usage of the dataset mentions than the kind of section title they originate from.



(a) SVM. (b) SciBERT.

Fig. 7: Confusion matrices for the validation set.



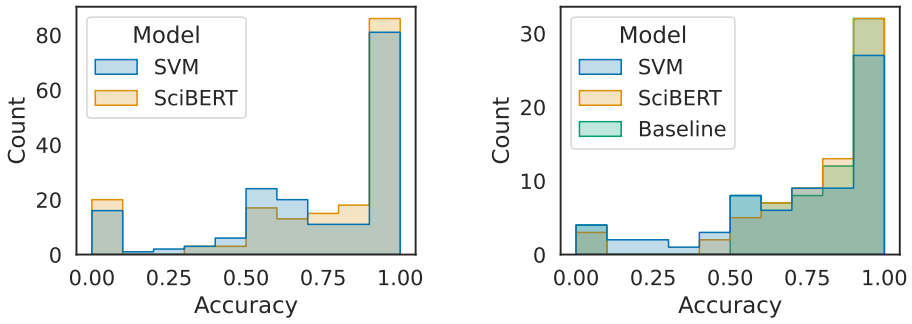
(a) Baseline. (b) SVM. (c) SciBERT.

Fig. 8: Confusion matrices for the test set.

Model Bias. Model bias towards certain dataset aliases in the evaluation sets was investigated through two methods of exploratory analysis.

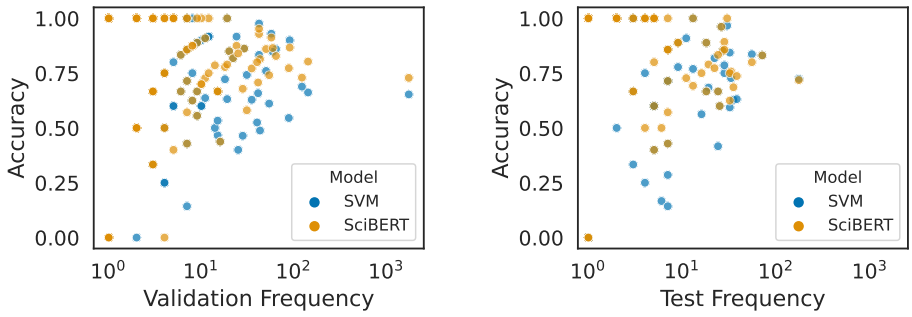
First, the distribution of accuracy scores for each unique alias in the two evaluation sets can be found in Figure 9. The histograms for both models reveal that the intent predictions for a lot of dataset aliases are fully correct. A large proportion of dataset aliases fall below 100% but above 50% accuracy. Since 50% accuracy in a binary classification task would be equivalent to random guessing, it makes sense that very few aliases have an accuracy below 50%.

Second, Figure 10 shows the relationship between the accuracy for each dataset alias and their frequency in the evaluation sets. Fully correct and fully wrong accuracies only comprise a small number of samples. However, Spearman's rank correlation coefficient [1] indicated a weak relationship for the validation set ($r = .194$) and a moderate relationship for the test set ($r = .362$).



(a) Validation set. (b) Test set.

Fig. 9: Distribution of model accuracies for every dataset alias.



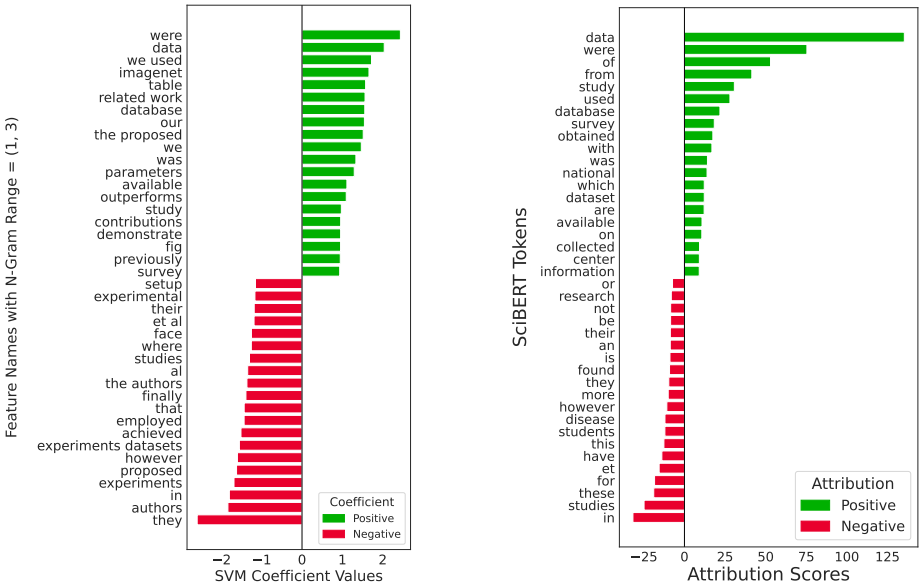
(a) Validation set, x-axis in log scale. (b) Test set, x-axis in log scale.

Fig. 10: Relationship between model accuracy and frequency in the respective evaluation set for every dataset alias.

Explainability. The most important features of the two models are displayed in Figure 11.

It can be seen that "were" and "data" are the two most important features towards the usage class for both models. As the feature "was" also occurs in the top 20 for both, it becomes clear that the verbs in the passive form are a strong indicator for the usage class. Surprisingly, the token "related work" also has a relatively high coefficient value for SVM, even though this category of sections is associated with the opposing class in this experiment. Other than that, most of the remaining features with a positive importance are related to phrases where authors describe the source of their data or their utilisation of it.

Regarding the most important features towards the background class, the tokens "in", "they", "their", "et al" and "studies" are important features for both models. Likewise, the features "however" and "not" are present, indicating that a contrast can be a strong indicator for the background class. Interestingly enough, especially given the presence of "related work" among the most positive features, the tokens "experimental" and "setup" occur in the top 20 most negative features for SVM. Other features of SVM include words that are typically used to refer to previous work, such as "the authors", "proposed" and "employed".



(a) Fitted SVM's most important features, based on model coefficients after training.

(b) Fine-tuned SciBERT's most imporant features, based on aggregated token attribute scores from the test set.

Fig. 11: The top 20 most positive-scoring and top 20 most negative-scoring features for SVM and SciBERT after training.

6 Discussion

The findings of the present study are discussed in this section. For both research questions, it is discussed what the implications of the findings are and how they relate to the findings of previous work. Furthermore, this section contains a reflection on the limitations of this study, while also providing suggestions for future research directions.

6.1 Dataset Mention Detection

Implications. The dataset mention detection task can be reflected upon from two broad perspectives, namely by looking at the binary classifiers on their own and the way in which they affect the extraction models.

Binary Classifiers. Results obtained from the binary classifiers on the publication-split indicate that RoBERTa outperforms the other models in terms of precision for positive samples and recall for negative samples. However, the three classifiers show similar performances regarding the precision for negative samples and the recall for positive samples. When considering the label-based split, it is evident that all models exhibit lower precision and recall for positive samples compared to the publication-based split. This indicates that the models struggle more when tasked to identify unseen dataset titles. RoBERTa seems to be more sensitive to this change, as its performance is noticeably affected compared to the other models. One speculative explanation for this is that the contextual information surrounding dataset mentions is not useful enough for the models to correctly recognise dataset mentions, causing the models to rely too much on the dataset titles alone.

These findings make for an interesting contrast to the research by Younes and Mathiak [38]. RoBERTa results on the publication-split were nearly identical, but for the label-split there was a clear drop of performance at certain metrics. This shows how the model slightly lacks the robustness to generalise to unseen dataset titles, something that was not examined by Younes and Mathiak [38]. However, the weak generalisability of BERT-based models to detect unseen datasets is in line with more recent studies on dataset mention detection [25].

Extraction Models. Results regarding the extraction task demonstrated that the incorporation of binary classifiers improves the $F_{0.5}$ score, with the exception of a few combinations. Increasing the application of filtered texts resulted in a greater improvement in precision. This was until a certain point for most models, where the $F_{0.5}$ score started to decrease due to the trade-off in recall. With the exception of KM3 on the label-split, RoBERTa demonstrated superior performance compared to the other models in relation to $F_{0.5}$.

Regarding the runtimes, it was observed that the advantages from utilising a binary classifier were contingent upon the specific extraction model employed, with KM2 demonstrating greater benefits compared to KM3. Combinations involving SVM and RoBERTa exhibited the shortest extraction runtimes, resulting

in a reduction of over 50%. Nonetheless, it should be noted that the chosen hardware configuration led to more time being required for both training and filtering processes in the case of RoBERTa, due to its reliance on GPU resources.

Interpretation. All in all, the experimental outcomes suggest that the incorporation of binary classifiers benefits the correctness as well as the efficiency of existing dataset mention detection models. RoBERTa in particular is a suitable model to improve the correctness of predicted dataset mentions. The drawbacks of the model come down to its relatively weaker ability to generalise to unseen instances and reliance on GPU resources with respect to efficiency. However, these drawbacks can be compensated by opting for a slightly lesser performing traditional machine learning model.

Limitations. To start, it is important to acknowledge that the benchmark used in this study does not possess perfect precision, resulting in false positives that can be considered correct. In some cases, the models successfully extract both the full title of a dataset and its corresponding acronym and submit both as predictions. However, due to the way the dataset was annotated, there were instances where only one of the two could be considered a correct prediction. This imperfection in the dataset introduces a level of ambiguity and may affect the overall reliability of the obtained results. In addition, the availability of unique dataset names in the SUTD dataset is limited, which may have implications for the generalisability of the findings beyond this specific dataset. As shown by the results on the label-split validation set, the models were not completely robust to detect mentions of unseen datasets.

Another important limitation to consider is the restricted scope of model selection. The study primarily focused on a particular set of classifiers and embedding techniques, which may not encompass the full spectrum of possibilities. This limitation raises the need for further exploration and comparison of alternative models to gain a more comprehensive understanding of their performance. Furthermore, it is worth noting that the tuning and validation process was relatively minimal, which may have affected the robustness and generalisability of the selected models as well. Besides, the absence of K-fold cross-validation on distinct data splits with respect to unseen datasets leaves room for potential bias and overfitting concerns.

Finally, the hardware setup employed in this study utilised a simple Sagemaker notebook instance for runtime measurements. While the decision to use a less resource-intensive hardware setup was driven by cost considerations, it is important to acknowledge that this choice may not accurately reflect the performance or behavior of the tested models in larger or more complex systems. The cost-driven hardware selection introduces a potential limitation when generalising the obtained results, considering the runtimes that were measured and the absence of hyperparameter tuning.

6.2 Intent Classification

Implications. A weakly-supervised learning setup was created to evaluate how well different types of classification models were able to recognise the intention behind mentioning a dataset. The results of the intent classification task revealed notable differences between the SVM and SciBERT models. SciBERT consistently outperformed SVM on both the validation and test sets, demonstrating superior performance across all metrics. These findings suggest that using a pre-trained transformer model, such as SciBERT, can lead to higher intent classification accuracy compared to traditional machine learning approaches.

Upon comparing the baseline results with the model performances, it becomes evident that neither SVM nor SciBERT achieved the same level of accuracy or precision as the baseline. However, both models exceeded the baseline's recall, with SciBERT achieving the highest recall scores overall. These findings indicate that the machine learning models are well capable of detecting signs of dataset usage that occur outside of methodology sections. At the same time, the models have the weakness that this also leads to a higher number of false positives.

The fact that SciBERT outperformed SVM at every metric demonstrates the power of pre-trained language models, as previously stated in research on citation intent classification [16,37]. However, related work on general text classification tasks showed that an SVM can offer a competitive counterpart to BERT-based models [34]. This is not the case for dataset mention intent classification, although there was no feature engineering besides the vectorisation of raw texts.

Limitations. Firstly, the test set exhibited a relatively small number of unique dataset aliases compared to the training and validation sets, raising concerns about potential bias. This discrepancy in dataset aliases could affect the model's generalisability and performance on unseen instances. Moreover, the test set was sourced from a different dataset, namely the SUTD data, causing dataset aliases to overlap with the training and validation sets. This overlap may introduce a source of bias, potentially inflating the model's performance due to its familiarity with the overlapping aliases. However, analysis of the relationship between alias frequency and accuracy did not suggest a strong influence to be present.

Annotation Procedure. Secondly, one of the main obstacles for this research was to create a dataset to perform experiments on. Due to practical constraints it was decided to work with a weakly-labelled dataset, where the training data was annotated automatically according to a set of rules while the relatively small test set was annotated by human annotators. The baseline approach of labelling by type of section title, that was used to annotate the train data, achieved an accuracy of 79% on the test set, revealing that its intent labels were wrong in at least 20% of the cases. Moreover, correlation tests for SVM and SciBERT showed that the models were in fact able to learn the intent recognition task from noisy training data. Nonetheless, differences between the baseline's logic and the goal of intent recognition suggest there is reason to explore how well models perform when trained on more appropriate data.

Moreover, limitations can arise from the annotation task's relatively rushed execution, resulting in potential conflicts within the annotation guidelines. The conflicts may have introduced subjectivity and inconsistency in the labeling process, potentially impacting the reliability and accuracy of the labeled data. Nonetheless, double annotations for each sample and evaluation of the inter-annotator agreement according to Cohen's kappa showed that there was a substantial agreement among annotators.

Experimental Scope. As with RQ1, the model selection process was limited, employing only SVM to represent traditional machine learning models and SciBERT to represent pre-trained language models. Although these choices offer reasonable benchmarks, the absence of a broader range of models could restrict the exploration of alternative architectures and their comparative performance. The absence of K-fold cross-validation with different validation splits is another limitation. This approach could have provided a more robust evaluation of the model's performance, by reducing the influence of specific data configurations and validating its effectiveness across various subsets of the dataset.

6.3 Future Work

This paper provides four broad suggestions for future research directions.

Direct Follow-up Experiments. One potential direction of future research would be to directly expand the conducted experiments regarding dataset mention detection. Since model selection was limited to two extraction models and three binary classifiers, the first step could be to explore other models. For example, the SUTD Kaggle competition winning submission was out of scope for the present study. It would be a logical step to evaluate the effects of different binary classifiers on this extraction model, or even to look beyond submissions from the SUTD competition. Not only are there more extraction models to examine, but a variety of different embedding techniques and binary classification models is also of interest to study further. The same can be said for the task of intent classification, where only two models have been examined. Likewise, further experimentation with certain model configurations, including the classification threshold, the amount of training data or the loss function, could lead to performance improvements.

Entity Masking and Linguistic Features. Future research can also focus on investigating the relationship between specific linguistic features and their association with dataset mentions. The differences in terms of performance between detecting seen and unseen dataset titles in dataset mention detection showed how reliant models can be on the dataset titles themselves, rather than the surrounding contextual information. A logical follow-up to this finding is to experiment with entity masking techniques, where dataset labels are replaced with general masking tokens, forcing the models to make use of contextual information.

Furthermore, inspection of the intent classification models revealed certain linguistic patterns were found to have notable associations with either of the two classes. For instance, features such as "we", "our" and verbs in the passive form were predominantly linked to the usage class, while features like "they", "their" and "proposed" were primarily associated with the background class. However, it is important to note that this study already examined the capabilities of SciBERT, a language model specifically designed for scientific texts. Further investigations can still explore alternative models or refine existing ones in order to better capture and utilise these linguistic cues for enhanced correctness.

Creating a Novel Dataset. One defining aspect of this work has been the weakly-supervised setup for the intent classification task. For future research on dataset mentions, the availability of a larger annotated dataset would greatly benefit the scientific field. The creation of such a dataset should be guided by stringent annotation guidelines, accompanied by well-defined definitions for various classes of intentions pertaining to dataset mentions. Moreover, future datasets do not need to be limited to the two classes that were studied in the present paper, but are even encouraged to be built on more broadly defined taxonomies as seen with previous work in citation intent classification.

The creation of a dataset like this would facilitate the development of more accurate and robust intent classification models. On top of that, an expanded dataset would enable secondary experiments, including the performance evaluation of models across different scientific domains of publications. Additionally, it would open avenues to explore the application of intent classification in a zero-shot setting, thereby enhancing the generalisability and practicality of models in real-world scenarios.

Entity Linking. Finally, future research in dataset mention detection should also shift focus towards entity linking, thereby associating dataset mentions with existing dataset entities. Drawing inspiration from related work on citation intent classification [3], dataset search [11] and author name disambiguation [29], we propose utilising meta-information of datasets and exploring the applicability of graph-based algorithms. By leveraging knowledge graphs, researchers can establish a comprehensive network of dataset entities with rich semantic relationships, facilitating the disambiguation of dataset mentions. This approach can advance the understanding of dataset usage, foster data integration across domains, and contribute to the development of automated tools for dataset retrieval and analysis in scientific literature.

7 Conclusion

In this study, we investigated the effectiveness of different models and classifiers for dataset mention detection and intent classification in scientific publications. Our research questions aimed to understand the behaviour of models, draw conclusions based on the obtained results and provide suggestions for future work.

Regarding the first research question, we explored various combinations between extraction models and preliminary classifiers for dataset mention detection. Findings revealed that RoBERTa outperforms the other models for most metrics. However, the models struggled when tasked with identifying unseen dataset titles, most notably RoBERTa, suggesting a limited ability to generalise. For such cases, it can still be beneficial to consider traditional machine learning models. These findings highlight the need for further investigation into the subject area, especially regarding the generalisability of BERT-based models to unseen instances.

As for the extraction task itself, results showed that the inclusion of binary classifiers improved the $F_{0.5}$ score and reduced the runtime. However, it should be noted that the efficiency of RoBERTa was negatively affected by the used hardware setup, due to the models reliance on GPU resources. Overall, the incorporation of binary classifiers yielded favorable outcomes, enhancing the correctness and efficiency of the existing extraction models. These findings contribute to the collective knowledge and ongoing advancements in machine learning models dedicated to this task.

With respect to the second research question, which focused on intent classification, we compared the performance of SVM and SciBERT models. Results demonstrated that SciBERT consistently outperforms SVM, showcasing the strength of pre-trained transformer models for various NLP tasks. Although the models detected signs of dataset usage occurring outside the normally associated sections, they also exhibited a higher number of false positives. These findings highlight the potential of pre-trained models for this task, but also emphasize the need for further exploration in the discipline.

To open new avenues for research, we propose several future directions. Firstly, expanding the range of models and configurations could directly lead to performance improvements. Additionally, investigating the relationship between dataset mentions and linguistic modeling, including entity masking and feature engineering, could lead to discovering new patterns, while the creation of a larger annotated dataset with well-defined annotation guidelines would enhance the development of more accurate models. Furthermore, exploring entity linking techniques could advance the understanding of dataset usage and facilitate dataset retrieval and analysis in scientific literature.

In conclusion, this study contributes to the field of dataset mention detection and intent classification in scientific literature by evaluating different models and classifiers. The findings underscore the strengths and limitations of these approaches and provide valuable insights for future research. By addressing the findings of this paper and the suggested avenues for future work, researchers can continue to improve the accuracy, robustness, and applicability of machine learning surrounding the usage of datasets. Ultimately, these contributions advance the scientific field, enhancing the transparency, efficiency and impact of scholarly discourse.

Acknowledgements. First, I express my sincere gratitude to Elsevier, a pioneering company in the field of academic publishing, for providing me with an invaluable internship opportunity. Second, I thank my internship supervisor Hosein Azarbonyad and academic supervisor Michael Cochez for their guidance and mentorship throughout the course of my thesis project. Furthermore, I would like to thank Dimitrios Alivanistos, Artemis Capari, Florian Kunneman and Noa van Mervennée for being there to discuss ideas with me and for giving me incredibly helpful advice about various aspects of the thesis. For their indispensable assistance in the process of annotating, I thank Zubair Afzal, Federico Andreetto, Savvas Chamezopoulos, Georgios Cheirmpos, Yury Kashnitsky, Dan Li, Janneke van de Loo, Nishant Mintri, Anke Otten, Matei Penca and Seyed Amin Tabatabaei. I also thank all others on Elsevier's Data Science Research Content team for their genuine interest and active engagement in the numerous stand-up meetings during my internship. Last but certainly not least, I extend my heartfelt appreciation to my family and friends for their unwavering support in maintaining my personal well-being over the past six months. Their constant presence, encouragement and understanding have been vital for me to successfully navigate through this significant phase of my life.

References

1. Spearman Rank Correlation Coefficient, pp. 502–505. Springer, New York, NY (2008). https://doi.org/10.1007/978-0-387-32833-1_379
2. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1371
3. Berrebbi, D., Huynh, N., Balalau, O.: Graphcite: Citation intent classification in scientific publications via graph embeddings. In: Companion Proceedings of the Web Conference 2022. p. 779–783. WWW '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3487553.3524657
4. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. (2009)
5. Callaghan, S.: On the importance of data transparency. Patterns **1**(4) (2020). https://doi.org/10.1016/j.patter.2020.100070
6. Cohan, A., Ammar, W., van Zuylen, M., Cady, F.: Structural scaffolds for citation intent classification in scientific publications. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3586–3596. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1361
7. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960). https://doi.org/10.1177/001316446002000104
8. Heddes, J., Meerdink, P., Pieters, M., Marx, M.: The automatic detection of dataset names in scientific articles. Data **6**(8) (2021). https://doi.org/10.3390/data6080084
9. Hou, Y., Jochim, C., Gleize, M., Bonin, F., Ganguly, D.: TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 707–714. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.59
10. Hugging Face: Roberta base model (2019), https://huggingface.co/roberta-base?
11. Irrera, O.: Data search in practice: How to find scientific datasets and to link them to the literature. In: FDIA2022: Future Directions in Information Access (2022)
12. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for pytorch (2020)
13. Kolyada, N., Potthast, M., Stein, B.: Beyond metadata: What paper authors say about corpora they use. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 5085–5090. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.451
14. Kumar, S., Ghosal, T., Ekbal, A.: Dataquest: An approach to automatically extract dataset mentions from scientific papers. In: Ke, H.R., Lee, C.S., Sugiyama, K. (eds.) Towards Open and Trustworthy Digital Societies. pp. 43–53. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-91669-5_4
15. Lafia, S., Ko, J.W., Moss, E., Kim, J., Thomer, A., Hemphill, L.: Detecting informal data references in academic literature. Deep Blue (2021). https://doi.org/10.7302/1671

16. Lahiri, A., Sanyal, D.K., Mukherjee, I.: Citeprompt: Using prompts to identify citation intent in scientific papers. In: ACM/IEEE Joint Conference on Digital Libraries 2023 (2023). https://doi.org/10.48550/arXiv.2304.12730
17. Lane, J., Gimeno, E., Levitskaya, E., Zhang, Z., Zigoni, A.: Data inventories for the modern age? using data science to open government data. Harvard Data Science Review **4**(2) (2022). https://doi.org/10.1162/99608f92.8a3f2336
18. Li, P., Liu, Q., Cheng, Q., Lu, W.: Data set entity recognition based on distant supervision. The Electronic Library **39(3)**, 435–449 (2021). https://doi.org/10.1108/EL-10-2020-0301
19. 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. CoRR **abs/1907.11692** (2019). https://doi.org/10.48550/arXiv.1907.11692
20. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in Adam. CoRR **abs/1711.05101** (2017). https://doi.org/10.48550/arXiv.1711.05101
21. Mercier, D., Rizvi, S.T.R., Rajashekar, V., Dengel, A., Ahmed, S.: Impactcite: An xlnet-based method for citation impact analysis. CoRR **abs/2005.06611** (2020), https://arxiv.org/abs/2005.06611
22. Mondal, I., Hou, Y., Jochim, C.: End-to-end NLP knowledge graph construction. CoRR **abs/2106.01167** (2021), https://arxiv.org/abs/2106.01167
23. Mullick, A., Pal, S., Nayak, T., Lee, S.C., Bhattacharjee, S., Goyal, P.: Using sentence-level classification helps entity extraction from material science literature. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 4540–4545. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.483
24. Oesterling, A., Ghosal, A., Yu, H., Xin, R., Baig, Y., Semenova, L., Rudin, C.: Multitask learning for citation purpose classification. CoRR **abs/2106.13275** (2021), https://arxiv.org/abs/2106.13275
25. Pan, H., Zhang, Q., Dragut, E., Caragea, C., Latecki, L.J.: DMDD: A large-scale dataset for dataset mentions detection (2023). https://doi.org/10.48550/arXiv.2305.11779
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
27. Pierse, C.: Transformers Interpret (Feb 2021), https://github.com/cdpierse/transformers-interpret
28. Plesa, N.: Dataset list - a list of the biggest machine learning datasets (2021), https://datasetlist.com
29. Rettig, L., Baumann, K., Sigloch, S., Cudré-Mauroux, P.: Leveraging knowledge graph embeddings to disambiguate author names in scientific data. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 5549–5557 (2022). https://doi.org/10.1109/BigData55660.2022.10020229
30. Roman, M., Shahid, A., Uddin, M.I., Hua, Q., Maqsood, S.: Exploiting contextual word embedding of authorship and title of articles for discovering citation intent classification. Complexity **2021**, 1–13 (04 2021). https://doi.org/10.1155/2021/5554874
31. Schwartz, A.S., Hearst, M.A.: A simple algorithm for identifying abbreviation definitions in biomedical text. In: Biocomputing 2003, pp. 451–462 (2002). https://doi.org/10.1142/9789812776303_0042

32. Singh, J., Wadhawan, A.: PublishInCovid19 at WNUT 2020 shared task-1: Entity recognition in wet lab protocols using structured learning ensemble and contextualised embeddings. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.wnut-1.35
33. Visser, R., Dunaiski, M.: Sentiment and intent classification of in-text citations using bert. In: Gerber, A. (ed.) Proceedings of 43rd Conference of the South African Institute of Computer Scientists and Information Technologists. EPiC Series in Computing, vol. 85, pp. 129–145. EasyChair (2022). https://doi.org/10.29007/wk21
34. Wahba, Y., Madhavji, N., Steinbacher, J.: A comparison of svm against pre-trained language models (PLMs) for text classification tasks. In: Machine Learning, Optimization, and Data Science: 8th International Conference, LOD 2022, Certosa Di Pontignano, Italy, September 18–22, 2022, Revised Selected Papers, Part II. p. 304–313. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-25891-6_23
35. Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K.A., Ceder, G., Jain, A.: Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. Journal of Chemical Information and Modeling **59**(9), 3692–3702 (2019). https://doi.org/10.1021/acs.jcim.9b00470
36. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)
37. Wright, D., Augenstein, I.: CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. In: Findings of ACL-IJCNLP. Association for Computational Linguistics (2021). https://doi.org/10.48550/arXiv.2105.10912
38. Younes, Y., Mathiak, B.: Handling class imbalance when detecting dataset mentions with pre-trained language models. In: Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022). pp. 79–88. Association for Computational Linguistics, Trento, Italy (Dec 2022), https://aclanthology.org/2022.icnlsp-1.9
39. Zaratiana, U., Holat, P., Tomeh, N., Charnois, T.: Hierarchical transformer model for scientific named entity recognition (2022). https://doi.org/10.48550/ARXIV.2203.14710