

Machine Learning for Subject Classification of Books Using Open Bibliographic Metadata

Livija Ružić

Osnovna škola Prelog
Trg bana Jelačića 2, 40323, Prelog
livija.ruzic@gmail.com

Dino Vlahek

Faculty of Organization and Informatics, University of Zagreb
Department of Theoretical and Applied Foundations of Information Sciences
Pavlinska 2, 42000 Varaždin
dvlahek@foi.hr

Abstract. *This study evaluates automatic subject classification of books using open bibliographic metadata from Open Library. A balanced dataset of 15,000 records was constructed across ten broad subject categories, using only titles and descriptions as predictive input, while original subject labels were used only to derive target classes. Classical TF-IDF-based classifiers were compared with MiniLM sentence embeddings followed by Logistic Regression. The transformer-based representation achieved the best performance, with a macro-F1 score of 0.6173, outperforming the strongest classical baseline. The results indicate that semantic embeddings can support broad subject-category prediction, although expert review remains necessary for ambiguous records.*

Keywords. automatic subject classification, bibliographic metadata, Open Library, machine learning, TF-IDF, transformer embeddings

1 Introduction

Libraries and digital repositories nowadays provide access to large and heterogeneous bibliographic collections. In such collections, metadata play a central role in discovery, retrieval, cataloguing, and user access. Among different types of bibliographic metadata, subject information is particularly important, as it allows users to identify relevant books even when they do not know the exact title, author, or terminology used in the catalogue. Subject classification can therefore be viewed as a key component of knowledge organization, since it connects bibliographic records with controlled vocabularies, classification systems, and user-oriented access points (Golub, 2019; Golub, Hagelbäck, & Ardö, 2021). However, subject assignment is a demanding process. It often requires expert interpretation of bibliographic descriptions, knowledge

of cataloguing practice, and the use of controlled subject vocabularies. This becomes increasingly difficult in large-scale digital collections, where records may be incomplete, inconsistently described, or available only through short textual fields. For this reason, automatic subject indexing and automatic subject classification have been studied as computational approaches for supporting metadata creation and enrichment (Golub et al., 2021; Suominen, 2019). These approaches are not intended to replace professional cataloguing, but to provide candidate subject categories that can be reviewed, corrected, or refined by experts (Mödden, 2022). Automatic classification of books is especially challenging because bibliographic records are usually short and semantically ambiguous. A title may contain only a small amount of information, while descriptions differ considerably in length, style, and quality. In addition, subject categories are not always mutually exclusive. For example, a book about a historical figure may be related to both biography and history, while a book about computing may overlap with science, technology, education, or business. Consequently, the task cannot be reduced only to keyword matching, since the same subject area may be expressed using different vocabularies, while different subject areas may share similar terminology.

Earlier approaches to automatic text classification commonly relied on lexical document representations, such as bag-of-words and TF-IDF (Term Frequency–Inverse Document Frequency), combined with classical classifiers (Salton & Buckley, 1988; Sebastiani, 2002). Such methods remain important baselines because they are computationally efficient, relatively transparent, and often effective when discriminative terms are present in the input text (Kowsari et al., 2019). In bibliographic classification, however, their performance depends strongly on the availability and quality of textual metadata. When records are short or

when subject information is expressed indirectly, lexical representations may fail to capture semantic similarity between texts that do not share the same vocabulary (Nordström, 2018). Transformer-based language models provide an alternative representation strategy by learning dense semantic embeddings of text. The transformer architecture introduced attention-based language modelling (Vaswani et al., 2017), while BERT demonstrated the effectiveness of deep bidirectional pre-training for language understanding tasks (Devlin, Chang, Lee, & Toutanova, 2019). Sentence-BERT further adapted transformer models for efficient sentence- and paragraph-level embeddings suitable for similarity comparison and downstream classification (Reimers & Gurevych, 2019). Such representations are relevant for bibliographic metadata because titles and descriptions are often short textual units whose subject meaning may be expressed indirectly. Recent work has also explored BERT-based models for assisted subject indexing, indicating their potential relevance for library metadata workflows (Chou, 2022). Despite these developments, the evaluation of automatic subject classification must be designed carefully. Bibliographic records often contain original subject labels or subject headings, which may be strongly related to the target category. If such fields are used as predictive input, the resulting evaluation may suffer from label leakage and produce overly optimistic results (Kaufman, Rosset, & Perlich, 2012). Therefore, a realistic experimental setting should distinguish between metadata used to define the target variable and metadata used as input to the classifier.

This study investigates whether broad subject categories of books can be predicted from open bibliographic metadata. The study is based on records obtained from the Open Library works data dump (Open Library, 2025). From these records, a balanced dataset of 15,000 books across ten broad subject categories is constructed: Art, Biography, Business, Education, Fiction, History, Medicine, Religion, Science, and Technology. The predictive input is limited to the book title and description, while the original subject labels are used only to construct the target category. The objective of this study is to compare classical lexical representations with transformer-based semantic embeddings for broad book subject classification. For this purpose, TF-IDF features are evaluated using Decision Tree, Random Forest, Linear Support Vector Machine, k-Nearest Neighbors, and a multilayer perceptron classifier. In addition, MiniLM sentence embeddings are combined with Logistic Regression in order to evaluate a compact transformer-based representation under the same classification setting. The contributions of this paper are as follows: (i) A reproducible dataset construction procedure is presented for deriving broad subject categories from open bibliographic metadata while avoiding the direct use of subject labels as predictive features. (ii) Classical TF-IDF-based machine learn-

ing models are compared with transformer-based semantic embeddings for book subject classification using the same input fields and target categories. (iii) The obtained results are discussed from the perspective of semi-automatic metadata enrichment, where automatic classification is used as support for expert cataloguing rather than as a fully autonomous replacement for it.

2 Related Work

As the problem addressed in this paper combines bibliographic metadata processing, automatic subject indexing, and supervised text classification, the related work must be considered from several perspectives. First, automatic subject indexing and metadata enrichment in library and information science are reviewed. Classical machine learning approaches for text classification are then discussed, since they represent important and interpretable baselines. Finally, transformer-based text representations are considered as a more recent alternative for capturing semantic relationships in short bibliographic descriptions.

2.1 Automatic Subject Indexing and Bibliographic Metadata

Automatic subject classification has been studied as part of broader efforts to support metadata creation, cataloguing, and knowledge organization in libraries. Traditional cataloguing relies on expert interpretation of bibliographic records and controlled vocabularies. However, the increasing scale of digital and digitized collections has made fully manual subject assignment difficult to maintain, especially when metadata are incomplete, inconsistent, or available at different levels of granularity. In order to address these challenges, automatic subject indexing approaches have been introduced to assist the assignment of subject labels, classification codes, and other descriptive metadata (Golub, 2019; Golub et al., 2021). Practical systems such as Annif further demonstrate that machine-learning-based subject indexing can be implemented as a tool for suggesting subject headings and supporting metadata creation workflows (Suominen, 2019). In this context, automatic classification is not usually considered a replacement for professional cataloguing, but rather as a decision-support mechanism. Machine learning models can suggest candidate subject categories that are then verified, corrected, or refined by domain experts. This is particularly important in bibliographic classification, where subject categories may be semantically overlapping, hierarchically organized, and dependent on cataloguing practice. Similar machine-learning approaches have also been investigated for assigning controlled subject headings to scholarly documents, further indicating the relevance of automatic classification for large-scale bibliographic organization (Kazi, Thompson, Choudhury, & Oard, 2021). Recent work

on artificial intelligence and machine learning in bibliographic control has therefore emphasized the role of such methods in supporting, rather than replacing, expert-based knowledge organization workflows (Mödden, 2022).

2.2 Classical Machine Learning for Text Classification

Earlier approaches to automatic text and bibliographic classification commonly relied on lexical representations of documents. In such approaches, textual fields are transformed into high-dimensional feature vectors using bag-of-words or TF-IDF representations, where terms are weighted according to their importance within a document and across a collection (Salton & Buckley, 1988). These representations are then used together with classical classifiers, such as Naive Bayes, Support Vector Machines, logistic regression, decision trees, and ensemble-based models. Such machine-learning-based text categorization approaches have been widely studied for assigning documents to predefined categories (Sebastiani, 2002). Although these methods are relatively simple, they remain important baselines because they are computationally efficient, robust on small and medium-sized datasets, and more interpretable than many neural approaches (Kowsari et al., 2019). In bibliographic classification, the usefulness of these methods depends strongly on the quality and richness of the available metadata. Titles, subtitles, descriptions, and abstracts may contain discriminative terms that indicate the subject of a work. However, bibliographic records are often short, noisy, or incomplete, while different books belonging to the same broad subject area may use substantially different vocabularies. As a result, lexical methods can perform well when subject categories are associated with distinctive terminology, but their performance is limited when the same subject is expressed indirectly or when different subjects share similar terms (Nordström, 2018). Another limitation of classical lexical approaches is that they usually represent documents according to surface-level word occurrence. Consequently, semantic similarity is captured only indirectly through shared vocabulary. This can be problematic in book classification, where short descriptions may refer to the same topic using different words, or where the subject of a book is implied by context rather than stated explicitly. Nevertheless, such methods are still useful in empirical studies because they provide transparent and computationally inexpensive baselines against which more complex representation-learning approaches can be compared.

2.3 Transformer-Based Text Representations

More recent approaches to text classification rely on neural language models that learn dense semantic representations of text. The transformer architecture introduced the attention-based modelling principle on which many modern language models are based (Vaswani et al., 2017). Building on this paradigm, BERT introduced deep bidirectional pre-training as a general framework for language understanding tasks (Devlin et al., 2019). Unlike TF-IDF representations, which encode documents primarily through weighted word occurrence, transformer-based models can capture contextual meaning and semantic similarity between texts even when they do not share the same vocabulary. In library and information science, BERT-based models have also been explored for assisted subject indexing, indicating their relevance for automatic metadata enrichment and subject classification tasks (Chou, 2022). Sentence-BERT further adapted transformer models for efficient sentence-level and paragraph-level representation by producing embeddings suitable for similarity comparison and downstream classification tasks (Reimers & Gurevych, 2019). Such representations are especially relevant for bibliographic metadata because book titles and descriptions are often short textual units. In these cases, the ability to encode semantic similarity beyond exact word overlap may improve classification performance, particularly for broad subject categories whose textual expressions vary across records. However, transformer-based approaches also introduce several practical limitations. They are generally less transparent than classical lexical models, require more computational resources, and their performance may depend on the choice of the pre-trained model and downstream classifier (Chou, 2022; Devlin et al., 2019; Reimers & Gurevych, 2019). In addition, when bibliographic subject labels are available in the metadata, care must be taken to avoid using them as input features, because this may lead to label leakage and produce overly optimistic classification results (Kaufman et al., 2012).

Although previous studies report encouraging results for automatic subject indexing and text classification, direct numerical comparison is difficult because datasets, metadata fields, target vocabularies, and evaluation protocols differ substantially. Some studies predict controlled subject headings, some use classification codes, while others classify documents into broad topical groups. Reported F1-scores are therefore not directly comparable to the present ten-category Open Library setting. Nevertheless, prior work generally indicates that performance decreases when the input is restricted to short metadata fields and when categories are broad or semantically overlapping. The present study contributes to this line of work by reporting a reproducible comparison on title–description metadata only, where the best model achieved a macro-F1 score

of 0.6173.

3 Methodology

This section describes the experimental procedure used for automatic subject classification of books from open bibliographic metadata. The methodology consists of five main steps: acquisition and parsing of Open Library records, extraction of relevant bibliographic fields, construction of broad subject categories, transformation of textual metadata into numerical representations, and evaluation of classification models.

3.1 Data Source and Record Extraction

The dataset used in this study was constructed from the Open Library works data dump. Open Library provides bibliographic records in a tab-separated format, where each row contains the record type, record key, revision, last modification date, and a JSON representation of the corresponding work (Open Library, 2025). In this study, the works dump was parsed line by line in order to extract records suitable for supervised subject classification.

For each valid JSON record, the following fields were extracted: the Open Library work key; the book title; the textual description; the original subject labels; the last modification date. The description field in Open Library records may appear either as a plain string or as a structured object containing a textual value. Both cases were handled during preprocessing. The title and description were cleaned by removing line breaks, tab characters, and repeated whitespace. They were then concatenated into a single textual input field. Records with insufficient textual information were excluded by requiring the resulting title–description text to contain at least five words.

Importantly, original subject labels were not used as input features. They were used only for constructing the target variable, as described in Section 3.2. This distinction was necessary in the present experimental design because the ten broad target categories were derived from the same subject-label field. Therefore, using subject labels as predictive input would introduce label leakage and lead to an unrealistic estimate of classification performance (Kaufman et al., 2012). This restriction is specific to the present setup. If the broad categories were obtained independently, for example through manual annotation by human experts, subject labels could represent a legitimate and valuable source of predictive metadata.

3.2 Construction of Target Categories

Open Library subject labels are highly diverse, fine-grained, and not always standardized. Therefore, direct classification using the original labels would result in a large and sparse label space. In order to obtain

Table 1: Dataset construction audit during Open Library parsing.

Filtering outcome	Records
Parsed JSON records	357,133
Missing/invalid subject list	174,688
No broad category matched	85,078
More than one category matched	23,291
Exactly one category matched	74,076
Fewer than five text words	36,648
Excluded after class cap	22,428
Final usable records	15,000

a controlled experimental setting, the original subject labels were mapped into ten broad subject categories: Art, Biography, Business, Education, Fiction, History, Medicine, Religion, Science, and Technology.

For each category, a manually defined set of representative keywords was used. For example, the Fiction category included terms such as *fiction*, *novel*, *stories*, *fantasy*, and *mystery*; the Technology category included terms such as *technology*, *engineering*, *computer*, *programming*, *software*, and *machine learning*; while the Science category included terms such as *science*, *physics*, *chemistry*, *biology*, *mathematics*, and *astronomy*.

All subject labels were normalized before matching by converting text to lowercase, removing non-alphanumeric characters, and reducing repeated whitespace. Keyword matching was performed using word-aware regular expressions in order to reduce accidental substring matches. A record was assigned to a broad category only if its subject labels matched exactly one of the predefined categories. Records that matched more than one category were excluded, as were records that could not be mapped reliably to any selected category. This procedure was used to construct a balanced dataset of 15,000 records, with 1,500 books per category. During parsing, at most 500,000 lines of the Open Library works dump were scanned, and the procedure stopped earlier if the required number of records was obtained for all categories. The balanced design was used to prevent frequent categories from dominating the training process and to make the comparison of classification performance across categories more reliable.

Because no independent manually annotated gold standard was available for the broad categories, precision and recall of the keyword mapping itself could not be estimated in the strict supervised sense. The mapping was therefore used as a deterministic dataset construction rule rather than as an evaluated classifier. To reduce obvious mapping noise, only records that matched exactly one broad category were retained. A manual validation of a stratified sample of mapped records would be required to estimate the mapping precision and recall directly.

Table 1 summarizes the main filtering outcomes observed during dataset construction. Out of 357,133 parsed JSON records, 85,078 records did not match

any of the selected broad categories, while 23,291 records matched more than one broad category and were therefore excluded from the single-label dataset. In addition, 36,648 records were excluded because the title–description text contained fewer than five words. Among the records that matched exactly one broad category, 36,648 out of 74,076 records were excluded because the title–description text contained fewer than five words, corresponding to 49.47%. The final dataset contained 15,000 records, corresponding to 1,500 records for each of the ten broad subject categories.

This sampling strategy creates a controlled and balanced benchmark, but it may also introduce selection bias. Records with clear and unambiguous subject labels are more likely to be retained, while sparse, noisy, or genuinely multi-subject records are more likely to be excluded. Therefore, the final dataset should not be interpreted as a representative sample of all Open Library records, but as a filtered benchmark for evaluating classification from title–description metadata under clean single-label conditions.

The task was formulated as single-label classification in order to obtain a controlled benchmark for comparing textual representations and classifiers under the same evaluation protocol. This design choice simplifies the bibliographic reality, since books may naturally belong to more than one subject area. However, retaining only records that matched exactly one broad category made it possible to construct a balanced dataset and to evaluate all models using standard multiclass metrics. The implications of this simplification are considered in the error analysis and in the discussion of future multi-label extensions.

3.3 Text Representation

Let x_i denote the textual input of the i -th book record, obtained by concatenating its title and description, and let $y_i \in \mathcal{C}$ denote the corresponding broad subject category. The dataset is therefore represented as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N,$$

where $N = 15000$.

Two types of textual representation were evaluated. First, a classical lexical representation was obtained using term frequency–inverse document frequency (TF-IDF). The TF-IDF vectorizer was configured to use lowercasing, English stop-word removal, unigrams and bigrams, a maximum vocabulary size of 50,000 terms, a minimum document frequency threshold of three documents ($\text{min_df}=3$), a maximum document frequency threshold of 90% of the training documents ($\text{max_df}=0.90$), and sublinear term-frequency scaling. This representation was used as input for the classical machine learning models.

Second, a transformer-based semantic representation was evaluated using the

`sentence-transformers/all-MiniLM-L6-v2` model. Each title–description text was encoded into a dense sentence embedding. The embeddings were normalized and then used as input to a Logistic Regression classifier. This setting does not perform full transformer fine-tuning; instead, the transformer is used as a fixed sentence encoder, followed by a conventional supervised classifier. Because the `sentence-transformers/all-MiniLM-L6-v2` encoder truncates inputs longer than 256 word pieces by default, the tokenized length of each title–description input was additionally inspected. This check was used to determine how many records may have been affected by truncation before embedding generation.

For k-nearest neighbors and the multilayer perceptron classifier, dimensionality reduction was applied before classification. This was not intended to imply that nearest-neighbor methods are generally unsuitable for TF-IDF representations. In fact, nearest-neighbor classification can be applied to sparse TF-IDF vectors when an appropriate similarity measure, such as cosine similarity, is used. In the present experiment, truncated singular value decomposition was used as a practical design choice to obtain a compact dense representation, reduce computational cost, and use the same reduced representation for both the KNN and MLP baselines. The reduced vectors were then standardized before classification.

This configuration was selected deliberately as a lightweight and reproducible setting. The objective was to evaluate whether frozen semantic embeddings can improve over classical lexical representations when computational resources and training data are limited. Full transformer fine-tuning was not included in the present comparison because it introduces additional computational cost, hyperparameter sensitivity, and a different experimental setting.

3.4 Classification Models

Six classification approaches were compared. Five approaches used TF-IDF-based lexical features, while one approach used transformer-based sentence embeddings:

1. TF-IDF with Decision Tree;
2. TF-IDF with Random Forest;
3. TF-IDF with Linear Support Vector Machine;
4. TF-IDF with truncated SVD and k-Nearest Neighbors;
5. TF-IDF with truncated SVD and multilayer perceptron neural network;
6. `sentence-transformers/all-MiniLM-L6-v2` sentence embeddings with Logistic Regression.

The Decision Tree classifier was trained with a maximum depth of 60 and a minimum of three samples per leaf. The Random Forest classifier used 300 trees, a minimum of two samples per leaf, and balanced subsampling of class weights. The Linear Support Vector Machine used a linear decision function with $C = 1.0$ and balanced class weights. The k-Nearest Neighbors classifier used $k = 7$, distance-based weighting, and cosine distance. The multilayer perceptron consisted of two hidden layers with 256 and 128 neurons, ReLU activation, L_2 regularization parameter $\alpha = 10^{-4}$, initial learning rate 10^{-3} , batch size 64, and a maximum of 50 training iterations. Finally, the transformer-based approach used Logistic Regression with balanced class weights and a maximum of 2000 iterations.

The selected models represent different classification principles. Decision Tree and Random Forest models represent tree-based learning, Linear Support Vector Machine is a strong linear baseline for sparse textual data, k-Nearest Neighbors represents similarity-based classification, and the multilayer perceptron provides a neural baseline on top of reduced TF-IDF features. The MiniLM-based approach was included to evaluate whether dense transformer embeddings provide an advantage over lexical representations when only title and description are available as input.

In addition to the main fair comparison, a separate diagnostic upper-bound experiment was conducted using the original Open Library subject labels as predictive text. This experiment was not treated as part of the main comparison, because the broad target categories were derived from the same metadata field. Its purpose was only to estimate how much performance may increase when the classifier receives metadata that is directly related to the target construction rule.

3.5 Training and Testing Procedure

The final dataset was randomly shuffled using a fixed random seed and divided into training and test subsets. A stratified split was used, with 75% of the records assigned to the training set and 25% assigned to the test set. Stratification preserved the same class distribution in both subsets. Since the dataset was balanced, each category was represented equally in the training and test data.

All models were trained only on the training subset. The test subset was used exclusively for final evaluation. This ensured that the reported results measured the ability of the models to generalize to previously unseen bibliographic records.

3.6 Evaluation Metrics

Model performance was evaluated using standard multiclass classification metrics. The following measures were reported: accuracy; balanced accuracy; macro-averaged precision; macro-averaged recall; macro-

averaged F1-score; weighted F1-score; and total runtime. Since the dataset was balanced, accuracy and balanced accuracy are directly comparable, and accuracy is not dominated by more frequent classes. Macro-F1 was nevertheless used as the primary comparison metric because it combines class-level precision and recall and is sensitive to uneven error patterns across categories. This is relevant in the present task because some broad categories are semantically more overlapping than others, even though all categories contain the same number of test examples. Weighted-F1 was reported additionally as an aggregate measure that accounts for the number of samples per class. In addition to aggregate metrics, a full classification report was generated for each model. This allowed the comparison of precision, recall, and F1-score at the category level.

3.7 Analysis of Misclassification Patterns

After model evaluation, the models were ranked according to macro-F1. The confusion matrix of the best-performing model was then generated and analyzed. This analysis was used to identify common misclassification patterns and subject categories that were frequently confused. Such an analysis is particularly important in bibliographic classification because not all errors have the same interpretation. Misclassifications between categories such as Biography and History, Science and Technology, or Business and Education may reflect genuine semantic overlap in the bibliographic descriptions. Therefore, the confusion matrix was interpreted not only as an error analysis tool, but also as an indicator of the conceptual difficulty of broad subject classification. Finally, example predictions of the best-performing model were saved for qualitative inspection. These examples included the original record information, the true category, the predicted category, and an indicator of whether the prediction was correct. This allowed a more detailed interpretation of individual correct and incorrect classifications.

4 Results

In this section, the results of the automatic subject classification experiment described in Section 3 are presented. The evaluation was conducted on the stratified test set containing 3,750 bibliographic records, with 375 records from each of the ten subject categories. Since the dataset was balanced, accuracy and balanced accuracy are directly comparable. Macro-F1 is used as the primary performance indicator because it summarizes class-level precision and recall and is sensitive to uneven error patterns across semantically overlapping categories.

The MiniLM truncation audit showed that 324 out of 15,000 records exceeded the 256 word-piece limit, corresponding to 2.16% of the dataset. The median input

length was 15 word pieces, the mean input length was 35.0 word pieces, and the 95th percentile was 166 word pieces. Therefore, truncation affected only a small subset of records, although some descriptions were substantially longer, with the maximum observed length reaching 1,732 word pieces. This should be considered when interpreting the transformer-based results for records with longer descriptions.

4.1 Model Comparison

Table 2 reports the performance of the evaluated classification models. The compared approaches include five TF-IDF-based classical machine learning models and one transformer-based semantic representation using MiniLM embeddings followed by Logistic Regression.

The reported runtime denotes the total measured execution time of the evaluated pipeline. For TF-IDF-based models, this includes TF-IDF feature construction, model training, and prediction on the test set. For the MiniLM-based model, it includes embedding generation for the training and test texts, Logistic Regression training, and prediction. Therefore, the runtime values should be interpreted as end-to-end pipeline times rather than isolated classifier training times.

The diagnostic subject-label experiment achieved an accuracy of 0.9995 and a macro-F1 score of 0.9995. This result should not be interpreted as a fair comparison with the title–description models, because the subject labels were also used to construct the broad target categories. However, the near-perfect performance confirms that subject labels contain direct category information. This supports their exclusion from the main predictive input and shows why the title–description setting provides a more realistic estimate of classification performance.

The best result was obtained by the transformer-based model, which achieved an accuracy of 0.6157 and a macro-F1 score of 0.6173. The strongest classical baseline was TF-IDF + Linear SVM, with a macro-F1 score of 0.5323, followed by TF-IDF + Random Forest with a macro-F1 score of 0.5238. Thus, the MiniLM-based representation improved macro-F1 by approximately 8.5 percentage points compared with the best TF-IDF-based baseline. This result indicates that dense semantic embeddings are more suitable for this task than purely lexical representations. The advantage is expected, since the input consists of relatively short bibliographic texts in which the subject category may be expressed indirectly. Nevertheless, the TF-IDF + Linear SVM and TF-IDF + Random Forest models achieved competitive results, especially considering their substantially lower execution time. Classical lexical models therefore remain relevant baselines for automatic bibliographic classification, particularly in settings where interpretability and computational efficiency are important. Figure 1 provides a visual com-

parison of the macro-F1 scores across all evaluated models.

4.2 Per-Class Performance

Since the MiniLM embeddings combined with Logistic Regression achieved the highest macro-F1 score, this model was analyzed further at the level of individual subject categories. Table 3 reports precision, recall, and F1-score for each class.

The highest F1-score was achieved for Technology, followed by Business, Medicine, Religion, and Education. These categories are characterized by relatively distinctive vocabularies, such as terms related to computing, engineering, management, health, theology, or teaching. In contrast, the lowest F1-scores were obtained for Biography and History. This is consistent with the structure of the task, since these two categories are often semantically close. Historical books may focus on individuals, while biographical books frequently contain historical context. The obtained class-level results therefore show that the classification difficulty is not uniform across categories. Categories with more specialized terminology are classified more reliably, while categories with broader or overlapping semantic scope produce more errors.

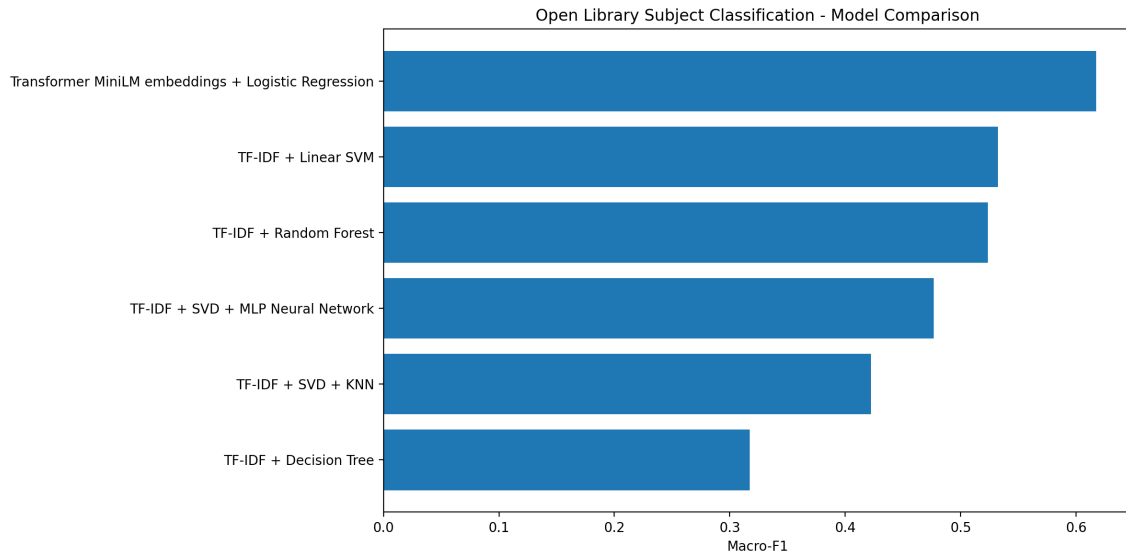
4.3 Confusion Matrix Analysis

In order to examine the structure of classification errors, the confusion matrix of the best-performing model was analyzed. Table 4 reports the confusion matrix, where rows correspond to true labels and columns correspond to predicted labels, while Figure 2 provides its graphical representation.

The confusion matrix confirms that the model correctly classified the majority of records in all categories. The highest number of correct classifications was obtained for Technology, with 274 correctly classified records, followed by Business with 259, Religion with 252, and Medicine with 245. These results are consistent with the per-class F1-scores reported in Table 3. The most evident confusion occurs between semantically related categories. Biography is often confused with History, with 58 Biography records classified as History. Conversely, 42 History records were classified as Biography. This confirms that the boundary between these categories is not always clear when only title and description are used. A similar pattern can be observed between Science and Technology. Technology achieved the highest overall class performance, but 36 Technology records were classified as Science, while 25 Science records were classified as Technology. This indicates that applied scientific and technical descriptions may share similar vocabulary. Additional confusions are also meaningful from a bibliographic perspective. Art is sometimes classified as History, which may correspond to art history records.

Table 2: Model comparison on the Open Library subject classification task.

Model	Accuracy	Balanced Acc.	Macro Prec.	Macro Recall	Macro-F1	Weighted-F1	Total runtime (s)
Transformer MiniLM embeddings + Logistic Regression	0.6157	0.6157	0.6205	0.6157	0.6173	0.6173	92.88
TF-IDF + Linear SVM	0.5301	0.5301	0.5359	0.5301	0.5323	0.5323	1.68
TF-IDF + Random Forest	0.5173	0.5173	0.5397	0.5173	0.5238	0.5238	2.99
TF-IDF + SVD + MLP Neural Network	0.4739	0.4739	0.4875	0.4739	0.4766	0.4766	36.92
TF-IDF + SVD + KNN	0.4200	0.4200	0.4294	0.4200	0.4222	0.4222	7.16
TF-IDF + Decision Tree	0.2877	0.2877	0.6171	0.2877	0.3172	0.3172	1.35

**Figure 1:** Comparison of the evaluated classification models according to macro-F1. The transformer-based model achieved the highest score, followed by TF-IDF + Linear SVM and TF-IDF + Random Forest.**Table 3:** Per-class performance of the best-performing model: Transformer MiniLM embeddings + Logistic Regression.

Class	Precision	Recall	F1-score	Support
Art	0.587	0.576	0.581	375
Biography	0.454	0.483	0.468	375
Business	0.687	0.691	0.689	375
Education	0.712	0.613	0.659	375
Fiction	0.608	0.608	0.608	375
History	0.462	0.525	0.492	375
Medicine	0.712	0.653	0.682	375
Religion	0.648	0.672	0.660	375
Science	0.624	0.605	0.614	375
Technology	0.712	0.731	0.721	375

Fiction is sometimes confused with Biography, which may reflect memoir-like, biographical, or narrative descriptions. Religion also shows confusions with Biography and History, which may occur in works about religious figures, traditions, or historical religious movements. These results suggest that a part of the observed error is not caused only by model limitations, but also by the inherent ambiguity of broad subject classification. In such cases, a single-label prediction may be too restrictive, since some books can reasonably belong to more than one subject area. Consequently, the reported single-label accuracy may underestimate the practical usefulness of the model in a decision-support setting. A prediction counted as incorrect under the

single-label protocol may still be useful if it represents a plausible secondary category for librarian review. For this reason, the present results should be interpreted as conservative evidence for broad category suggestion rather than as a complete evaluation of multi-subject cataloguing.

4.4 Summary of Results

The experimental results show that transformer-based sentence embeddings achieved the best overall performance on the Open Library subject classification task. The MiniLM embeddings combined with Logistic Regression achieved a macro-F1 score of 0.6173, outperforming the strongest classical baseline, TF-IDF + Linear SVM, by approximately 8.5 percentage points.

At the same time, the results show that the task remains challenging. Even the best model achieved moderate, rather than near-perfect, classification performance. This is expected because the input was intentionally restricted to title and description, while original subject labels were excluded from the predictive representation. Therefore, the task reflects a realistic setting in which the model must infer the subject category from incomplete and sometimes ambiguous bibliographic text.

Overall, the results support the use of semantic text representations for automatic subject-category predic-

Table 4: Confusion matrix for the best-performing model. Rows represent true labels and columns represent predicted labels.

True / Pred.	Art	Biography	Business	Education	Fiction	History	Medicine	Religion	Science	Technology
Art	216	34	6	11	17	43	7	12	9	20
Biography	36	181	11	5	38	58	8	28	10	0
Business	6	13	259	12	3	23	16	4	15	24
Education	15	10	12	230	16	22	18	19	15	18
Fiction	14	49	4	4	228	19	9	25	17	6
History	35	42	19	9	23	197	10	20	13	7
Medicine	8	24	21	14	19	8	245	14	13	9
Religion	18	34	6	7	18	23	6	252	9	2
Science	9	12	17	21	13	22	17	12	227	25
Technology	11	0	22	10	0	11	8	3	36	274

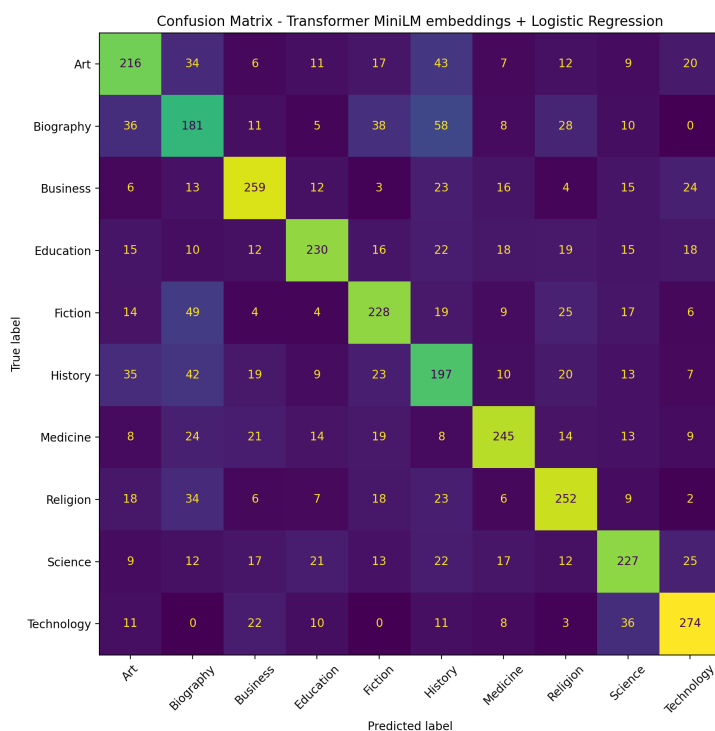


Figure 2: Confusion matrix for the best-performing model, Transformer MiniLM embeddings + Logistic Regression. Diagonal values represent correct classifications, while off-diagonal values represent misclassifications.

tion from bibliographic metadata. However, the observed misclassification patterns also show that automatic classification should be interpreted as a decision-support mechanism. The model can provide useful candidate categories, but expert review remains important, especially for records located near the semantic boundaries between records subject areas. In a practical cataloguing workflow, the model should be used as a decision-support tool rather than as an autonomous cataloguing system. For each new record, the model could provide the most likely broad category together with alternative high-ranking categories for librarian review. Since the present study did not perform probability calibration or threshold optimization, no fixed confidence threshold for automatic acceptance can be recommended. In deployment, such a threshold should be selected on a validation set according to institutional requirements. For example, high-

confidence predictions could be accepted automatically only when precision exceeds a predefined operational target, while lower-confidence and semantically borderline cases should be routed to expert review.

5 Conclusion

This study presented a reproducible experiment for automatic subject classification of books based on open bibliographic metadata from Open Library. The predictive input was intentionally limited to title and description, while original subject labels were used only to construct the target categories. In this way, the evaluation avoided label leakage and reflected a more realistic setting for metadata enrichment. The results showed that MiniLM sentence embeddings combined with Logistic Regression achieved the best overall per-

formance, outperforming classical TF-IDF-based models. However, the obtained macro-F1 score of 0.6173 also shows that the task remains challenging. Many errors occurred between semantically related categories, such as Biography and History or Science and Technology, indicating that some misclassifications reflect genuine ambiguity in bibliographic subject assignment.

Overall, the findings suggest that semantic text representations can support broad subject-category prediction from limited bibliographic descriptions. Nevertheless, automatic classification should be understood as a decision-support mechanism rather than a replacement for expert cataloguing. The dataset construction procedure may favor records with clearer and less ambiguous subject metadata, since ambiguous, sparse, and unmapped records were excluded. In addition, the transformer model was used only as a fixed sentence encoder. Fine-tuning MiniLM or another transformer model on the same training data may improve performance, but it would require a separate experimental design and additional hyperparameter control. Future work should reformulate the task as multi-label category recommendation, where several plausible subject categories are ranked and presented to librarians for review.

References

- Chou, C.-Y. (2022). An analysis of bert (nlp) for assisted subject indexing. *Cataloging & Classification Quarterly*, 60(8), 683–702. doi: 10.1080/01639374.2022.2138666
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1423
- Golub, K. (2019). Automatic subject indexing of text. In B. Hjørland & C. Gnoli (Eds.), *Isko encyclopedia of knowledge organization*. International Society for Knowledge Organization. Retrieved from <https://www.isko.org/cyclo/automatic>
- Golub, K., Hagelbäck, J., & Ardö, A. (2021). Automated subject indexing: An overview. *Cataloging & Classification Quarterly*, 59(8), 702–719. doi: 10.1080/01639374.2021.2012311
- Kaufman, S., Rosset, S., & Perlich, C. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 15. doi: 10.1145/2382577.2382579
- Kazi, N., Thompson, B., Choudhury, S. R., & Oard, D. W. (2021). Automatically cataloging scholarly articles using library of congress subject headings. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Student research workshop* (pp. 69–75). Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-srw.7
- Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150. doi: 10.3390/info10040150
- Mödden, E. (2022). Artificial intelligence, machine learning and bibliographic control: Ddc short numbers: Towards machine-based classifying. *JLIS: Italian Journal of Library, Archives and Information Science= Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione*: 13, 1, 2022, 256–264.
- Nordström, J. (2018). *Automated classification of bibliographic data using svm and naive bayes* (Unpublished master's thesis). Uppsala University.
- Open Library. (2025). *Open library data dumps*. <https://openlibrary.org/developers/dumps>. (Accessed: 2026-05-17)
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3982–3992). doi: 10.18653/v1/D19-1410
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. doi: 10.1016/0306-4573(88)90021-0
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47. doi: 10.1145/505282.505283
- Suominen, O. (2019). Annif: Diy automated subject indexing using multiple algorithms. *LIBER quarterly*, 29(1), 1–25.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30).